Differential-Algebraic Equations Forum

DAE-F

Achim Ilchmann
Timo Reis *Editors*

Surveys in Differential-Algebraic Equations IV

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 IV

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 978-3-319-46617-0 ISBN 978-3-319-46618-7 (eBook)
DOI 10.1007/978-3-319-46618-7

Library of Congress Control Number: 2017930698

Mathematics Subject Classification (2010): 15A22, 15A24, 34A09, 34A12, 65D05, 65F30, 65F50, 65L05, 65L06, 65L80, 65M20, 70E55, 93B07, 93B10, 93B25, 93B27, 93B05, 93B07, 93B10, 93B25, 93B27, 93C05, 01-02, 34-03

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

We are pleased to present the fourth volume of survey articles in various fields of differential-algebraic equations (DAEs).

In the present 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 "On the History of Differential-Algebraic Equations: A Retrospective with Personal Side Trips" gives an overview of the timeline and achievements on theory and practice of differential-algebraic equations in the past few decades. In "DAE Aspects of Multibody System Dynamics", the contributions of DAE theory and numerical analysis for modelling and simulation of systems in mechanical multibody dynamics are highlighted. In "Model Reduction for DAEs: A Survey", the state of the art in approximation of large-scale DAEs by low-dimensional ones is presented. The chapter "Observability of Linear Differential-Algebraic Systems: A Survey" treats observability for linear time-invariant DAEs. The fifth chapter is a survey of numerical methods for DAEs.

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 differential-algebraic equations are now an established field in applied mathematics.

Ilmenau, Germany — Achim Ilchmann
Hamburg, Germany — Timo Reis
August 2016

Contents

On the History of Differential-Algebraic Equations

A Retrospective with Personal Side Trips

Bernd Simeon

Abstract The present article takes an off-the-wall approach to the history of Differential-Algebraic Equations and uses personal side trips and memories of conferences, workshops, and summer schools to highlight some of the milestones in the field. Emphasis is in particular placed on the application fields that set the ball rolling and on the development of numerical methods.

Keywords Differential-algebraic equations • Historical remarks • Index notions • BDF methods • Runge–Kutta methods • Partial differential-algebraic equations • Constrained mechanical system • Electric circuit analysis

Mathematics Subject Classification (2010): 34A09, 65L80, 65M20, 01-02, 34-03

1 Introduction

To write about the history of a subject is a challenge that grows with the number of pages as the original goal of completeness becomes more and more impossible. With this in mind, the present article takes a very narrow approach and uses personal side trips and memories of conferences, workshops, and summer schools as the stage for highlighting some of the most important protagonists and their contributions to the field of Differential-Algebraic Equations (DAEs).

Completeness is thus out of the question, and instead it is my intention to provide a storyline that intersperses facts and results with background information. The latter is particularly important in teaching. In my experience students love personal stories about those who first found the theorem they are confronted with. For this reason, I hope that this work will not only be of interest for colleagues and researchers in

B. Simeon (✉)
Felix-Klein-Zentrum, TU Kaiserslautern, D-67663, Kaiserslautern, Germany
e-mail: simeon@mathematik.uni-kl.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7_1

general, but also for the next generations of motivated PhD students who choose the rich topic of differential-algebraic equations as their subject.

The paper is organized as follows. Under the heading *The Early Days* I recall my first encounter with DAEs way back in 1987 and then go further back in time, with particular focus on the application fields in mechanics and electric circuits that finally would trigger an avalanche of research in applied mathematics and in the engineering sciences. The second section is called *The Boom Days* and covers essentially the period from 1989 to 1996 when DAEs had become a hot topic and attracted more and more researchers. Finally, the last section has the title *Consolidation* and highlights the developments of the following 10 years until 2006 when an Oberwolfach Workshop celebrated *25 Years of DAEs*.

As pointed out above, this essay does not aim at completeness, and it has a bias towards numerical analysis. Those readers who would like to know more about the topic of DAEs and the rich oeuvre that has accumulated over the years are referred to the monographs of Brenan, Campbell and Petzold [17], Griepentrog and März [37], Hairer and Wanner [41], Kunkel and Mehrmann [56], Lamour, März and Tischendorf [58], and to the survey of Rabier and Rheinboldt [74].

2 The Early Days

Who were the pioneers that first studied the subject of differential-algebraic equations? And what was the motivation to look into such systems? This section starts at the end of the *Early Days* when I personally happened to learn about DAEs and then goes further back in time, arriving finally at the works of Kirchhoff [52] and Lagrange [57] who introduced differential equations with constraints in order to model electric circuits and mechanical systems.

2.1 First Encounter

It was the winter of 1986/87 when I first encountered the topic of DAEs. At that time, I was a math student at TU München, and I took part in a seminar on *Numerical Methods for Electric Circuit Analysis* organized by Claus Führer, Albert Gilg, and Peter Lory, under the guidance of Roland Bulirsch. Several of the student presentations in the seminar dealt with the transient analysis of electric circuits and the quest for the development of appropriate time integration methods. Since my own presentation, however, was concerned with sparsity considerations and the efficient solution of linear systems of equations, DAEs did not really attract my attention.

More than one year passed before this attitude would eventually changed. Meanwhile, Claus Führer had completed his PhD at TU München and was back at the German Aerospace Center (DLR) in Oberpfaffenhofen, and he offered me

```
      SUBROUTINE DDASSL (RES,NEQ,T,Y,YPRIME,TOUT,INFO,RTOL,ATOL,
     +     IDID,RWORK,LRW,IWORK,LIW,RPAR,IPAR,JAC)
C***BEGIN PROLOGUE  DDASSL
C***PURPOSE  This code solves a system of differential/algebraic
C            equations of the form G(T,Y,YPRIME) = 0.
```

Fig. 1 Calling sequence of the DASSL code [17, 72] that has had an enormous impact on the subject of DAEs and that is still in wide use today. The original code is written in FORTRAN77 in double precision. A recent implementation in C is part of the SUNDIALS suite of codes [50]

an interesting topic for my Diploma Thesis, with Peter Rentrop as supervisor at TU München. The topic was concerned with the computation of quasi-stationary solutions of DAEs arising in mechanical multibody systems, with special focus on wheel-rail dynamics. In order to be able to draw on the expertise of the engineers and practitioners at DLR, I got a contract to work there as a student assistant and wrote most of the thesis at the lab in Oberpfaffenhofen.

In June 1988, a couple of weeks after I had started at DLR, Claus Führer asked me to help him with the preparation of a 3-day workshop on *Numerical Time Integration Methods for ODEs and DAEs* that was hosted by the Carl-Cranz-Gesellschaft e.V., a society that provides continuing education and training in the engineering sciences. The main speaker of the workshop was Linda Petzold, and thus I had the great opportunity to attend her lessons and also to run simulations with the DASSL code [72], see Fig. 1.

In her talks, Linda Petzold typically began with *fully implicit systems*

$$\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \boldsymbol{0} \tag{2.1}$$

with state variables $\boldsymbol{x}(t) \in \mathbb{R}^{n_x}$ and a nonlinear, vector-valued function $\boldsymbol{F}$ of corresponding dimension. Clearly, if the $n_x \times n_x$ Jacobian $\partial \boldsymbol{F}/\partial \dot{\boldsymbol{x}}$ is invertible, then by the implicit function theorem, it is theoretically possible to transform (2.1), at least locally, to an explicit system of ordinary differential equations. If $\partial \boldsymbol{F}/\partial \dot{\boldsymbol{x}}$ is singular, however, (2.1) constitutes the most general form of a *differential-algebraic equation*.

At that time, DAEs were becoming a hot topic, in particular in numerical analysis, and Linda Petzold was one of the leading pioneers who set the pace and laid the foundation for what was to come in the years thereafter. In particular, the development of the DASSL code that she had started in the early 1980s [17, 72] set a corner-stone that still exists today.

Conceptually, it is intriguingly simple to replace the differential operator $\mathrm{d}/\mathrm{d}t$ in (2.1) by the Backward Differentiation Formula (BDF)

$$\varrho \boldsymbol{x}_{n+k} := \sum_{i=0}^{k} \alpha_i \boldsymbol{x}_{n+i} = \tau \dot{\boldsymbol{x}}(t_{n+k}) + \mathcal{O}(\tau^{k+1}) \tag{2.2}$$

where $\boldsymbol{x}_{n+i}$ stands for the discrete approximation of $\boldsymbol{x}(t_{n+i})$ with stepsize τ and where the α_i, $i = 0, \ldots, k$, denote the method coefficients that constitute the difference operator ϱ. Using the finite difference approximation $\varrho\boldsymbol{x}_{n+k}/\tau$ of the time derivative, the numerical solution of the DAE (2.1) then boils down to solving the nonlinear system

$$\boldsymbol{F}\left(\frac{\varrho\boldsymbol{x}_{n+k}}{\tau}, \boldsymbol{x}_{n+k}, t_{n+k}\right) = \boldsymbol{0} \tag{2.3}$$

for $\boldsymbol{x}_{n+k}$ in each time step, and this is exactly the underlying idea of DASSL.

I still recall the atmosphere of departure at that workshop in Oberpfaffenhofen, and over the following years, at various other meetings, I had the chance to become part of a scientific community in this field that was growing steadily. Below, I will come back to this point by interspersing further personal side trips.

2.2 Who Coined the Term DAEs?

Linda Petzolds's academic teacher was Bill Gear, who is widely recognized as the first mathematician of modern time who turned his attention to the field of DAEs. The first occurrence of the term *Differential-Algebraic Equation* can be found in the title of Gear's paper *Simultaneous numerical solution of differential-algebraic equations* [33] from 1971, and in the same year his famous book *Numerical Initial Value Problems in Ordinary Differential Equations* [32] appeared where he considers examples from electric circuit analysis in the form

$$\boldsymbol{E}\dot{\boldsymbol{x}} = \boldsymbol{\phi}(\boldsymbol{x}, t) \tag{2.4}$$

with singular capacitance matrix $\boldsymbol{E} \in \mathbb{R}^{n_x \times n_x}$ and right-hand side function $\boldsymbol{\phi}$.

Moreover, it was also Gear who made the BDF methods popular for solving stiff ODE systems and who wrote one of the first sophisticated codes with variable order and variable stepsize, the DIFSUB routine, for this purpose. The extension of this BDF method to linear-implicit systems (2.4) by means of the difference operator ϱ from (2.2) is straightforward and provided the first available DAE solver.

Two application fields, namely electric circuit analysis and constrained mechanical systems, are among the major driving forces for the development of DAEs. Below, this statement will be made more explicit by looking at the corresponding modeling concepts. Bill Gear had the farsightedness to perceive very early the importance of these modeling approaches for today's simulation software. During an Oberwolfach workshop in 1981, he suggested to study the *mathematical pendulum in Cartesian coordinates*

$$\ddot{q}_1 = -2q_1\lambda, \tag{2.5a}$$

$$\ddot{q}_2 = -\gamma - 2q_2\lambda, \tag{2.5b}$$

$$0 = q_1^2 + q_2^2 - 1 \tag{2.5c}$$

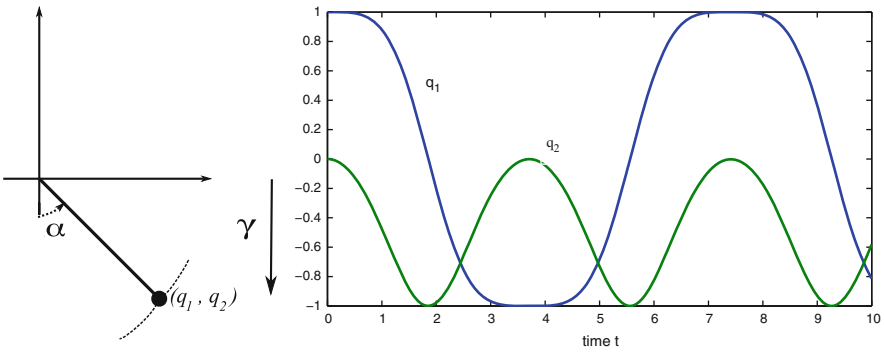


Fig. 2 The"inevitable" pendulum

that describes the motion of a mass point with coordinates (q_1, q_2) in the plane subject to a constraint. The constraint models the massless rod of length 1 that connects the mass point to a pivot placed in the origin of the coordinate system, Fig. 2. The motion of the mass point is then determined by the gravity (parameter γ) and by the constraint forces that are expressed in terms of the unknown Lagrange multiplier λ.

The DAE (2.5) is an example of Lagrange equations of the first kind that we will discuss below. By introducing velocity variables, it can be easily converted to a system of first order that fits into the class of linear-implicit DAEs (2.4).

In retrospective, the applied mathematics community in 1981 was not ready to understand the importance of this new paradigm for modeling technical systems, and the engineering disciplines still preferred to manually transform the models to ordinary differential equations. It would take several more years until the growing use of sophisticated modeling software necessitated a different viewpoint, see the sections below on electric circuit analysis and constrained mechanical systems.

The notion of an index of the DAE (2.1) goes also back to Gear [34, 35]. He introduced what we call today the *differentiation index*. This non-negative integer k is defined by

$k = 0$: If $\partial \boldsymbol{F}/\partial \dot{\boldsymbol{x}}$ is non-singular, the index is 0.
$k > 0$: Otherwise, consider the system of equations

$$\begin{aligned} \boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) &= \boldsymbol{0}, \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \frac{\partial}{\partial \dot{\boldsymbol{x}}}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t)\,\boldsymbol{x}^{(2)} + \ldots &= \boldsymbol{0}, \\ &\vdots \\ \frac{\mathrm{d}^s}{\mathrm{d}t^s}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \frac{\partial}{\partial \dot{\boldsymbol{x}}}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t)\,\boldsymbol{x}^{(s+1)} + \ldots &= \boldsymbol{0} \end{aligned} \tag{2.6}$$

as a system in the separate dependent variables $\dot{\boldsymbol{x}}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(s+1)}$, with $\boldsymbol{x}$ and t as independent variables. Then the index k is the smallest s for which it is possible, using algebraic manipulations only, to extract an ordinary differential equation $\dot{\boldsymbol{x}} = \boldsymbol{\psi}(\boldsymbol{x}, t)$ (the underlying ODE) from (2.6).

Meanwhile other notions of an index have emerged, but despite its ambiguity with respect to the algebraic manipulations, the differentiation index is still the most popular and widespread tool to classify DAEs.

In the next section, other index concepts and their relation to the differential index will be addressed, and also more protagonists will enter the stage. This first section on the early days of DAEs closes now with a look at the application fields that set the ball rolling.

2.3 *Kirchhoff, Weierstrass, and Kronecker*

In 1847, Kirchhoff first published his *circuit laws* that describe the conservation properties of electric circuits [52]. These laws consist of the current law and the voltage law, which both follow from Maxwell's equations of electro-dynamics. When these laws are applied to circuits with time-dependent behavior, the corresponding equations are typically given as a linear-implicit system (2.4). Often, the structure even turns out to be a linear constant coefficient DAE

$$\boldsymbol{E}\dot{\boldsymbol{x}} + \boldsymbol{H}\boldsymbol{x} = \boldsymbol{c}(t) \tag{2.7}$$

with matrices $\boldsymbol{E}, \boldsymbol{H} \in \mathbb{R}^{n_x \times n_x}$ and a time-dependent source term $\boldsymbol{c}(t) \in \mathbb{R}^{n_x}$.

An example of such an electric circuit is the *differentiator* [40] shown in Fig. 3. It consists of a resistance R, an inductance L, an ideal operational amplifier $A = \infty$, and a given voltage source $V(t)$. The $n_x = 6$ unknowns read here $\boldsymbol{x} = (V_1, V_2, V_3, I, I_L, I_V)$ with voltages V_i and currents I, I_L, I_V. From Kirchhoff's laws and the properties of the amplifier and the inductance one obtains the relations

$$\begin{aligned} I + (V_1 - V_2)/R &= 0, \\ -(V_1 - V_2)/R + I_L &= 0, \\ -I_L + I_V &= 0, \end{aligned}$$

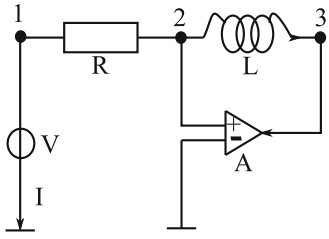


Fig. 3 Differentiator circuit

$$V_1 = V(t),$$
$$V_2 = 0,$$
$$V_2 - V_3 = L \cdot \dot{I}_L.$$

This linear system has the form (2.7) with singular inductance matrix

$$\boldsymbol{E} = \begin{pmatrix} 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & L & 0 \end{pmatrix}. \tag{2.8}$$

If the matrix $\boldsymbol{E}$ was regular, it could be brought to the right-hand side by formal inversion, ending up in a system of ODEs. Here however, it is singular and thus we face a DAE problem.

Weierstrass and Kronecker were in Berlin at the same time as Kirchhoff, and it is quite possible to suppose that they knew his work.[1] Weierstrass and later Kronecker were thus inspired to study such singular systems and provided an elegant theory that is still fundamental today in order to understand the specific properties of DAEs.

We assume that the *matrix pencil* $\mu\boldsymbol{E} + \boldsymbol{H} \in \mathbb{R}^{n_x \times n_x}[\mu]$ is regular. That is, there exists $\mu \in \mathbb{C}$ such that the matrix $\mu\boldsymbol{E} + \boldsymbol{H}$ is regular. Otherwise, the pencil is singular, and (2.7) has either no or infinitely many solutions. This latter case was first studied by Kronecker [54], see also [21, 30].

If $\mu\boldsymbol{E} + \boldsymbol{H}$ is regular, there exist nonsingular matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ such that

$$\boldsymbol{UEV} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{N} \end{pmatrix}, \quad \boldsymbol{UHV} = \begin{pmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \tag{2.9}$$

where $\boldsymbol{N}$ is a nilpotent matrix, $\boldsymbol{I}$ the identity matrix, and $\boldsymbol{C}$ a matrix that can be assumed to be in Jordan canonical form. Note that the dimensions of these square blocks in (2.9) are uniquely determined. The transformation (2.9) is called the *Weierstrass canonical form* [86]. It is a generalization of the Jordan canonical form and contains the essential structure of the linear system (2.7).

In the Weierstrass canonical form (2.9), the singularity of the DAE is represented by the nilpotent matrix $\boldsymbol{N}$. Its degree of nilpotency, i.e., the smallest positive integer k such that $\boldsymbol{N}^k = \boldsymbol{0}$, plays a key role when studying closed-form solutions of the linear system (2.7) and is identical to the differentiation index of (2.7).

[1] The relation of the work of Weierstrass and Kronecker to Kirchhoff's circuit laws was pointed out to me by Volker Mehrmann when we met in September 2014 during a Summer School on DAEs in Elgersburg, Germany.

To construct a solution of (2.7), we introduce new variables and right-hand side vectors

$$\boldsymbol{V}^{-1}\boldsymbol{x} =: \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{pmatrix}, \quad \boldsymbol{U}\boldsymbol{c} =: \begin{pmatrix} \boldsymbol{\delta} \\ \boldsymbol{\theta} \end{pmatrix}. \tag{2.10}$$

Premultiplying (2.7) by $\boldsymbol{U}$ then leads to the *decoupled system*

$$\dot{\boldsymbol{y}} + \boldsymbol{C}\boldsymbol{y} = \boldsymbol{\delta}\,, \tag{2.11a}$$

$$\boldsymbol{N}\dot{\boldsymbol{z}} + \boldsymbol{z} = \boldsymbol{\theta}\,. \tag{2.11b}$$

While the solution of the ODE (2.11a) follows by integrating and results in an expression based on the matrix exponential $\exp(-\boldsymbol{C}(t-t_0))$, the Eq. (2.11b) for $\boldsymbol{z}$ can be solved recursively by differentiating. More precisely, it holds that

$$\boldsymbol{N}\ddot{\boldsymbol{z}} + \dot{\boldsymbol{z}} = \dot{\boldsymbol{\theta}} \quad \Rightarrow \quad \boldsymbol{N}^2\ddot{\boldsymbol{z}} = -\boldsymbol{N}\dot{\boldsymbol{z}} + \boldsymbol{N}\dot{\boldsymbol{\theta}} = \boldsymbol{z} - \boldsymbol{\theta} + \boldsymbol{N}\dot{\boldsymbol{\theta}}\,.$$

Repeating the differentiation and multiplication by $\boldsymbol{N}$, we can eventually exploit the nilpotency and get

$$\boldsymbol{0} = \boldsymbol{N}^k\boldsymbol{z}^{(k)} = (-1)^k\boldsymbol{z} + \sum_{\ell=0}^{k-1}(-1)^{k-1-\ell}\boldsymbol{N}^\ell\boldsymbol{\theta}^{(\ell)}.$$

This implies the explicit representation

$$\boldsymbol{z} = \sum_{\ell=0}^{k-1}(-1)^\ell\boldsymbol{N}^\ell\boldsymbol{\theta}^{(\ell)}. \tag{2.12}$$

The above solution procedure illustrates several crucial points about DAEs and how they differ from ODEs. Remarkably, the linear constant coefficient case also displays these points, and thus the work of Weierstrass and Kronecker still represents the foundation of DAE theory today.

We highlight two crucial points:

1. The solution of (2.7) rests on $k-1$ differentiation steps. This requires that the derivatives of certain components of $\boldsymbol{\theta}$ exist up to $\ell = k-1$. Furthermore, some components of $\boldsymbol{z}$ may be continuous but not differentiable depending on the smoothness of $\boldsymbol{\theta}$.
2. The components of $\boldsymbol{z}$ are directly given in terms of the right-hand side data $\boldsymbol{\theta}$ and its derivatives. Accordingly, the initial value $\boldsymbol{z}(t_0) = \boldsymbol{z}_0$ is fully determined by (2.12) and, in contrast to $\boldsymbol{y}_0$, cannot be chosen arbitrarily. Initial values $(\boldsymbol{y}_0, \boldsymbol{z}_0)$ where $\boldsymbol{z}_0$ satisfies (2.12) are called *consistent*. The same terminology applies to the initial value $\boldsymbol{x}_0$, which is consistent if, after the transformation (2.10), $\boldsymbol{z}_0$ satisfies (2.12).

Today, more than 150 years after the discoveries of Kirchhoff, electric circuit analysis remains one of the driving forces in the development of DAEs. The interplay of modeling and mathematical analysis is particularly important in this field, and the interested reader is referred to Günther and Feldmann [39] and März and Tischendorf [66] as basic works. The first simulation code that generated a model in differential-algebraic form was the SPICE package [70].

2.4 Euler and Lagrange

Even older than the DAEs arising from Kirchhoff's laws are the Euler–Lagrange equations. They were first published in Lagrange's famous work *Mécanique analytique* [57] in 1788.

Consider a mechanical system that consists of rigid bodies interacting via springs, dampers, joints, and actuators, Fig. 4. The bodies possess a certain geometry and mass while the interconnection elements are massless. Let $\boldsymbol{q}(t) \in \mathbb{R}^{n_q}$ denote a vector that comprises the coordinates for position and orientation of all bodies in the system. Revolute, translational, universal, and spherical joints are examples of bondings in such a multibody system. They may constrain the motion $\boldsymbol{q}$ and hence determine its kinematics.

If constraints are present, we express the resulting conditions on $\boldsymbol{q}$ in terms of n_λ constraint equations

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q}) \,. \tag{2.13}$$

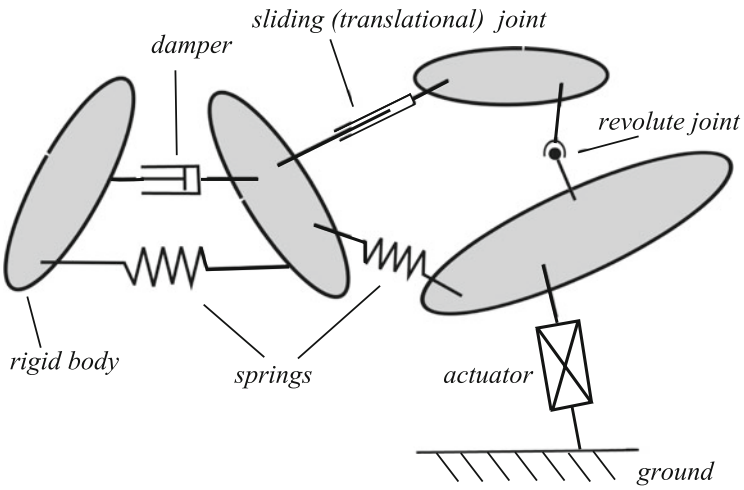


Fig. 4 Sketch of a multibody system with rigid bodies and typical interconnections

Obviously, a meaningful model requires $n_\lambda < n_q$. The Eqs. (2.13) that restrict the motion $\boldsymbol{q}$ are called *holonomic constraints*, and the rectangular matrix

$$\boldsymbol{G}(\boldsymbol{q}) := \frac{\partial \boldsymbol{g}(\boldsymbol{q})}{\partial \boldsymbol{q}} \in \mathbb{R}^{n_\lambda \times n_q}$$

is called the *constraint Jacobian*.

Using both the redundant position variables $\boldsymbol{q}$ and additional Lagrange multipliers $\boldsymbol{\lambda}$ to describe the dynamics leads to the *equations of constrained mechanical motion*, also called the *Lagrange equations of the first kind* or the *Euler–Lagrange equations*

$$\boldsymbol{M}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} = \boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) - \boldsymbol{G}(\boldsymbol{q})^T \boldsymbol{\lambda}\,, \tag{2.14a}$$

$$\boldsymbol{0} = \boldsymbol{g}(\boldsymbol{q})\,, \tag{2.14b}$$

where $\boldsymbol{M}(\boldsymbol{q}) \in \mathbb{R}^{n_q \times n_q}$ stands for the *mass matrix* and $\boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}, t) \in \mathbb{R}^{n_q}$ for the vector of *applied and internal forces*.

The standard example for such a constrained mechanical system are the Eqs. (2.5) of the mathematical pendulum. For a long time, it was common sense that the Euler–Lagrange equations should be transformed to the *state space form*, also called the *Lagrange equations of the second kind*. In case of the pendulum, this means that the Cartesian coordinates can be expressed as $q_1 = \sin\alpha$, $q_2 = -\cos\alpha$ with the angle α as minimal coordinate, Fig. 2. By inserting these relations into (2.5), the constraints and the Lagrange multiplier cancel, and one arrives at the second order ODE

$$\ddot{\alpha} = -\gamma \sin\alpha \tag{2.15}$$

as state space form.

It seems obvious that a state space form such as (2.15) constitutes a more appropriate and easier model than the differential-algebraic system (2.14), or (2.5), respectively, in redundant coordinates. In practice, however, the state space form suffers from serious drawbacks.

The analytical complexity of the constraint equations (2.13) makes it in various applications impossible to obtain a set of minimal coordinates that is valid for all configurations of the multibody system. Moreover, although we know from the theorem on implicit functions that such a set exists in a neighborhood of the current configuration, it might lose its validity when the configuration changes. This holds in particular for multibody systems with so-called *closed kinematic loops*.

Even more, the modeling of subsystems like electrical and hydraulic feedback controls, which are essential for the performance of modern mechanical systems, is limited. The differential-algebraic model, on the other hand, bypasses topological analysis and offers the choice of using a set of coordinates $\boldsymbol{q}$ that possess physical significance.

This reasoning in favor of the differential-algebraic model (2.14) became more and more widespread in the 1980s, driven by the development of sophisticated software packages, so-called *multibody formalisms*. One of the first packages that fully exploited this new way of modelling is due to Haug [48].

A look at the leading software tools in the field today shows a clear picture. Some of the codes generate a differential-algebraic model whenever a constraint is present, while others try to generate a state space form as long as it is convenient. But the majority of commercial products rely on the differential-algebraic approach as the most general way to handle complex technical applications [31, 80].

The main difference between DAEs arising from electric circuit analysis and DAEs that model constrained mechanical systems is the richer structure of the latter. For example, for *conservative multibody systems*, i.e., systems where the applied forces can be written as the gradient of a potential U, the Euler–Lagrange equations (2.14) result from Hamilton's principle of least action

$$\int_{t_0}^{t_1} \left(T - U - \boldsymbol{g}(\boldsymbol{q})^T \boldsymbol{\lambda} \right) \mathrm{d}t \rightarrow \text{ stationary !} \tag{2.16}$$

where the kinetic energy possesses a representation as quadratic form

$$T(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \frac{1}{2} \dot{\boldsymbol{q}}^T \boldsymbol{M}(\boldsymbol{q}) \dot{\boldsymbol{q}} \,.$$

In the least action principle (2.16), we observe the fundamental Lagrange multiplier technique for coupling constraints and dynamics [19]. Extensions of the multiplier technique exist in various more general settings such as dissipative systems or even inequality constraints.

The pendulum equations (2.5) are an example of a constrained mechanical system. Though they simply describe the motion of a single mass point, several key properties of the Euler–Lagrange equations can also be studied: the differential equations are of second order, the constraint equations are mostly nonlinear, and one observes a clear semi-explicit structure with differential variables $\boldsymbol{q}$ and algebraic variables $\boldsymbol{\lambda}$.

The Euler–Lagrange equations are of index 3 and form the prototype for a system of higher index. Index-reduction techniques are thus required and in fact, in 1972 this issue was addressed by Baumgarte [12]. He observed that in (2.14), the Lagrange multipliers can be eliminated by differentiating the constraints twice. The first differentiation leads to the *constraints at velocity level*

$$\boldsymbol{0} = \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{G}(\boldsymbol{q})\, \dot{\boldsymbol{q}}. \tag{2.17}$$

A second differentiation step yields the *constraints at acceleration level*

$$\boldsymbol{0} = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{G}(\boldsymbol{q})\, \ddot{\boldsymbol{q}} + \boldsymbol{\kappa}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \,, \quad \boldsymbol{\kappa}(\boldsymbol{q}, \dot{\boldsymbol{q}}) := \frac{\partial \boldsymbol{G}(\boldsymbol{q})}{\partial \boldsymbol{q}} (\dot{\boldsymbol{q}}, \dot{\boldsymbol{q}}) \,, \tag{2.18}$$

where the two-form $\boldsymbol{\kappa}$ comprises additional derivative terms. The combination of the dynamic equation

$$\boldsymbol{M}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} = \boldsymbol{f}(\boldsymbol{q},\dot{\boldsymbol{q}},t) - \boldsymbol{G}(\boldsymbol{q})^T\boldsymbol{\lambda}$$

with (2.18) results in a linear system for $\ddot{\boldsymbol{q}}$ and $\boldsymbol{\lambda}$ with the saddle point matrix

$$\begin{pmatrix} \boldsymbol{M}(\boldsymbol{q}) & \boldsymbol{G}(\boldsymbol{q})^T \\ \boldsymbol{G}(\boldsymbol{q}) & \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{(n_q+n_\lambda)\times(n_q+n_\lambda)}. \tag{2.19}$$

For a well-defined multibody system, this matrix is invertible in a neighborhood of the solution, and in this way, the Lagrange multiplier can be computed as a function of $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$.

However, the well-known drift-off phenomenon requires additional stabilization measures, and Baumgarte came up with the idea to combine original and differentiated constraints as

$$\boldsymbol{0} = \boldsymbol{G}(\boldsymbol{q})\,\ddot{\boldsymbol{q}} + \boldsymbol{\kappa}(\boldsymbol{q},\dot{\boldsymbol{q}}) + 2\alpha\boldsymbol{G}(\boldsymbol{q})\dot{\boldsymbol{q}} + \beta^2\boldsymbol{g}(\boldsymbol{q}) \tag{2.20}$$

with scalar parameters α and β. The free parameters α and β should be chosen in such a way that

$$\boldsymbol{0} = \ddot{\boldsymbol{w}} + 2\alpha\dot{\boldsymbol{w}} + \beta^2\boldsymbol{w} \tag{2.21}$$

becomes an asymptotically stable equation, with $\boldsymbol{w}(t) := \boldsymbol{g}(\boldsymbol{q}(t))$.

From today's perspective, the crucial point in Baumgarte's approach is the choice of the parameters. Nevertheless, it was the very beginning of a long series of works that tried to reformulate the Euler–Lagrange equations in such a way that the index is lowered while still maintaining the information of all constraint equations. For a detailed analysis of this stabilization and related techniques we refer to Ascher et al. [8, 10].

Another—very early—stabilization of the Euler–Lagrange equations is due to Gear, Gupta and Leimkuhler [36]. This formulation still represents the state-of-the-art in multibody dynamics. It uses a formulation of the equations of motion as system of first order with velocity variables $\boldsymbol{v} = \dot{\boldsymbol{q}}$ and simultaneously enforces the constraints at velocity level (2.17) and the position constraints (2.13), where the latter are interpreted as invariants and appended by means of extra Lagrange multipliers.

In this way, one obtains an enlarged system

$$\begin{aligned} \dot{\boldsymbol{q}} &= \boldsymbol{v} - \boldsymbol{G}(\boldsymbol{q})^T\boldsymbol{\mu}\,, \\ \boldsymbol{M}(\boldsymbol{q})\,\dot{\boldsymbol{v}} &= \boldsymbol{f}(\boldsymbol{q},\boldsymbol{v},t) - \boldsymbol{G}(\boldsymbol{q})^T\boldsymbol{\lambda}\,, \\ \boldsymbol{0} &= \boldsymbol{G}(\boldsymbol{q})\,\boldsymbol{v}\,, \\ \boldsymbol{0} &= \boldsymbol{g}(\boldsymbol{q}) \end{aligned} \tag{2.22}$$

with additional multipliers $\boldsymbol{\mu}(t) \in \mathbb{R}^{n_\lambda}$. A straightforward calculation shows

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{G}(\boldsymbol{q})\dot{\boldsymbol{q}} = \boldsymbol{G}(\boldsymbol{q})\,\boldsymbol{v} - \boldsymbol{G}(\boldsymbol{q})\boldsymbol{G}^T(\boldsymbol{q})\boldsymbol{\mu} = -\boldsymbol{G}(\boldsymbol{q})\boldsymbol{G}^T(\boldsymbol{q})\boldsymbol{\mu}$$

and one concludes that $\boldsymbol{\mu} = \mathbf{0}$ since $\boldsymbol{G}(\boldsymbol{q})$ is of full rank and hence $\boldsymbol{G}(\boldsymbol{q})\boldsymbol{G}^T(\boldsymbol{q})$ is invertible. With the additional multipliers $\boldsymbol{\mu}$ vanishing, (2.22) and the original equations of motion (2.14) coincide along any solution. Yet, the index of the *GGL formulation* (2.22) is 2 instead of 3. Some authors refer to (2.22) as a *stabilized index-2 system*.

In the fall of 1988—when I was finishing my master thesis at DLR Oberpfaffenhofen, Claus Führer and Ben Leimkuhler then showed that the GGL formulation in combination with a BDF discretization is basically equivalent to solving the equations of constrained mechanical motion as an overdetermined system by means of a certain generalized inverse [29]. The result became one of the most highly cited papers of those years, which demonstrates that the DAEs and their numerical analysis had attracted wide attention by then.

The above paragraphs on stabilized formulations of the Euler–Lagrange equations demonstrate that the development of theory and numerical methods for DAEs was strongly intertwined with the mathematical models. This holds for all application fields where DAEs arise. We leave this point as a loose end and turn now to what one could call the *Golden Age of DAEs.*

3 The Boom Days

Between 1989 and 1996, both theory and numerical analysis of DAEs were booming, and many groups in mathematics and engineering started to explore this new research topic. Driven by the development of powerful simulation packages in the engineering sciences, the demand for efficient and robust integration methods was growing steadily while at the same time it had become apparent that higher index problems require stabilization measures or appropriate reformulations.

This trend was reflected by a series of workshops and conferences dedicated to DAEs, and three such occasions will serve here as the stage for showcasing a—rather personal—selection of hot topics.

3.1 The Paderborn Workshops

After finishing my diploma degree, I worked for a couple of months for the DLR (German Aerospace Center) until the end of 1989. Sponsored by the Volkswagen Foundation, an interdisciplinary project on *Identifizierungs-, Analyse- und Entwurfsmethoden für mechanische Mehrkörpersysteme in Deskriptorform (Iden-*

tification, Analysis and Design for Mechanical Multibody Systems in Descriptor Form) gave me the opportunity to do a PhD at TU München, with Peter Rentrop as supervisor. Our partners were the DLR lab in Oberpfaffenhofen with Claus Führer and Willi Kortüm and the University of Wuppertal with Peter C. Müller, who acted as coordinator of the joint project.

A part of the project plan was the organization of two workshops that should bring together the leading experts in control theory, engineering, and mathematics and thus foster the further development of DAEs. The first workshop took place in March 1992 in the Liborianum Monastery in Paderborn, and this marked the outset of a bi-annual series of workshops that would last until 2005. A recent revival meeting was organized by Sebastian Schöps and colleagues in March 2013.

Those who have attended one or more of the Paderborn Workshops recall the vivid atmosphere that was full of stimulating discussions. Confusion and misunderstandings in the communication between mathematicians and engineers happened quite often in these early days, and the distinction between a capacitor and a capacitance or the explanation of an error message *'corrector could not converge'* could result in a controversial and simultaneously entertaining discussion, see the snapshot of a slide in Fig. 5 that was presented at the first Paderborn Workshop in 1992. Looking back, these workshops gave me a great chance to get in touch with various leading researchers in the field.

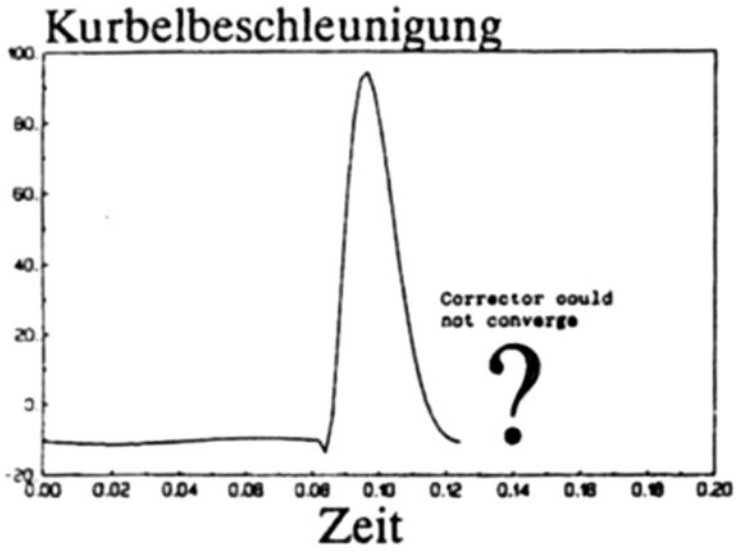


Fig. 5 Snapshot of a slide presented at the Paderborn Workshop in March 1992. Courtesy of Günter Leister, Daimler AG Sindelfingen

3.1.1 The Geneva School

At the first Paderborn meeting in 1992, Ernst Hairer gave a talk on *half-explicit Runge–Kutta methods* for semi-explicit DAEs of index 2 [15, 16]. Jointly with Christian Lubich and Michel Roche, he had written the groundbreaking monograph on *The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods* [44] 3 years before. In this rich work, several new method classes, a new paradigm for the construction of convergence proofs, a new index concept, and the new RADAU5 code are presented. From then on, the Geneva School played a very strong role in the further development of DAEs and corresponding numerical methods.

The *perturbation index* as defined in [44] sheds a different light on DAEs and adopts the idea of a well-posed mathematical model. While the differential index is based on successively differentiating the original DAE (2.1) until the obtained system can be solved for $\dot{\boldsymbol{x}}$, the perturbation index measures the sensitivity of the solutions to perturbations in the equation.

The system $\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \boldsymbol{0}$ has perturbation index $k \geq 1$ along a solution $\boldsymbol{x}(t)$ on $[t_0, t_1]$ if k is the smallest integer such that, for all functions $\hat{\boldsymbol{x}}$ having a defect

$$\boldsymbol{F}(\dot{\hat{\boldsymbol{x}}}, \hat{\boldsymbol{x}}, t) = \boldsymbol{\delta}(t)\,,$$

there exists on $[t_0, t_1]$ an estimate

$$\|\hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)\| \leq c\Big(\|\hat{\boldsymbol{x}}(t_0) - \boldsymbol{x}(t_0)\| + \max_{t_0 \leq \xi \leq t} \|\boldsymbol{\delta}(\xi)\| + \ldots + \max_{t_0 \leq \xi \leq t} \|\boldsymbol{\delta}^{(k-1)}(\xi)\|\Big)$$

whenever the expression on the right-hand side is sufficiently small. Note that the constant c depends only on $\boldsymbol{F}$ and on the length of the interval, but not on the perturbation $\boldsymbol{\delta}$. The perturbation index is $k = 0$ if

$$\|\hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)\| \leq c\Big(\|\hat{\boldsymbol{x}}(t_0) - \boldsymbol{x}(t_0)\| + \max_{t_0 \leq \xi \leq t} \Big\| \int_{t_0}^{\xi} \boldsymbol{\delta}(\tau)\, \mathrm{d}\tau \Big\|\Big)\,,$$

which is satisfied for ordinary differential equations.

If the perturbation index exceeds $k = 1$, derivatives of the perturbation show up in the estimate and indicate a certain degree of ill-posedness. For example, if $\boldsymbol{\delta}$ contains a small high-frequency term $\epsilon \sin \omega t$ with $\epsilon \ll 1$ and $\omega \gg 1$, the resulting derivatives will induce a severe amplification in the bound for $\hat{\boldsymbol{x}}(t) - \boldsymbol{x}(t)$.

Unfortunately, the differential and the perturbation index are not equivalent in general and may even differ substantially [23]. The story of this discovery is connected with another personal side trip in the following chapter.

The definition of the perturbation index is solely a prelude in [44]. As the title says, most of the monograph deals with Runge–Kutta methods, in particular implicit ones. These are extended to linear-implicit systems $\boldsymbol{E}\dot{\boldsymbol{x}} = \boldsymbol{\phi}(\boldsymbol{x}, t)$ by assuming for a moment that the matrix $\boldsymbol{E}$ is invertible and discretizing $\dot{\boldsymbol{x}} = \boldsymbol{E}^{-1}\boldsymbol{\phi}(\boldsymbol{x}, t)$. Multiplying

the resulting scheme by $\boldsymbol{E}$, one gets the method definition

$$\boldsymbol{E}\boldsymbol{X}_i = \boldsymbol{E}\boldsymbol{x}_0 + \tau \sum_{j=1}^{s} a_{ij}\boldsymbol{\phi}(\boldsymbol{X}_j, t_0 + c_j\tau), \quad i = 1, \ldots, s; \tag{3.1a}$$

$$\boldsymbol{x}_1 = \left(1 - \sum_{i,j=1}^{s} b_i\gamma_{ij}\right)\boldsymbol{x}_0 + \tau \sum_{i,j=1}^{s} b_i\gamma_{ij}\boldsymbol{X}_j. \tag{3.1b}$$

Here, the method coefficients are denoted by $(a_{ij})_{i,j=1}^{s}$ and $b_1, \ldots, b_s$ while $(\gamma_{ij}) = (a_{ij})^{-1}$ is the inverse of the coefficient matrix, with s being the number of stages. Obviously, (3.1) makes sense also in the case where $\boldsymbol{E}$ is singular.

Using *stiffly accurate methods* for differential-algebraic equations is advantageous, which becomes evident if we consider the discretization of the semi-explicit system

$$\dot{\boldsymbol{y}} = \boldsymbol{a}(\boldsymbol{y}, \boldsymbol{z}), \tag{3.2a}$$

$$\boldsymbol{0} = \boldsymbol{b}(\boldsymbol{y}, \boldsymbol{z}) \tag{3.2b}$$

with differential variables $\boldsymbol{y}$ and algebraic variables $\boldsymbol{z}$. The method (3.1) then reads

$$\boldsymbol{Y}_i = \boldsymbol{y}_0 + \tau \sum_{j=1}^{s} a_{ij}\boldsymbol{a}(\boldsymbol{Y}_j, \boldsymbol{Z}_j), \quad i = 1, \ldots, s, \tag{3.3a}$$

$$\boldsymbol{0} = \boldsymbol{b}(\boldsymbol{Y}_i, \boldsymbol{Z}_i), \tag{3.3b}$$

for the internal stages and

$$\boldsymbol{y}_1 = \boldsymbol{y}_0 + \tau \sum_{j=1}^{s} b_j\boldsymbol{a}(\boldsymbol{Y}_j, \boldsymbol{Z}_j), \tag{3.4a}$$

$$z_1 = \left(1 - \sum_{i,j=1}^{s} b_i\gamma_{ij}\right)z_0 + \tau \sum_{i,j=1}^{s} b_i\gamma_{ij}\boldsymbol{Z}_j \tag{3.4b}$$

as update for the numerical solution after one step. For stiffly accurate methods, we have $\sum_{i,j=1}^{s} b_i\gamma_{ij} = 1$ and $\boldsymbol{y}_1 = \boldsymbol{Y}_s, z_1 = \boldsymbol{Z}_s$. The update (3.4) is hence superfluous and furthermore, the constraint $\boldsymbol{0} = \boldsymbol{b}(\boldsymbol{y}_1, z_1)$ is satisfied by construction.

It is not the purpose of this article to dive further into the world of Runge–Kutta methods, but like in numerical ODEs, the rivalry between multistep methods and Runge–Kutta methods also characterizes the situation for DAEs. While Linda Petzold's DASSL code is the most prominent multistep implementation, the RADAU5 and RADAU codes [41, 42] represent the one-step counterparts and have also become widespread in various applications.

The competition for the best code was a major driving force in the numerical analysis of DAEs, and from time to time those in favor of multistep methods looked also at one-step methods, e.g., in [9], and vice versa. Nevertheless, I would like to quote from a statement of Linda Petzold that nicely reflects the different communities: *'The BDFs are so beautiful, why would anybody consider a different method?'*

Simultaneously to the joint work with Ernst Hairer and Michel Roche, Christian Lubich investigated a different class of discretization schemes, the *half-explicit methods* [61]. These methods are tailored for semi-explicit DAEs and discretize the differential equations explicitly while the constraint equations are enforced in an implicit fashion. As an example, consider the Euler–Lagrange equations (2.14) with velocity constraint (2.17). The half-explicit Euler method as a generic algorithm for the method class reads

$$\begin{aligned} \boldsymbol{q}_{n+1} &= \boldsymbol{q}_n + \tau \boldsymbol{v}_n \,, \\ \boldsymbol{M}(\boldsymbol{q}_n)\, \boldsymbol{v}_{n+1} &= \boldsymbol{M}(\boldsymbol{q}_n)\, \boldsymbol{v}_n + \tau \boldsymbol{f}(\boldsymbol{q}_n, \boldsymbol{v}_n, t_n) - \tau \boldsymbol{G}(\boldsymbol{q}_n)^T \boldsymbol{\lambda}_n \,, \\ \boldsymbol{0} &= \boldsymbol{G}(\boldsymbol{q}_{n+1})\, \boldsymbol{v}_{n+1} \,. \end{aligned} \tag{3.5}$$

Only a linear system of the form

$$\begin{pmatrix} \boldsymbol{M}(\boldsymbol{q}_n) & \boldsymbol{G}(\boldsymbol{q}_n)^T \\ \boldsymbol{G}(\boldsymbol{q}_{n+1}) & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{v}_{n+1} \\ \tau \boldsymbol{\lambda}_n \end{pmatrix} = \begin{pmatrix} \boldsymbol{M}(\boldsymbol{q}_n)\boldsymbol{v}_n + \tau \boldsymbol{f}(\boldsymbol{q}_n, \boldsymbol{v}_n, t_n) \\ \boldsymbol{0} \end{pmatrix}$$

arises here in each step. The scheme (3.5) forms the basis for a class of extrapolation methods [61, 63], and also for half-explicit Runge–Kutta methods as introduced in [44] and then further enhanced by Brasey and Hairer [16] and Arnold and Murua [6].

These methods have in common that only information about the velocity constraints is required. As a remedy for the drift-off, which grows only linearly but might still be noticeable, the following projection, which is also due to Lubich [61], can be applied: Let $\boldsymbol{q}_{n+1}$ and $\boldsymbol{v}_{n+1}$ denote the numerical solution of the system, obtained by integration from consistent values $\boldsymbol{q}_n$ and $\boldsymbol{v}_n$. Then, the projection consists of the following steps:

$$\textit{solve} \begin{cases} \boldsymbol{0} = \boldsymbol{M}(\tilde{\boldsymbol{q}}_{n+1})(\tilde{\boldsymbol{q}}_{n+1} - \boldsymbol{q}_{n+1}) + \boldsymbol{G}(\tilde{\boldsymbol{q}}_{n+1})^T \boldsymbol{\mu}, \\ \boldsymbol{0} = \boldsymbol{g}(\tilde{\boldsymbol{q}}_{n+1}) \end{cases} \textit{for } \tilde{\boldsymbol{q}}_{n+1}, \boldsymbol{\mu} \,; \tag{3.6a}$$

$$\textit{solve} \begin{cases} \boldsymbol{0} = \boldsymbol{M}(\tilde{\boldsymbol{q}}_{n+1})(\tilde{\boldsymbol{v}}_{n+1} - \boldsymbol{v}_{n+1}) + \boldsymbol{G}(\tilde{\boldsymbol{q}}_{n+1})^T \boldsymbol{\eta}, \\ \boldsymbol{0} = \boldsymbol{G}(\tilde{\boldsymbol{q}}_{n+1})\, \tilde{\boldsymbol{v}}_{n+1} \end{cases} \textit{for } \tilde{\boldsymbol{v}}_{n+1}, \boldsymbol{\eta} \,. \tag{3.6b}$$

A simplified Newton method can be used to solve the nonlinear system (3.6a) while (3.6b) represents a linear system for $\tilde{\boldsymbol{v}}_{n+1}$ and $\boldsymbol{\eta}$ with similar structure.

The projection can also be employed for stabilizing the equations of motion with acceleration constraint (2.18) where the position and velocity constraints are invariants and not preserved by the time integration, see Eich [27] and von

Schwerin [81]. Such projection methods are particularly attractive in combination with explicit ODE integrators.

3.1.2 DAEs and Control Theory

Control theory had and still has a considerable impact on DAEs. This holds both for interesting applications and for theoretical work. At the Paderborn Workshops, this was reflected by the participation of groups from Engineering Control and from Mathematical Control.

Peter C. Müller, organizer of the Paderborn Workshops and with degrees in mathematics and engineering perfectly suited for bringing together the different communities, was one of the first in control theory who realized that many such problems lead to DAEs in a natural way. But the traditional approach had always been to manually transform these models into ODEs, which at the time had become more and more tedious or even impossible. Classical concepts such as controllability and observability for DAEs were addressed by Müller and his co-workers in the early 1990s and regularly presented at the Paderborn Workshops [68, 69].

From the very beginning, Volker Mehrmann came quite often to Paderborn. With his background in numerical linear algebra and mathematical control theory, he brought in a completely new perspective. As he told me recently, Volker Mehrmann first got in touch with DAEs when working for the IBM Scientific Center in Heidelberg from 1988 to 1989, jointly with Peter Kunkel. They were confronted with a differential-algebraic Riccati equation

$$\begin{aligned} -\boldsymbol{E}(t)^T\dot{\boldsymbol{X}}(t)\boldsymbol{E}(t) = \boldsymbol{E}(t)^T\boldsymbol{X}(t)\boldsymbol{A}(t) + \boldsymbol{A}(t)^T\boldsymbol{X}^T\boldsymbol{E}(t) + \boldsymbol{Q}(t) \\ -\boldsymbol{E}(t)^T\boldsymbol{X}(t)\boldsymbol{W}(t)\boldsymbol{X}(t)\boldsymbol{E}(t) \end{aligned} \tag{3.7}$$

with matrices $\boldsymbol{X}(t), \boldsymbol{A}(t), \boldsymbol{Q}(t), \boldsymbol{W}(t) \in \mathbb{R}^{n_x \times n_x}$ and singular matrix $\boldsymbol{E}(t)$ of the same dimension. Such equations arise for example from optimal regulator problems or from optimal filters with DAE models involved, and a straightforward strategy is to rewrite the symmetric unknown matrix $\boldsymbol{X}$ into a long vector of size $n_x(n_x+1)/2$ and then to convert (3.7) to a DAE.

In the particular application Kunkel and Mehrmann were considering, however, this turned out to be more challenging than expected [55]. Even more the numerical solution of the final DAE by DASSL produced trajectories that did not match with the results of the code LIMEX, an extrapolation method that had just before been released by Deuflhard, Hairer and Zugck [25]. This surprising behavior woke the interest of Kunkel and Mehrmann for the problem class. Later on, it turned out that both codes had computed correct solutions but the equation itself admitted multiple solutions.

3.1.3 The Berlin School

Two and a half years after the Fall of the Berlin Wall, the first Paderborn Workshop provided an opportunity to get in touch with Roswitha März and her co-workers from the Humboldt Universität zu Berlin. A long time before, März had already started to work on DAEs, with the book by her and Griepentrog [37] being the very first monograph on the topic, and over the years her projector-based analysis became the distinguishing mark of what I call here the Berlin School.

This approach is characterized by striving for a rigorous mathematical treatment of DAEs. Following [58], the projector construction can be easily illustrated by means of the constant coefficient DAE (2.7) that read

$$\boldsymbol{E}\dot{\boldsymbol{x}} + \boldsymbol{H}\boldsymbol{x} = \boldsymbol{c}.$$

While the Weierstrass canonical form (2.9) leads to a transformed system in new variables and is hard to compute in practice, the projector-based analysis maintains the original state variables $\boldsymbol{x}$ and proceeds as follows.

In the first step, one sets $\boldsymbol{G}_0 := \boldsymbol{E}$, $\boldsymbol{B}_0 := \boldsymbol{H}$ and determines the subspace $\mathcal{N}_0 := \ker \boldsymbol{G}_0$. For singular $\boldsymbol{G}_0$, this kernel will be non-trivial, and a projector onto $\mathcal{N}_0$ is denoted by $\boldsymbol{Q}_0$. The complementary projector is

$$\boldsymbol{P}_0 := \boldsymbol{I} - \boldsymbol{Q}_0\,.$$

For the projectors $\boldsymbol{P}_0$ and $\boldsymbol{Q}_0$, important properties hold such as $\boldsymbol{P}_0\boldsymbol{Q}_0 = \boldsymbol{Q}_0\boldsymbol{P}_0 = \boldsymbol{0}$ and $\boldsymbol{G}_0 = \boldsymbol{G}_0(\boldsymbol{P}_0 + \boldsymbol{Q}_0) = \boldsymbol{G}_0\boldsymbol{P}_0$. The original DAE system $\boldsymbol{G}_0\dot{\boldsymbol{x}} + \boldsymbol{B}_0\boldsymbol{x} = \boldsymbol{c}$ is then equivalent to

$$\boldsymbol{G}_0\boldsymbol{P}_0\dot{\boldsymbol{x}} + \boldsymbol{B}_0(\boldsymbol{P}_0 + \boldsymbol{Q}_0)\boldsymbol{x} = \boldsymbol{c} \tag{3.8a}$$

$$\Leftrightarrow \qquad \boldsymbol{G}_1(\boldsymbol{P}_0\dot{\boldsymbol{x}} + \boldsymbol{Q}_0\boldsymbol{x}) + \boldsymbol{B}_1\boldsymbol{x} = \boldsymbol{c} \tag{3.8b}$$

where

$$\boldsymbol{G}_1 := \boldsymbol{G}_0 + \boldsymbol{B}_0\boldsymbol{Q}_0, \qquad \boldsymbol{B}_1 := \boldsymbol{B}_0\boldsymbol{P}_0\,.$$

This step is repeated in terms of

$$\boldsymbol{G}_{i+1} := \boldsymbol{G}_i + \boldsymbol{B}_i\boldsymbol{Q}_i, \qquad \boldsymbol{B}_{i+1} := \boldsymbol{B}_i\boldsymbol{P}_i\,,$$

and it can be shown that the corresponding sequence of matrices $\boldsymbol{G}_i$ in front of the derivative $\dot{\boldsymbol{x}}$ has the property

$$\operatorname{im}\boldsymbol{G}_0 \subseteq \operatorname{im}\boldsymbol{G}_1 \subseteq \ldots \subseteq \operatorname{im}\boldsymbol{G}_i\,.$$

In other words, the regularity of the leading matrix grows, and in the end $\boldsymbol{G}_i$ will become a regular matrix for some i. This is guaranteed for regular matrix pencils

$(\boldsymbol{E}, \boldsymbol{H})$ where the process stops when the nilpotency index of the Weierstrass form k equals the step number i. For singular pencils, the projector-based approach also provides a means to detect and analyze the singularity.

The real power of this procedure unfolds in particular for time-variant systems

$$\boldsymbol{E}(t)\dot{\boldsymbol{x}}(t) + \boldsymbol{H}(t)\boldsymbol{x}(t) = \boldsymbol{c}(t)\,. \tag{3.9}$$

Again, the construction of the matrix chain is the main ingredient and motivates the concept of the *tractability index*, see [58] for a recent comprehensive exposition.

Similar to the competition for the best numerical method where mostly either the BDF or the Runge–Kutta schemes have been favored by the different research groups, the projector-based analysis has contended with several other approaches over the years. Among these are the derivative array technique and the interpretation of DAEs as differential equations on manifolds that will be discussed below.

3.2 The Oberwolfach Workshop in 1993

The Oberwolfach Workshop *Differential-Algebraic Equations: Theory and Applications in Technical Simulation* organized by Hans-Georg Bock, Peter Rentrop, and Werner C. Rheinboldt in June 1993 shone a flashlight on the dynamic development in the field in those years. I recall the atmosphere as very stimulating and full of momentum, and for a PhD student it was both encouraging—the research field was hot and one was part of a rapidly growing community—and discouraging—so many brilliant minds were already working in the field.

A particular challenge emerged during the first day as several speakers presented new time integration methods and proved their efficiency by showing results where the DASSL code was beaten when solving Andrews' squeezer, also known as the seven-body mechanism [4]. In this way, a benchmark was set, and during the following days a kind of horse race took place where the speakers tried to further push their integration schemes.

3.2.1 Differential Equations on Manifolds

In 1984, Werner Rheinboldt investigated DAEs from the viewpoint of differential geometry [78]. While the approaches discussed so far are mainly inspired by differential calculus and algebraic considerations, a fundamentally different aspect is brought into play by his idea of *differential equations on manifolds*.

Referring to [1, 5] for the theoretical underpinnings, we briefly illustrate this approach by considering the semi-explicit system

$$\dot{\boldsymbol{y}} = \boldsymbol{a}(\boldsymbol{y}, \boldsymbol{z}), \tag{3.10a}$$

$$\boldsymbol{0} = \boldsymbol{b}(\boldsymbol{y}) \tag{3.10b}$$

under the assumption

$$\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{y}}(\boldsymbol{y}) \cdot \frac{\partial \boldsymbol{a}}{\partial \boldsymbol{z}}(\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{n_z \times n_z} \quad \text{is invertible} \tag{3.11}$$

in a neighborhood of the solution. Clearly, (3.10) is of index 2 where the constraint $\mathbf{0} = \boldsymbol{b}(\boldsymbol{y})$, assuming sufficient differentiability, defines the manifold

$$\mathcal{M} := \{\boldsymbol{y} \in \mathbb{R}^{n_y} : \boldsymbol{b}(\boldsymbol{y}) = \mathbf{0}\}. \tag{3.12}$$

The full rank condition (3.11) for the matrix product $\partial \boldsymbol{b}/\partial \boldsymbol{y} \cdot \partial \boldsymbol{a}/\partial \boldsymbol{z}$ implies that the Jacobian $\boldsymbol{B}(\boldsymbol{y}) = \partial \boldsymbol{b}(\boldsymbol{y})/\partial \boldsymbol{y} \in \mathbb{R}^{n_z \times n_y}$ also possesses full rank n_z. Hence, for fixed $\boldsymbol{y} \in \mathcal{M}$, the tangent space

$$\mathcal{T}_y\mathcal{M} := \{\boldsymbol{v} \in \mathbb{R}^{n_y} : \boldsymbol{B}(\boldsymbol{y})\boldsymbol{v} = \mathbf{0}\} \tag{3.13}$$

is the kernel of $\boldsymbol{B}$ and has the same dimension $n_y - n_z$ as the manifold $\mathcal{M}$. Figure 6 depicts $\mathcal{M}$, $\mathcal{T}_y\mathcal{M}$, and a solution of the DAE (3.10), which, starting from a consistent initial value, is required to proceed on the manifold.

The differential equation on the manifold $\mathcal{M}$ that is equivalent to the DAE (3.10) is obtained as follows: The hidden constraint

$$\mathbf{0} = \boldsymbol{B}(\boldsymbol{y})\boldsymbol{a}(\boldsymbol{y}, \boldsymbol{z})$$

can be solved for $\boldsymbol{z}(\boldsymbol{y})$ according to the rank condition (3.11) and the implicit function theorem. Moreover, for $\boldsymbol{y} \in \mathcal{M}$ it holds that $\boldsymbol{a}(\boldsymbol{y}, \boldsymbol{z}(\boldsymbol{y})) \in \mathcal{T}_y\mathcal{M}$, which defines a vector field on the manifold $\mathcal{M}$. Overall,

$$\dot{\boldsymbol{y}} = \boldsymbol{a}(\boldsymbol{y}, \boldsymbol{z}(\boldsymbol{y})) \quad \text{for } \boldsymbol{y} \in \mathcal{M} \tag{3.14}$$

then represents a differential equation on the manifold [78].

In theory, and also computationally [79], it is possible to transform the differential equation (3.14) from the manifold to an ordinary differential equation in a linear

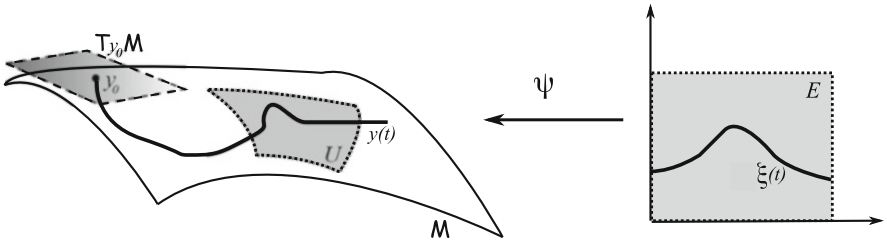


Fig. 6 Manifold $\mathcal{M}$, tangent space $\mathcal{T}_y\mathcal{M}$, and local parametrization

space of dimension $n_y - n_z$. For this purpose, one introduces a *local parametrization*

$$\boldsymbol{\psi} : E \to \mathcal{U} \tag{3.15}$$

where E is an open subset of $\mathbb{R}^{n_y - n_z}$ and $\mathcal{U} \subset \mathcal{M}$, see Fig. 6. Such a parametrization is not unique and holds only locally in general. It is, however, possible to extend it to a family of parametrizations such that the whole manifold is covered. For $\boldsymbol{y} \in \mathcal{U}$ and local coordinates $\boldsymbol{\xi} \in E$ we thus get the relations

$$\boldsymbol{y} = \boldsymbol{\psi}(\boldsymbol{\xi}), \qquad \dot{\boldsymbol{y}} = \boldsymbol{\Psi}(\boldsymbol{\xi})\dot{\boldsymbol{\xi}}, \quad \boldsymbol{\Psi}(\boldsymbol{\xi}) := \frac{\partial \boldsymbol{\psi}}{\partial \boldsymbol{\xi}}(\boldsymbol{\xi}) \in \mathbb{R}^{n_y \times (n_y - n_z)}.$$

Premultiplying (3.14) by the transpose of the Jacobian $\boldsymbol{\Psi}(\boldsymbol{\xi})$ of the parametrization and substituting $\boldsymbol{y}$ by $\boldsymbol{\psi}(\boldsymbol{\xi})$, we arrive at

$$\boldsymbol{\Psi}(\boldsymbol{\xi})^T \boldsymbol{\Psi}(\boldsymbol{\xi})\dot{\boldsymbol{\xi}} = \boldsymbol{\Psi}(\boldsymbol{\xi})^T \boldsymbol{a}(\boldsymbol{\psi}(\boldsymbol{\xi}), z(\boldsymbol{\psi}(\boldsymbol{\xi}))) \,. \tag{3.16}$$

Since the Jacobian $\boldsymbol{\Psi}$ has full rank for a valid parametrization, the matrix $\boldsymbol{\Psi}^T\boldsymbol{\Psi}$ is invertible, and (3.16) constitutes the desired ordinary differential equation in the local coordinates $\boldsymbol{\xi}$. In analogy to a mechanical system in minimal coordinates, we call (3.16) a *local state space form*.

The process of transforming a differential equation on a manifold to a local state space form constitutes a *push forward* operator, while the reverse mapping is called a *pull back* operator [1]. It is important to realize that the previously defined concept of an index does not appear in the theory of differential equations on manifolds. Finding hidden constraints by differentiation, however, is also crucial for the classification of DAEs from a geometric point of view.

The geometrical viewpoint was considered very early by Sebastian Reich [75], but its full potential became clear only a couple of years later when the topic of geometric numerical integration emerged, cf. [46].

3.2.2 Singularly Perturbed Problems and Regularization

In the early days of DAEs, regularization was a quite popular means to convert the algebraic part into a differential equation. Motivated by physical examples such as stiff springs or parasitic effects in electric circuits, a number of authors have looked into this topic. Furthermore, it is also interesting to start with a singularly perturbed ODE, discretize it, and then to analyze the behavior of the exact and numerical solutions in the limit case.

To study an example for a semi-explicit system, we consider Van der Pol's equation

$$\epsilon\ddot{q} + (q^2 - 1)\dot{q} + q = 0 \tag{3.17}$$

with parameter $\epsilon > 0$. This is an oscillator equation with a nonlinear damping term that acts as a controller. For large amplitudes $q^2 > 1$, the damping term introduces dissipation into the system while for small values $q^2 < 1$, the sign changes and the damping term is replaced by an excitation, leading thus to a self-exciting oscillator. Introducing Liénhard's coordinates [45]

$$z := q, \quad y := \epsilon \dot{z} + (z^3/3 - z),$$

we transform (3.17) into the first order system

$$\dot{y} = -z\,, \tag{3.18a}$$

$$\epsilon \dot{z} = y - \frac{z^3}{3} + z\,. \tag{3.18b}$$

The case $\epsilon \ll 1$ is of special interest. In the limit $\epsilon = 0$, the Eq. (3.18b) turns into a constraint and we arrive at the semi-explicit system

$$\dot{y} = -z\,, \tag{3.19a}$$

$$0 = y - \frac{z^3}{3} + z\,. \tag{3.19b}$$

In other words, Van der Pol's equation (3.18) in Liénhard's coordinates is an example of a *singularly perturbed system* which tends to the semi-explicit (3.19) when $\epsilon \to 0$.

Such a close relation between a singularly perturbed system and a differential-algebraic equation is quite common and can be found in various application fields. Often, the parameter ϵ stands for an almost negligible physical quantity or the presence of strongly different time scales. Analyzing the *reduced system*, in this case (3.19), usually proves successful to gain a better understanding of the original perturbed equation [71]. In the context of regularization methods, this relation is also exploited, but in the reverse direction [47]. One starts with a DAE such as (3.19) and replaces it by a singularly perturbed ODE, in this case (3.18).

In numerical analysis, the derivation and study of integration schemes via a singularly perturbed ODE has been termed the *indirect approach* [44] and has led to much additional insight [43, 60, 62], both for the differential-algebraic equation as the limit case and for the stiff ODE case. A particularly interesting method class for the indirect approach are Rosenbrock methods as investigated by Rentrop, Roche and Steinebach [77].

3.2.3 General Fully Implicit DAEs

At the Oberwolfach Workshop of 1993, I met another of the pioneers in the field of DAEs, Steve Campbell. In the late 1970s, he had worked on singular systems of

differential equations and applications in control theory, which led to the book [20]. In the first phase of my PhD, this book became a valuable source and inspiration for me when working on the Drazin inverse in multibody dynamics [85].

Before further discussing the solution of general fully implicit DAEs (2.1), which was the topic of Steve Campbell's talk in Oberwolfach, it makes sense to recall the solution theory in the linear constant coefficient case. While the Weierstrass transformation (2.9) provides the complete structural information of a linear DAE system in new coordinates and decouples the solution, the Drazin inverse represents an elegant means to express the solution in the original coordinates.

To this end, we define

$$\hat{\boldsymbol{E}} := (\mu \boldsymbol{E} + \boldsymbol{H})^{-1}\boldsymbol{E}, \quad \hat{\boldsymbol{H}} := (\mu \boldsymbol{E} + \boldsymbol{H})^{-1}\boldsymbol{H}$$

where $\mu \in \mathbb{C}$ is chosen such that the inverse of $\mu \boldsymbol{E} - \boldsymbol{H}$ exists, which is possible for a regular matrix pencil. Let $\hat{\boldsymbol{E}}$ be decomposed in Jordan canonical form, i.e.

$$\hat{\boldsymbol{E}} = \boldsymbol{T} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{N} \end{bmatrix} \boldsymbol{T}^{-1} \tag{3.20}$$

where $\boldsymbol{R}$ is associated with the non-zero eigenvalues and $\boldsymbol{N}$ is associated with the zero eigenvalues and therefore is nilpotent, as in the Weierstrass canonical form (2.9). The Drazin inverse $\hat{\boldsymbol{E}}^D$ is defined by [26]

$$\hat{\boldsymbol{E}}^D := \boldsymbol{T} \begin{pmatrix} \boldsymbol{R}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} \boldsymbol{T}^{-1} \tag{3.21}$$

or, equivalently, by the axioms

(D1) $\hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^D = \hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}$,
(D2) $\hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}\hat{\boldsymbol{E}}^D = \hat{\boldsymbol{E}}^D$,
(D3) $\hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}^{k+1} = \hat{\boldsymbol{E}}^k$, where k is the nilpotency index of $\boldsymbol{N}$.

The inverse $\hat{\boldsymbol{E}}^D$ always exists, is uniquely determined, and is equal to $\hat{\boldsymbol{E}}^{-1}$ for regular $\hat{\boldsymbol{E}}$. The product $\hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}$ is a projector which can be used to guarantee consistent initial values. Overall, if the matrix pencil $\mu \boldsymbol{E} + \boldsymbol{H} \in \mathbb{R}^{n_x \times n_x}[\mu]$ is regular, the homogeneous linear constant coefficient DAE $\boldsymbol{E}\dot{\boldsymbol{x}} + \boldsymbol{H}\boldsymbol{x} = \boldsymbol{0}$ possesses the solution [20]

$$\boldsymbol{x}(t) = \exp(\hat{\boldsymbol{E}}^D\hat{\boldsymbol{H}}t)\hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}\boldsymbol{x}_0\,. \tag{3.22}$$

The initial vector x_0 is consistent if and only if $\hat{\boldsymbol{E}}^D\hat{\boldsymbol{E}}\boldsymbol{x}_0 = \boldsymbol{x}_0$. For the inhomogeneous case see also [20].

In contrast to the solution theory in the linear constant coefficient case, the treatment of fully implicit DAEs without a given internal structure is still challenging, even from today's perspective. For this purpose, Campbell [22] introduced the derivative array as a key concept that carries all the information of the DAE system. The derivative array is constructed from the definition of the differential index, i.e., one considers the equations

$$\begin{aligned} \boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) &= \boldsymbol{0}, \\ \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \frac{\partial}{\partial \dot{\boldsymbol{x}}}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t)\,\boldsymbol{x}^{(2)} + \ldots &= \boldsymbol{0}, \\ &\vdots \\ \frac{\mathrm{d}^k}{\mathrm{d}t^k}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t) = \frac{\partial}{\partial \dot{\boldsymbol{x}}}\boldsymbol{F}(\dot{\boldsymbol{x}}, \boldsymbol{x}, t)\,\boldsymbol{x}^{(k+1)} + \ldots &= \boldsymbol{0} \end{aligned} \tag{3.23}$$

for a DAE of index k. Upon discretization, (2.6) becomes an overdetermined system that can be tackled by least squares techniques. The challenge in this procedure, however, lies in the normally unknown index k and its determination.

Algorithms based on the derivative array are a powerful means for general unstructured DAE systems, and this holds even for the linear constant coefficient case since the computation of the Weierstrass form or the Drazin inverse is very sensitive to small perturbations and thus problematic in finite precision arithmetic. For the derivative array, in contrast, so-called staircase algorithms have been developed that rely on orthogonal matrix multiplications and are much more stable [13].

The Oberwolfach Workshop of 1993 presented various other new developments that would be worthwhile for an exposition. An example is the dummy derivatives technique by Mattson and Söderlind [67] that provides a method to lower the index of an unstructured DAE and that is in use in today's general modeling languages.

3.2.4 The Flying Wheelset

Besides Andrew's squeezer, other benchmark examples in multibody dynamics have been established over the years. A single wheelset running on a straight track is such an example, and for a while it became well known due to the paper [27] by Edda Eich, who had discovered a strong drift-off for the formulation of index 1 and presented this result at the Oberwolfach Workshop of 1993. There is a lesson to be learnt from this example, and I like to tell the following story when teaching to PhD and master students.

In 1994, one year after the *flying wheelset* had taken off, Sebastian Reich contacted me to send him the source code so that he could run some numerical tests with it. At that time, he was working on a class of stabilization methods that are related to Baumgarte's approach, jointly with Uri Ascher, H. Chin and Linda Petzold

[10]. The wheelset seemed a perfect example for these methods, but Sebastian Reich discovered some strange results that contradicted the theory and made him suspicious. So I received an e-mail where he described the results and questioned our Fortran code. This request made Claus Führer and me cross-check the code that we had written 5 years earlier when working on the survey paper [84]. And it turned out that the constraints on the acceleration level had a flaw that had been introduced when merging two blocks of output from a computer algebra program. A simple − sign was false, and after having corrected it, the drift-off was drastically reduced—the wheelset had landed.

In conclusion, the cross-checking of numerical results and the exchange of codes and benchmark problems are absolutely essential for our field in order to reproduce results and eliminate human errors (Fig. 7).

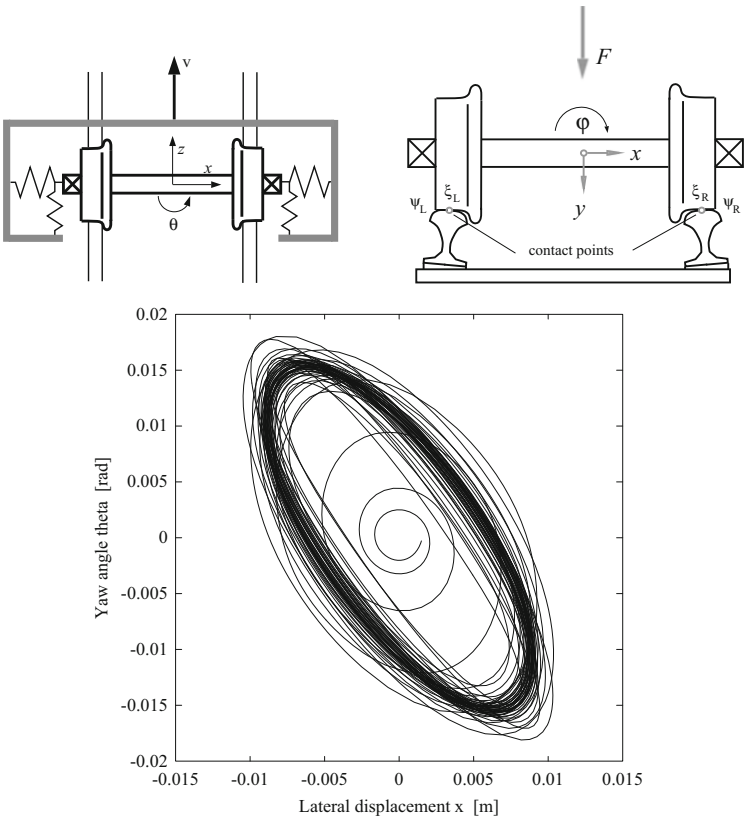


Fig. 7 Wheelset on track (*top*) and phase diagram (*bottom*) of the hunting motion [84]

3.3 The Oberwolfach Workshop in 1996

In some sense, the Oberwolfach Workshop in 1996, organized by Roswitha März and Linda Petzold, marks the end of the "Boom Days". At that time, several groups were heading into new fields such as geometric integration and partial diffferential-algebraic equations. In various respects, a solid body of knowledge had emerged by then, which is reflected by the books of Brenan, Campbell and Petzold [17] and Hairer and Wanner [41] that both appeared in 1996. Though primarily targeting numerical methods, both works have meanwhile become standard references on DAEs in general.

3.3.1 A Famous Inequality and Why One Should Never Trust Authorities

At the Oberwolfach Workshop in 1996, Steve Campbell gave a talk on the relation between the differential and the perturbation index and showed that these notions are not equivalent in general and may even differ substantially [23].

This surprising revelation brings me to another side trip that I love to tell master and PhD students. In 1990, Bill Gear had written a paper [35] in which he addressed the new perturbation index and proved the inequality

$$\mathrm{DI} \leq \mathrm{PI} \leq \mathrm{DI} + 1. \tag{3.24}$$

Here, DI stands for the differential index and PI for the perturbation index. I had the pleasure of attending a summer school on DAEs in Paris in 1992 where Gear talked about DAEs in general and the index notions in particular. The school had been organized by Linda Petzold, and besides her and Gear, Claus Führer and Christian Lubich were also among the speakers.

After Gear's talk, it seemed that everybody in the audience was convinced that (3.24) was right and another milestone in the development of DAEs had been reached. Back home in Munich, I had a master student working on the paper [35] in order to prepare a seminar talk. The student was bright and repeatedly came to my office with questions about the proof of (3.24). In the end, we both were not able to completely follow the lines of reasoning, but I wiped away any doubts by declaring that the great Bill Gear would definitely be right. But he was not.

The counterexample found by Steve Campbell is simple. It reads

$$\begin{pmatrix} 0 & y_3 & 0 \\ 0 & 0 & y_3 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \dot{y}_3 \end{pmatrix} + \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \mathbf{0}. \tag{3.25}$$

The last equation is $y_3 = 0$, which immediately implies $y_1 = 0$ and $y_2 = 0$. Differentiating these equations once yields the underlying ordinary differential equation, and accordingly the differential index equals 1. If the right-hand side, on

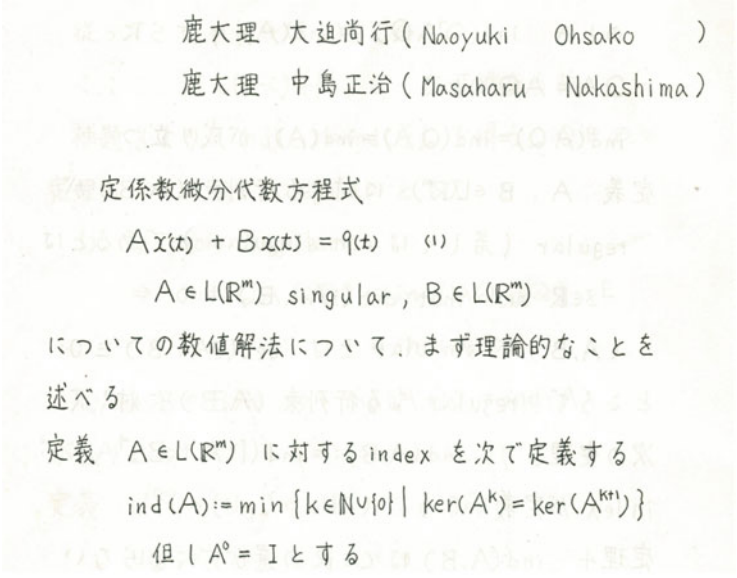
the other hand, is perturbed by $\boldsymbol{\delta} = (\delta_1, \delta_2, \delta_3)^T$, we can compute the perturbed solution in a way similar to the derivation of (2.12), obtaining eventually an expression for $\hat{y}_1$ that involves the second derivative $\delta_3^{(2)}$. The perturbation index is hence 3.

The example (3.25) extends easily to arbitrary dimension n_y. While the perturbation index equals n_y and grows with the dimension, the differential index stays at 1. In case of semi-explicit systems, however, such an inconsistency does not arise, and both indices can be shown to be equivalent.

The bottom line of this story is clear. As we all are human, our results might be wrong and call for validation by colleagues and students. If I had really put the result (3.24) into question, I would have had the chance to work on a counterexample on my own, at a time when I was still a PhD student. But I missed the chance since I had too much trust in authorities.

4 Consolidation

By the mid-1990s, the boom days had slowly turned into a constant and stable flow of ongoing work. Furthermore, DAEs could be found all around the world in different languages and scientific contexts. To illustrate this, Fig. 8 displays the first page of a Japanese text on DAEs by Naoyuki Ohsako and Masaharu Nakashima.


鹿大理 大迫尚行 (Naoyuki Ohsako)
鹿大理 中島正治 (Masaharu Nakashima)

定係数微分代数方程式

$A\dot{x}(t) + Bx(t) = q(t)$ (1)

$A \in L(\mathbb{R}^m)$ singular, $B \in L(\mathbb{R}^m)$

についての数値解法について、まず理論的なことを述べる

定義 $A \in L(\mathbb{R}^m)$ に対する index を次で定義する

$\mathrm{ind}(A) := \min\{k \in \mathbb{N} \cup \{0\} \mid \ker(A^k) = \ker(A^{k+1})\}$

但し $A^0 = I$ とする

Fig. 8 Title page of a Japanese text on linear constant coefficient DAEs by N. Ohsako and M. Nakashima, Dept. of Mathematics, Kagoshima University, Japan

4.1 The NUMDIFF Conference in 1997

The series of NUMDIFF seminars goes back to the 1980s when Karl Strehmel initiated a conference format that brought eastern and western mathematicians together in the venerable city of Halle in East Germany. Concentrating on time-dependent problems and specific integration methods, NUMDIFF filled a gap and soon became a well-established conference that still takes place today.

In the early 1990s, the DAEs were also a prominent topic at the NUMDIFF seminars, but it was only in 1997, when the conference moved to Alexisbad in the Harz Mountains, that NUMDIFF really focused on DAEs and offered the stage for a new and long-lasting development. This conference marks the outset of the topic of Partial Differential Algebraic Equations (PDAEs).

The Halle group was one of the driving forces in this emerging field. An example of a linear PDAE in the unknown $\boldsymbol{u}(x,t) \in \mathbb{R}^{n_x}$ is given by

$$\boldsymbol{E}\boldsymbol{u}_t + \boldsymbol{H}\boldsymbol{u}_{xx} + \boldsymbol{C}\boldsymbol{u} = \boldsymbol{c} \tag{4.1}$$

where at least one of the square matrices $\boldsymbol{E}, \boldsymbol{H} \in \mathbb{R}^{n_x \times n_x}$ is singular, see Lucht et al. [64]. Obviously, (4.1) is a generalization of the linear constant coefficient DAE (2.7) with given right-hand side $\boldsymbol{c} = \boldsymbol{c}(x,t)$, and one is tempted to directly transfer the already available concepts and techniques to this problem field. However, as the theory of partial differential equations is much more heterogeneous than the one for ordinary differential equations, a general methodology for PDAEs is more than a hard task, and it is more rewarding to study special classes.

One specific aspect concerns the influence of the discretization of the spatial variable x and its derivative $\boldsymbol{u}_{xx}$, which leads to a finite-dimensional DAE in time t. This discretization clearly has an influence on the structure and may even affect the index of the resulting system. At the time when the PDAEs began to attract attention, there was much expectation that such fundamental questions could be answered in some generality. Over the years, however, it turned out that it is more advantageous to look into particular application fields. Moreover, the well-established PDE and numerical PDE communities were quite reluctant to accept the viewpoint of differential-algebraic equations and considered it to be a game that is not worth the candle.

4.2 Examples of PDAEs

We need some convincing examples to demonstrate the benefits of a differential-algebraic viewpoint in the PDE context. Two such examples are sketched next.

A classical example of a PDAE is given by the Navier–Stokes equations

$$\dot{\boldsymbol{u}} + (\boldsymbol{u}\cdot\nabla)\boldsymbol{u} + \frac{1}{\rho}\nabla p = \nu\boldsymbol{\Delta}\boldsymbol{u} + \boldsymbol{l}, \tag{4.2a}$$

$$0 = \nabla\cdot\boldsymbol{u} \tag{4.2b}$$

for the velocity field $\boldsymbol{u}(x,t)$ and the pressure $p(x,t)$ in a d-dimensional domain Ω, with mass density ρ, viscosity ν, and source term $\boldsymbol{l}(x,t)$. The second Eq. (4.2b) models the incompressibility of the fluid and defines a constraint for the velocity field. For simplification, the convection term $(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}$ in (4.2a) can be omitted, which makes the overall problem linear and more amenable for the analysis. In an abstract notation, the resulting Stokes problem then reads

$$\dot{\boldsymbol{u}} + \mathcal{A}\boldsymbol{u} + \mathcal{B}'p = \boldsymbol{l}, \tag{4.3a}$$

$$\mathcal{B}\boldsymbol{u} = 0, \tag{4.3b}$$

with differential operators $\mathcal{A}$ and $\mathcal{B}$ expressing the Laplacian and the divergence, respectively. The notation $\mathcal{B}'$ stands for the conjugate operator of $\mathcal{B}$, which here is the gradient.

The discretization, e.g., by a Galerkin-projection

$$\boldsymbol{u}(x,t) \doteq \boldsymbol{N}(x)\boldsymbol{q}(t), \qquad p(x,t) \doteq \boldsymbol{Q}(x)\boldsymbol{\lambda}(t)$$

with ansatz functions $\boldsymbol{N}$ and $\boldsymbol{Q}$ in some finite element spaces, transforms the infinite-dimensional PDAE (4.3) to the DAE

$$\boldsymbol{M}\dot{\boldsymbol{q}} + \boldsymbol{A}\boldsymbol{q} + \boldsymbol{B}^T\boldsymbol{\lambda} = \boldsymbol{l}, \tag{4.4a}$$

$$\boldsymbol{B}\boldsymbol{q} = \boldsymbol{0}. \tag{4.4b}$$

While the mass matrix $\boldsymbol{M}$ and stiffness matrix $\boldsymbol{A}$ are symmetric positive definite and symmetric positive semi-definite, respectively, and easy to handle, the constraint matrix $\boldsymbol{B}$ is generated by mixing the discretizations for the velocity field and the pressure. It is well known in mixed finite elements [18] that a bad choice for the discretization will either result in a rank-deficient matrix $\boldsymbol{B}$ or in a situation where the smallest singular value of $\boldsymbol{B}$ is approaching zero for a decreasing mesh size. This means that the DAE (4.4) may become singular or almost singular due to the spatial discretization. The famous LBB-condition by Ladyshenskaja, Babŭska, and Brezzi [18] gives a means to classify the discretization pairs for $\boldsymbol{u}$ and p. If the matrix $\boldsymbol{B}$ has full rank, the index of the DAE (4.4) is $k = 2$.

To summarize, PDEs with constraints such as the Navier–Stokes equations often feature a rich structure that should be exploited, and building on the available PDE methodology reveals interesting cross-connections with the differential-algebraic viewpoint. In this context, the abstract formulation (4.3) as a *transient saddle point problem* defines a rather broad problem class where many application fields can be subsumed [82].

By combining the state-of-the-art in DAEs with advanced PDE methodology and numerics, powerful algorithms can then be developed that break new ground. Time-space adaptivity for PDAEs is one such topic where many different aspects are put together in order to set up numerical schemes with sophisticated error control. The work by Lang [59] defines a cornerstone in this field.

A time-space adaptive solver for the Navier–Stokes equations (4.2) can be constructed in the following way. Discretization in time by the implicit midpoint rule with stepsize τ yields

$$\frac{\boldsymbol{u}_{i+1} - \boldsymbol{u}_i}{\tau} + (\boldsymbol{u}_{i+\frac{1}{2}} \cdot \nabla)\boldsymbol{u}_{i+\frac{1}{2}} + \frac{1}{\rho}\nabla p_{i+1} = \nu \Delta \boldsymbol{u}_{i+\frac{1}{2}} + \boldsymbol{l}_{i+\frac{1}{2}}, \tag{4.5a}$$

$$\nabla \cdot \boldsymbol{u}_{i+1} = 0 \tag{4.5b}$$

where $\boldsymbol{u}_{i+1/2} = (\boldsymbol{u}_i + \boldsymbol{u}_{i+1})/2$, $\boldsymbol{l}_{i+1/2} = \boldsymbol{l}(x, t_i + \tau/2)$. Note that (4.5b) is evaluated at time t_{i+1}, which is a typical technique in DAE time integration and enforces the constraint at the next time step. The discrete pressure p_{i+1} is interpreted in the same way.

In this form, the system (4.5) represents a sequence of stationary nonlinear PDE problems, which is the backbone of the reverse method of lines. This method has mainly been investigated in the context of parabolic partial differential equations [14, 28] but also plays a role in various other applications [11].

The key idea for adaptivity in time and space is now that, like in ODE and DAE time integration, the basic time stepping scheme (4.5) is combined with a second method to obtain an error estimator in time. The error estimation in space, on the other hand, is performed while solving (4.5) by an adaptive finite element method.

As a computational example, taken from [73], Fig. 9 shows a snapshot of the flow over an obstacle in a pipe at Reynolds number RE = 1000. Here, P_1 finite elements for both velocity field and pressure are employed, stabilized by Streamline Galerkin Least Squares [51]. For the time integration, the implicit midpoint scheme (4.5) is combined with a simple implicit Euler step. At the bottom the current mesh is displayed and on top the vorticity of the corresponding velocity field is shown. The adaptive algorithm captures the solution details by placing additional grid points in areas where the vorticity is high. On the other hand, unnecessary grid points in other areas are automatically removed. In this example, the adaption criterion kept the number of unknowns at around 21,000 per time step.

As a final remark, it should be stressed that the simulation for the results in Fig. 9 requires profound numerical skills from different fields and is not straightforward

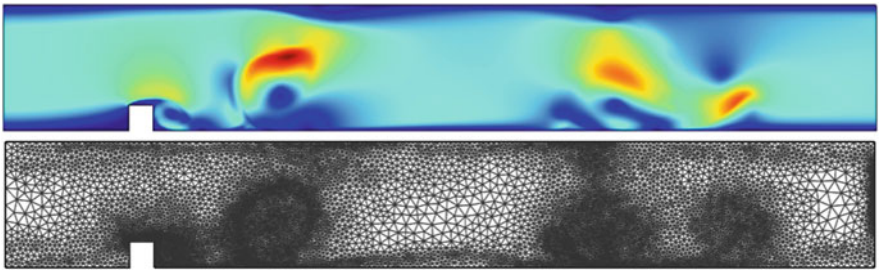


Fig. 9 Time-space adaptive solution of flow over obstacle, RE = 1000

to set up. For more details on time-space adaptivity and the reverse method of lines, the reader is referred to the above references.

4.3 Pantograph and Catenary

While the Navier–Stokes equations are a PDAE system that features an explicit constraint defined over the whole domain, many PDAEs actually arise from coupling subsystems with a different level of mathematical modeling. In computational mechanics, flexible multibody systems are a typical member of this problem class. While the rigid body dynamics results in ODEs and DAEs, the inclusion of elastic or flexible bodies leads to the PDEs of elasto-dynamics where the interaction with the rigid bodies leads to additional coupling equations and constraints.

The system of pantograph and catenary [7, 83] is a nice example for a flexible multibody system and, moreover, illustrates the differential-algebraic methodology for setting up the equations of motion. The following unknowns are used in this simplified model, Fig. 10:

$$
\begin{aligned}
&r_1(t): \text{ vertical motion of body 1 (pantograph head),}\\
&r_2(t): \text{ vertical motion of body 2 (pantograph base),}\\
&w_3(x,t): \text{ vertical displacement of carrier wire,}\\
&w_4(x,t): \text{ vertical displacement of contact wire.}
\end{aligned}
$$

In the first step, we neglect the constraints and consider the equations of unconstrained motion that read

$$m_1\ddot{r}_1 = -d_1(\dot{r}_1 - \dot{r}_2) - c_1(r_1 - r_2)\,, \tag{4.6a}$$

$$m_2\ddot{r}_2 = -d_2\dot{r}_2 + d_1(\dot{r}_1 - \dot{r}_2) - c_2 r_2 + c_1(r_1 - r_2) + F_0\,, \tag{4.6b}$$

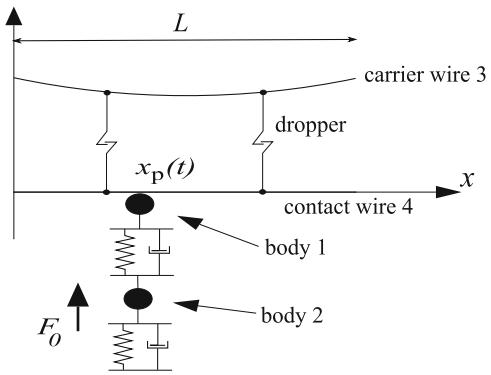


Fig. 10 Pantograph and catenary, simplified benchmark problem of [7]

$$\rho_3 \mathtt{A}_3 \ddot{w}_3 = -\beta_3 \dot{w}_3 + \mathtt{T}_3 w_3'' - \rho_3 \mathtt{A}_3 \gamma \,, \tag{4.6c}$$

$$\rho_4 \mathtt{A}_4 \ddot{w}_4 = -\beta_4 \dot{w}_4 + \mathtt{T}_4 w_4'' - \mathtt{E}_4 \mathtt{I}_4 w_4'''' - \rho_4 \mathtt{A}_4 \gamma \,. \tag{4.6d}$$

Here, the first two equations describe the pantograph motion with damper and spring constants d_1, d_2, c_1, c_2 and a constant force F_0 which includes the influence of gravity. The carrier is expressed by the equation of a vibrating string with tensile force $\mathtt{T}_3$ and viscous damping factor β_3. Finally, the beam equation for the contact wire includes both a pre-stress term due to the tensile force $\mathtt{T}_4$ as well as a bending stiffness term with factor $\mathtt{E}_4\mathtt{I}_4$. The notation of the other parameters is straightforward with $\mathtt{A}$ standing for the cross-section area and γ for the gravity constant.

The next step integrates the coupling conditions where we assume bilateral contact to simplify the discussion. Contact wire and carrier are interconnected by two massless droppers with relative distances l_1 and l_2 and positions $x_{p,1}$ and $x_{p,2}$. The third constraint results from the coupling of contact wire and body 1 in the moving contact point $x_p(t)$. In strong or pointwise form, we require thus

$$w_3(x_{p,1}, t) - w_4(x_{p,1}, t) + l_1 = 0 \,, \tag{4.7a}$$

$$w_3(x_{p,2}, t) - w_4(x_{p,2}, t) + l_2 = 0 \,, \tag{4.7b}$$

$$w_4(x_p(t), t) - r_1(t) = 0 \,. \tag{4.7c}$$

To include these constraints by appropriate Lagrange multipliers in the equations of motion (4.6) is a bit tricky since the constraints are formulated in isolated points of a one-dimensional continuum and result in Dirac δ-distributions in the PDEs (4.6c) and (4.6d). If one passes to a weak formulation where the δ-distribution is multiplied by a test function, however, a well-defined model with a total of 3 discrete Lagrange multipliers is obtained that are associated with the 3 constraints (4.7).

Overall, the resulting model can then be written in a similar fashion as the transient saddle point problem (4.3) where the main difference lies in the second time derivative. More precisely, by introducing suitable operators, the pantograph and catenary model can be written as [82]

$$\ddot{\boldsymbol{u}} + \mathcal{A}\boldsymbol{u} + \mathcal{B}'\boldsymbol{\lambda} = \boldsymbol{l}, \tag{4.8a}$$

$$\mathcal{B}\boldsymbol{u} = \boldsymbol{m}, \tag{4.8b}$$

where $\boldsymbol{u}$ comprises all unknown discrete and continuous displacements and $\boldsymbol{\lambda}$ stands for the Lagrange multipliers. For a related approach in elasto-dynamics see [3].

In electrical circuit simulation, the inclusion of heating effects or semiconductors also results in PDAE models where ODEs, DAEs, and PDEs are coupled via network approaches, see, e.g., [2, 38].

4.4 The Oberwolfach Workshop in 2006

In the spring of 2006, Oberwolfach offered again to be the showcase for the latest developments in DAEs—25 years after the workshop where Bill Gear had first investigated the mathematical pendulum (2.5) in Cartesian coordinates. The organizers were Stephen Campbell, Roswitha März, Linda Petzold, and Peter Rentrop. Among the participants from all over the world was Bill Gear himself, and during the week it became evident that DAEs were now well established in many fields.

In the same year, the book by Kunkel and Mehrmann [56] appeared, which shed new light on topics such as boundary value problems in differential-algebraic equations and the numerical treatment of fully implicit systems (2.1).

Most talks at the meeting addressed the field of PDAEs, but among the other prominent topics were optimization and optimal control problems with constraints described by DAEs , see, e.g., [24, 53], and model order reduction for descriptor systems [76]. It moreover turned out that even those topics which had seemed to be mature and fully understood were still producing surprises. An example of such a surprise is the properly stated leading term in linear time-variant DAEs [65] where (3.9) is replaced by

$$\boldsymbol{E}(t)\frac{\mathrm{d}}{\mathrm{d}t}\big(\boldsymbol{D}(t)\boldsymbol{x}(t)\big) + \boldsymbol{H}(t)\boldsymbol{x}(t) = \boldsymbol{c}(t)\,, \tag{4.9}$$

together with the transversality condition

$$\ker \boldsymbol{E}(t) \oplus \operatorname{im} \boldsymbol{D}(t) = \mathbb{R}^{n_x}.$$

In this way, the matrix $\boldsymbol{D}$ precisely determines the relevant derivatives and adds additional structure to the system, which is beneficial in the analysis.

An emerging topic at the time was *stochastic differential-algebraic equations* or SDAEs for short. Since many models in science and engineering contain uncertain quantities, it is natural to extend the methodology for DAEs by corresponding random terms. This could either be a parameter or coefficient that is only known approximately or even an extra diffusion term in the differential equation that is expressed in terms of a Wiener process. For the constant coefficient system (2.7), such a diffusion term leads to the linear SDAE

$$\boldsymbol{E}\mathrm{d}\boldsymbol{x}(t) + \boldsymbol{H}\boldsymbol{x}(t) = \boldsymbol{c}(t) + \boldsymbol{C}\mathrm{d}\boldsymbol{W}(t) \tag{4.10}$$

with a Wiener process $\boldsymbol{W}$ in $\mathbb{R}^{n_x}$ and a square matrix $\boldsymbol{C}$. For work in this field and applications in electrical circuit analysis we refer to [49, 87].

Looking back, the Oberwolfach Workshop of 2006 showed several lines for future research on DAEs but also left the impression that a process of diversification had started that is still going on today. Figure 11 shows the participants of this memorable conference.

Fig. 11 Participants of the Oberwolfach Workshop 2006, Bildarchiv des Mathematischen Forschungsinstituts Oberwolfach

At this point, the retrospective stops, with various loose ends and with lots of interesting topics left out as completeness has been beyond question from the very beginning of this undertaking. The story of differential-algebraic equations keeps going on, and many new results and entertaining stories are still to be found and told.

Acknowledgements Over the years, I have had the privilege to meet so many colleagues working in the field of differential-algebraic equations. Our discussions and the stories that were told are an integral part of this survey article, and I would like to thank them all for their invisible but highly acknowledged contribution. It was my academic teacher Peter Rentrop who gave me the chance to do a PhD in this exciting field and who made my participation at various conferences and summer schools possible during the *Boom Days*. I am more than grateful for his inspiration and support.

Moreover, I wish to sincerely thank all my master and PhD students who worked in this or related fields for their collaboration, their effort, and their patience. Special thanks, finally, goes to Ernst Hairer who read an early version of this manuscript, and to Achim Ilchmann who always encouraged me to continue with this effort.

References

1. Abraham, R., Marsden, J.E., Ratiu, T.: Manifolds, Tensor Analysis, and Applications. Springer, New York (1988)
2. Ali, G., Bartel, A., Günther, M., Tischendorf, C.: Elliptic partial differential-algebraic multiphysics models in electrical network design. Math. Models Methods Appl. Sci. **13**(09), 1261–1278 (2003)
3. Altmann, R.: Index reduction for operator differential-algebraic equations in elastodynamics. Z. Angew. Math. Mech. **93**(9), 648–664 (2013)
4. Andrews, G.C., Ormrod, M.K.: Advent: a simulation program for constrained planar kinematic and dynamic systems. In: Presented at the Design Engineering Technical Conference, Columbus, Ohio, 5–8 October 1986. Departement of Mechanical Engineering, University of Waterloo, Ontario, Canada, N2L 3G1 (1986)
5. Arnold, V.I.: Ordinary Differential Equations. MIT Press, Cambridge (1981)
6. Arnold, M., Murua, A.: Non-stiff integrators for differential–algebraic systems of index 2. Numer. Algorithms **19**(1–4), 25–41 (1998)
7. Arnold, M., Simeon, B.: Pantograph and catenary dynamics: a benchmark problem and its numerical solution. Appl. Numer. Math. **34**, 345–362 (2000)
8. Ascher, U., Lin, P.: Sequential regularization methods for nonlinear higher index DAEs. SIAM J. Sci. Comput. **18**, 160–181 (1997)
9. Ascher, U.M., Petzold, L.R.: Projected implicit Runge-Kutta methods for differential-algebraic equations. SIAM J. Numer. Anal. **28**, 1097–1120 (1991)
10. Ascher, U., Chin, H., Petzold, L., Reich, S.: Stabilization of constrained mechanical systems with DAEs and invariant manifolds. J. Mech. Struct. Mach. 23: 135–158 (1995)
11. Bangerth, W., Rannacher, R.: Adaptive Finite Element Methods for Differential Equations. Birkhäuser, Basel (2013)
12. Baumgarte, J.: Stabilization of constraints and integrals of motion in dynamical systems. Comput. Methods Appl. Mech. **1**, 1–16 (1972)
13. Benner, P., Losse, P., Mehrmann, V., Voigt, M.: Numerical linear algebra methods for linear differential-algebraic equations. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations III. DAE-Forum, pp. 117–175. Springer, Cham (2015)
14. Bornemann, F.A.: An adaptive multilevel approach to parabolic equations: II. Variable-order time discretization based on a multiplicative error correction. IMPACT Comput. Sci. Eng. **3**(2), 93–122 (1991)
15. Brasey, V.: A half-explicit method of order 5 for solving constrained mechanical systems. Computing **48**, 191–201 (1992)
16. Brasey, V., Hairer, E.: Half-explicit Runge–Kutta methods for differential-algebraic systems of index 2. SIAM J. Numer. Anal. **30**, 538–552 (1993)
17. Brenan, K.E., Campbell, S.L., Petzold, L.R.: The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations. SIAM, Philadelphia (1996)
18. Brezzi, F., Fortin, M.: Mixed and Hybrid Finite Element Methods. Springer, New York (1991)
19. Brizard, A.: An Introduction to Lagrangian Mechanics. World Scientific, Singapore (2008)
20. Campbell, S.L.: Singular Systems of Differential Equations. Pitman, London (1980)
21. Campbell, S.L.: Singular Systems of Differential Equations II. Research Notes in Mathematics, vol. 61. Pitman, London (1982)
22. Campbell, S.L.: Least squares completions for nonlinear differential-algebraic equations. Numer. Math. **65**, 77–94 (1993)
23. Campbell, S., Gear, C.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)
24. Callies, R., Rentrop, P.: Optimal control of rigid-link manipulators by indirect methods. GAMM-Mitteilungen **31**(1), 27–58 (2008)
25. Deuflhard, P., Hairer, E., Zugck, J.: One-step and extrapolation methods for differential-algebraic systems. Numer. Math. **51**(5), 501–516 (1987)

26. Drazin, M.: Pseudo inverses in associative rays and semigroups. Am. Math. Mon. **65**, 506–514 (1958)
27. Eich, E.: Convergence results for a coordinate projection method applied to constrained mechanical systems. SIAM J. Numer. Anal. **30**(5), 1467–1482 (1993)
28. Franzone, P.C., Deuflhard, P., Erdmann, B., Lang, J., Pavarino, L.F.: Adaptivity in space and time for reaction-diffusion systems in electrocardiology. SIAM J. Sci. Comput. **28**(3), 942–962 (2006)
29. Führer, C., Leimkuhler, B.: Numerical solution of differential-algebraic equations for constrained mechanical motion. Numer. Math. **59**, 55–69 (1991)
30. Gantmacher, F.: Matrizenrechnung, Teil 2. VEB Deutscher Verlag der Wissenschaften, Berlin (1959)
31. Garcia de Jalón, J., Bayo, E.: Kinematic and Dynamic Simulation of Multibody Systems. Springer, New York (1994)
32. Gear, C.W.: Numerical Initial Value Problems in Ordinary Differential Equations. Prentice-Hall, Upper Saddle River (1971)
33. Gear, C.W.: Simultaneous numerical solution of differential-algebraic equations. IEEE Trans. Circuit Theory **CT-18**(1), 89–95 (1971)
34. Gear, C.W.: Differential-algebraic equation index transformation. SIAM J. Sci. Stat. Comput. **9**, 39–47 (1988)
35. Gear, C.W.: Differential-algebraic equations, indices, and integral algebraic equations. SIAM J. Numer. Anal. **27**, 1527–1534 (1990)
36. Gear, C.W., Gupta, G., Leimkuhler, B.: Automatic integration of the Euler-Lagrange equations with constraints. J. Comput. Appl. Math. **12** & **13**, 77–90 (1985)
37. Griepentrog, E., März, R.: Differential-Algebraic Equations and Their Numerical Treatment. Teubner-Texte zur Mathematik, vol. 88. Teubner Verlagsgesellschaft, Leipzig (1986)
38. Günther, M.: Partielle differential-algebraische Systeme in der numerischen Zeitbereichsanalyse elektrischer Schaltungen. VDI-Verlag, Reihe 20, Düsseldorf (2001)
39. Günther, M., Feldmann, U.: CAD based electric circuit modeling in industry I: mathematical structure and index of network equations. Surv. Math. Ind. **8**, 97–129 (1999)
40. Günther, M., Hoschek, M., Rentrop, P.: Differential-algebraic equations in electric circuit simulation. Int. J. Electron. Commun. **54**, 101–107 (2000)
41. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer, Berlin (1996)
42. Hairer, E., Wanner, G.: Stiff differential equations solved by Radau methods. J. Comput. Appl. Math. **111**, 93–111 (1999)
43. Hairer, E., Lubich, C., Roche, M.: Error of Runge–Kutta methods for stiff problems studied via differential algebraic equations. BIT Numer. Math. **28**(3), 678–700 (1988)
44. Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods. Lecture Notes in Mathematics, vol. 1409. Springer, Heidelberg (1989)
45. Hairer, E., Nørsett, S., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, Berlin (1993)
46. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Springer, Berlin (2002)
47. Hanke, M.: On the regularization of index 2 differential-algebraic equations. J. Math. Anal. Appl. **151**(1), 236–253 (1990)
48. Haug, E.: Computer-Aided Kinematics and Dynamics of Mechanical Systems. Allyn and Bacon, Boston (1989)
49. Higham, D.J., Mao, X., Stuart, A.M.: Strong convergence of Euler-type methods for nonlinear stochastic differential equations. SIAM J. Numer. Anal. **40**(3), 1041–1063 (2002)
50. 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**(3), 363–396 (2005)
51. Hughes, T.J., Franca, L.P., Hulbert, G.M.: A new finite element formulation for computational fluid dynamics: VIII. The Galerkin/least-squares method for advective-diffusive equations. Comput. Methods Appl. Mech. Eng. **73**(2), 173–189 (1989)

52. Kirchhoff, G.: Ueber die Auflösung der Gleichungen, auf welche man bei der Untersuchung der linearen Vertheilung galvanischer Ströme geführt wird. Ann. Phys. **148**(12), 497–508 (1847)
53. Körkel, S., Kostina, E., Bock, H.G., Schlöder, J.P.: Numerical methods for optimal control problems in design of robust optimal experiments for nonlinear dynamic processes. Optim. Methods Softw. **19**(3–4), 327–338 (2004)
54. Kronecker, L.: Algebraische Reduktion der Schaaren bilinearer Formen. Akademie der Wissenschaften Berlin **III**, 141–155 (1890)
55. Kunkel, P., Mehrmann, V.: Numerical solution of differential algebraic Riccati equations. Linear Algebra Appl. **137**, 39–66 (1990)
56. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations – Analysis and Numerical Solution. EMS Publishing House, Zürich (2006)
57. Lagrange, J.L.: Méchanique analytique. Libraire chez la Veuve Desaint, Paris (1788)
58. Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Differential-Algebraic Equations Forum. Springer, Berlin (2013)
59. Lang, J.: Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems: Theory, Algorithm, and Applications, vol. 16. Springer, Berlin (2013)
60. Lötstedt, P., Petzold, L.: Numerical solution of nonlinear differential equations with algebraic constraints i: convergence results for BDF. Math.Comput. **46**, 491–516 (1986)
61. Lubich, C.: h^2 extrapolation methods for differential-algebraic equations of index-2. Impact Comput. Sci. Eng. **1**, 260–268 (1989)
62. Lubich, C.: Integration of stiff mechanical systems by Runge-Kutta methods. ZAMP **44**, 1022–1053 (1993)
63. Lubich, C., Engstler, C., Nowak, U., Pöhle, U.: Numerical integration of constrained mechanical systems using MEXX. Mech. Struct. Mach. **23**, 473–495 (1995)
64. Lucht, W., Strehmel, K., Eichler-Liebenow, C.: Indexes and special discretization methods for linear partial differential algebraic equations. BIT Numer. Math. **39**(3), 484–512 (1999)
65. März, R.: Differential algebraic systems anew. Appl. Numer. Math. **42**(1), 315–335 (2002)
66. März, R., Tischendorf, C.: Recent results in solving index-2 differential-algebraic equations in circuit simulation. SIAM J. Sci. Comput. **18**, 139–159 (1997)
67. Mattson, S., Söderlind, G.: Index reduction in differential-algebraic equations using dummy derivatives. SIAM J. Sci. Comput. **14**(3), 677–692 (1993)
68. Müller, P.C.: Stability of linear mechanical systems with holonomic constraints. Appl. Mech. Rev. **46**(11S), S160–S164 (1993)
69. Müller, P.C.: Stability and optimal control of nonlinear descriptor systems: a survey. Appl. Math. Comput. Sci. **8**, 269–286 (1998)
70. Nagel, L.W., Pederson, D.: Spice (simulation program with integrated circuit emphasis). Technical Report UCB/ERL M382, EECS Department, University of California, Berkeley (1973)
71. O'Malley, R.E.: Introduction to Singular Perturbations. Academic, New York (1974)
72. Petzold, L.: A description of DASSL: a differential/algebraic system solver. In: Proceedings of 10th IMACS World Congress, Montreal, 8–13 August 1982
73. Plinninger, T., Simeon, B.: Adaptivity in space and time for solving transient problems in COMSOL. In: Proceedings COMSOL Conference Hannover (2008)
74. Rabier, P., Rheinboldt, W.: Theoretical and numerical analysis of differential-algebraic equations. In: Ciarlet, P., Lions, J. (eds.) Handbook of Numerical Analysis, vol. VIII. Elsevier, Amsterdam (2002)
75. Reich, S.: On a geometric interpretation of DAEs. Circ. Syst. Signal Process. **9**, 367–382 (1990)
76. Reis, T., Stykel, T.: Stability analysis and model order reduction of coupled systems. Math. Comput. Model. Dyn. Syst. **13**(5), 413–436 (2007)
77. Rentrop, P., Roche, M., Steinebach, G.: The application of Rosenbrock–Wanner type methods with stepsize control in differential-algebraic equations. Numer. Math. **55**, 545–563 (1989)
78. Rheinboldt, W.: Differential - algebraic systems as differential equations on manifolds. Math. Comput. **43**(168), 2473–482 (1984)

79. Rheinboldt, W.: Manpak: a set of algorithms for computations on implicitly defined manifolds. Comput. Math. Appl. **32**, 15–28 (1996)
80. Schiehlen, W. (ed.): Multibody System Handbook. Springer, Heidelberg (1990)
81. Schwerin, R.: Multibody System Simulation. Springer, Berlin (1999)
82. Simeon, B.: Computational Flexible Multibody Dynamics: A Differential-Algebraic Approach. Springer, Heidelberg (2013)
83. Simeon, B., Arnold, M.: Coupling DAE's and PDE's for simulating the interaction of pantograph and catenary. Math. Comput. Model. Syst. **6**, 129–144 (2000)
84. Simeon, B., Führer, C., Rentrop, P.: Differential-algebraic equations in vehicle system dynamics. Surv. Math. Ind. **1**, 1–37 (1991)
85. Simeon, B., Führer, C., Rentrop, P.: The Drazin inverse in multibody system dynamics. Numer. Math. **64**, 521–539 (1993)
86. Weierstrass, K.: Zur Theorie der bilinearen und quadratischen Formen, pp. 310–338. Monatsber. Akad. Wiss., Berlin (1868)
87. Winkler, R.: Stochastic differential algebraic equations of index 1 and applications in circuit simulation. J. Comput. Appl. Math. **157**(2), 477–505 (2003)

DAE Aspects of Multibody System Dynamics

Martin Arnold

Abstract The dynamical simulation of mechanical multibody systems has stimulated the development of theory and numerical methods for higher index differential-algebraic equations (DAEs) for more than three decades. The equations of motion are linearly implicit second order differential equations. For constrained systems, they form an index-3 DAE with a specific structure that is exploited in theoretical investigations as well as in the numerical solution. In the present survey paper, we give an introduction to this field of research with focus on classical and more recent solution techniques for the time integration of constrained mechanical systems in multibody system dynamics. Part of the material is devoted to topics of current research like multibody system models with nonlinear configuration spaces or systems with redundant constraints.

Keywords Constrained mechanical systems • DAE time integration • Multibody formalisms • Rank-deficient mass matrix • Redundant constraints • Stabilized index-2 formulation

Mathematics Subject Classification (2010): 34A09, 34A12, 65F50, 65L05, 65L80, 70E55

1 Introduction

Multibody system dynamics is a branch of technical mechanics that considers the dynamical interaction of rigid and flexible bodies in complex engineering systems [75]. Multibody system models are frequently used in such diverse fields of application like robotics, vehicle system dynamics, biomechanics, aerospace engineering and wind turbine design. They are composed of a finite number of rigid

M. Arnold (✉)
Institute of Mathematics, Martin Luther University Halle-Wittenberg, 06099 Halle (Saale), Germany
e-mail: martin.arnold@mathematik.uni-halle.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7_2

or flexible bodies and their connecting elements that are assumed to be massless [75, 76].

In engineering, the modelling of mechanical multibody systems follows a generic network approach [52] with basic elements like rigid bodies, flexible bodies, force elements and joints being available in model libraries. The interaction of these basic elements is described by equations of motion resulting from the principles of classical mechanics [74]. The separate modelling of system components in this network approach is attractive from the viewpoint of model setup but results systematically in a redundant system description [52]. Constraints have to be added to guarantee a consistent state of the overall multibody system model.

Formally, these constrained systems could always be transformed to an analytically equivalent ordinary differential equation (ODE) introducing appropriate generalized coordinates [1]. The progress in analysis and numerical solution of differential-algebraic equations (DAEs) allows, however, to solve the constrained systems directly in terms of the original redundant coordinates which proves to be much more efficient than (semi-)analytical solution techniques being based on a minimum set of independent coordinates. A short historical review of these developments has recently been published in [83, Sect. 2.4].

Constrained multibody system models are challenging from the viewpoint of DAE theory since their index is three and index reduction techniques are mandatory for a numerically stable time integration by error controlled variable step size solvers. These index reduction techniques rely on time derivatives of the constrained equations that have a direct physical interpretation as hidden constraints at the level of velocity or acceleration coordinates [40]. Classical approaches like Baumgarte stabilization [21] or the stabilized index-2 formulation of the equations of motion in the sense of Gear, Gupta and Leimkuhler [44] have been developed a long time before the "boom days" of DAE theory [83] that started in the late 1980s.

There is a rich literature on numerical methods in multibody dynamics [37, 88], in particular on time integration methods for constrained systems. The comprehensive survey in [50, Chap. VII] is an excellent reference in this field.

Multibody system dynamics is, however, much more than just the simulation of constrained N-body systems, see Fig. 1. In engineering, the methods and software tools of multibody system dynamics are used as integration platform for multidisciplinary simulation in nonlinear system dynamics [11]. The analysis of flexible bodies provides a close link to structural mechanics. Specific aspects of such flexible multibody systems have been discussed recently from a mathematical viewpoint [82] and from the viewpoint of engineering [20]. The monograph of Géradin and Cardona [45] was an early attempt to bridge the gap between both disciplines.

The present paper considers some DAE aspects of multibody numerics being relevant to applications in engineering. It starts in Sect. 2 with an introduction to constrained systems studying systematically conditions for the existence and uniqueness of solutions for a large problem class of practical interest including systems with rank-deficient mass matrix and redundant constraints.

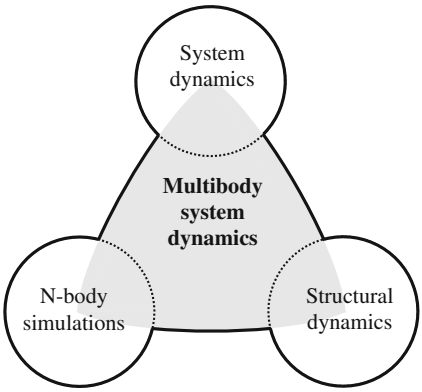


Fig. 1 Multibody system dynamics and related fields of dynamical analysis

In Sect. 3, we consider systems with nonlinear configuration spaces representing the orientation of (rigid or flexible) bodies in space. The resulting model equations are substantially more complex than the ones that are typically discussed in the mathematical literature on multibody numerics. This section ends with a compact introduction to multibody formalisms that exploit the model topology for an efficient evaluation of the equations of motion in large scale engineering applications.

Section 4 provides a consistent introduction to DAE time integration methods in multibody dynamics that covers ODE based solution techniques like Runge–Kutta or linear multi-step methods [50] as well as Newmark type integrators from structural dynamics [45]. There is a special focus on the stabilized index-2 formulation of the equations of motion that may be considered as a quasi-standard in industrial multibody system simulation [14].

2 Constrained Mechanical Systems

In Lagrangian mechanics, the motion of a conservative mechanical system is characterized by a variational principle that takes into account the potential energy $U(\boldsymbol{q})$ and the kinetic energy

$$T(\boldsymbol{q}, \dot{\boldsymbol{q}}) := \frac{1}{2} \dot{\boldsymbol{q}}^\top \mathbf{M}(\boldsymbol{q}) \dot{\boldsymbol{q}} \,.$$

The potential energy results in potential forces $-\nabla U(\boldsymbol{q})$. It is formulated in terms of *position coordinates* $\boldsymbol{q}(t) \in \mathbb{R}^{n_q}$ that describe the configuration of the system and define *velocity coordinates* $\dot{\boldsymbol{q}}(t) := (\mathrm{d}\boldsymbol{q}/\mathrm{d}t)(t)$. Mass and inertia terms are summarized in the symmetric, positive semi-definite *mass matrix* $\mathbf{M}(\boldsymbol{q}) \in \mathbb{R}^{n_q \times n_q}$.

In the present section, we consider constrained systems and derive in Sect. 2.1 their equations of motion. These are classical results that may be found in any textbook on mechanics, e.g., [1]. Sufficient conditions for the unique solvability of

initial value problems are discussed in Sect. 2.2, see also [50, Sect. VII.1]. A more refined analysis is necessary for systems with rank-deficient mass matrix or rank-deficient constraint matrix that have recently found new interest in the literature [42] and will be studied in Sect. 2.3.

2.1 Equations of Motion

The motion of a mechanical system may be subject to constraints in form of equations (*bilateral* constraints) or inequalities (*unilateral* constraints). In the present section, we consider *holonomic* constraints

$$\boldsymbol{g}(t, \boldsymbol{q}(t)) = \boldsymbol{0}\,, \quad (t \in [t_0, t_{\text{end}}]) \tag{2.1}$$

that have to be satisfied in the whole time interval of interest. For more general types of constraints, we refer to Sect. 3.2 below.

To derive the equations of motion from a variational principle, we summarize kinetic and potential energy in the Lagrangian

$$L(\boldsymbol{q}, \dot{\boldsymbol{q}}) := T(\boldsymbol{q}, \dot{\boldsymbol{q}}) - U(\boldsymbol{q})\,.$$

In the constrained case, we introduce *Lagrange multipliers* $\boldsymbol{\lambda}(t) \in \mathbb{R}^{n_\lambda}$ to couple $n_\lambda \leq n_q$ holonomic constraints (2.1) to $L(\boldsymbol{q}, \dot{\boldsymbol{q}})$ and consider the augmented action integral

$$\int_{t_0}^{t_{\text{end}}} \Big(L(\boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)) - \big(\boldsymbol{g}(t, \boldsymbol{q}(t))\big)^{\top} \boldsymbol{\lambda}(t) \Big)\, \mathrm{d}t\,.$$

According to Hamilton's principle of least action, the extremals of this functional coincide with the motion of the mechanical system. The Euler equations for this variational problem are given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\Big(\frac{\partial L}{\partial \dot{q}_k}(\boldsymbol{q}, \dot{\boldsymbol{q}})\Big) - \frac{\partial L}{\partial q_k}(\boldsymbol{q}, \dot{\boldsymbol{q}}) + \Big(\frac{\partial \boldsymbol{g}}{\partial q_k}(\boldsymbol{q})\Big)^{\top} \boldsymbol{\lambda} = \boldsymbol{0}\,, \quad (k = 1, \ldots, n_q) \tag{2.2}$$

with $\boldsymbol{g}(t, \boldsymbol{q}(t)) = \boldsymbol{0}$, see (2.1). In vector form, they may be summarized to

$$\mathbf{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} = \boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}) - \mathbf{G}^{\top}(t, \boldsymbol{q})\, \boldsymbol{\lambda}\,, \tag{2.3a}$$

$$\boldsymbol{0} = \boldsymbol{g}(t, \boldsymbol{q}) \tag{2.3b}$$

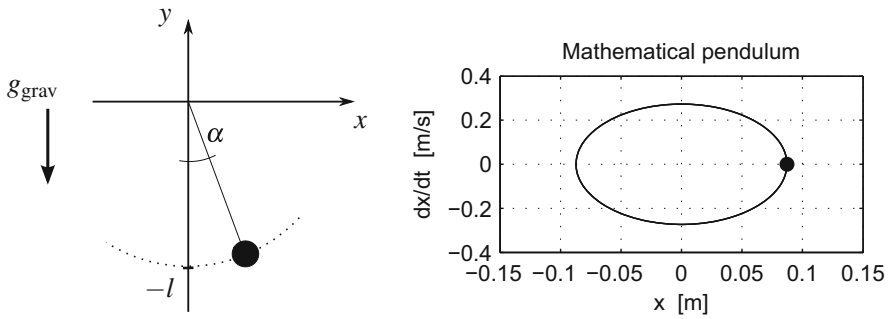


Fig. 2 Configuration and phase plot of the mathematical pendulum, Cartesian coordinates

with the *constraint matrix* $\mathbf{G}(t,\boldsymbol{q}) := (\partial \boldsymbol{g}/\partial \boldsymbol{q})(t,\boldsymbol{q}) \in \mathbb{R}^{n_\lambda \times n_q}$ and the *force vector*

$$\boldsymbol{f}(\boldsymbol{q},\dot{\boldsymbol{q}}) := -\nabla_{\boldsymbol{q}}\, U(\boldsymbol{q}) + \nabla_{\boldsymbol{q}}\, T(\boldsymbol{q},\dot{\boldsymbol{q}}) - \Big(\frac{\partial}{\partial \dot{\boldsymbol{q}}}\big(\nabla_{\boldsymbol{q}}\, T(\boldsymbol{q},\dot{\boldsymbol{q}})\big)\Big)^{\top}\dot{\boldsymbol{q}}\,. \tag{2.4}$$

For systems with constant mass matrix $\mathbf{M}$, we just have $\boldsymbol{f}(\boldsymbol{q},\dot{\boldsymbol{q}}) = -\nabla_{\boldsymbol{q}}\, U(\boldsymbol{q})$ since $\nabla_{\boldsymbol{q}}\, T(\boldsymbol{q},\dot{\boldsymbol{q}}) \equiv \mathbf{0}$.

Example 2.1 The mathematical pendulum is a rather simple model problem that has been used already in the 1980s to study constrained mechanical systems from the viewpoint of DAE theory [40, 47]. It consists of a point mass $m > 0$ that moves under the influence of gravity and is forced by a massless rod of length $l > 0$ to keep a fixed distance to the origin, see Fig. 2.

The pendulum has one degree of freedom that is given by the angle α between rod and y-axis with $\alpha^* = 0$ denoting the equilibrium position, see Fig. 2. Taking into account that $x = l\sin\alpha$, $y = -l\cos\alpha$ implies $\dot{x} = -l\dot{\alpha}\cos\alpha$ and $\dot{y} = -l\dot{\alpha}\sin\alpha$, we may express the kinetic energy $T = m(\dot{x}^2 + \dot{y}^2)/2$ and the potential energy $U = mg_{\text{grav}}\, y$ in terms of α and $\dot{\alpha}$:

$$T(\alpha,\dot{\alpha}) = \frac{ml^2}{2}\dot{\alpha}^2\,, \quad U(\alpha) = -mg_{\text{grav}}\, l\cos\alpha$$

with g_{grav} denoting the gravitational acceleration constant. The equations of motion (2.3) are given by the second order ordinary differential equation (ODE)

$$ml^2\ddot{\alpha} = -mg_{\text{grav}}\, l\sin\alpha \quad \Rightarrow \quad \ddot{\alpha} = -\frac{g_{\text{grav}}}{l}\sin\alpha \tag{2.5}$$

since the position coordinates $\boldsymbol{q} = \alpha \in \mathbb{R}$ are not subject to constraints. All solutions of (2.5) are periodic. As a typical example, we show in Fig. 2 the phase plot $(x,\dot{x})$ for initial values $\alpha_0 = 5°$, $\dot{\alpha}_0 = 0\,\text{rad/s}$ that are marked in the diagram by the dot

at $x_0 = l\sin(5\pi/180)$, $\dot{x}_0 = 0\,\mathrm{m/s}$. The physical model parameters are $m = 1.0\,\mathrm{kg}$, $l = 1.0\,\mathrm{m}$ and $g_{\mathrm{grav}} = 9.81\,\mathrm{m/s^2}$.

An analytically equivalent description of the mathematical pendulum is given by the Cartesian coordinates $\boldsymbol{q} = (x, y)^\top \in \mathbb{R}^2$ that are redundant and have to satisfy $x^2 + y^2 = l^2$ (Pythagorean theorem). Scaling this holonomic constraint by a factor of $1/2$, we get the equations of motion

$$m\ddot{x} = \qquad\quad - x\lambda\,, \tag{2.6a}$$

$$m\ddot{y} = -mg_{\mathrm{grav}} - y\lambda\,, \tag{2.6b}$$

$$0 = \frac{1}{2}(x^2 + y^2 - l^2)\,, \tag{2.6c}$$

see (2.3). The mass matrix $\mathbf{M} = m\mathbf{I}_2$ is a constant multiple of the identity matrix $\mathbf{I}_2 \in \mathbb{R}^{2\times 2}$. Force vector and constraint matrix are given by $\boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}) = (\,0,\, -mg_{\mathrm{grav}}\,)^\top$ and $\mathbf{G}(\boldsymbol{q}) = (\,x,\, y\,) \in \mathbb{R}^{1\times 2}$.

Example 2.1 illustrates that one and the same mechanical system may be represented by different sets of coordinates resulting in unconstrained systems like (2.5) or constrained systems like (2.6). Obviously, the mathematical structure of the constrained equations (2.6) is more complex. On the other hand, the Cartesian coordinate approach is more flexible in the modelling of more complex systems as can be seen already from the model of a chain of $N \geq 2$ mathematical pendulums:

Example 2.2 Consider a chain of $N \geq 2$ point masses m being connected by massless rods of length l and attach the first point mass by another massless rod of length l to the origin $(x_0, y_0) = (0, 0)$. This chain of mathematical pendulums moves under the influence of gravity.

In the special case $N = 2$ we obtain the double pendulum that is depicted by the left plot of Fig. 3. Phase plots $(x_1, \dot{x}_1)$ and $(x_2, \dot{x}_2)$ illustrate the complex dynamical behaviour that is known to be chaotic. We started with zero initial velocities $\dot{\boldsymbol{q}}_0 = \mathbf{0}$ and an initial position $\boldsymbol{q}_0 = (x_1(t_0), y_1(t_0), x_2(t_0), y_2(t_0))^\top$ that is defined by initial values for the angles α_i between rod "i" and the y-axis, $(\,i = 1, 2\,)$, see Fig. 3. The physical parameter values are the same as in Example 2.1 and the initial values are set to $\alpha_1(t_0) = 5°$, $\alpha_2(t_0) = 0°$.

To set up the equations of motion in the general case, we consider $N \geq 2$ point masses with Cartesian coordinates $\boldsymbol{q}_i = (x_i, y_i)^\top$, $(\,i = 1, \ldots, N\,)$, and obtain a constrained system in $n_{\boldsymbol{q}} = 2N$ position coordinates $\boldsymbol{q} = (\boldsymbol{q}_1^\top, \ldots, \boldsymbol{q}_N^\top)^\top$ that are subject to $n_{\boldsymbol{\lambda}} = N$ constraints $(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 = l^2$, $(\,i = 1, \ldots, N\,)$. Following step-by-step the analysis in Example 2.1, we get the kinetic energy $T(\boldsymbol{q}, \dot{\boldsymbol{q}}) = \sum_i m(\dot{x}_i^2 + \dot{y}_i^2)/2$, the potential energy $U(\boldsymbol{q}) = \sum_i mg_{\mathrm{grav}}\, y_i$ and the

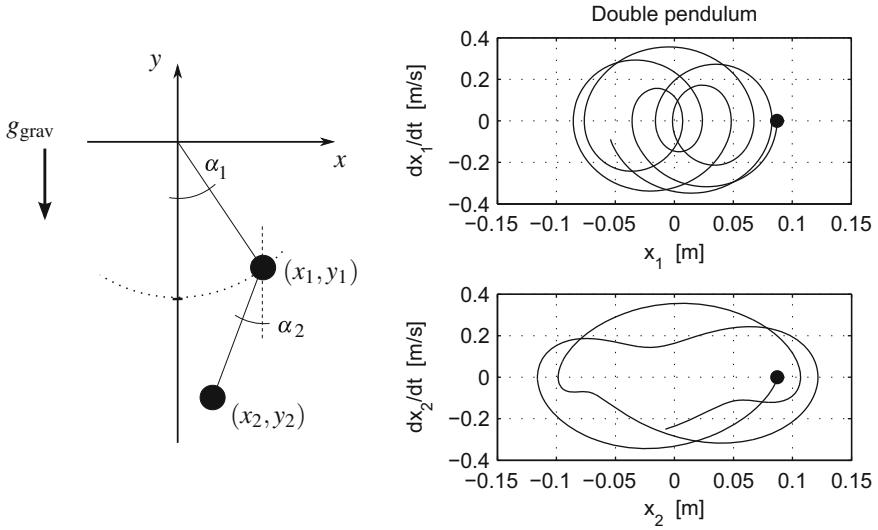


Fig. 3 Configuration and phase plots of a double pendulum, Cartesian coordinates

equations of motion

$$m\ddot{x}_i = -(x_i - x_{i-1})\lambda_i + (x_{i+1} - x_i)\lambda_{i+1}\,,\ (i = 1, \ldots, N-1)\,, \tag{2.7a}$$

$$m\ddot{x}_N = -(x_N - x_{N-1})\lambda_N\,, \tag{2.7b}$$

$$m\ddot{y}_i = -mg_{\text{grav}} - (y_i - y_{i-1})\lambda_i + (y_{i+1} - y_i)\lambda_{i+1}\,,\ (i = 1, \ldots, N-1)\,, \tag{2.7c}$$

$$m\ddot{y}_N = -mg_{\text{grav}} - (y_N - y_{N-1})\lambda_N\,, \tag{2.7d}$$

$$0 = \frac{1}{2}\left((x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 - l^2\right)\,,\ (i = 1, \ldots, N) \tag{2.7e}$$

with Lagrange multipliers $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)^\top \in \mathbb{R}^N$. Comparing (2.7) with the equations of motion in compact form (2.3), we see that the constraints (2.7e) define a vector valued function $\boldsymbol{g} = (g_1, \ldots, g_N)^\top$ in (2.3b) that yields a sparse constraint matrix $\mathbf{G}(\boldsymbol{q}) = \left(G_{ij}(\boldsymbol{q})\right)_{i,j} \in \mathbb{R}^{N\times 2N}$ with non-zero elements

$$G_{i,2i-1}(\boldsymbol{q}) = x_i - x_{i-1}\,, \qquad G_{i,2i}(\boldsymbol{q}) = y_i - y_{i-1}\,, \qquad (i = 1, \ldots, N)\,,$$
$$G_{i,2i+1}(\boldsymbol{q}) = -(x_{i+1} - x_i)\,, \quad G_{i,2i+2}(\boldsymbol{q}) = -(y_{i+1} - y_i)\,, \quad (i = 1, \ldots, N-1)\,.$$

The mass matrix $\mathbf{M}(\boldsymbol{q})$ and the force vector $\boldsymbol{f}(\boldsymbol{q},\dot{\boldsymbol{q}})$ in the dynamical equations (2.3a) are given by $\mathbf{M} = \text{blockdiag}\,(\mathbf{M}_1, \ldots, \mathbf{M}_N), \boldsymbol{f} = (\boldsymbol{f}_1^\top, \ldots, \boldsymbol{f}_N^\top)^\top$ with $\mathbf{M}_i = m\mathbf{I}_2$ and $\boldsymbol{f}_i(\boldsymbol{q},\dot{\boldsymbol{q}}) = (0, -mg_{\text{grav}})^\top, (i = 1,\ldots,N)$.

Cartesian coordinates are favourable to derive the equations of motion (2.7) since kinetic energy and potential energy are given in terms of $(x_i, y_i, \dot{x}_i, \dot{y}_i)$, $(i = 1,\ldots,N)$. Mass matrix and constraint matrix are sparse. Furthermore, the mass matrix $\mathbf{M}$ is constant and block-diagonal. The sparsity pattern of the constraint matrix $\mathbf{G}(\boldsymbol{q})$ corresponds to the coordinates of direct neighbours in the chain.

Example 2.2 illustrates that redundant position coordinates $\boldsymbol{q}$ may help to speed up the modelling process of complex systems. In principle such redundant coordinates $\boldsymbol{q}$ and the corresponding constraints $\boldsymbol{g}(t,\boldsymbol{q}) = \mathbf{0}$ in (2.3) could be avoided choosing appropriate generalized coordinates. For larger systems, the use of such generalized coordinates is, however, often technically much more complicated than for simple model problems like the mathematical pendulum with equations of motion (2.5). As a typical example, we consider the double pendulum with the configuration being depicted in Fig. 3.

Example 2.3 Let α_i $(i = 1, 2)$ denote the angle between rod "i" and the y-axis and use position coordinates $\boldsymbol{q} = (\alpha_1, \alpha_2)^\top \in \mathbb{R}^2$. We get

$$x_i = x_{i-1} + l\sin\alpha_i\,, \quad y_i = y_{i-1} - l\cos\alpha_i\,, \quad (i = 1, 2)\,.$$

with $(x_0, y_0) = (0, 0)$ and may express the kinetic and potential energy in terms of $\boldsymbol{q}$, $\dot{\boldsymbol{q}}$ using $\dot{x}_1 = l\dot{\alpha}_1\cos\alpha_1, \dot{x}_2 = \sum_i l\dot{\alpha}_i\cos\alpha_i, \dot{y}_1 = -l\dot{\alpha}_1\sin\alpha_1, \dot{y}_2 = -\sum_i l\dot{\alpha}_i\sin\alpha_i$:

$$T(\boldsymbol{q},\dot{\boldsymbol{q}}) = \sum_{i=1}^{2}\frac{m}{2}(\dot{x}_i^2 + \dot{y}_i^2) = \frac{ml^2}{2}\big(\dot{\alpha}_1^2 + 2\cos(\alpha_2 - \alpha_1)\dot{\alpha}_1\dot{\alpha}_2 + \dot{\alpha}_2^2\big)\,,$$

$$U(\boldsymbol{q}) = \sum_{i=1}^{2} mg_{\text{grav}}\, y_i = -mg_{\text{grav}}\, l\,(2\cos\alpha_1 + \cos\alpha_2)\,.$$

Evaluating the force vector according to (2.4), we have to take into account the state dependent mass matrix $\mathbf{M}(\boldsymbol{q})$ that results in $\nabla_{\boldsymbol{q}} T(\boldsymbol{q},\dot{\boldsymbol{q}}) \neq \mathbf{0}$. Then, the equations of motion are obtained in form of a linearly implicit second order system of ordinary differential equations with state dependent mass matrix $\mathbf{M}(\boldsymbol{q})$:

$$\begin{pmatrix} 2 & \cos(\alpha_2 - \alpha_1) \\ \cos(\alpha_2 - \alpha_1) & 1 \end{pmatrix}\begin{pmatrix} \ddot{\alpha}_1 \\ \ddot{\alpha}_2 \end{pmatrix} = \begin{pmatrix} -2\frac{g_{\text{grav}}}{l}\sin\alpha_1 + \sin(\alpha_2 - \alpha_1)\dot{\alpha}_2^2 \\ -\frac{g_{\text{grav}}}{l}\sin\alpha_2 - \sin(\alpha_2 - \alpha_1)\dot{\alpha}_1^2 \end{pmatrix}.$$

For the double pendulum, these algebraic manipulations may still be performed by hand but for larger systems the use of computer algebra programs becomes mandatory. As an alternative, we will consider in Sect. 3.3 below a mixed coordinate formulation that allows to evaluate the accelerations $\ddot{\boldsymbol{q}}(t)$ numerically by a block Gauss elimination for a large sparse system of linear equations.

2.2 Existence and Uniqueness

Holonomic constraints (2.1) restrict the configuration space at the level of position coordinates. They imply *hidden constraints* at the level of velocity coordinates $\dot{\boldsymbol{q}}$ that are obtained by differentiation of (2.1) w.r.t. t:

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(t,\boldsymbol{q}(t)) = \frac{\partial \boldsymbol{g}}{\partial t}(t,\boldsymbol{q}(t)) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(t,\boldsymbol{q}(t))\,\dot{\boldsymbol{q}}(t) = \boldsymbol{g}_t(t,\boldsymbol{q}) + \mathbf{G}(t,\boldsymbol{q})\dot{\boldsymbol{q}}\,. \tag{2.8}$$

The second time derivative of the holonomic constraints (2.1) defines hidden constraints at the level of acceleration coordinates $\ddot{\boldsymbol{q}}$:

$$\mathbf{0} = \frac{\mathrm{d}^2}{\mathrm{d}t^2}\boldsymbol{g}(t,\boldsymbol{q}(t)) = \boldsymbol{g}_{tt}(t,\boldsymbol{q}) + 2\boldsymbol{g}_{tq}(t,\boldsymbol{q})\dot{\boldsymbol{q}} + \mathbf{G}(t,\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{g}_{qq}(t,\boldsymbol{q})(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}) \tag{2.9}$$

with $\boldsymbol{g}_{tq}(t,\boldsymbol{q}) = \mathbf{G}_t(t,\boldsymbol{q})$. The curvature term $\boldsymbol{g}_{qq}(t,\boldsymbol{q})(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}})$ represents the second partial derivatives of the vector valued function $\boldsymbol{g}(t,\boldsymbol{q})$ w.r.t. its vector valued argument $\boldsymbol{q}$ in the sense that

$$\boldsymbol{g}_{qq}(t,\boldsymbol{q})(\boldsymbol{w},\boldsymbol{z}) = \frac{\partial}{\partial \boldsymbol{q}}\big(\mathbf{G}(t,\boldsymbol{q})\boldsymbol{w}\big)\boldsymbol{z}\,,\ (\boldsymbol{w},\boldsymbol{z}\in\mathbb{R}^{n_q})\,. \tag{2.10}$$

Here we assume tacitly that the constraint function $\boldsymbol{g}$ is as often continuously differentiable as it is necessary to define the constraint matrix $\mathbf{G}(t,\boldsymbol{q})$ and to derive the hidden constraints (2.8) and (2.9). Appropriate smoothness assumptions will be specified in Theorem 2.2 below.

The hidden constraints (2.8) are part of the *derivative array* of DAE (2.3), see [26]. But they are not just the result of an abstract mathematical transformation but have a reasonable physical interpretation as well [40]. To discuss this aspect in more detail, we focus on *scleronomic* constraints

$$\boldsymbol{g}(\boldsymbol{q}) = \mathbf{0} \tag{2.11}$$

that do not depend explicitly on time t and restrict the configuration of the constrained system to the manifold

$$\mathfrak{M} := \{\,\boldsymbol{q} \,:\, \boldsymbol{g}(\boldsymbol{q}) = \mathbf{0}\,\}\,. \tag{2.12}$$

For scleronomic constraints, the hidden constraints (2.8) and (2.9) are simplified because the partial derivatives w.r.t. t vanish identically:

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q})\dot{\boldsymbol{q}}\,, \tag{2.13}$$

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{g}_{qq}(\boldsymbol{q})(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}})\,. \tag{2.14}$$

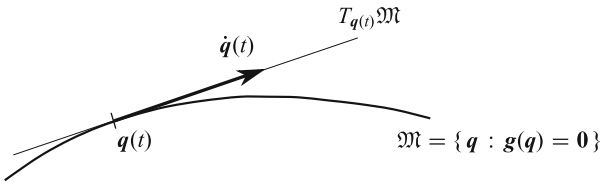


Fig. 4 Constraint manifold $\mathfrak{M} := \{ q : g(q) = 0 \}$ with tangent space $T_q\mathfrak{M}$

Since ker $\mathbf{G}(q)$ spans the tangent space $T_q\mathfrak{M}$ of the manifold at point $q \in \mathfrak{M}$, the hidden constraints (2.13) indicate that the velocity vector $\dot{q}(t)$ is in the tangent space $T_{q(t)}\mathfrak{M}$, see Fig. 4. Therefore, the solution $q(t)$ remains in manifold $\mathfrak{M}$ for all $t \in [t_0, t_{\text{end}}]$, see [71, 72].

Example 2.4 The mathematical pendulum is a model in the (x, y)-plane with a point mass moving in the one-dimensional manifold $\mathfrak{M} = \{ q = (x, y)^\top : x^2 + y^2 = l^2 \}$, see Example 2.1. Manifold $\mathfrak{M}$ is a circle and its tangent space $T_q\mathfrak{M} \subset \mathbb{R}^2$ consists of all vectors being orthogonal to q.

The trajectory $q(t)$ will follow the circle iff $\dot{q}(t) \in T_{q(t)}\mathfrak{M}$, i.e., iff $0 = (q(t))^\top \dot{q}(t) = x(t)\dot{x}(t) + y(t)\dot{y}(t)$. This is exactly the hidden constraint (2.13) at the level of velocity coordinates that results from formal differentiation of constraint (2.6c). A second differentiation step yields the hidden constraint (2.14) at the level of acceleration coordinates:

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}(x\dot{x} + y\dot{y}) = x\ddot{x} + y\ddot{y} + \dot{x}^2 + \dot{y}^2 .$$

This equation may be solved w.r.t. the Lagrange multiplier λ since $\ddot{x} = -x\lambda/m$, $\ddot{y} = -g_{\text{grav}} - y\lambda/m$, see (2.6a,b):

$$\lambda = \lambda(x, \dot{x}, y, \dot{y}) := m\,\frac{-g_{\text{grav}}\, y + \dot{x}^2 + \dot{y}^2}{x^2 + y^2} = m\,\frac{-g_{\text{grav}}\, y + \dot{x}^2 + \dot{y}^2}{l^2} . \tag{2.15}$$

The dynamical equations (2.6a,b) with λ being substituted by $\lambda(x, \dot{x}, y, \dot{y})$ according to (2.15) define a system of second order ODEs for variables x and y that is analytically equivalent to the constrained system.

Initial values $(x_0, \dot{x}_0, y_0, \dot{y}_0, \lambda_0)$ for the constrained system (2.6) have to be consistent with the constraint (2.6c) at position level and with its counterparts (2.13) and (2.14) at the level of velocity and acceleration coordinates:

$$x_0^2 + y_0^2 = l^2 , \quad x_0\dot{x}_0 + y_0\dot{y}_0 = 0 , \quad \lambda_0 = \lambda(x_0, \dot{x}_0, y_0, \dot{y}_0) .$$

Example 2.4 shows that holonomic constraints (2.1) and the corresponding hidden constraints (2.8), (2.9) define conditions on initial values $\boldsymbol{q}_0 = \boldsymbol{q}(t_0)$, $\dot{\boldsymbol{q}}_0 = \dot{\boldsymbol{q}}(t_0)$, $\boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(t_0)$. We will discuss these conditions for constrained systems

$$\mathbf{M}(t,\boldsymbol{q})\ddot{\boldsymbol{q}} = \boldsymbol{f}(t,\boldsymbol{q},\dot{\boldsymbol{q}}) - \mathbf{G}^\top(t,\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{2.16a}$$

$$\mathbf{0} = \boldsymbol{g}(t,\boldsymbol{q}) \tag{2.16b}$$

with $\mathbf{G}(t,\boldsymbol{q}) = (\partial \boldsymbol{g}/\partial \boldsymbol{q})(t,\boldsymbol{q})$. This problem class is slightly more general than (2.3) and covers time dependent force terms $\boldsymbol{f}$ as well as condensed mass matrices $\mathbf{M}(t,\boldsymbol{q})$ that result from the application of multibody formalisms to systems with rheonomic joint equations, see Sect. 3.3.

Remark 2.1 In some textbooks, the argument t in the equations of motion (2.3) and (2.16) is omitted to keep the notation compact. In the ODE case, this is justified by the observation that any second order system $\ddot{\boldsymbol{x}} = \boldsymbol{f}(t,\boldsymbol{x},\dot{\boldsymbol{x}})$ in $\mathbb{R}^k$ is equivalent to an autonomous system $\ddot{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{f}}(\bar{\boldsymbol{x}},\dot{\bar{\boldsymbol{x}}})$ in $\mathbb{R}^{k+1}$ with $\bar{\boldsymbol{x}} := (t, \boldsymbol{x}^\top)^\top, \bar{\boldsymbol{f}} := (0, \boldsymbol{f}^\top)^\top$, $\bar{\boldsymbol{x}}_0 := (t_0, \boldsymbol{x}_0^\top)^\top$, $\dot{\bar{\boldsymbol{x}}}_0 := (1, \dot{\boldsymbol{x}}_0^\top)^\top$, see, e.g., [48, Sect. II.2] for the corresponding transformation in the case of first order ODEs.

Applying this transformation formally to constrained systems (2.16) with *rheonomic* constraints $\mathbf{0} = \boldsymbol{g}(t,\boldsymbol{q})$, we obtain $\bar{\boldsymbol{q}} = (t, \boldsymbol{q}^\top)^\top$, scleronomic constraints $\mathbf{0} = \bar{\boldsymbol{g}}(\bar{\boldsymbol{q}}) := \boldsymbol{g}(t,\boldsymbol{q})$ and a constraint Jacobian $(\partial \bar{\boldsymbol{g}}/\partial \bar{\boldsymbol{q}})(\bar{\boldsymbol{q}})$ that is composed of the constraint matrix $\mathbf{G}(t,\boldsymbol{q}) = (\partial \boldsymbol{g}/\partial \boldsymbol{q})(t,\boldsymbol{q})$ and the partial derivatives $(\partial \boldsymbol{g}/\partial t)(t,\boldsymbol{q})$ that do not appear in (2.16). Therefore, the structure of the equations of motion (2.16) gets lost by the transformation to an autonomous system in coordinates $\bar{\boldsymbol{q}} = (t, \boldsymbol{q}^\top)^\top$ if $\partial \boldsymbol{g}/\partial t \not\equiv \mathbf{0}$. That's why we will consider the equations of motion in their original non-autonomous form (2.16).

The dynamical equations (2.16a) and the hidden constraints (2.9) may be summarized to a system of $n_q + n_\lambda$ linear equations in $\ddot{\boldsymbol{q}}$ and $\boldsymbol{\lambda}$:

$$\begin{pmatrix} \mathbf{M}(t,\boldsymbol{q}) & \mathbf{G}^\top(t,\boldsymbol{q}) \\ \mathbf{G}(t,\boldsymbol{q}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \ddot{\boldsymbol{q}} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}(t,\boldsymbol{q},\dot{\boldsymbol{q}}) \\ -\boldsymbol{g}_{qq}(t,\boldsymbol{q})(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}) - 2\boldsymbol{g}_{tq}(t,\boldsymbol{q})\dot{\boldsymbol{q}} - \boldsymbol{g}_{tt}(t,\boldsymbol{q}) \end{pmatrix}. \tag{2.17}$$

For any given arguments $t, \boldsymbol{q}, \dot{\boldsymbol{q}}$ the Lagrange multipliers $\boldsymbol{\lambda} = \boldsymbol{\lambda}(t,\boldsymbol{q},\dot{\boldsymbol{q}})$ are uniquely defined by this block structured system if the 2×2 block matrix at the left-hand side of (2.17) is non-singular. The non-zero blocks $\mathbf{M} = \mathbf{M}(t,\boldsymbol{q})$ and $\mathbf{G} = \mathbf{G}(t,\boldsymbol{q})$ of this 2×2 matrix are known from the definition of the kinetic energy $T(\dot{\boldsymbol{q}})$ and from the hidden constraints (2.8) at the level of velocity coordinates. For physical reasons, we assume as before that $\mathbf{M}$ is symmetric, positive semi-definite to get a positive semi-definite quadratic form $T(\dot{\boldsymbol{q}}) = 0.5\,\dot{\boldsymbol{q}}^\top \mathbf{M}\dot{\boldsymbol{q}}$.

Lemma 2.1 *Consider a symmetric, positive semi-definite matrix* $\mathbf{M} \in \mathbb{R}^{k\times k}$ *and a matrix* $\mathbf{G} \in \mathbb{R}^{m\times k}$ *with* rank $\mathbf{G} = m \le k$. *If* $\mathbf{M}$ *is positive definite at the null space*

of $\mathbf{G}$*, then matrix*

$$\begin{pmatrix} \mathbf{M} & \mathbf{G}^\top \\ \mathbf{G} & \mathbf{0} \end{pmatrix} \tag{2.18}$$

is non-singular.

Proof The terms $\boldsymbol{\xi}^\top \mathbf{M}\boldsymbol{\xi}$ and $\|\mathbf{G}\boldsymbol{\xi}\|_2^2$ are non-negative for all vectors $\boldsymbol{\xi} \in \mathbb{R}^k$ since matrix $\mathbf{M}$ is positive semi-definite and $\|\mathbf{G}\boldsymbol{\xi}\|_2 \geq 0$. Furthermore, $\|\mathbf{G}\boldsymbol{\xi}\|_2 = 0$ implies $\mathbf{G}\boldsymbol{\xi} = \mathbf{0}$ and $\boldsymbol{\xi} \in \ker \mathbf{G}$, i.e., $\boldsymbol{\xi} = \mathbf{0}$ or $\boldsymbol{\xi}^\top \mathbf{M}\boldsymbol{\xi} > 0$ because $\mathbf{M}$ is positive definite at $\ker \mathbf{G}$. Taking into account that

$$\boldsymbol{\xi}^\top \mathbf{M}\boldsymbol{\xi} + \|\mathbf{G}\boldsymbol{\xi}\|_2^2 = \boldsymbol{\xi}^\top (\mathbf{M} + \mathbf{G}^\top \mathbf{G})\boldsymbol{\xi}$$

we see that the symmetric matrix $\mathbf{M} + \mathbf{G}^\top \mathbf{G} \in \mathbb{R}^{k\times k}$ is positive definite. Therefore, its inverse is well defined and matrix $\mathbf{G}(\mathbf{M} + \mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top \in \mathbb{R}^{m\times m}$ is symmetric, positive definite for any matrix $\mathbf{G}$ of full rank $m \leq k$. The assertion of the lemma follows from a block factorization of the 2×2 block matrix in three non-singular factors:

$$\begin{pmatrix} \mathbf{M} & \mathbf{G}^\top \\ \mathbf{G} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & -\mathbf{G}^\top \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{M} + \mathbf{G}^\top \mathbf{G} & \mathbf{0} \\ \mathbf{G} & -\mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & (\mathbf{M} + \mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top \\ \mathbf{0} & \mathbf{G}(\mathbf{M} + \mathbf{G}^\top \mathbf{G})^{-1}\mathbf{G}^\top \end{pmatrix}.$$

■

Theorem 2.2 *Consider vectors* $\boldsymbol{q}_0, \dot{\boldsymbol{q}}_0 \in \mathbb{R}^{n_q}$ *that satisfy at* $t = t_0$ *the (hidden) constraints at the levels of position and velocity coordinates:*

$$\mathbf{0} = \boldsymbol{g}(t_0, \boldsymbol{q}_0) = \mathbf{G}(t_0, \boldsymbol{q}_0)\dot{\boldsymbol{q}}_0 + \boldsymbol{g}_t(t_0, \boldsymbol{q}_0)\,. \tag{2.19}$$

We assume that functions $\mathbf{M}(t, \boldsymbol{q})$*,* $\boldsymbol{f}(t, \boldsymbol{q}, \dot{\boldsymbol{q}})$ *and* $\boldsymbol{g}(t, \boldsymbol{q})$ *are well defined and continuous in a neighbourhood of* $(t_0, \boldsymbol{q}_0, \dot{\boldsymbol{q}}_0)$ *with* $\boldsymbol{g}(t, \boldsymbol{q})$ *being two times continuously differentiable. Furthermore, functions* $\mathbf{M}$*,* $\boldsymbol{f}$ *and the second (partial) derivatives of* $\boldsymbol{g}$ *are assumed to satisfy Lipschitz conditions w.r.t. arguments* $\boldsymbol{q}$ *and* $\dot{\boldsymbol{q}}$*.*

If the constraint matrix $\mathbf{G}(t_0, \boldsymbol{q}_0)$ *has full rank* n_λ *and the mass matrix* $\mathbf{M}(t_0, \boldsymbol{q}_0)$ *is symmetric, positive semi-definite and positive definite at* $\ker \mathbf{G}(t_0, \boldsymbol{q}_0)$*, then there is a uniquely defined vector* $\boldsymbol{\lambda}_0 \in \mathbb{R}^{n_\lambda}$ *such that the initial value problem*

$$\boldsymbol{q}(t_0) = \boldsymbol{q}_0\,, \quad \dot{\boldsymbol{q}}(t_0) = \dot{\boldsymbol{q}}_0\,, \quad \boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}_0 \tag{2.20}$$

for the constrained system (2.16) is locally uniquely solvable.

Proof The assumptions on $\mathbf{M}(t_0, \boldsymbol{q}_0)$ and $\mathbf{G}(t_0, \boldsymbol{q}_0)$ imply that the 2×2 block matrix at the left-hand side of (2.17) is non-singular for arguments $t = t_0$, $\boldsymbol{q} = \boldsymbol{q}_0$, see Lemma 2.1. Therefore, this block matrix is non-singular for any arguments $(t, \boldsymbol{q})$ in a neighbourhood of $(t_0, \boldsymbol{q}_0)$ since functions $\mathbf{M}$ and $\mathbf{G}$ are continuous w.r.t. t and $\boldsymbol{q}$,

see [46, Lemma 2.3.3]. In this neighbourhood, the system of linear equations (2.17) is uniquely solvable w.r.t. $\ddot{\boldsymbol{q}}$ and $\boldsymbol{\lambda}$ and defines continuous functions $\boldsymbol{a}$ and $\boldsymbol{\lambda}$ such that

$$\ddot{\boldsymbol{q}} = \boldsymbol{a}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}) \,, \quad \boldsymbol{\lambda} = \boldsymbol{\lambda}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}) \,.$$

The initial value problem $\boldsymbol{q}(t_0) = \boldsymbol{q}_0$, $\dot{\boldsymbol{q}}(t_0) = \dot{\boldsymbol{q}}_0$ for the second order ODE $\ddot{\boldsymbol{q}}(t) = \boldsymbol{a}(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t))$ is locally uniquely solvable since the right-hand side $\boldsymbol{a}$ satisfies a Lipschitz condition w.r.t. $\boldsymbol{q}$ and $\dot{\boldsymbol{q}}$. The solution $\boldsymbol{q}(t)$ of this ODE initial value problem satisfies the dynamical equations (2.16a) with $\boldsymbol{\lambda} := \boldsymbol{\lambda}(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t))$ since these equations are represented by the first block row in (2.17). The initial value for the Lagrange multipliers is uniquely defined by $\boldsymbol{\lambda}_0 := \boldsymbol{\lambda}(t_0, \boldsymbol{q}_0, \dot{\boldsymbol{q}}_0)$.

To verify the constraint equations (2.16b), we consider the constraint residual $\boldsymbol{\gamma}(t) := \boldsymbol{g}(t, \boldsymbol{q}(t))$ and its time derivatives

$$\dot{\boldsymbol{\gamma}}(t) = \boldsymbol{g}_t\big(t, \boldsymbol{q}(t)\big) + \mathbf{G}\big(t, \boldsymbol{q}(t)\big)\dot{\boldsymbol{q}}(t) \,,$$
$$\ddot{\boldsymbol{\gamma}}(t) = \boldsymbol{g}_{tt}\big(t, \boldsymbol{q}(t)\big) + 2\boldsymbol{g}_{tq}\big(t, \boldsymbol{q}(t)\big)\dot{\boldsymbol{q}}(t) + \mathbf{G}\big(t, \boldsymbol{q}(t)\big)\ddot{\boldsymbol{q}}(t) + \boldsymbol{g}_{qq}\big(t, \boldsymbol{q}(t)\big)\big(\dot{\boldsymbol{q}}(t), \dot{\boldsymbol{q}}(t)\big) \,,$$

see (2.1), (2.8), (2.9). The second block row of (2.17) shows that the residual $\ddot{\boldsymbol{\gamma}}(t)$ in the hidden constraints (2.9) at the level of acceleration coordinates vanishes identically. Hence, $\boldsymbol{\gamma}(t)$ solves the second order ODE $\ddot{\boldsymbol{\gamma}}(t) = \mathbf{0}$ with initial values $\boldsymbol{\gamma}(t_0) = \boldsymbol{g}(t_0, \boldsymbol{q}_0) = \mathbf{0}$ and $\dot{\boldsymbol{\gamma}}(t_0) = \mathbf{G}(t_0, \boldsymbol{q}_0)\dot{\boldsymbol{q}}_0 + \boldsymbol{g}_t(t_0, \boldsymbol{q}_0) = \mathbf{0}$, see (2.19). Since this solution is unique, we get $\boldsymbol{\gamma}(t) \equiv \mathbf{0}$ and therefore also $\boldsymbol{g}(t, \boldsymbol{q}(t)) \equiv \mathbf{0}$. I.e., the constraint equations (2.16b) are satisfied in the whole time interval of interest and functions $\boldsymbol{q}(t)$, $\boldsymbol{\lambda}(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t))$ solve the initial value problem $\boldsymbol{q}(t_0) = \boldsymbol{q}_0$, $\dot{\boldsymbol{q}}(t_0) = \dot{\boldsymbol{q}}_0$, $\boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}_0 = \boldsymbol{\lambda}(t_0, \boldsymbol{q}_0, \dot{\boldsymbol{q}}_0)$ for the constrained system (2.16). ■

Definition 2.1 Initial values $\boldsymbol{q}_0, \dot{\boldsymbol{q}}_0 \in \mathbb{R}^{n_q}$, $\boldsymbol{\lambda}_0 \in \mathbb{R}^{n_\lambda}$ are *consistent* with the equations of motion (2.16) if $\boldsymbol{q}_0$ and $\dot{\boldsymbol{q}}_0$ satisfy the (hidden) constraints at the levels of position and velocity coordinates, see (2.19), and there is a vector $\ddot{\boldsymbol{q}}_0$ such that $\ddot{\boldsymbol{q}} = \ddot{\boldsymbol{q}}_0$, $\boldsymbol{\lambda} = \boldsymbol{\lambda}_0$ solve the system of linear equations (2.17) with $t := t_0$, $\boldsymbol{q} := \boldsymbol{q}_0$, $\dot{\boldsymbol{q}} := \dot{\boldsymbol{q}}_0$.

Remark 2.2

(a) For any consistent initial values $\boldsymbol{q}_0$, $\dot{\boldsymbol{q}}_0$, $\boldsymbol{\lambda}_0$, the initial value problem $\boldsymbol{q}(t_0) = \boldsymbol{q}_0$, $\dot{\boldsymbol{q}}(t_0) = \dot{\boldsymbol{q}}_0$, $\boldsymbol{\lambda}(t_0) = \boldsymbol{\lambda}_0$ for DAE (2.16) is locally uniquely solvable if rank $\mathbf{G}(t_0, \boldsymbol{q}_0) = n_\lambda$, $\mathbf{M}(t_0, \boldsymbol{q}_0)$ is symmetric positive semi-definite and positive definite at ker $\mathbf{G}(t_0, \boldsymbol{q}_0)$ and functions $\mathbf{M}, \boldsymbol{f}$ and $\boldsymbol{g}$ satisfy appropriate smoothness assumptions, see Theorem 2.2.

(b) Following a *coordinate partitioning* approach [89], consistent initial values $\boldsymbol{q}_0$, $\dot{\boldsymbol{q}}_0$, $\boldsymbol{\lambda}_0$ may be obtained from any pair of vectors $\bar{\boldsymbol{q}}_0$, $\bar{\dot{\boldsymbol{q}}}_0 \in \mathbb{R}^{n_q}$ provided that $\|\boldsymbol{g}(t_0, \bar{\boldsymbol{q}}_0)\| \le \delta$ with a sufficiently small constant $\delta > 0$: The full rank assumption on the constraint matrix $\mathbf{G}$ allows to select in a first step n_λ linearly independent column vectors of $\mathbf{G}(t_0, \bar{\boldsymbol{q}}_0)$. There is a matrix $\hat{\mathbf{P}} \in \mathbb{R}^{n_q \times n_\lambda}$ being composed of n_λ unit vectors such that $\mathbf{G}(t_0, \bar{\boldsymbol{q}}_0)\hat{\mathbf{P}} \in \mathbb{R}^{n_\lambda \times n_\lambda}$ is non-singular.

In the second step, vector $\boldsymbol{q}_0 \in \mathbb{R}^{n_q}$ is decomposed into $n_q - n_\lambda$ *independent* coordinates $\bar{\mathbf{P}}^\top \boldsymbol{q}_0 \in \mathbb{R}^{n_q - n_\lambda}$ and n_λ *dependent* coordinates $\hat{\boldsymbol{q}}_0 := \hat{\mathbf{P}}^\top \boldsymbol{q}_0 \in \mathbb{R}^{n_\lambda}$ with a matrix $\bar{\mathbf{P}} \in \mathbb{R}^{n_q \times (n_q - n_\lambda)}$ that is defined such that $\mathbf{P} := \begin{pmatrix} \bar{\mathbf{P}} & \hat{\mathbf{P}} \end{pmatrix} \in \mathbb{R}^{n_q \times n_q}$ forms a permutation matrix, i.e., $\mathbf{I}_{n_q} = \mathbf{P}\mathbf{P}^\top = \bar{\mathbf{P}}\bar{\mathbf{P}}^\top + \hat{\mathbf{P}}\hat{\mathbf{P}}^\top$. Finally, we fix $\bar{\mathbf{P}}^\top \boldsymbol{q}_0 := \bar{\mathbf{P}}^\top \bar{\boldsymbol{q}}_0$ and get consistent position coordinates $\boldsymbol{q}_0 = \mathbf{P}\mathbf{P}^\top \boldsymbol{q}_0 := \bar{\mathbf{P}}\bar{\mathbf{P}}^\top \bar{\boldsymbol{q}}_0 + \hat{\mathbf{P}}\hat{\boldsymbol{q}}_0$ solving

$$\mathbf{0} = \boldsymbol{g}(t_0, \bar{\mathbf{P}}\bar{\mathbf{P}}^\top \bar{\boldsymbol{q}}_0 + \hat{\mathbf{P}}\hat{\boldsymbol{q}}_0) \tag{2.21}$$

w.r.t. $\hat{\boldsymbol{q}}_0 \in \mathbb{R}^{n_\lambda}$. According to the Implicit function theorem, Eq. (2.21) are locally uniquely solvable if $\|\boldsymbol{g}(t_0, \bar{\boldsymbol{q}}_0)\| \leq \delta \ll 1$ since

$$\frac{\partial \boldsymbol{g}}{\partial \hat{\boldsymbol{q}}_0}(t_0, \bar{\boldsymbol{q}}_0) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(t_0, \bar{\boldsymbol{q}}_0) \frac{\partial \boldsymbol{q}_0}{\partial \hat{\boldsymbol{q}}_0}(\bar{\boldsymbol{q}}_0) = \mathbf{G}(t_0, \bar{\boldsymbol{q}}_0)\hat{\mathbf{P}}$$

is non-singular by construction.

In the same way, consistent initial values $\dot{\boldsymbol{q}}_0 = \bar{\mathbf{P}}\bar{\mathbf{P}}^\top \dot{\bar{\boldsymbol{q}}}_0 + \hat{\mathbf{P}}\hat{\dot{\boldsymbol{q}}}_0$ with $\hat{\dot{\boldsymbol{q}}}_0 \in \mathbb{R}^{n_\lambda}$ are obtained from the system of n_λ *linear* equations

$$\mathbf{0} = \mathbf{G}(t_0, \boldsymbol{q}_0)\dot{\boldsymbol{q}}_0 + \boldsymbol{g}_t(t_0, \boldsymbol{q}_0) = \mathbf{G}(t_0, \boldsymbol{q}_0)\bar{\mathbf{P}}\bar{\mathbf{P}}^\top \dot{\bar{\boldsymbol{q}}}_0 + \mathbf{G}(t_0, \boldsymbol{q}_0)\hat{\mathbf{P}}\hat{\dot{\boldsymbol{q}}}_0 + \boldsymbol{g}_t(t_0, \boldsymbol{q}_0)$$

provided that $\mathbf{G}(t_0, \boldsymbol{q}_0)\hat{\mathbf{P}}$ is non-singular as well. At the end, the 2×2 block system (2.17) yields consistent initial values $\boldsymbol{\lambda}_0$ for the Lagrange multipliers.

Remark 2.3

(a) For the index analysis, the equations of motion (2.16) are transformed to an equivalent first order DAE introducing velocity coordinates $\boldsymbol{v}(t) := \dot{\boldsymbol{q}}(t)$ and substituting $\dot{\boldsymbol{q}} \to \boldsymbol{v}$, $\ddot{\boldsymbol{q}} \to \dot{\boldsymbol{v}}$. With the assumptions of Theorem 2.2, functions $\dot{\boldsymbol{v}}(t) = \ddot{\boldsymbol{q}}(t)$ and $\boldsymbol{\lambda}(t)$ are obtained from the system of linear equations (2.17) that contains the second time derivative of the holonomic constraints (2.16b).

The 2×2 block matrix in (2.17) is non-singular and does not depend on $\boldsymbol{v}$, $\dot{\boldsymbol{v}}$ and $\boldsymbol{\lambda}$. Therefore, the time derivative of (2.17) may be solved w.r.t. $\ddot{\boldsymbol{v}}$ and $\dot{\boldsymbol{\lambda}}$ providing an explicit expression for $\dot{\boldsymbol{\lambda}}$ that utilizes the *third* time derivative of (2.16b). Consequently, the differentiation index of the equivalent first order system is (at most) three [47, 60].

(b) For positive definite mass matrices $\mathbf{M}(t, \boldsymbol{q})$, the dynamical equations (2.16a) may formally be solved w.r.t. $\ddot{\boldsymbol{q}} = \dot{\boldsymbol{v}}$ resulting in the first order DAE

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{2.22a}$$

$$\dot{\boldsymbol{v}} = [\mathbf{M}^{-1}\boldsymbol{f}](t, \boldsymbol{q}, \boldsymbol{v}) - [\mathbf{M}^{-1}\mathbf{G}^\top](t, \boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{2.22b}$$

$$\mathbf{0} = \boldsymbol{g}(t, \boldsymbol{q}) \tag{2.22c}$$

that is of Hessenberg form [26]. For full rank matrices $\mathbf{G}$ and symmetric, positive definite matrices $\mathbf{M}$, matrix $\mathbf{G}\mathbf{M}^{-1}\mathbf{G}^\top$ is non-singular and (2.22b)

implies

$$\boldsymbol{\lambda} = \boldsymbol{f}_0(t, \boldsymbol{q}, \boldsymbol{v}) - [(\mathbf{G}\mathbf{M}^{-1}\mathbf{G}^\top)^{-1}](t, \boldsymbol{q}) \cdot \mathbf{G}(t, \boldsymbol{q})\dot{\boldsymbol{v}} \tag{2.23}$$

with an appropriate function $\boldsymbol{f}_0$. The time derivative of (2.23) shows that $\dot{\boldsymbol{\lambda}}(t)$ is composed of functions depending on t, $\boldsymbol{q}$, $\boldsymbol{v}$ and $\dot{\boldsymbol{v}} = [\mathbf{M}^{-1}\boldsymbol{f}] - [\mathbf{M}^{-1}\mathbf{G}^\top]\boldsymbol{\lambda}$ and of the vector $\mathbf{G}(t, q)\ddot{\boldsymbol{v}}$ that is pre-multiplied by the non-singular matrix $-(\mathbf{G}\mathbf{M}^{-1}\mathbf{G}^\top)^{-1}$.

Since $\mathbf{G}(t, q)\ddot{\boldsymbol{v}}$ cannot be obtained from DAE (2.22) and its first two time derivatives, the differentiation index of DAE (2.22) is larger than two. Taking into account the upper bound from part (a) of this remark, we see that the equations of motion (2.16) form an index-3 DAE if $\mathbf{M}(t, \boldsymbol{q})$ is symmetric and positive definite. Note that differentiation index and perturbation index of (2.16) coincide since the equivalent first order system is of Hessenberg form [30, 47].

The analytical transformation of the equations of motion (2.16) to the Hessenberg form index-3 DAE (2.22) is a common approach in DAE theory. This transformation is essentially based on the assumption that the mass matrix $\mathbf{M}$ is symmetric, positive definite [26, 50, 58]. However, the existence and uniqueness result in Theorem 2.2 is not restricted to this problem class but applies as well to a class of model equations (2.16) with rank-deficient mass matrix $\mathbf{M}$. In this more general setting, the structure of (2.16) is more complex and its index may be less than three [16]:

Example 2.5 A (pathological) example of problems with rank-deficient mass matrix $\mathbf{M}$ are constrained systems (2.16) with $\mathbf{M}(t, \boldsymbol{q}) = \mathbf{0}_{n_q \times n_q}$. This matrix is positive semi-definite and it is positive definite at $\ker \mathbf{G}(t, \boldsymbol{q})$ if $n_q = n_\lambda$ and $\mathbf{G}(t, \boldsymbol{q})$ is non-singular. For such systems, there is no need to consider the 2×2 block system (2.17) since the Lagrange multipliers $\boldsymbol{\lambda}(t) = [\mathbf{G}^{-\top}\boldsymbol{f}]\big(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)\big)$ are directly defined by the dynamical equations (2.16a).

The differentiation index of the corresponding first order system in variables $\boldsymbol{q}$, $\boldsymbol{v} := \dot{\boldsymbol{q}}$ and $\boldsymbol{\lambda}$ is two [16]. If $\mathbf{G}$ is non-singular, $\mathbf{M} \equiv \mathbf{0}$ and $\boldsymbol{f}$ is independent of $\dot{\boldsymbol{q}}$, then (2.16) defines even an index-1 DAE (in variables $\boldsymbol{q}$ and $\boldsymbol{\lambda}$):

$$\mathbf{0} = \boldsymbol{f}(t, \boldsymbol{q}) - \mathbf{G}^\top(t, \boldsymbol{q})\,\boldsymbol{\lambda}\,, \quad \mathbf{0} = \boldsymbol{g}(t, \boldsymbol{q})\,.$$

2.3 *Positive Semi-Definite Mass Matrices, Rank-Deficient Constraint Matrices*

In engineering applications, there are certain types of position coordinates $\boldsymbol{q}$ that result systematically in constrained systems (2.16) with rank-deficient mass matrix, see [42, 64, 86] and the references therein. From the viewpoint of physics, the kinetic energy $T = 0.5\,\dot{\boldsymbol{q}}^\top\mathbf{M}\dot{\boldsymbol{q}}$ should define a positive semi-definite quadratic form and any non-zero velocity increment being compatible with the hidden constraints (2.8)

should result in a positive contribution to T, see [42]. Both properties of T are achieved by the assumptions of Lemma 2.1 that considers symmetric, positive semi-definite mass matrices $\mathbf{M}$ being positive definite at $\ker \mathbf{G}$.

These assumptions imply that the augmented matrix $\mathbf{M} + \mathbf{G}^\top \mathbf{G}$ with $\operatorname{rank} \mathbf{G} = n_\lambda$ is symmetric, positive definite [45, Sect. 10.2] and the 2×2 block matrix in (2.18) is non-singular, see Lemma 2.1. For a more detailed analysis, we decouple in the present section the nullspace of $\mathbf{M}$ from its orthogonal complement and consider furthermore systems with rank-deficient constraint matrix $\mathbf{G}$ resulting from redundant constraints (2.16b) that are typical of some algorithms for computer-aided setup of complex, three dimensional multibody system models [39, 42].

Lötstedt [59] pointed out that equations of motion (2.16) with consistent, but redundant constraints (2.16b) do not define unique Lagrange multipliers $\boldsymbol{\lambda}(t)$. Nevertheless, the constraint forces $-\mathbf{G}^\top \boldsymbol{\lambda}$ and the position coordinates $\boldsymbol{q}(t)$ are well defined. Modelling aspects and analytical aspects of equations of motion (2.16) with rank-deficient mass matrix or rank-deficient constraint matrix have recently been studied in great detail by García de Jalón and Gutiérrez-López [42]. They also refer to the work of Frączek and Wojtyra [39] who have shown that the uniqueness of $\boldsymbol{q}(t)$ cannot longer be guaranteed if the dynamical equations (2.16a) depend nonlinearly on $\boldsymbol{\lambda}$ (and the constraints (2.16b) are redundant), see also the more general and more abstract analysis of overdetermined and underdetermined DAEs by Kunkel and Mehrmann [58].

The internal structure of equations of motion (2.16) with rank-deficient mass matrix $\mathbf{M}$ or rank-deficient constraint matrix $\mathbf{G}$ may be studied conveniently by a decomposition of the 2×2 block matrix in (2.18) that takes into account nontrivial nullspaces $\ker \mathbf{M}$ and $\ker \mathbf{G}$:

Lemma 2.3 *Consider matrices $\mathbf{M} \in \mathbb{R}^{k\times k}$ and $\mathbf{G} \in \mathbb{R}^{m\times k}$ with $\operatorname{rank} \mathbf{M} = r \le k$ and $\operatorname{rank} \mathbf{G} = s \le m \le k$. If $\mathbf{M}$ is symmetric, positive semi-definite and positive definite at $\ker \mathbf{G}$, then there are non-singular matrices $\mathbf{U} \in \mathbb{R}^{k\times k}$ and $\mathbf{Q} \in \mathbb{R}^{m\times m}$ such that*

$$\left(\begin{array}{c|c} \mathbf{M} & \mathbf{G}^\top \\ \hline \mathbf{G} & \mathbf{0} \end{array}\right) = \left(\begin{array}{c|c} \mathbf{U} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Q} \end{array}\right) \left(\begin{array}{cc|ccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{k-r} \\ \mathbf{0} & \bar{\mathbf{M}} & \bar{\mathbf{G}}^\top & \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \bar{\mathbf{G}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{k-r} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{array}\right) \left(\begin{array}{c|c} \mathbf{U}^\top & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{Q}^\top \end{array}\right) \tag{2.24}$$

with a non-singular matrix $\bar{\mathbf{M}} \in \mathbb{R}^{r\times r}$ and a matrix $\bar{\mathbf{G}} \in \mathbb{R}^{(s-(k-r))\times r}$ that has full rank $s - (k - r)$.

Proof If $r = \operatorname{rank} \mathbf{M} < k$, then the nullspace of $\mathbf{M}$ is non-trivial and there is an orthonormal basis $\{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-r}\}$ of $\ker \mathbf{M}$. Summarizing these basis vectors in a matrix $\bar{\bar{\mathbf{U}}} := \left(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_{k-r}\right) \in \mathbb{R}^{k\times(k-r)}$, we may define a matrix $\bar{\mathbf{U}} \in \mathbb{R}^{k\times r}$ such that $\hat{\mathbf{U}} := \left(\bar{\bar{\mathbf{U}}} \;\; \bar{\mathbf{U}}\right) \in \mathbb{R}^{k\times k}$ is orthogonal. Since $\mathbf{M}\bar{\bar{\mathbf{U}}} = \mathbf{0}_{k\times(k-r)}$ and $\hat{\mathbf{U}}^\top \mathbf{M} \hat{\mathbf{U}}$ is

symmetric, we get

$$\hat{\mathbf{U}}^\top \mathbf{M} \hat{\mathbf{U}} = \begin{pmatrix} \bar{\bar{\mathbf{U}}}^\top \\ \bar{\mathbf{U}}^\top \end{pmatrix} \begin{pmatrix} \mathbf{0} & \mathbf{M}\bar{\mathbf{U}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}} \end{pmatrix} \tag{2.25}$$

with the matrix $\bar{\mathbf{M}} := \bar{\mathbf{U}}^\top \mathbf{M} \bar{\mathbf{U}} \in \mathbb{R}^{r \times r}$ that is non-singular because of rank $\bar{\mathbf{M}} =$ rank $\hat{\mathbf{U}}^\top \mathbf{M} \hat{\mathbf{U}} =$ rank $\mathbf{M} = r$.

The column vectors of $\mathbf{G}\bar{\bar{\mathbf{U}}} \in \mathbb{R}^{m \times (k-r)}$ are linearly independent since otherwise there would be a vector $\boldsymbol{\zeta} \in \mathbb{R}^{k-r}$ with $\boldsymbol{\zeta} \neq \mathbf{0}$ and $\mathbf{0} = (\mathbf{G}\bar{\bar{\mathbf{U}}})\boldsymbol{\zeta} = \mathbf{G}(\bar{\bar{\mathbf{U}}}\boldsymbol{\zeta})$, i.e., $\boldsymbol{\xi} := \bar{\bar{\mathbf{U}}}\boldsymbol{\zeta} \in \ker \mathbf{G} \setminus \{\mathbf{0}\}$. Since $\mathbf{M}$ is positive definite at $\ker \mathbf{G}$, we would get $0 < \boldsymbol{\xi}^\top \mathbf{M} \boldsymbol{\xi} = \boldsymbol{\zeta}^\top \bar{\bar{\mathbf{U}}}^\top \mathbf{M} \bar{\bar{\mathbf{U}}} \boldsymbol{\zeta}$ which contradicts span $\bar{\bar{\mathbf{U}}} = \ker \mathbf{M}$.

Because of rank $\mathbf{G}\bar{\bar{\mathbf{U}}} = k - r \leq m$, there is a QR factorization

$$\mathbf{G}\bar{\bar{\mathbf{U}}} = \bar{\bar{\mathbf{Q}}} \begin{pmatrix} \bar{\bar{\mathbf{R}}} \\ \mathbf{0} \end{pmatrix}$$

with an orthogonal matrix $\bar{\bar{\mathbf{Q}}} \in \mathbb{R}^{m \times m}$ and a non-singular matrix $\bar{\bar{\mathbf{R}}} \in \mathbb{R}^{(k-r) \times (k-r)}$, see, e.g., [46]. We get

$$\bar{\bar{\mathbf{Q}}}^\top \mathbf{G} \hat{\mathbf{U}} = \begin{pmatrix} \bar{\bar{\mathbf{R}}} & \bar{\bar{\mathbf{G}}} \\ \mathbf{0} & \hat{\mathbf{G}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \bar{\bar{\mathbf{R}}} \\ \mathbf{I}_{m-(k-r)} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{0} & \hat{\mathbf{G}} \\ \mathbf{I}_{k-r} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{k-r} & \bar{\bar{\mathbf{R}}}^{-1} \bar{\bar{\mathbf{G}}} \\ \mathbf{0} & \mathbf{I}_r \end{pmatrix} \tag{2.26}$$

with matrices $\hat{\mathbf{G}} \in \mathbb{R}^{(m-(k-r)) \times r}$ and $\bar{\bar{\mathbf{G}}} \in \mathbb{R}^{(k-r) \times r}$. The right-hand side of (2.26) is a product of three block matrices. Since the first and the last factor are non-singular, we get

$$\text{rank } \hat{\mathbf{G}} + \text{rank } \mathbf{I}_{k-r} = \text{rank } \bar{\bar{\mathbf{Q}}}^\top \mathbf{G} \hat{\mathbf{U}} = \text{rank } \mathbf{G} = s\,,$$

i.e., rank $\hat{\mathbf{G}} = s - (k - r) \leq m - (k - r)$. If matrix $\mathbf{G}$ has full rank m, then $\hat{\mathbf{G}}$ has full rank as well and we define $\bar{\mathbf{G}} := \hat{\mathbf{Q}}\hat{\mathbf{G}}$ with the identity matrix $\hat{\mathbf{Q}} := \mathbf{I}_{m-(k-r)}$, see [16]. Otherwise, matrix $\hat{\mathbf{G}}$ is rank deficient and $s - (k - r)$ linearly independent row vectors may be selected by some pivoting strategy that results in a decomposition

$$\hat{\mathbf{G}} = \hat{\mathbf{Q}} \begin{pmatrix} \bar{\mathbf{G}} \\ \mathbf{0} \end{pmatrix}$$

with non-singular $\hat{\mathbf{Q}} \in \mathbb{R}^{(m-(k-r))\times(m-(k-r))}$ and a matrix $\bar{\mathbf{G}} \in \mathbb{R}^{(s-(k-r))\times r}$ of full rank $s-(k-r)$. Inserting this expression in (2.26), we get

$$\mathbf{G} = \mathbf{Q}\begin{pmatrix} \mathbf{0} & \bar{\mathbf{G}} \\ \mathbf{0} & \mathbf{0} \\ \mathbf{I}_{k-r} & \mathbf{0} \end{pmatrix}\mathbf{U}^\top$$

and non-singular transformation matrices

$$\mathbf{Q} := \bar{\bar{\mathbf{Q}}}\begin{pmatrix} \mathbf{0} & \bar{\bar{\mathbf{R}}} \\ \hat{\mathbf{Q}} & \mathbf{0} \end{pmatrix} \in \mathbb{R}^{m\times m}, \quad \mathbf{U} := \hat{\mathbf{U}}\begin{pmatrix} \mathbf{I}_{k-r} & \mathbf{0} \\ (\bar{\bar{\mathbf{R}}}^{-1}\bar{\bar{\mathbf{G}}})^\top & \mathbf{I}_r \end{pmatrix} \in \mathbb{R}^{k\times k}.$$

To complete the proof, we observe that (2.25) implies

$$\mathbf{M} = \hat{\mathbf{U}}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}} \end{pmatrix}\hat{\mathbf{U}}^\top = \mathbf{U}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}} \end{pmatrix}\mathbf{U}^\top$$

since the second factor in the definition of $\mathbf{U}$ is block lower triangular and satisfies

$$\begin{pmatrix} \mathbf{I}_{k-r} & \mathbf{0} \\ (\bar{\bar{\mathbf{R}}}^{-1}\bar{\bar{\mathbf{G}}})^\top & \mathbf{I}_r \end{pmatrix}\begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}} \end{pmatrix}\begin{pmatrix} \mathbf{I}_{k-r} & \bar{\bar{\mathbf{R}}}^{-1}\bar{\bar{\mathbf{G}}} \\ \mathbf{0} & \mathbf{I}_r \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \bar{\mathbf{M}} \end{pmatrix}. \quad \blacksquare$$

Remark 2.4 Consider equations of motion (2.16) with linear holonomic constraints $\mathbf{0} = \mathbf{G}\boldsymbol{q} - \boldsymbol{z}(t)$ and constant matrices $\mathbf{M}, \mathbf{G}$ that satisfy the assumptions of Lemma 2.3 with $k = n_{\boldsymbol{q}}$, $m = n_{\boldsymbol{\lambda}}$. The matrix factorization (2.24) suggests to multiply the dynamical equations (2.16a) and the constraint equations (2.16b) by $\mathbf{U}^{-1}$ and $\mathbf{Q}^{-1}$, respectively, to decompose the 2×2 block system (2.17) into

$$\bar{\bar{\boldsymbol{\lambda}}} = \bar{\bar{\boldsymbol{f}}}, \tag{2.27a}$$

$$\begin{pmatrix} \bar{\mathbf{M}} & \bar{\mathbf{G}}^\top \\ \bar{\mathbf{G}} & \mathbf{0} \end{pmatrix}\begin{pmatrix} \ddot{\bar{\boldsymbol{q}}} \\ \bar{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{f}} \\ -\bar{\boldsymbol{g}}_{\boldsymbol{qq}}(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}) - 2\bar{\boldsymbol{g}}_{t\boldsymbol{q}}\dot{\boldsymbol{q}} - \bar{\boldsymbol{g}}_{tt} \end{pmatrix} = \begin{pmatrix} \bar{\boldsymbol{f}} \\ \ddot{\bar{\boldsymbol{z}}} \end{pmatrix}, \tag{2.27b}$$

$$\mathbf{0} = -\hat{\boldsymbol{g}}_{\boldsymbol{qq}}(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}) - 2\hat{\boldsymbol{g}}_{t\boldsymbol{q}}\dot{\boldsymbol{q}} - \hat{\boldsymbol{g}}_{tt} = \ddot{\hat{\boldsymbol{z}}}, \tag{2.27c}$$

$$\ddot{\bar{\bar{\boldsymbol{q}}}} = -\bar{\bar{\boldsymbol{g}}}_{\boldsymbol{qq}}(\dot{\boldsymbol{q}},\dot{\boldsymbol{q}}) - 2\bar{\bar{\boldsymbol{g}}}_{t\boldsymbol{q}}\dot{\boldsymbol{q}} - \bar{\bar{\boldsymbol{g}}}_{tt} = \dddot{\bar{\bar{\boldsymbol{z}}}} \tag{2.27d}$$

with

$$\mathbf{U}^{-1}\boldsymbol{f} = \begin{pmatrix} \bar{\bar{\boldsymbol{f}}} \\ \bar{\boldsymbol{f}} \end{pmatrix}, \quad \mathbf{Q}^{-1}\boldsymbol{g} = \begin{pmatrix} \bar{\boldsymbol{g}} \\ \hat{\boldsymbol{g}} \\ \bar{\bar{\boldsymbol{g}}} \end{pmatrix}, \quad \mathbf{Q}^{-1}\boldsymbol{z} = \begin{pmatrix} \bar{\boldsymbol{z}} \\ \hat{\boldsymbol{z}} \\ \bar{\bar{\boldsymbol{z}}} \end{pmatrix}, \quad \mathbf{U}^\top\boldsymbol{q} = \begin{pmatrix} \bar{\bar{\boldsymbol{q}}} \\ \bar{\boldsymbol{q}} \end{pmatrix}, \quad \mathbf{Q}^\top\boldsymbol{\lambda} = \begin{pmatrix} \bar{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\lambda}} \\ \bar{\bar{\boldsymbol{\lambda}}} \end{pmatrix},$$

functions $\bar{\boldsymbol{q}}, \bar{\boldsymbol{f}} \in \mathbb{R}^r$, functions $\bar{\bar{\boldsymbol{q}}}, \bar{\bar{\boldsymbol{\lambda}}}, \bar{\bar{\boldsymbol{f}}}, \bar{\bar{\boldsymbol{g}}}, \bar{\bar{\boldsymbol{z}}} \in \mathbb{R}^{k-r}$, functions $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{g}}, \bar{\boldsymbol{z}} \in \mathbb{R}^{s-(k-r)}$, functions $\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{g}}, \hat{\boldsymbol{z}} \in \mathbb{R}^{m-s}$ and $r = \text{rank } \mathbf{M}$, $s = \text{rank } \mathbf{G}$.

If the mass matrix $\mathbf{M}$ is symmetric, positive definite and $\mathbf{G}$ has full rank, then we have $\boldsymbol{q} = \bar{\boldsymbol{q}}$, $\boldsymbol{\lambda} = \bar{\boldsymbol{\lambda}}$ and the 2×2 block system (2.17) coincides with (2.27b). If $\mathbf{M}$ is rank deficient, then $k - r$ components of the Lagrange multipliers $\boldsymbol{\lambda}$ are explicitly defined by the $k-r$ algebraic equations (2.27a) that do not depend on any derivatives of the constraint function $\boldsymbol{g}$, see Example 2.5. Furthermore, there are $k - r$ second order ODEs (2.27d) for $k - r$ components of $\boldsymbol{q}$. The solution components $\bar{\boldsymbol{q}} \in \mathbb{R}^r$ and $\bar{\boldsymbol{\lambda}} \in \mathbb{R}^{s-(k-r)}$ are defined by the 2×2 block system (2.27b) with the symmetric, positive definite reduced mass matrix $\bar{\mathbf{M}}$ and a reduced constraint matrix $\bar{\mathbf{G}}$ that has full rank $s - (k - r)$.

A rank-deficient constraint matrix $\mathbf{G}$ indicates holonomic constraints (2.16b) that are either redundant or inconsistent. In (2.27), this fact is reflected by $m-s$ equations $\ddot{\hat{\boldsymbol{z}}}(t) = \mathbf{0}$, see (2.27c). If the compatibility conditions (2.27c) are violated, then there is no solution of the equations of motion since the holonomic constraints $\mathbf{0} = \mathbf{G}\boldsymbol{q} - \boldsymbol{z}(t)$ are not consistent.

For redundant constraints, the position coordinates $\boldsymbol{q}$ are uniquely defined by the solution $(\bar{\boldsymbol{q}}, \bar{\bar{\boldsymbol{q}}})$ of (2.27b,d) and the compatibility conditions (2.27c) are satisfied in the whole time interval of interest. Equation (2.27a,b) define $s = \text{rank } \mathbf{G}$ components of the Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^m$ with $m = n_{\boldsymbol{\lambda}}$. The remaining $m-s$ components are summarized in the vector $\hat{\boldsymbol{\lambda}} \in \mathbb{R}^{m-s}$ that does not at all appear in the decoupled equations of motion (2.27).

In the nonlinear case, the characterization of (consistent) redundant constraints (2.16b) is technically more challenging than in the linear setting of Remark 2.4. To avoid state dependent transformation matrices $\mathbf{U}(t, \boldsymbol{q})$, $\mathbf{Q}(t, \boldsymbol{q})$, we follow a local approach that is tailored to the existence and uniqueness result in Theorem 2.4 below:

Definition 2.2 Consider equations of motion (2.16) with $n_{\boldsymbol{\lambda}}$ holonomic constraints $\boldsymbol{g}(t, \boldsymbol{q}) = \mathbf{0}$ and a constraint matrix $\mathbf{G}(t, \boldsymbol{q}) := (\partial \boldsymbol{g}/\partial \boldsymbol{q})(t, \boldsymbol{q}) \in \mathbb{R}^{n_{\boldsymbol{\lambda}} \times n_{\boldsymbol{q}}}$ that has constant rank in a neighbourhood $\mathscr{U}(t^*, \boldsymbol{q}^*)$ of a given point $(t^*, \boldsymbol{q}^*) \in [t_0, t_{\text{end}}] \times \mathbb{R}^{n_{\boldsymbol{q}}}$:

$$\text{rank } \mathbf{G}(t, \boldsymbol{q}) = s \leq n_{\boldsymbol{\lambda}} \leq n_{\boldsymbol{q}}\,, \quad ((t, \boldsymbol{q}) \in \mathscr{U}(t^*, \boldsymbol{q}^*))\,.$$

The constraints $\boldsymbol{g}(t, \boldsymbol{q}) = \mathbf{0}$ are said to be *redundant* (in $\mathscr{U}(t^*, \boldsymbol{q}^*)$) if $\tilde{\mathbf{Q}}\boldsymbol{g}(t, \boldsymbol{q}) = \mathbf{0}$ implies $\boldsymbol{g}(t, \boldsymbol{q}) = \mathbf{0}$ for any constant matrix $\tilde{\mathbf{Q}} \in \mathbb{R}^{s \times n_{\boldsymbol{\lambda}}}$ with $\text{rank } \left(\tilde{\mathbf{Q}}\mathbf{G}(t^*, \boldsymbol{q}^*)\right) = s$.

Theorem 2.4 *Consider equations of motion (2.16) with functions $\mathbf{M}, \boldsymbol{f}, \boldsymbol{g}$ satisfying all assumptions of Theorem 2.2 except the full rank assumption on $\mathbf{G}(t_0, q_0)$.*

(a) If the holonomic constraints (2.16b) are redundant in a neighbourhood $\mathscr{U}_0$ of $(t_0, \boldsymbol{q}_0)$, then there is a vector $\boldsymbol{\lambda}_0 \in \mathbb{R}^{n_{\boldsymbol{\lambda}}}$ such that the initial value problem (2.20) for the constrained system (2.16) is locally solvable. The solution $\boldsymbol{q}(t)$ is locally uniquely defined and independent of the choice of $\boldsymbol{\lambda}_0$.

(b) With these assumptions, the differentiation index and the perturbation index of (2.16) are bounded by three. For symmetric, positive definite mass matrices $\mathbf{M}(t_0, \boldsymbol{q}_0)$*, the variables* $\boldsymbol{q}$ *and* $\dot{\boldsymbol{q}}$ *are solutions of an equivalent index-3 DAE in Hessenberg form.*

Proof Applying Lemma 2.3 with (constant) matrices $\mathbf{M} := \mathbf{M}(t_0, \boldsymbol{q}_0)$, $\mathbf{G} := \mathbf{G}(t_0, \boldsymbol{q}_0)$, we get the matrix decomposition (2.24) and (constant) non-singular transformation matrices $\mathbf{U}$ and $\mathbf{Q}$.

The idea of the proof is to delete in (2.16) all terms corresponding to the fourth block row and to the fourth block column of the 5×5 block matrix in (2.24) and to show that the solution of this reduced system solves the original equations of motion (2.16) as well. We define

$$\tilde{\mathbf{Q}} := \begin{pmatrix} \mathbf{I}_{s-(k-r)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_{k-r} \end{pmatrix} \mathbf{Q}^{-1} \in \mathbb{R}^{s \times m}, \quad \tilde{\boldsymbol{g}}(t, \boldsymbol{q}) := \tilde{\mathbf{Q}} \boldsymbol{g}(t, \boldsymbol{q})$$

with $k = n_{\boldsymbol{q}}$, $m = n_{\boldsymbol{\lambda}}$, $r = \text{rank } \mathbf{M}(t_0, \boldsymbol{q}_0)$, $s = \text{rank } \mathbf{G}(t_0, \boldsymbol{q}_0)$ and get

$$\tilde{\mathbf{G}}(t_0, \boldsymbol{q}_0) := \frac{\partial \tilde{\boldsymbol{g}}}{\partial \boldsymbol{q}}(t_0, \boldsymbol{q}_0) = \tilde{\mathbf{Q}} \mathbf{G}(t_0, \boldsymbol{q}_0) = \begin{pmatrix} \mathbf{0} & \bar{\mathbf{G}}(t_0, \boldsymbol{q}_0) \\ \mathbf{I}_{k-r} & \mathbf{0} \end{pmatrix} \mathbf{U}^{\top} \tag{2.28}$$

with a matrix $\bar{\mathbf{G}}(t_0, \boldsymbol{q}_0) \in \mathbb{R}^{(s-(k-r)) \times (k-r)}$ of full rank $s - (k - r)$, see Lemma 2.3.

Equation (2.28) shows that the left-multiplication by $\tilde{\mathbf{Q}}$ selects $s = \text{rank } \mathbf{G}$ linearly independent row vectors of $\mathbf{G}(t_0, \boldsymbol{q}_0)$, i.e., all $m = n_{\boldsymbol{\lambda}}$ row vectors of $\mathbf{G}(t_0, \boldsymbol{q}_0)$ may be represented by a linear combination of the row vectors of matrix $\tilde{\mathbf{G}}(t_0, \boldsymbol{q}_0) \in \mathbb{R}^{s \times k}$ and there is a matrix $\tilde{\tilde{\mathbf{Q}}}(t_0, \boldsymbol{q}_0) \in \mathbb{R}^{m \times s}$ such that

$$\mathbf{G}(t_0, \boldsymbol{q}_0) = \tilde{\tilde{\mathbf{Q}}}(t_0, \boldsymbol{q}_0) \tilde{\mathbf{G}}(t_0, \boldsymbol{q}_0) \,. \tag{2.29}$$

The continuity of the matrix valued functions $\mathbf{G}(t, \boldsymbol{q})$ and $\tilde{\mathbf{G}}(t, \boldsymbol{q})$ implies that there is a (sufficiently small) neighbourhood $\mathscr{U}_0$ of $(t_0, \boldsymbol{q}_0)$ such that $\text{rank } \tilde{\mathbf{G}}(t, \boldsymbol{q}) = \text{rank } \mathbf{G}(t, \boldsymbol{q}) = s$ and

$$\mathbf{G}(t, \boldsymbol{q}) = \tilde{\tilde{\mathbf{Q}}}(t, \boldsymbol{q}) \tilde{\mathbf{G}}(t, \boldsymbol{q}) \tag{2.30}$$

with $\tilde{\tilde{\mathbf{Q}}}(t, \boldsymbol{q}) \in \mathbb{R}^{m \times s}$ for all $(t, \boldsymbol{q}) \in \mathscr{U}_0$. This matrix $\tilde{\tilde{\mathbf{Q}}}(t, \boldsymbol{q})$ has to have full rank s since the left-multiplication of (2.30) by $\tilde{\mathbf{Q}}$ results in a matrix of rank s.

The matrix factorization (2.30) allows to express the constraint forces $-\mathbf{G}^{\top}(t, \boldsymbol{q})\, \boldsymbol{\lambda}$ in terms of $-\tilde{\mathbf{G}}^{\top}(t, \boldsymbol{q})\, \tilde{\boldsymbol{\lambda}}$ with $\tilde{\boldsymbol{\lambda}} = \tilde{\tilde{\mathbf{Q}}}^{\top}(t, \boldsymbol{q})\, \boldsymbol{\lambda} \in \mathbb{R}^s$. On the other hand, we have $-\mathbf{G}^{\top}(t, \boldsymbol{q})\, \boldsymbol{\lambda} = -\tilde{\mathbf{G}}^{\top}(t, \boldsymbol{q})\, \tilde{\boldsymbol{\lambda}}$ for all $\boldsymbol{\lambda} \in \mathbb{R}^m$ satisfying

$$\boldsymbol{\lambda} = [\tilde{\tilde{\mathbf{Q}}}(\tilde{\tilde{\mathbf{Q}}}^{\top} \tilde{\tilde{\mathbf{Q}}})^{-1}](t, \boldsymbol{q})\, \tilde{\boldsymbol{\lambda}} + \hat{\boldsymbol{\lambda}} \tag{2.31}$$

with some $\hat{\boldsymbol{\lambda}} \in \ker \tilde{\tilde{\mathbf{Q}}}^{\top}(t,\boldsymbol{q})$. The nullspace of $\tilde{\tilde{\mathbf{Q}}}^{\top}(t,\boldsymbol{q})$ has dimension $m-s$. It is non-trivial if the constraint matrix $\mathbf{G}(t,\boldsymbol{q})$ is rank deficient. In that case, the variables $\hat{\boldsymbol{\lambda}}$ are left undefined by the constrained system (2.16), see also the corresponding discussion for systems with constant matrices $\mathbf{M}$ and $\mathbf{G}$ in Remark 2.4.

Because of (2.29), we have $\ker \tilde{\mathbf{G}}(t_0,\boldsymbol{q}_0) \subset \ker \mathbf{G}(t_0,\boldsymbol{q}_0)$ and the mass matrix $\mathbf{M}(t_0,\boldsymbol{q}_0)$ is positive definite at the nullspace of $\tilde{\mathbf{G}}(t_0,\boldsymbol{q}_0)$. Furthermore, function $\tilde{\boldsymbol{g}}(t,\boldsymbol{q})$ satisfies the smoothness assumptions of Theorem 2.2 since the matrix decomposition (2.24) was evaluated for matrices $\mathbf{M}$, $\mathbf{G}$ with *fixed* arguments $t=t_0$, $\boldsymbol{q}=\boldsymbol{q}_0$. Therefore, we may apply Theorem 2.2 to the reduced system

$$\mathbf{M}(t,\boldsymbol{q})\ddot{\boldsymbol{q}} = \boldsymbol{f}(t,\boldsymbol{q},\dot{\boldsymbol{q}}) - \tilde{\mathbf{G}}^{\top}(t,\boldsymbol{q})\,\tilde{\boldsymbol{\lambda}}\,, \tag{2.32a}$$

$$\mathbf{0} = \tilde{\boldsymbol{g}}(t,\boldsymbol{q}) \tag{2.32b}$$

and get a locally uniquely defined solution $\boldsymbol{q}(t)$ with initial values $\boldsymbol{q}(t_0)=\boldsymbol{q}_0$, $\dot{\boldsymbol{q}}(t_0)=\dot{\boldsymbol{q}}_0$. In $\mathscr{U}_0$, the s linearly independent constraints (2.32b) of the reduced system imply the $m \geq s$ redundant constraints (2.16b) of the original equations of motion, see Definition 2.2. Furthermore, the reduced system (2.32) defines unique Lagrange multipliers $\tilde{\boldsymbol{\lambda}}(t) \in \mathbb{R}^s$ and the set of all solutions $\boldsymbol{\lambda}(t) \in \mathbb{R}^m$ according to (2.31).

To prove part (b) of the Theorem, we apply the index analysis of Remark 2.3 to the reduced system (2.32). ■

3 From Constrained Mechanical Systems to Multibody System Dynamics

Mechanical multibody systems are composed of a finite number of rigid or flexible bodies being connected by *joints* that restrict the relative motion of bodies w.r.t. each other and by *force elements* like springs, dampers or actuators that cause forces and momenta acting on the interconnected bodies but do not restrict the degrees of freedom of their relative motion. The mass of a multibody system is concentrated in the bodies and the connecting elements are idealized to be massless. After space discretization of the flexible components, the mechanical state of the system may be characterized by elements of a finite dimensional configuration space that describe the position and orientation of all bodies and the elastic deformation of the flexible parts.

The equations of motion follow systematically from principles of classical mechanics that result in linearly implicit systems of second order differential equations. Efficient time integration methods in multibody numerics are essentially based on the specific mathematical structure of these model equations. Discussing this structure, we started in Sect. 2 at a rather basic level with constrained systems of point masses. The modelling of rigid body systems is substantially more complex

since the orientation of the bodies in 2-D or 3-D has to be taken into account which may result in nonlinear configuration spaces, see Sect. 3.1. There is a rich literature on the general structure of model equations in multibody system dynamics that is shortly summarized in Sect. 3.2. Finally, we consider in Sect. 3.3 some specific algorithms of multibody dynamics that exploit the topology of a multibody system model to speed up the evaluation of the model equations.

3.1 Configuration of Rigid Body Systems

The configuration of rigid bodies is characterized by their position and orientation in space. For simplicity, we restrict ourselves in the present section to the discussion of systems in $\mathbb{R}^3$ (*spatial* systems). *Planar* systems may be considered as a special case of this general setting with position coordinates being restricted to a two-dimensional subspace.

In $\mathbb{R}^3$, the position of body $(\bullet)^{(i)}$ is described by coordinates $\boldsymbol{x}^{(i)} \in \mathbb{R}^3$ and its orientation may be represented conveniently by a rotation matrix

$$\mathbf{R}^{(i)} \in \mathrm{SO}(3) = \{\, \mathbf{R} \in \mathbb{R}^{3\times 3} \,:\, \mathbf{R}^\top \mathbf{R} = \mathbf{I}_3 \,,\ \det \mathbf{R} = +1 \,\} \,.$$

The special orthogonal group SO(3) is a subgroup of the general linear group $\mathrm{GL}(3) = \{\, \mathbf{A} \,:\, \mathbf{A} \in \mathbb{R}^{3\times 3} \,:\, \det \mathbf{A} \neq 0 \,\}$ and forms a three-dimensional differentiable manifold in $\mathbb{R}^9$. *Lie group* theory provides the analytical framework for differential equations on such manifolds with group structure. The interested reader is referred to [49, Chap. IV] for a compact introduction and to [53] for a comprehensive survey of analytical and numerical aspects of differential equations on finite dimensional Lie groups.

Remark 3.1

(a) The Lie group structure of configuration spaces may be exploited explicitly in the time integration of the equations of motion, see, e.g., [22, 27, 32, 85]. Position vector $\boldsymbol{x} \in \mathbb{R}^3$ and rotation matrix $\mathbf{R} \in \mathrm{SO}(3)$ are either combined in the direct product $G = \mathrm{SO}(3) \times \mathbb{R}^3$ with group operation

$$(\mathbf{R}_a, \boldsymbol{x}_a) \circ (\mathbf{R}_b, \boldsymbol{x}_b) = (\mathbf{R}_a \mathbf{R}_b, \boldsymbol{x}_a + \boldsymbol{x}_b)$$

or in the semi-direct product $G = \mathrm{SE}(3) := \mathrm{SO}(3) \ltimes \mathbb{R}^3$ with group operation

$$(\mathbf{R}_a, \boldsymbol{x}_a) \circ (\mathbf{R}_b, \boldsymbol{x}_b) = (\mathbf{R}_a \mathbf{R}_b, \mathbf{R}_a \boldsymbol{x}_b + \boldsymbol{x}_a) \,,$$

see [29] and the more detailed discussions in [17] and [65]. With these notations, the configuration space of a rigid N-body system is given by the direct products $(\mathrm{SO}(3) \times \mathbb{R}^3)^N$ or $(\mathrm{SE}(3))^N$, respectively.

(b) The inherent nonlinear structure of the configuration space results in nontrivial kinematic relations that express the time derivatives of the position coordinates $q = (\boldsymbol{x}, \mathbf{R}) \in G$ in terms of velocity coordinates $\boldsymbol{v}$. The Lie group structure of G implies $\dot{q}(t) \in T_{q(t)}G$ with T_qG denoting the tangent space. Taking into account the linear structure of T_qG, the velocity coordinates $\boldsymbol{v}$ are defined by elements of a *linear* space $\mathbb{R}^k$. For a single rigid body, we get

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{u}(t) = \mathbf{R}(t)\mathbf{U}(t) \tag{3.1a}$$

with $\boldsymbol{u}(t)$ and $\mathbf{U}(t)$ denoting the translation velocity w.r.t. an inertial and a body-attached frame, respectively. The corresponding angular velocities $\boldsymbol{\omega}$ (inertial frame) and $\boldsymbol{\Omega}$ (body-attached frame) are related by

$$\tilde{\boldsymbol{\omega}}(t) = \mathbf{R}(t)\tilde{\boldsymbol{\Omega}}(t)\mathbf{R}^\top(t)$$

with $\widetilde{(\bullet)} : \mathbb{R}^3 \to \mathfrak{so}(3) = \{\mathbf{A} \in \mathbb{R}^{3\times 3} : \mathbf{A} + \mathbf{A}^\top = \mathbf{0}\}$ denoting the *tilde operator* that maps $\boldsymbol{\Omega} \in \mathbb{R}^3$ to the skew-symmetric matrix

$$\widetilde{\boldsymbol{\Omega}} := \begin{pmatrix} 0 & -\Omega_3 & \Omega_2 \\ \Omega_3 & 0 & -\Omega_1 \\ -\Omega_2 & \Omega_1 & 0 \end{pmatrix}$$

and represents the vector product $\boldsymbol{p} \times \boldsymbol{q}$ in $\mathbb{R}^3$ in the sense that $\tilde{\boldsymbol{p}}\boldsymbol{q} = \boldsymbol{p} \times \boldsymbol{q}$ for any vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^3$. The kinematic relations for $\mathbf{R}$ are given by

$$\dot{\mathbf{R}}(t) = \tilde{\boldsymbol{\omega}}(t)\mathbf{R}(t) = \mathbf{R}(t)\tilde{\boldsymbol{\Omega}}(t)\,. \tag{3.1b}$$

Equation (3.1) allow to represent the time derivative of $q = (\boldsymbol{x}, \mathbf{R}) \in G$ by a velocity vector $\boldsymbol{v} \in \mathbb{R}^6$ being composed of translation velocity and angular velocity (either in the inertial or in the body-attached frame).

The structural difference between the kinematic relations (3.1) and the more classical setting $\dot{\boldsymbol{q}}(t) = \boldsymbol{v}(t)$ in linear spaces, see (2.22a), is given by the Lie group ODE (3.1b) on SO(3). In the following, we will discuss analytical and numerical aspects of these equations and will assume that the angular velocities are defined in the body-attached frame. As a typical model problem, we consider a slowly rotating heavy top with its tip being fixed to the origin:

Example 3.1 In the gravity field, the kinetic and potential energy of a spinning top of mass m are given by [29]

$$T = \frac{1}{2}m\dot{\boldsymbol{x}}^\top\dot{\boldsymbol{x}} + \frac{1}{2}\boldsymbol{\Omega}^\top\mathbf{J}\boldsymbol{\Omega}\,, \quad U = -\boldsymbol{x}^\top m\boldsymbol{\gamma} \quad \text{with} \quad \boldsymbol{\gamma} = \begin{pmatrix} 0 \\ 0 \\ -g_{\text{grav}} \end{pmatrix}$$

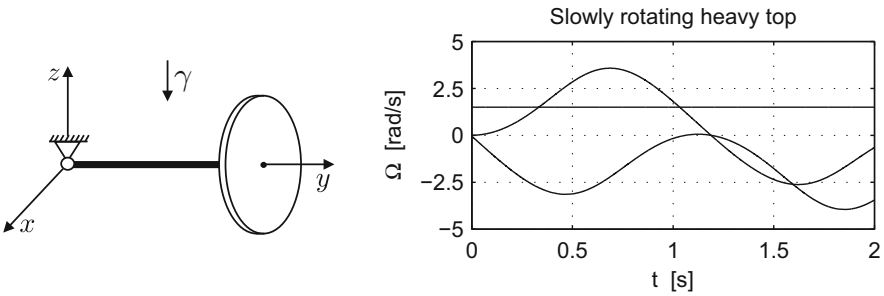


Fig. 5 Configuration and angular velocity of a slowly rotating heavy top [27], see also [45]

and the gravitational acceleration constant g_{grav}. Here, the tensor of inertia $\mathbf{J}$ is defined w.r.t. the center of mass in the body-attached frame. If this center of mass has position $\mathbf{X} \in \mathbb{R}^3$ for the reference configuration $\mathbf{R} = \mathbf{I}_3$, then its current position in the inertial frame is given by $\boldsymbol{x}(t) = \mathbf{R}(t)\mathbf{X}$ since the tip of the top is fixed at the origin, see Fig. 5. This constraint implies $\dot{\boldsymbol{x}}(t) = \dot{\mathbf{R}}(t)\mathbf{X} = \mathbf{R}(t)\tilde{\boldsymbol{\Omega}}(t)\mathbf{X}$, see (3.1b), and we get $\dot{\boldsymbol{x}} = -\mathbf{R}\tilde{\mathbf{X}}\boldsymbol{\Omega}, \dot{\boldsymbol{x}}^\top = -\boldsymbol{\Omega}^\top\tilde{\mathbf{X}}^\top\mathbf{R}^\top = \boldsymbol{\Omega}^\top\tilde{\mathbf{X}}\mathbf{R}^\top$ and

$$T = \frac{1}{2}\boldsymbol{\Omega}^\top(\mathbf{J} - m\tilde{\mathbf{X}}\tilde{\mathbf{X}})\boldsymbol{\Omega}\,, \quad U = -\mathbf{X}^\top\mathbf{R}^\top m\boldsymbol{\gamma}\,.$$

In Sect. 2.1, we discussed the derivation of the equations of motion in linear configuration spaces using Hamilton's principle of least action. For nonlinear configuration spaces, the nonlinear kinematic relations (3.1b) have to be taken into account [29]. For the heavy top problem we obtain equations of motion

$$\dot{\mathbf{R}} = \mathbf{R}\tilde{\boldsymbol{\Omega}}\,, \tag{3.2a}$$

$$\bar{\mathbf{J}}\dot{\boldsymbol{\Omega}} + \boldsymbol{\Omega} \times \bar{\mathbf{J}}\boldsymbol{\Omega} = \mathbf{X} \times \mathbf{R}^\top m\boldsymbol{\gamma} \tag{3.2b}$$

with $\bar{\mathbf{J}} := \mathbf{J} - m\tilde{\mathbf{X}}\tilde{\mathbf{X}}$ denoting the moment of inertia w.r.t. the origin [27]. The right plot of Fig. 5 shows the angular velocity $\boldsymbol{\Omega}(t)$ for model parameters $m = 15.0\,\text{kg}$, $\bar{\mathbf{J}} = \text{diag}(15.234375,\, 0.46875,\, 15.234375)\,\text{kg}\,\text{m}^2$, $g_{\text{grav}} = 9.81\,\text{m/s}^2$ and initial values

$$\mathbf{R}(0) = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{pmatrix}\,, \quad \boldsymbol{\Omega}(0) = \begin{pmatrix} 0 \\ 1.5 \\ -0.0461538 \end{pmatrix} \frac{\text{rad}}{\text{s}}\,.$$

The direct time discretization of Lie group ODEs by Lie group integrators is a challenging topic of active research. In practical applications it is, however, more common to use parametrizations of the rotation matrix by elements of a linear space.

Remark 3.2

(a) There is no *global* parametrization of SO(3) by elements of $\mathbb{R}^3$ but small deviations from a nominal state may be described very efficiently by three Euler angles [76]. Euler angles define a decomposition of the rotation matrix into a sequence of three *elementary* rotations about axes of coordinates. A common sequence of such elementary rotations is given by

$$\mathbf{R}(\boldsymbol{q}_\mathbf{R}) = \begin{pmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{pmatrix} \begin{pmatrix} \cos\phi & -\sin\phi & 0 \\ \sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

with angles ϕ (precession), θ (nutation) and ψ (spin) that are summarized in a parameter vector $\boldsymbol{q}_\mathbf{R} = (\phi, \theta, \psi)^\top \in \mathbb{R}^3$.

For $\theta = \theta^* = 0$, this parametrization gets singular since only the sum $\phi + \psi$ is well defined in this case and there is a continuum of parameter vectors $\boldsymbol{q}_\mathbf{R}$ yielding one and the same rotation matrix $\mathbf{R}(\boldsymbol{q}_\mathbf{R})$. In engineering applications, such singular configurations are avoided switching to an alternative sequence of elementary rotations whenever $|\theta|$ gets too small [76].

(b) Beyond the singularities, we may insert the parametrization $\mathbf{R}(\boldsymbol{q}_\mathbf{R}(t))$ into the kinematic relations (3.1b) to get a linear relation between $\dot{\boldsymbol{q}}_\mathbf{R}$ and the angular velocity $\boldsymbol{\Omega}$:

$$\sum_{j=1}^{3} \frac{\partial \mathbf{R}}{\partial q_{\mathbf{R},j}}\big(\boldsymbol{q}_\mathbf{R}(t)\big)\dot{q}_{\mathbf{R},j}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\mathbf{R}\big(\boldsymbol{q}_\mathbf{R}(t)\big) = \mathbf{R}\big(\boldsymbol{q}_\mathbf{R}(t)\big)\tilde{\boldsymbol{\Omega}}(t)\,.$$

This equation be summarized in matrix–vector form

$$\mathbf{H}_0\big(\boldsymbol{q}_\mathbf{R}(t)\big)\dot{\boldsymbol{q}}_\mathbf{R}(t) = \boldsymbol{\Omega}(t) \tag{3.3}$$

using the matrix valued function $\mathbf{H}_0(\boldsymbol{q}_\mathbf{R}) = (\,h_{ij}(\boldsymbol{q}_\mathbf{R})\,)_{i,j} \in \mathbb{R}^{3\times 3}$ that is defined by its elements

$$h_{ij}(\boldsymbol{q}_\mathbf{R}) := \frac{1}{2}\Big(\big(\mathbf{R}^\top(\boldsymbol{q}_\mathbf{R})\frac{\partial \mathbf{R}}{\partial q_{\mathbf{R},j}}(\boldsymbol{q}_\mathbf{R})\big)_{l_{i+2},l_{i+1}} - \big(\mathbf{R}^\top(\boldsymbol{q}_\mathbf{R})\frac{\partial \mathbf{R}}{\partial q_{\mathbf{R},j}}(\boldsymbol{q}_\mathbf{R})\big)_{l_{i+1},l_{i+2}}\Big)$$

with indices $l_1 = l_4 = 1$, $l_2 = l_5 = 2$, $l_3 = 3$. Straightforward computations yield [76]

$$\mathbf{H}_0(\boldsymbol{q}_\mathbf{R}) = \mathbf{H}_0(\phi, \theta, \psi) = \begin{pmatrix} -\cos\phi\sin\theta & \sin\phi & 0 \\ \sin\phi\sin\theta & \cos\phi & 0 \\ \cos\theta & 0 & 1 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

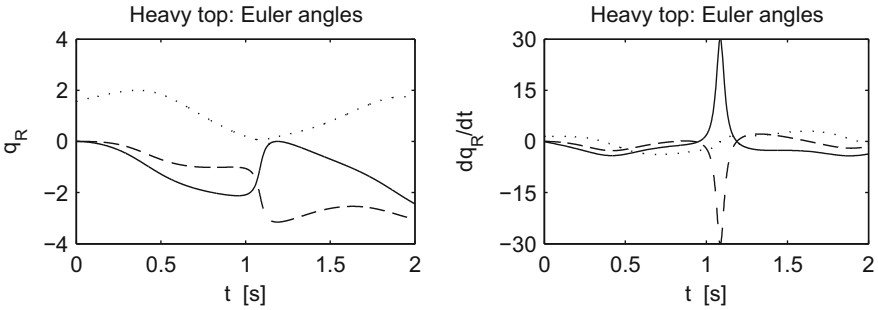


Fig. 6 Parametrization of the heavy top model by Euler angles: precession ϕ (*dashed line*), nutation θ (*dotted line*), spin ψ (*solid line*)

(c) The linear relation $\mathbf{H}_0(\boldsymbol{q}_\mathbf{R})\dot{\boldsymbol{q}}_\mathbf{R} = \boldsymbol{\Omega}$, see (3.3), may be used to eliminate for all bodies $(\bullet)^{(i)}$ the angular velocity $\boldsymbol{\Omega}^{(i)}$ and its time derivative in the equations of motion resulting in a second order system (2.3) with configuration variables $\boldsymbol{q} \in \mathbb{R}^{6N}$ being composed of the position coordinates $\boldsymbol{x}^{(i)}$ and the vectors of Euler angles $\boldsymbol{q}_\mathbf{R}^{(i)}$ for all N bodies in the rigid body system.

(d) Alternatively, the kinematic relations (3.1b) may be substituted by

$$\dot{\boldsymbol{q}}_\mathbf{R}(t) = \mathbf{H}_0^{-1}\big(\boldsymbol{q}_\mathbf{R}(t)\big)\boldsymbol{\Omega}(t) \tag{3.4}$$

resulting in a system of first order differential equations in terms of position coordinates $\boldsymbol{q} \in \mathbb{R}^{n_q}$ and velocity coordinates $\boldsymbol{v} \in \mathbb{R}^{n_v}$.

For the heavy top model of Example 3.1, these coordinates are given by $\boldsymbol{q} = \boldsymbol{q}_R$, $\boldsymbol{v} = \boldsymbol{\Omega}$ with $n_q = n_v = 3$ and position coordinates $\boldsymbol{q} = \boldsymbol{q}_\mathbf{R} = (\phi, \theta, \psi)^\top$ that are shown in the left plot of Fig. 6. The nutation $\theta(t)$ has its minimum value $\theta(t^*) = 0.059\,\text{rad} = 3.4°$ at $t = t^* \approx 1.1\,\text{s}$ without reaching the singular configuration at $\theta^* = 0$. The rapid changes of ϕ and ψ in a neighbourhood of $t = t^*$ may, however, result in (very) small time step sizes in an error controlled variable step size solver. The right plots of Figs. 5 and 6 illustrate that $\max_t \|\dot{\boldsymbol{q}}_\mathbf{R}(t)\|$ is larger by a factor of 10 than the corresponding maximum value $\max_t \|\boldsymbol{\Omega}(t)\|$ of the angular velocity $\boldsymbol{\Omega}$.

(e) For a rigid body system with N bodies, the kinematic equations (3.1a) and (3.4) may be summarized to $\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}$ with position coordinates $\boldsymbol{q}$ being defined by $\boldsymbol{x}^{(i)}$, $\boldsymbol{q}_\mathbf{R}^{(i)}$, $(i = 1, \ldots, N)$, and velocity coordinates $\boldsymbol{v}$ that summarize the corresponding velocity terms $\mathbf{U}^{(i)}$ and $\boldsymbol{\Omega}^{(i)}$ (or their counterparts $\boldsymbol{u}^{(i)}$, $\boldsymbol{\omega}^{(i)}$ in the inertial frame). The equations of motion get the form (2.22) with (2.22a) being substituted by

$$\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}\,. \tag{3.5}$$

The parametrization by Euler angles is quite popular in multibody dynamics but fails systematically for systems with large rotations. In that case, parametrizations without singularities prove to be favourable. According to [85], " ... it is now well established that the optimal singularity free parametrization is defined in terms of the (four) unit quaternion parameters."

Remark 3.3

(a) Unit quaternions may be interpreted as normalized elements of $\mathbb{R}^4$:

$$\mathbb{Q} = \{\, \mathsf{p} = (p_0, p_1, p_2, p_3)^\top \in \mathbb{R}^4 \,:\, \|\mathsf{p}\|_2 = 1 \,\}.$$

Identifiying the *scalar* part p_0 of quaternion p with the quaternion $(p_0, 0, 0, 0)^\top$ and its *vector* part $\boldsymbol{p} = (p_1, p_2, p_3)^\top$ with the quaternion $(0, p_1, p_2, p_3)^\top$, we get $\mathsf{p} = p_0 + \boldsymbol{p}$ and its conjugate $\mathsf{p}^* := p_0 - \boldsymbol{p}$.

(b) The multiplication of two quaternions $\mathsf{p} = p_0 + \boldsymbol{p}$ and $\mathsf{q} = q_0 + \boldsymbol{q}$ is defined by

$$\mathsf{q} * \mathsf{p} = q_0 p_0 - \boldsymbol{q} \cdot \boldsymbol{p} + q_0 \boldsymbol{p} + p_0 \boldsymbol{q} + \boldsymbol{q} \times \boldsymbol{p}$$

and allows a very compact and computationally efficient representation of rotations in terms of unit quaternions [76]. Identifying a given vector $\boldsymbol{w} \in \mathbb{R}^3$ with the quaternion $0 + \boldsymbol{w}$, we get $\boldsymbol{w}^{\mathsf{p}} := \mathbf{R}(\mathsf{p})\boldsymbol{w}$ by

$$\begin{pmatrix} 0 \\ \boldsymbol{w}^{\mathsf{p}} \end{pmatrix} = \begin{pmatrix} 0 \\ \boldsymbol{w} \end{pmatrix}^{\mathsf{p}} := \mathsf{p} * \begin{pmatrix} 0 \\ \boldsymbol{w} \end{pmatrix} * \mathsf{p}^*$$

and the parametrization

$$\mathbf{R}(\mathsf{p}) = \begin{pmatrix} p_0^2 + p_1^2 - p_2^2 - p_3^2 & 2p_1p_2 - 2p_0p_3 & 2p_1p_3 + 2p_0p_2 \\ 2p_1p_2 + 2p_0p_3 & p_0^2 - p_1^2 + p_2^2 - p_3^2 & 2p_2p_3 - 2p_0p_1 \\ 2p_1p_3 - 2p_0p_2 & 2p_2p_3 + 2p_0p_1 & p_0^2 - p_1^2 - p_2^2 + p_3^2 \end{pmatrix}$$

of rotation matrices $\mathbf{R}(\mathsf{p})$ in terms of unit quaternions $\mathsf{p} = (p_0, p_1, p_2, p_3)^\top \in \mathbb{Q}$.

(c) As in Remark 3.2(b), we may express the angular velocity $\boldsymbol{\Omega}$ in terms of the time derivative of the parameter vector, see (3.3):

$$\boldsymbol{\Omega} = \mathbf{H}_0(\mathsf{p})\dot{\mathsf{p}} \quad \text{with} \quad \mathbf{H}_0(\mathsf{p}) = \mathbf{H}_0(p_0, \boldsymbol{p}) = (\,-2\boldsymbol{p},\; 2p_0\mathbf{I} - 2\tilde{\boldsymbol{p}}\,) \in \mathbb{R}^{3\times 4}. \tag{3.6}$$

In that way, the equations of motion are obtained as second order system (2.3) with configuration variables $\boldsymbol{q} \in \mathbb{R}^{7N}$ being composed of the position coordinates $\boldsymbol{x}^{(i)}$ and the vectors of unit quaternions $\mathsf{p}^{(i)}$ for all N bodies in the rigid body system, see also Remark 3.2(c). The normalization of the unit quaternions may be guaranteed by N constraints (2.3b) with $g_i(\boldsymbol{q}) := \|\mathsf{p}^{(i)}\|_2^2 - 1$, $(\, i = 1, \ldots, N\,)$.

(d) The normalization condition for a unit quaternion p implies a hidden constraint

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\left((\mathsf{p}(t))^\top \mathsf{p}(t) - 1\right) = 2(\mathsf{p}(t))^\top \dot{\mathsf{p}}(t) = 2p_0(t)\dot{p}_0(t) + 2(\boldsymbol{p}(t))^\top \dot{\boldsymbol{p}}(t)$$

that may be combined with (3.6) to

$$\begin{pmatrix} 2p_0 & 2\boldsymbol{p}^\top \\ -2\boldsymbol{p} & 2p_0\mathbf{I} - 2\tilde{\boldsymbol{p}} \end{pmatrix} \dot{\mathsf{p}} = \begin{pmatrix} 0 \\ \boldsymbol{\Omega} \end{pmatrix}.$$

These four linear equations in terms of $\dot{\mathsf{p}} = (\dot{p}_0, \dot{p}_1, \dot{p}_2, \dot{p}_3)^\top$ yield kinematic relations

$$\dot{\mathsf{p}}(t) = \mathbf{H}\big(\mathsf{p}(t)\big)\boldsymbol{\Omega}(t) \quad \text{with} \quad \mathbf{H}(\mathsf{p}) := \frac{1}{2}\begin{pmatrix} -\boldsymbol{p}^\top \\ p_0\mathbf{I} + \tilde{\boldsymbol{p}} \end{pmatrix} \in \mathbb{R}^{4\times 3}, \tag{3.7}$$

position coordinates $\mathsf{p} \in \mathbb{R}^4$ and velocity coordinates $\boldsymbol{\Omega} \in \mathbb{R}^3$, see [76].

Figure 7 shows simulation results for the heavy top model of Example 3.1. The components of p vary smoothly and without singularities in time. The comparison of the right plots in Figs. 5 and 7 shows that the maximum amplitude of $\dot{\mathsf{p}}$ is of the same size as the one of $\boldsymbol{\Omega}$. In time integration, the normalization condition $\|\mathsf{p}\|_2 = 1$ may be enforced conveniently re-normalizing the numerical solution $\mathsf{p}_n \approx \mathsf{p}(t_n)$ after each successful time step.

(e) For a rigid body system with N bodies, the kinematic equations (3.1a) and (3.7) are again summarized in compact form: $\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}$. Note that the different dimensions of p and $\boldsymbol{\Omega}$ result in a *rectangular* matrix $\mathbf{H}(\boldsymbol{q}) \in \mathbb{R}^{7N\times 6N}$ since the position coordinates $\boldsymbol{q}$ are defined by $\boldsymbol{x}^{(i)}$, $\mathsf{p}^{(i)}$, $(i = 1, \ldots, N)$, and the velocity coordinates $\boldsymbol{v}$ are composed of the velocity terms $\mathbf{U}^{(i)}$ and $\boldsymbol{\Omega}^{(i)}$ (or their counterparts $\boldsymbol{u}^{(i)}$, $\boldsymbol{\omega}^{(i)}$ in the inertial frame). As in Remark 3.2(e), we get equations of motion of the form (2.22) with (2.22a) being substituted by $\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}$, see (3.5).

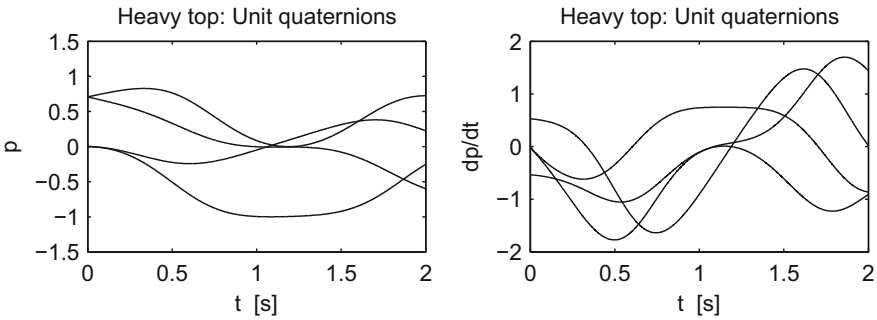


Fig. 7 Parametrization of the heavy top model by unit quaternions. *Left plot*: $\mathsf{p}(t)$, *right plot* $\dot{\mathsf{p}}(t)$

The mathematical structure of the configuration space for flexible bodies is very similar to the one for rigid body systems if the flexible body is discretized in space by finite elements (or finite differences). Following the *finite element approach* to flexible multibody dynamics [45], the configuration variables describe the nodal translations and rotations.

An alternative approach is based on (modal) model reduction and considers small elastic deformations w.r.t. a *floating frame of reference* that describes large translations and rotations of the flexible body in space [77, 78]. Here, the configuration variables of each flexible body are composed of coordinates describing the position and orientation of its (floating) frame of reference and modal coordinates describing the (small) deformations w.r.t. this reference frame. As before, the basic mathematical structure of configuration space and equations of motion is the one that is known from rigid body systems.

For a more detailed discussion of flexible multibody systems, we refer to the rich literature in this field including monographs like [20, 45, 77, 79, 82].

3.2 Model Equations in Multibody System Dynamics

The state variables of a mechanical multibody system model describe the position and orientation of all bodies, the elastic deformation of the flexible components and the internal state of all force elements. Parametrizing the rotation matrices by elements of a linear space we get position coordinates $\boldsymbol{q} \in \mathbb{R}^{n_q}$ with time derivatives that depend linearly on velocity coordinates $\boldsymbol{v} \in \mathbb{R}^{n_v}$, see Sect. 3.1. Position and velocity coordinates have either one and the same dimension $n_q = n_v$ or the dimension of $\boldsymbol{q}$ exceeds the one of $\boldsymbol{v}$ and the position coordinates are subject to $n_q - n_v > 0$ invariants

$$\mathbf{0} = \boldsymbol{\gamma}(\boldsymbol{q}) \tag{3.8}$$

representing, e.g., the normalization of unit quaternions.

The internal state of force elements is characterized by continuous state variables $\boldsymbol{c}(t) \in \mathbb{R}^{n_c}$ and by time-discrete state variables $\boldsymbol{r}_j \in \mathbb{R}^{n_r}$ that remain constant in each sampling interval $[T_j, T_{j+1}) \in [t_0, t_{\text{end}}]$. The state variables represent, e.g., hydraulic and electronic system components or control structures [14, 37]. They are subject to changes according to first order ODEs

$$\dot{\boldsymbol{c}} = \boldsymbol{d}(t, \boldsymbol{q}, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{r}_j, \boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\eta})$$

and (time-)discrete state equations

$$\boldsymbol{r}_{j+1} = \boldsymbol{a}(\boldsymbol{r}_j, \boldsymbol{r}_{j-1}, \ldots, T_{j+1}, \boldsymbol{q}, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \,. \tag{3.9}$$

The right-hand sides $\boldsymbol{d}$ and $\boldsymbol{a}$ depend on t, $\boldsymbol{q}$, $\boldsymbol{v}$, $\boldsymbol{c}$, $\boldsymbol{r}_j$ and on Lagrange multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\eta}$ that correspond to holonomic and to nonholonomic constraints, respectively.

They may depend furthermore on additional algebraic variables $\boldsymbol{s}$ and $\boldsymbol{w}$ that are introduced for a more convenient model setup in industrial applications [14, 73, 84]. Contact point coordinates $\boldsymbol{s} \in \mathbb{R}^{n_s}$ are used in the modelling of contact conditions to determine the position of contact points on the surfaces of contacting bodies. They are implicitly defined by a system of n_s nonlinear equations

$$\boldsymbol{0} = \boldsymbol{h}(t, \boldsymbol{q}, \boldsymbol{s}) \tag{3.10a}$$

with non-singular Jacobian $\partial \boldsymbol{h}/\partial \boldsymbol{s}$. In the same way, coordinates $\boldsymbol{w} \in \mathbb{R}^{n_w}$ are implicitly defined by a system of n_w nonlinear equations

$$\boldsymbol{0} = \boldsymbol{b}(t, \boldsymbol{q}, \boldsymbol{s}, \boldsymbol{v}, \boldsymbol{c}, \boldsymbol{r}_j, \boldsymbol{w}, \boldsymbol{\lambda}, \boldsymbol{\eta}) \tag{3.10b}$$

with non-singular Jacobian $\partial \boldsymbol{b}/\partial \boldsymbol{w}$. Variables of this type are used, e.g., in the modelling of joint friction that results in force vectors $\boldsymbol{f}$ depending nonlinearly on the constraint forces [73].

If there are bodies in the multibody system model that are permanently in contact, then their relative motion is restricted by contact conditions that contribute to the holonomic constraints

$$\boldsymbol{0} = \boldsymbol{g}(t, \boldsymbol{q}, \boldsymbol{s}) \,, \tag{3.11}$$

see [14, 84]. The structure of these constraint equations is slightly more complex than in the classical setting of Sect. 2.1, see (2.1). Formally, the contact point coordinates $\boldsymbol{s}$ in (3.11) could be eliminated applying the implicit function theorem to (3.10a) resulting in

$$\boldsymbol{0} = \bar{\boldsymbol{g}}(t, \boldsymbol{q}) := \boldsymbol{g}\big(t, \boldsymbol{q}, \boldsymbol{s}(t, \boldsymbol{q})\big)$$

with $\boldsymbol{s} = \boldsymbol{s}(t, \boldsymbol{q})$ being implicitly defined by

$$\boldsymbol{0} = \boldsymbol{h}\big(t, \boldsymbol{q}, \boldsymbol{s}(t, \boldsymbol{q})\big) \,.$$

Implicit differentiation yields

$$\boldsymbol{0} = \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{q}}(t, \boldsymbol{q}, \boldsymbol{s}) + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{s}}(t, \boldsymbol{q}, \boldsymbol{s}) \, \frac{\partial \boldsymbol{s}}{\partial \boldsymbol{q}}(t, \boldsymbol{q})$$

and

$$\boldsymbol{0} = \frac{\partial \boldsymbol{h}}{\partial t}(t, \boldsymbol{q}, \boldsymbol{s}) + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{q}}(t, \boldsymbol{q}, \boldsymbol{s}) \, \dot{\boldsymbol{q}}(t) + \frac{\partial \boldsymbol{h}}{\partial \boldsymbol{s}}(t, \boldsymbol{q}, \boldsymbol{s}) \, \dot{\boldsymbol{s}}(t) \,.$$

Therefore, the constraint matrix is given by

$$\begin{aligned}\mathbf{G}(t,\boldsymbol{q},\boldsymbol{s}) = \bar{\mathbf{G}}(t,\boldsymbol{q}) &= \frac{\partial \bar{\boldsymbol{g}}}{\partial \boldsymbol{q}}(t,\boldsymbol{q}) = \frac{\mathrm{D}\boldsymbol{g}}{\mathrm{D}\boldsymbol{q}}\big(t,\boldsymbol{q},\boldsymbol{s}(t,\boldsymbol{q})\big) \\ &= \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(t,\boldsymbol{q},\boldsymbol{s}) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{s}}(t,\boldsymbol{q},\boldsymbol{s})\,\frac{\partial \boldsymbol{s}}{\partial \boldsymbol{q}}(t,\boldsymbol{q}) = \Big[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}} - \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{s}}\Big(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{s}}\Big)^{-1}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{q}}\Big](t,\boldsymbol{q},\boldsymbol{s})\end{aligned}$$

and the hidden constraints at the level of velocity coordinates get the form

$$\begin{aligned}\mathbf{0} &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(t,\boldsymbol{q}(t),\boldsymbol{s}(t)) = \frac{\partial \boldsymbol{g}}{\partial t}(t,\boldsymbol{q},\boldsymbol{s}) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(t,\boldsymbol{q},\boldsymbol{s})\,\dot{\boldsymbol{q}}(t) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{s}}(t,\boldsymbol{q},\boldsymbol{s})\,\dot{\boldsymbol{s}}(t) \\ &= \Big[\frac{\partial \boldsymbol{g}}{\partial t} - \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{s}}\Big(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{s}}\Big)^{-1}\frac{\partial \boldsymbol{h}}{\partial t}\Big](t,\boldsymbol{q},\boldsymbol{s}) + \Big[\frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}} - \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{s}}\Big(\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{s}}\Big)^{-1}\frac{\partial \boldsymbol{h}}{\partial \boldsymbol{q}}\Big](t,\boldsymbol{q},\boldsymbol{s})\,\dot{\boldsymbol{q}}(t)\,, \\ &= \boldsymbol{g}^{(I)}(t,\boldsymbol{q},\boldsymbol{s}) + \mathbf{G}(t,\boldsymbol{q},\boldsymbol{s})\,\dot{\boldsymbol{q}}(t) = \bar{\boldsymbol{g}}^{(I)}(t,\boldsymbol{q},\boldsymbol{s}) + \mathbf{G}(t,\boldsymbol{q},\boldsymbol{s})\,\mathbf{H}(\boldsymbol{q})\,\boldsymbol{v}\,,\end{aligned}$$

with $\boldsymbol{g}^{(I)}$ summarizing the partial derivatives of $\boldsymbol{g}$ and $\boldsymbol{h}$ w.r.t. t, see (2.8) and (3.5). In the dynamical equations, the holonomic constraints (3.11) result in constraint forces $-\mathbf{H}^{\top}(\boldsymbol{q})\,\mathbf{G}^{\top}(t,\boldsymbol{q},\boldsymbol{s})\,\boldsymbol{\lambda}$ with Lagrange multipliers $\boldsymbol{\lambda} \in \mathbb{R}^{n_\lambda}$. Additional constraint forces $-\mathbf{K}^{\top}(t,\boldsymbol{q},\boldsymbol{s})\,\boldsymbol{\eta}$ with Lagrange multipliers $\boldsymbol{\eta} \in \mathbb{R}^{n_k}$ correspond to n_k nonholonomic constraints that are assumed to be in Pfaffian form $\mathbf{0} = \mathbf{K}(t,\boldsymbol{q},\boldsymbol{s})\,\boldsymbol{v} + \boldsymbol{k}_0(t,\boldsymbol{q},\boldsymbol{s})$, see [20].

With all these notations, the multibody system model equations may be summarized in a hybrid system of discrete state equations (3.9) and differential-algebraic equations

$$\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}\,, \tag{3.12a}$$

$$\mathbf{M}(t,\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(t,\boldsymbol{q},\boldsymbol{s},\boldsymbol{v},\boldsymbol{c},\boldsymbol{r}_j,\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\eta}) - \mathbf{H}^{\top}(\boldsymbol{q})\,\mathbf{G}^{\top}(t,\boldsymbol{q},\boldsymbol{s})\,\boldsymbol{\lambda} - \mathbf{K}^{\top}(t,\boldsymbol{q},\boldsymbol{s})\,\boldsymbol{\eta}\,, \tag{3.12b}$$

$$\dot{\boldsymbol{c}} = \boldsymbol{d}(t,\boldsymbol{q},\boldsymbol{s},\boldsymbol{v},\boldsymbol{c},\boldsymbol{r}_j,\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\eta})\,, \tag{3.12c}$$

$$\mathbf{0} = \boldsymbol{b}(t,\boldsymbol{q},\boldsymbol{s},\boldsymbol{v},\boldsymbol{c},\boldsymbol{r}_j,\boldsymbol{w},\boldsymbol{\lambda},\boldsymbol{\eta})\,, \tag{3.12d}$$

$$\mathbf{0} = \boldsymbol{h}(t,\boldsymbol{q},\boldsymbol{s})\,, \tag{3.12e}$$

$$\mathbf{0} = \boldsymbol{g}(t,\boldsymbol{q},\boldsymbol{s})\,, \tag{3.12f}$$

$$\mathbf{0} = \mathbf{K}(t,\boldsymbol{q},\boldsymbol{s})\boldsymbol{v} + \boldsymbol{k}_0(t,\boldsymbol{q},\boldsymbol{s}) \tag{3.12g}$$

that describe the evolution of all time-continuous state variables for $t \in [T_j, T_{j+1})$.

Existence and uniqueness of solutions for DAE (3.12) may be studied along the lines of the analysis in Sect. 2.2 provided that the terms

$$\frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\lambda}} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{w}}\Big(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{w}}\Big)^{-1}\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\lambda}} \quad \text{and} \quad \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{\eta}} - \frac{\partial \boldsymbol{f}}{\partial \boldsymbol{w}}\Big(\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{w}}\Big)^{-1}\frac{\partial \boldsymbol{b}}{\partial \boldsymbol{\eta}}$$

are sufficiently small [62]. Essential assumptions for the existence of a locally uniquely defined solution are known from Theorem 2.2: The symmetric, positive semi-definite mass matrix $\mathbf{M}(t,\boldsymbol{q})$ is assumed to have full rank at the nullspace of the extended constraint matrix

$$\begin{pmatrix} \mathbf{G}(t,\boldsymbol{q},\boldsymbol{s})\,\mathbf{H}(\boldsymbol{q}) \\ \mathbf{K}(t,\boldsymbol{q},\boldsymbol{s}) \end{pmatrix}$$

and this matrix has to have full rank. With these assumptions, the index of DAE (3.12) is bounded by three, see Remark 2.3(a).

Note that the full rank assumption on $\mathbf{G}(t,\boldsymbol{q},\boldsymbol{s})\ \mathbf{H}(\boldsymbol{q})$ would be violated if the invariants (3.8) would be considered in the holonomic constraints (3.12f) since $\mathbf{0} \equiv \boldsymbol{\gamma}(\boldsymbol{q}(t))$ and the kinematic equations $\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\boldsymbol{v}$, see (3.12a), imply

$$\mathbf{0}_{n_q-n_v} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}\big(\boldsymbol{q}(t)\big) = \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}\big(\boldsymbol{q}(t)\big)\,\dot{\boldsymbol{q}}(t) = \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\,\mathbf{H}(\boldsymbol{q})\,\boldsymbol{v}$$

for any velocity coordinates $\boldsymbol{v} \in \mathbb{R}^{n_v}$, i.e., $(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})(\boldsymbol{q})\,\mathbf{H}(\boldsymbol{q}) \equiv \mathbf{0}_{(n_q-n_v)\times n_v}$.

As an alternative, $n_q - n_v$ linearly independent invariants (3.8) with a Jacobian $(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})(\boldsymbol{q})$ of full rank could be enforced in time integration substituting the kinematic equations (3.12a) by

$$\dot{\boldsymbol{q}} = \mathbf{H}(\boldsymbol{q})\,\boldsymbol{v} - \Big(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\Big)^{\top}\boldsymbol{\mu}\,, \tag{3.13a}$$

$$\mathbf{0} = \boldsymbol{\gamma}(\boldsymbol{q}) \tag{3.13b}$$

with artificial multipliers $\boldsymbol{\mu} \in \mathbb{R}^{n_q-n_v}$, see [43]. These new variables vanish identically for the analytical solution since $(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})(\boldsymbol{q})\,\mathbf{H}(\boldsymbol{q}) = \mathbf{0}$ implies

$$\begin{aligned}\mathbf{0} &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}\big(\boldsymbol{q}(t)\big) = \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}\big(\boldsymbol{q}(t)\big)\,\dot{\boldsymbol{q}}(t) = \frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\Big(\mathbf{H}(\boldsymbol{q})\,\boldsymbol{v} - \Big(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\Big)^{\top}\boldsymbol{\mu}\Big) \\ &= -\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\Big(\frac{\partial \boldsymbol{\gamma}}{\partial \boldsymbol{q}}(\boldsymbol{q})\Big)^{\top}\boldsymbol{\mu}\,.\end{aligned}$$

and $(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})^{\top}$ is non-singular by assumption. For the numerical solution, the correction term $-(\partial\boldsymbol{\gamma}/\partial\boldsymbol{q})^{\top}\boldsymbol{\mu}$ in (3.13a) remains typically in the size of the discretization error [43].

3.3 Multibody Formalisms and Topological Solvers

In Sect. 2.1, we considered conservative systems being characterized by potential forces $-\nabla U(\boldsymbol{q})$ and used Hamilton's principle of least action to derive the equations

of motion (2.3). Formally, this approach may be generalized to non-conservative systems including, e.g., dissipative terms and actuator forces. In engineering applications it is, however, more common to use equilibrium conditions for forces and momenta for deriving the equations of motion of complex multibody systems [74, 76].

These *Newton–Euler equations* are formulated most conveniently in an inertial frame using absolute coordinates. To simplify the notation, we restrict ourselves in the present section to linear configuration spaces and consider (absolute) position coordinates $\boldsymbol{p}_i(t) \in \mathbb{R}^{d_i}$, $(i = 1, \ldots, N)$, for the N bodies of the multibody system. Position and orientation of a rigid body $(\bullet)^{(i)}$ are described by $\boldsymbol{p}_i \in \mathbb{R}^6$ for 3-D models ($d_i = 6$, see Sect. 3.1) and by $\boldsymbol{p}_i \in \mathbb{R}^3$ in the 2-D case. For point masses, the (absolute) position may be characterized by Cartesian coordinates $\boldsymbol{p}_i \in \mathbb{R}^{d_i}$ with $d_i = 3$ in 3-D and $d_i = 2$ in 2-D, see, e.g., Example 2.1.

In this body-oriented modelling framework, the interaction of bodies may be described by *force elements* and by *joints* [52, 74]. Force elements represent, e.g., spring-damper elements and actuators and contribute in the mathematical model to the force vector $\boldsymbol{f}$.

Joints restrict the relative motion of (two) bodies w.r.t. each other and result in (holonomic) constraints (2.1). Therefore, the basic internal structure of the DAE model equations (2.3) is characterized by the *topology* of the multibody system in terms of bodies and joints. The topology of a system with N bodies is represented by a labelled graph with $N + 1$ vertices for the (rigid or flexible) bodies $(\bullet)^{(i)}$, $(i = 1, \ldots, N)$, and an extra (virtual) body $(\bullet)^{(0)}$ that is inertially fixed and stands for the inertial system. Two vertices of the graph are connected by an edge if and only if the corresponding bodies in the multibody system model are connected by a joint restricting their relative motion, see Fig. 8.

In dynamical simulation, the topology of the multibody system model is exploited to evaluate the equations of motion efficiently. An early reference in this field is the work of Featherstone who developed an algorithm to evaluate the equations of motion for a tree structured system of N bodies with $\mathcal{O}(N)$ complexity [38], see also [24]. Such *multibody formalisms* may be interpreted as a block Gauss elimination for an augmented set of equations of motion, see, e.g., [61, 90] and the references therein. From the viewpoint of numerical linear algebra, these algorithms define *topological solvers* [88] for large systems of linear equations (2.17) with sparse matrices $\mathbf{M}$ and $\mathbf{G}$.

As a typical example, we consider in the present section a multibody formalism for tree structured systems that is based on a mixed coordinate formulation. The equations of motion are reduced to a second order ODE in joint coordinates $\boldsymbol{q}$ with a right-hand side that may be evaluated with $\mathcal{O}(N)$ complexity. These results have recently been published in a slightly more general setting in [7, 8]. They are essentially based on the work of Lubich et al. [61] and Eich-Soellner and Führer [37].

The graph of a tree structured multibody system is acyclic, i.e., it is free of loops. Furthermore, it is connected and may be ordered such that there is a *root vertex* and all vertices except this root vertex have a uniquely defined predecessor. It is assumed

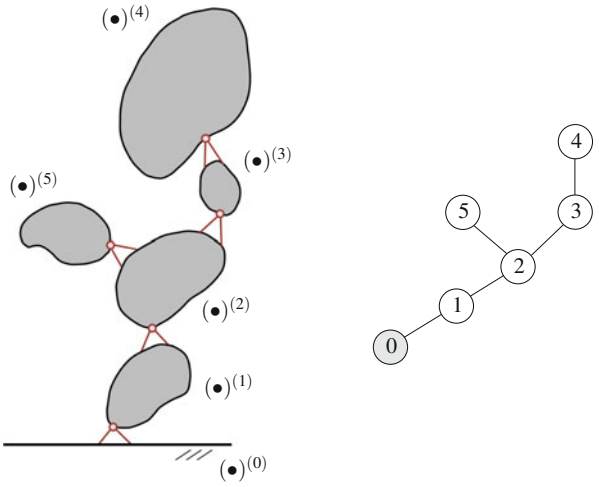


Fig. 8 Topology and labelled graph of a multibody system model with $N = 5$ bodies, see [8]

that the root vertex corresponds to the (inertially fixed) *root body* $(\bullet)^{(0)}$ and that the other vertices are labelled such that the labels are monotonically increasing along each branch of the kinematic tree.

With these assumptions, all bodies $(\bullet)^{(i)}$, $(\, i = 1, \ldots, N\,)$, have a uniquely defined predecessor $(\bullet)^{(\pi_i)}$ and the labels satisfy $\pi_i < i$. Each body $(\bullet)^{(i)}$ may have (direct) successors $(\bullet)^{(j)}$ being characterized by $\pi_j = i$ or, equivalently, by $j \in I_i := \{\, k : \pi_k = i \,\}$ with an index set I_i that represents the set of all successors of a given body $(\bullet)^{(i)}$ in the multibody system model. Bodies without successors ($I_i = \emptyset$) correspond to leafs of the kinematic tree and are therefore called "leaf bodies". The tree structured system in Fig. 8 has the two leaf bodies $(\bullet)^{(4)}$ and $(\bullet)^{(5)}$ and we have $I_1 = \{\, 2 \,\}, I_2 = \{\, 3,\ 5 \,\}, I_3 = \{\, 4 \,\}$ since $\pi_1 = 0$, $\pi_2 = 1$, $\pi_3 = \pi_5 = 2$ and $\pi_4 = 3$.

Position and orientation of body $(\bullet)^{(i)}$ are characterized by the (absolute) position coordinates $\boldsymbol{p}_i(t) \in \mathbb{R}^{d_i}$. The *relative* position and orientation of body $(\bullet)^{(i)}$ w.r.t. its predecessor $(\bullet)^{(\pi_i)}$ is characterized by joint coordinates $\boldsymbol{q}_i(t) \in \mathbb{R}^{n_i}$ representing the n_i degrees of freedom of the joint connecting $(\bullet)^{(i)}$ with $(\bullet)^{(\pi_i)}$:

$$\mathbf{0} = \boldsymbol{k}_i(\boldsymbol{p}_i, \boldsymbol{p}_{\pi_i}, \boldsymbol{q}_i, t)\,. \tag{3.14}$$

Here and in the following we assume that (3.14) is locally uniquely solvable w.r.t. $\boldsymbol{p}_i$ and that the Jacobian $\mathbf{K}_i = \partial \boldsymbol{k}_i / \partial \boldsymbol{p}_i$ is non-singular along the solution. In its most simple form, Eq. (3.14) defines $\boldsymbol{p}_i$ explicitly by $\boldsymbol{p}_i(t) = \boldsymbol{r}_i(\boldsymbol{p}_{\pi_i}(t), \boldsymbol{q}_i(t), t)$ resulting in $\mathbf{K}_i = \mathbf{I}_{d_i}$.

The kinematic relations (3.14) at the level of position coordinates imply relations at the level of velocity and acceleration coordinates that may formally be obtained

by (total) differentiation of (3.14) w.r.t. time t, see (2.8) and (2.9):

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{k}_i(\boldsymbol{p}_i(t), \boldsymbol{p}_{\pi_i}(t), \boldsymbol{q}_i(t), t) = \mathbf{K}_i\dot{\boldsymbol{p}}_i + \mathbf{H}_i\dot{\boldsymbol{p}}_{\pi_i} + \mathbf{J}_i\dot{\boldsymbol{q}}_i + \boldsymbol{k}_i^{(\mathrm{I})}(\boldsymbol{p}_0, \boldsymbol{p}, \boldsymbol{q}, t)\,, \tag{3.15}$$

$$\mathbf{0} = \mathbf{K}_i\ddot{\boldsymbol{p}}_i + \mathbf{H}_i\ddot{\boldsymbol{p}}_{\pi_i} + \mathbf{J}_i\ddot{\boldsymbol{q}}_i + \boldsymbol{k}_i^{(\mathrm{II})}(\boldsymbol{p}_0, \dot{\boldsymbol{p}}_0, \boldsymbol{p}, \dot{\boldsymbol{p}}, \boldsymbol{q}, \dot{\boldsymbol{q}}, t) \tag{3.16}$$

with

$$\mathbf{K}_i := \frac{\partial \boldsymbol{k}_i}{\partial \boldsymbol{p}_i} \in \mathbb{R}^{d_i \times d_i}\,, \quad \mathbf{H}_i := \frac{\partial \boldsymbol{k}_i}{\partial \boldsymbol{p}_{\pi_i}} \in \mathbb{R}^{d_i \times d_i}\,, \quad \mathbf{J}_i := \frac{\partial \boldsymbol{k}_i}{\partial \boldsymbol{q}_i} \in \mathbb{R}^{d_i \times n_i}\,. \tag{3.17}$$

It is assumed that the joint coordinates $\boldsymbol{q}_i(t)$ are defined such that all Jacobians $\mathbf{J}_i$ have full column rank: $\operatorname{rank}\mathbf{J}_i = n_i \leq d_i$. Functions $\boldsymbol{k}_i^{(\mathrm{I})} := \partial \boldsymbol{k}_i/\partial t$ and $\boldsymbol{k}_i^{(\mathrm{II})}$ summarize partial time derivatives and all lower order terms in the first and second time derivative of (3.14), respectively. They may depend on the (absolute) coordinates $\boldsymbol{p}_0$ of the root body, on the absolute coordinates $\boldsymbol{p} := (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_N)$ of the remaining N bodies in the system, on the corresponding joint coordinates $\boldsymbol{q} := (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_N)$ and on $\dot{\boldsymbol{p}}_0, \dot{\boldsymbol{p}}$ and $\dot{\boldsymbol{q}}$.

In recursive multibody formalisms, the position and velocity of the root body $(\boldsymbol{p}_0(t), \dot{\boldsymbol{p}}_0(t))$ as well as all joint coordinates $\boldsymbol{q}_i(t)$, $\dot{\boldsymbol{q}}_i(t)$, $(i = 1, \ldots, N)$, at a current time t are assumed to be given. Starting from the root body, the absolute position and velocity coordinates $\boldsymbol{p}_i(t)$, $\dot{\boldsymbol{p}}_i(t)$ of all N bodies $(\bullet)^{(i)}$, $(i = 1, \ldots, N)$, may then be computed recursively using (3.14) and (3.15), respectively (*forward recursion*).

The equilibrium conditions for forces and momenta are formulated for each individual body $(\bullet)^{(i)}$ using its absolute coordinates $\boldsymbol{p}_i$:

$$\mathbf{M}_i\ddot{\boldsymbol{p}}_i + \mathbf{K}_i^{\top}\boldsymbol{\mu}_i + \sum_{j \in I_i} \mathbf{H}_j^{\top}\boldsymbol{\mu}_j = \boldsymbol{f}_i\,, \quad (i = 1, \ldots, N)\,. \tag{3.18}$$

They contain the reaction forces of the joints connecting body $(\bullet)^{(i)}$ with its predecessor $(\mathbf{K}_i^{\top}\boldsymbol{\mu}_i)$ and with its successors in the kinematic tree $(\mathbf{H}_j^{\top}\boldsymbol{\mu}_j,\ j \in I_i)$. All remaining forces and momenta acting on body $(\bullet)^{(i)}$ are summarized in the force vector $\boldsymbol{f}_i = \boldsymbol{f}_i(\boldsymbol{p}, \dot{\boldsymbol{p}}, \boldsymbol{q}, \dot{\boldsymbol{q}}, t) \in \mathbb{R}^{d_i}$. The body mass matrix $\mathbf{M}_i \in \mathbb{R}^{d_i \times d_i}$ contains mass and inertia tensor of body $(\bullet)^{(i)}$ and is assumed to be symmetric, positive definite. For a discussion of rank-deficient body mass matrices $\mathbf{M}_i$ we refer to [7].

The specific structure of the joint reaction forces with Lagrange multipliers $\boldsymbol{\mu}_i(t) \in \mathbb{R}^{d_i}$ that satisfy

$$\mathbf{J}_i^{\top}\boldsymbol{\mu}_i = \mathbf{0}\,, \quad (i = 1, \ldots, N)\,, \tag{3.19}$$

results from the joint equations (3.14) and from d'Alembert's principle since the virtual work of constraint forces vanishes for all (virtual) displacements being compatible with (3.14). In (3.19), matrix $\mathbf{J}_i$ denotes the Jacobian of the constraint function $\boldsymbol{k}_i$ w.r.t. joint coordinates $\boldsymbol{q}_i \in \mathbb{R}^{n_i}$, see (3.17).

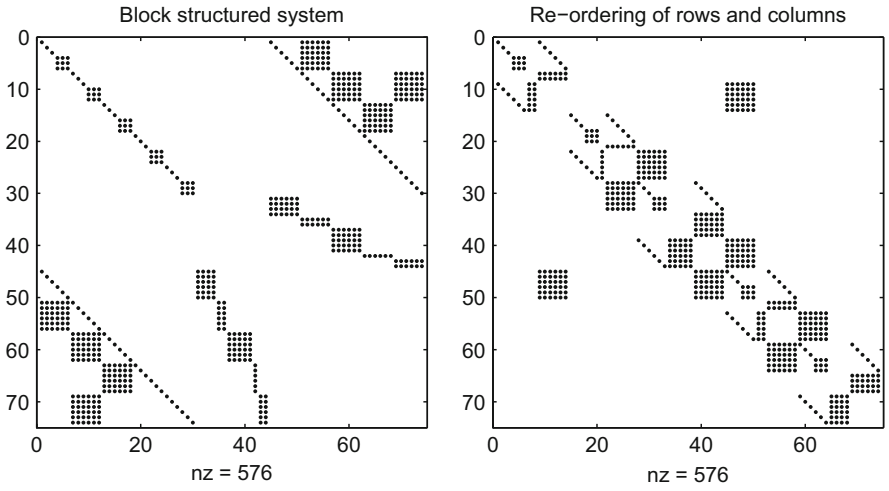


Fig. 9 Sparsity pattern of matrix (2.18) for a mixed coordinate formulation of the tree structured system of Fig. 8. *Left plot*: Original structure (2.18). *Right plot*: Structure after re-ordering of rows and columns according to the system's topology, see Example 3.2

Equations (3.16), (3.19) and the equilibrium conditions (3.18) are linear in $\ddot{\boldsymbol{p}}$, $\ddot{\boldsymbol{q}}$ and $\boldsymbol{\mu}$. They may be summarized in a large sparse system of the form (2.17) with $(\ddot{\boldsymbol{q}}, \boldsymbol{\lambda})$ being substituted by $\big((\ddot{\boldsymbol{p}}^\top, \ddot{\boldsymbol{q}}^\top)^\top, \boldsymbol{\mu}\big)$. The block diagonal mass matrix $\mathbf{M}$ is of size $(n_p + n_q) \times (n_p + n_q)$. It has rank n_p since the non-zero blocks on the main diagonal are given by the symmetric, positive definite body mass matrices $\mathbf{M}_i$, $(i = 1, \ldots, N)$. The non-zero blocks of the constraint matrix $\mathbf{G}$ result from the Jacobians $\mathbf{K}_i$, $\mathbf{H}_i$, $\mathbf{J}_i$, $(i = 1, \ldots, N)$, see (3.17).

Example 3.2 The left plot of Fig. 9 shows the sparsity pattern of matrix (2.18) for a 3-D version of the tree structured system in Fig. 8 with $N = 5$ rigid bodies ($d_i = 6$) and joint coordinates $\boldsymbol{q}_i$ of dimension $n_1 = 4$, $n_2 = 2$, $n_3 = 5$, $n_4 = 1$, $n_5 = 2$. The body mass matrices are given by $\mathbf{M}_i = \text{blockdiag}\,(m_i\mathbf{I}_3, \boldsymbol{\Theta}_i)$ with $m_i \in \mathbb{R}$ and $\boldsymbol{\Theta}_i \in \mathbb{R}^{3\times 3}$ denoting mass and inertia tensor of body $(\bullet)^{(i)}$, $(i = 1, \ldots, N)$. With kinematic relations $\boldsymbol{p}_i(t) = \boldsymbol{r}_i(\boldsymbol{p}_{\pi_i}(t), \boldsymbol{q}_i(t), t)$, we get Jacobians $\mathbf{K}_i = \mathbf{I}_6$, $(i = 1, \ldots, N)$.

To reduce the bandwidth of this sparse symmetric matrix, rows and columns are re-ordered according to the system's topology. This may be achieved by the vector of unknowns $\boldsymbol{x} = (\boldsymbol{x}_N^\top, \boldsymbol{x}_{N-1}^\top, \ldots, \boldsymbol{x}_1^\top)^\top$ with $\boldsymbol{x}_i \in \mathbb{R}^{12+n_i}$ summarizing the unknowns $\ddot{\boldsymbol{p}}_i$, $\ddot{\boldsymbol{q}}_i$ and $\boldsymbol{\mu}_i$ that correspond to body $(\bullet)^{(i)}$. Then we obtain an $N \times N$ block structured system with non-singular diagonal blocks $\mathbf{A}_i \in \mathbb{R}^{(12+n_i)\times(12+n_i)}$, $(i = N, N-1, \ldots, 1)$:

$$\boldsymbol{x}_i := \begin{pmatrix} \ddot{\boldsymbol{p}}_i \\ \ddot{\boldsymbol{q}}_i \\ \boldsymbol{\mu}_i \end{pmatrix}, \quad \mathbf{A}_i := \begin{pmatrix} \mathbf{M}_i & \mathbf{0} & \mathbf{K}_i^\top \\ \mathbf{0} & \mathbf{0} & \mathbf{J}_i^\top \\ \mathbf{K}_i & \mathbf{J}_i & \mathbf{0} \end{pmatrix}.$$

For chain structured systems, this re-ordered matrix is block-tridiagonal. In a tree structured system, each ramification yields an extra non-zero block below this block-tridiagonal band (accompanied by its transposed in the upper triangle).

This sparsity structure is illustrated by the right plot of Fig. 9 that shows a 5×5 block structure with diagonal blocks of dimension 14×14, 13×13, 17×17, 14×14 and 16×16. The non-zero off-diagonal blocks in block row 4, block column 1 and in block row 1, block column 4 correspond to the ramification of the kinematic tree at body $(\bullet)^{(2)}$ that has *two* successors $(\bullet)^{(3)}$ and $(\bullet)^{(5)}$. (Note that block column i is multiplied by vector $\boldsymbol{x}_{N+1-i}$ since $\boldsymbol{x} = (\boldsymbol{x}_N^\top, \boldsymbol{x}_{N-1}^\top, \ldots, \boldsymbol{x}_1^\top)^\top$).

The mixed coordinate formulation results in sparse systems (2.17) for the accelerations $\ddot{\boldsymbol{p}}, \ddot{\boldsymbol{q}}$ and the Lagrange multipliers $\boldsymbol{\mu}$. Example 3.2 shows how to re-arrange rows and columns of matrix (2.18) to get a sparse $N \times N$ block structure reflecting the system's topology. Lubich et al. [61] combine this approach with a block Gauss elimination to compute $\ddot{\boldsymbol{p}}, \ddot{\boldsymbol{q}}$ and $\boldsymbol{\mu}$ with $\mathcal{O}(N)$ complexity.

In engineering, such structure exploiting algorithms have been formulated such that all intermediate results have a straightforward physical interpretation ($\mathcal{O}(N)$-*formalisms*): We start with the observation that the equilibrium conditions (3.18) get a simpler form for leaf bodies $(\bullet)^{(i)}$ since $I_i = \{j : \pi_j = i\} = \emptyset$ in that case. We obtain

$$\bar{\mathbf{M}}_i \mathbf{K}_i \ddot{\boldsymbol{p}}_i + \boldsymbol{\mu}_i = \bar{\boldsymbol{f}}_i \tag{3.20}$$

with $\bar{\boldsymbol{f}}_i := \mathbf{K}_i^{-\top} \boldsymbol{f}_i$, $\mathbf{K}_i^{-\top} := (\mathbf{K}_i^\top)^{-1}$ and the symmetric, positive definite mass matrix $\bar{\mathbf{M}}_i := \mathbf{K}_i^{-\top} \mathbf{M}_i \mathbf{K}_i^{-1}$. Equations (3.16), (3.19) and (3.20) define a system of $2d_i + n_i$ linear equations that may be solved w.r.t. $\ddot{\boldsymbol{p}}_i, \ddot{\boldsymbol{q}}_i$ and $\boldsymbol{\mu}_i$:

Lemma 3.1 *Consider the system of linear equations*

$$\bar{\mathbf{M}}_i \mathbf{K}_i \ddot{\boldsymbol{p}}_i + \boldsymbol{\mu}_i = \bar{\boldsymbol{f}}_i \,, \tag{3.21a}$$

$$\mathbf{J}_i^\top \boldsymbol{\mu}_i = \mathbf{0} \,, \tag{3.21b}$$

$$\mathbf{K}_i \ddot{\boldsymbol{p}}_i + \mathbf{H}_i \ddot{\boldsymbol{p}}_{\pi_i} + \mathbf{J}_i \ddot{\boldsymbol{q}}_i + \boldsymbol{k}_i^{(\mathrm{II})} = \mathbf{0} \tag{3.21c}$$

with matrices $\bar{\mathbf{M}}_i, \mathbf{K}_i, \mathbf{H}_i \in \mathbb{R}^{d \times d}$, $\mathbf{J}_i \in \mathbb{R}^{d \times n}$ *and vectors* $\ddot{\boldsymbol{p}}_i, \boldsymbol{\mu}_i, \bar{\boldsymbol{f}}_i, \ddot{\boldsymbol{p}}_{\pi_i}, \boldsymbol{k}_i^{(\mathrm{II})} \in \mathbb{R}^d$, $\ddot{\boldsymbol{q}}_i \in \mathbb{R}^n$. *If* $\bar{\mathbf{M}}_i$ *is symmetric, positive definite,* $\mathbf{K}_i$ *is non-singular and* $\mathbf{J}_i$ *has full rank* $n \le d$, *then (3.21) may be solved w.r.t.* $\ddot{\boldsymbol{p}}_i$, $\boldsymbol{\mu}_i$, $\ddot{\boldsymbol{q}}_i$ *resulting in*

$$\ddot{\boldsymbol{q}}_i = -(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^\top \bar{\boldsymbol{f}}_i - (\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^\top \bar{\mathbf{M}}_i (\mathbf{H}_i \ddot{\boldsymbol{p}}_{\pi_i} + \boldsymbol{k}_i^{(\mathrm{II})}) \,, \tag{3.22a}$$

$$\ddot{\boldsymbol{p}}_i = -\mathbf{K}_i^{-1} (\bar{\mathbf{H}}_i \ddot{\boldsymbol{p}}_{\pi_i} + \bar{\boldsymbol{k}}_i^{(\mathrm{II})}) \,, \tag{3.22b}$$

$$\boldsymbol{\mu}_i = \bar{\boldsymbol{f}}_i + \bar{\mathbf{M}}_i \bar{\mathbf{H}}_i \ddot{\boldsymbol{p}}_{\pi_i} + \bar{\mathbf{M}}_i \bar{\boldsymbol{k}}_i^{(\mathrm{II})} \tag{3.22c}$$

with

$$\bar{\mathbf{H}}_i := \big(\mathbf{I}_d - \mathbf{J}_i(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\mathbf{M}}_i\big)\mathbf{H}_i\,, \tag{3.23a}$$

$$\bar{\boldsymbol{k}}_i^{(\mathrm{II})} := \big(\mathbf{I}_d - \mathbf{J}_i(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\mathbf{M}}_i\big)\boldsymbol{k}_i^{(\mathrm{II})} - \mathbf{J}_i(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\boldsymbol{f}}_i\,. \tag{3.23b}$$

Proof If $\bar{\mathbf{M}}_i \in \mathbb{R}^{d\times d}$ is symmetric, positive definite and $\mathbf{J}_i \in \mathbb{R}^{d\times n}$ has full rank, then matrix $\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i \in \mathbb{R}^{n\times n}$ is symmetric, positive definite as well and left-multiplication of (3.21c) by $(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\mathbf{M}}_i$ yields

$$\ddot{\boldsymbol{q}}_i = -(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{K}_i \ddot{\boldsymbol{p}}_i - (\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\mathbf{M}}_i (\mathbf{H}_i \ddot{\boldsymbol{p}}_{\pi_i} + \boldsymbol{k}_i^{(\mathrm{II})})\,. \tag{3.24}$$

Taking into account that left-multiplication of (3.21a) by $\mathbf{J}_i^\top$ results in

$$\mathbf{J}_i^\top \bar{\boldsymbol{f}}_i = \mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{K}_i \ddot{\boldsymbol{p}}_i + \mathbf{J}_i^\top \boldsymbol{\mu}_i = \mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{K}_i \ddot{\boldsymbol{p}}_i\,,$$

see (3.21b), we may substitute the first term in the right-hand side of (3.24) by $-(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1}\mathbf{J}_i^\top \bar{\boldsymbol{f}}_i$ and get the explicit expression (3.22a) for $\ddot{\boldsymbol{q}}_i$. This explicit expression is used to obtain assertion (3.22b) multiplying (3.21c) from the left by matrix $\mathbf{K}_i^{-1}$. Finally, assertion (3.22c) is seen to be a straightforward consequence of (3.21a) and (3.22b). ■

For leaf bodies $(\bullet)^{(i)}$, the equilibrium conditions (3.18) were transformed straightforwardly to the simpler form (3.20). Lemma 3.1 allows to get by induction these condensed equilibrium conditions with suitable $\bar{\mathbf{M}}_i, \bar{\boldsymbol{f}}_i$ for *all* bodies $(\bullet)^{(i)}$ of the tree structured system: Let us assume that the equilibrium conditions of all direct successors $(\bullet)^{(j)}$ of body $(\bullet)^{(i)}$ are given in form (3.20), i.e.,

$$\bar{\mathbf{M}}_j \mathbf{K}_j \ddot{\boldsymbol{p}}_j + \boldsymbol{\mu}_j = \bar{\boldsymbol{f}}_j\,,\ (j \in I_i)\,.$$

Applying Lemma 3.1 to body $(\bullet)^{(j)}$ we obtain

$$\boldsymbol{\mu}_j = \bar{\boldsymbol{f}}_j + \bar{\mathbf{M}}_j \bar{\mathbf{H}}_j \ddot{\boldsymbol{p}}_i + \bar{\mathbf{M}}_j \bar{\boldsymbol{k}}_j^{(\mathrm{II})}$$

since $\pi_j = i$ if $(\bullet)^{(j)}$ is a direct successor of $(\bullet)^{(i)}$, see (3.22c). Inserting this expression in (3.18), we get after left-multiplication by $\mathbf{K}_i^{-\top}$ the condensed equilibrium conditions (3.20) with

$$\begin{aligned}\bar{\mathbf{M}}_i &:= \mathbf{K}_i^{-\top}\mathbf{M}_i\mathbf{K}_i^{-1} + \sum_{j\in I_i} \mathbf{K}_i^{-\top}\mathbf{H}_j^\top \bar{\mathbf{M}}_j \bar{\mathbf{H}}_j K_i^{-1} \\ &= \mathbf{K}_i^{-\top}\mathbf{M}_i\mathbf{K}_i^{-1} + \sum_{j\in I_i} \mathbf{K}_i^{-\top}\mathbf{H}_j^\top \big(\bar{\mathbf{M}}_j - \bar{\mathbf{M}}_j \mathbf{J}_j (\mathbf{J}_j^\top \bar{\mathbf{M}}_j \mathbf{J}_j)^{-1}\mathbf{J}_j^\top \bar{\mathbf{M}}_j\big)\mathbf{H}_j \mathbf{K}_i^{-1}\,,\end{aligned} \tag{3.25a}$$

$$\bar{\boldsymbol{f}}_i := \mathbf{K}_i^{-\top} \boldsymbol{f}_i + \sum_{j \in I_i} \mathbf{K}_i^{-\top} \mathbf{H}_j^{\top} (\bar{\boldsymbol{f}}_j + \bar{\mathbf{M}}_j \boldsymbol{k}_j^{(\mathrm{II})})$$

$$= \mathbf{K}_i^{-\top} \boldsymbol{f}_i - \sum_{j \in I_i} \mathbf{K}_i^{-\top} \mathbf{H}_j^{\top} \left(\mathbf{I}_{d_i} - \bar{\mathbf{M}}_j \mathbf{J}_j (\mathbf{J}_j^{\top} \bar{\mathbf{M}}_j \mathbf{J}_j)^{-1} \mathbf{J}_j^{\top} \right) (\bar{\boldsymbol{f}}_j + \bar{\mathbf{M}}_j \boldsymbol{k}_j^{(\mathrm{II})}) \,. \tag{3.25b}$$

The condensed mass matrix $\bar{\mathbf{M}}_i$ in (3.25a) is symmetric, positive definite since it is composed of the symmetric, positive definite matrix $\mathbf{K}_i^{-\top} \mathbf{M}_i \mathbf{K}_i^{-1}$ and a finite sum of symmetric, positive semi-definite matrices [7, Lemma 1]. Starting from the leaf bodies and following all branches of the kinematic tree to the root, the compact form (3.20) of the equilibrium conditions may be obtained recursively for all N bodies $(\bullet)^{(i)}$ of the multibody system (*backward recursion*).

From the viewpoint of numerical linear algebra we may interpret the transformation of the equilibrium conditions (3.18) to their condensed form (3.20) as a block Gauss elimination that transforms sparse block structured matrices like the one in the right plot of Fig. 9 to upper block triangular form. From the viewpoint of physics, we observe that the condensed mass matrix $\bar{\mathbf{M}}_i$ in (3.20) summarizes in compact form the mass and inertia terms of body $(\bullet)^{(i)}$ and all its successors in the kinematic tree. The condensed force vector $\bar{\boldsymbol{f}}_i$ represents the corresponding forces and momenta.

Since the backward recursion results in condensed equilibrium conditions (3.20) for all N bodies of the multibody system, we may use Lemma 3.1 to verify that $\ddot{\boldsymbol{p}}_i = \boldsymbol{a}_i$, $(i = 1, \ldots, N)$, with vector valued functions $\boldsymbol{a}_i$ that are recursively defined by

$$\boldsymbol{a}_0 := \ddot{\boldsymbol{p}}_0 = \mathbf{0} \,, \tag{3.26a}$$

$$\boldsymbol{a}_i := -\mathbf{K}_i^{-1} (\bar{\mathbf{H}}_i \boldsymbol{a}_{\pi_i} + \bar{\boldsymbol{k}}_i^{(\mathrm{II})}) \,, \ (i = 1, \ldots, N) \,, \tag{3.26b}$$

see (3.22b). This *second forward recursion* exploits the assumption that the root body $(\bullet)^{(0)}$ is inertially fixed such that the sequence $(\boldsymbol{a}_i)_i$ may be initialized by (3.26a).

The recursive multibody formalism is completed using the explicit expression (3.22a) for the accelerations $\ddot{\boldsymbol{q}}_i$, $(i = 1, \ldots, N)$, from Lemma 3.1:

$$\ddot{\boldsymbol{q}}_i = -(\mathbf{J}_i^{\top} \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^{\top} \bar{\boldsymbol{f}}_i - (\mathbf{J}_i^{\top} \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^{\top} \bar{\mathbf{M}}_i (\mathbf{H}_i \boldsymbol{a}_{\pi_i} + \boldsymbol{k}_i^{(\mathrm{II})}) \,. \tag{3.27}$$

The recursive multibody formalism is summarized in Table 1. It is an *explicit* $\mathcal{O}(N)$-formalism since the right-hand side of an explicit second order ODE

$$\ddot{\boldsymbol{q}}(t) = \boldsymbol{\varphi}(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t)) \tag{3.28}$$

for the joint coordinates $\boldsymbol{q}(t)$ is evaluated with a complexity that grows linearly with the number N of bodies in the tree structured multibody system.

Table 1 Explicit $\mathcal{O}(N)$-formalism

Data	$t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t), \boldsymbol{p}_0(t), \dot{\boldsymbol{p}}_0(t)$
Result	$\ddot{\boldsymbol{q}} = \ddot{\boldsymbol{q}}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{p}_0, \dot{\boldsymbol{p}}_0)$
Step 1	*First forward recursion* Start at the root body $(\bullet)^{(0)}$ and follow the branches of the kinematic tree to evaluate recursively the absolute position and velocity coordinates $\boldsymbol{p}_i = \boldsymbol{p}_i(t, \boldsymbol{q}_i, \boldsymbol{p}_{\pi_i})$ and $\dot{\boldsymbol{p}}_i = \dot{\boldsymbol{p}}_i(t, \boldsymbol{q}_i, \dot{\boldsymbol{q}}_i, \boldsymbol{p}_{\pi_i}, \dot{\boldsymbol{p}}_{\pi_i})$, $(i = 1, \ldots, N)$, according to (3.14) and (3.15)
Step 2	*Backward recursion* Start at the leaf bodies and proceed along the branches of the kinematic tree to evaluate recursively the condensed mass matrices $\bar{\mathbf{M}}_i$ and the condensed force vectors $\bar{\boldsymbol{f}}_i$, $(i = N, N-1, \ldots, 1)$, according to (3.25)
Step 3	*Second forward recursion* Set $\boldsymbol{a}_0 := \mathbf{0}$ and follow the branches of the kinematic tree to evaluate recursively the acceleration terms $\boldsymbol{a}_i = \boldsymbol{a}_i(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \boldsymbol{p}_0, \dot{\boldsymbol{p}}_0)$, $(i = 1, \ldots, N)$, according to (3.26b)
Step 4	*Function evaluation* $\ddot{\boldsymbol{q}}_i = -(\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^\top \bar{\boldsymbol{f}}_i - (\mathbf{J}_i^\top \bar{\mathbf{M}}_i \mathbf{J}_i)^{-1} \mathbf{J}_i^\top \bar{\mathbf{M}}_i (\mathbf{H}_i \boldsymbol{a}_{\pi_i} + \boldsymbol{k}_i^{(\mathrm{II})})$, $(i = 1, \ldots, N)$

Alternatively, the equations of motion may be evaluated in residual form

$$\boldsymbol{r}(t, \boldsymbol{q}(t), \dot{\boldsymbol{q}}(t), \ddot{\boldsymbol{q}}(t)) = \mathbf{0}$$

with

$$\boldsymbol{r}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \ddot{\boldsymbol{q}}) := \mathbf{M}(t, \boldsymbol{q}) \ddot{\boldsymbol{q}} - \boldsymbol{f}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}) \,. \tag{3.29}$$

Residual formalisms [36] evaluate the residual $\boldsymbol{r}(t, \boldsymbol{q}, \dot{\boldsymbol{q}}, \ddot{\boldsymbol{q}})$ for given arguments t, $\boldsymbol{q}$, $\dot{\boldsymbol{q}}$ and for a given estimate of $\ddot{\boldsymbol{q}}$ more efficiently than the explicit formalism of Table 1. In time integration, they have to be combined with implicit integrators like DASSL[26] that are tailored to implicit differential equations in residual form. Since the linearly implicit structure of the residual in (3.29) may result in frequent re-evaluations of the iteration matrix in the implicit integrator, the overall performance of explicit formalisms in time integration is, however, often superior [12, 73].

Explicit formalisms and residual formalisms are tailored to tree structured systems. Multibody system models with closed kinematical loops are beyond this problem class since the loops result in cycles in the corresponding graph. Formally, such a more complex model may be transformed to tree structure cutting virtually the loop-closing joints to get a simplified model with tree structure [19]. For this simplified model, the right-hand side $\boldsymbol{\varphi}$ in (3.28) and the residual $\boldsymbol{r}$ in (3.29) are evaluated with $\mathcal{O}(N)$ complexity by multibody formalisms for tree structured systems. Finally, the virtually cut joints are considered in the equations of motion (2.16) by holonomic constraints (2.16b).

4 Time Integration Methods for Constrained Mechanical Systems

The time integration of constrained mechanical systems was a topic of very active research in the 1980s and 1990s. The interested reader may find a comprehensive introduction to this subject in [50, Chap. VII].

Early approaches in this field were based on the direct application of ODE time discretization methods to the constrained equations of motion (2.16), see [31, 67]. We will discuss time integration methods of this type and their limitations and shortcomings in Sect. 4.1. The robustness and numerical stability of numerical methods for higher index DAEs may be improved substantially by an analytical index reduction before time discretization. In Sect. 4.2, we will consider index reduction and projection techniques for constrained mechanical systems.

To omit technical and implementation details we focus in the present section on equations of motion of the form

$$\mathbf{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} = \boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}) - \mathbf{G}^{\top}(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.1a}$$

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q}) \tag{4.1b}$$

with a constraint matrix $\mathbf{G}(\boldsymbol{q}) = (\partial/\partial \boldsymbol{q})\boldsymbol{g}(\boldsymbol{q})$ of full rank and a symmetric, positive semi-definite mass matrix $\mathbf{M}(\boldsymbol{q})$ that is positive definite on the nullspace of $\mathbf{G}(\boldsymbol{q})$. With these assumptions, the index of (4.1) is less than or equal to three. For positive definite mass matrices $\mathbf{M}(\boldsymbol{q})$, the system is analytically equivalent to an index-3 DAE in Hessenberg form, see Sect. 2.2.

4.1 Direct Time Discretization of the Constrained Equations of Motion

For time discretization, the equations of motion (4.1) may either be considered as a second order DAE in terms of $\boldsymbol{q}$ and $\boldsymbol{\lambda}$ or as a first order DAE in terms of $\boldsymbol{q}$, $\boldsymbol{v}$ and $\boldsymbol{\lambda}$ with $\boldsymbol{v}(t) := \dot{\boldsymbol{q}}(t)$ denoting the velocity coordinates, see Sect. 2.2.

4.1.1 Time Integration of Second Order Systems by Newmark Type Methods

The numerical solution of unconstrained systems

$$\mathbf{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} = \boldsymbol{f}(\boldsymbol{q}, \dot{\boldsymbol{q}}) \tag{4.2}$$

by Newmark type methods is quite popular in structural dynamics and flexible multibody dynamics [45]. In its most general form this class of integrators is given by the *generalized-α method* that was originally introduced for linear systems $\mathbf{M}\ddot{\boldsymbol{q}} + \mathbf{D}\dot{\boldsymbol{q}} + \mathbf{K}\boldsymbol{q} = \boldsymbol{r}(t)$ with $\mathbf{M}$, $\mathbf{D}$, $\mathbf{K}$ denoting the (constant) mass, damping and stiffness matrix [33].

For nonlinear systems (4.2) with state dependent mass matrix $\mathbf{M}(\boldsymbol{q})$, we follow the approach of Brüls [9] who proposed to update the numerical solution $(\boldsymbol{q}_n, \dot{\boldsymbol{q}}_n, \ddot{\boldsymbol{q}}_n) \approx \big(\boldsymbol{q}(t_n), \dot{\boldsymbol{q}}(t_n), \ddot{\boldsymbol{q}}(t_n)\big)$ in time step $t_n \to t_{n+1} = t_n + h$ by

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + h\dot{\boldsymbol{q}}_n + h^2(0.5-\beta)\boldsymbol{a}_n + h^2\beta\boldsymbol{a}_{n+1}\,, \tag{4.3a}$$

$$\dot{\boldsymbol{q}}_{n+1} = \dot{\boldsymbol{q}}_n + h(1-\gamma)\boldsymbol{a}_n + h\gamma\boldsymbol{a}_{n+1} \tag{4.3b}$$

with acceleration like vectors $\boldsymbol{a}_n$ that are defined by a weighted linear combination

$$(1-\alpha_m)\boldsymbol{a}_{n+1} + \alpha_m\boldsymbol{a}_n = (1-\alpha_f)\ddot{\boldsymbol{q}}_{n+1} + \alpha_f\ddot{\boldsymbol{q}}_n \tag{4.3c}$$

such that the equilibrium conditions

$$\mathbf{M}(\boldsymbol{q}_{n+1})\ddot{\boldsymbol{q}}_{n+1} = \boldsymbol{f}(\boldsymbol{q}_{n+1}, \dot{\boldsymbol{q}}_{n+1}) \tag{4.4}$$

at $t = t_{n+1}$ are satisfied.

The method is characterized by the algorithmic parameters α_f, α_m, β and γ. It has local truncation errors of size $\mathscr{O}(h^3)$ in the update formulae (4.3a,b) for position and velocity coordinates if $\gamma = 0.5 - (\alpha_m - \alpha_f)$. In structural dynamics, the remaining free parameters α_f, α_m and β are adjusted such that the numerical solution for the scalar linear test equation $\ddot{q} + \omega^2 q = 0$ is stable for all time step sizes $h > 0$ and a user prescribed damping ratio $\rho_\infty \in [0, 1]$ is achieved in the limit case $h\omega \to \infty$, see [33].

An alternative definition of algorithmic parameters goes back to the work of Hilber et al. [51] who considered method (4.3), (4.4) for systems (4.2) with constant mass matrix $\mathbf{M}$. In these *HHT−α methods*, the parameters α_f, α_m are given by $\alpha_f = -\alpha \in [0, 1/3]$ and $\alpha_m = 0$ and the update of vectors $\boldsymbol{a}_n$ is simplified to

$$\mathbf{M}\boldsymbol{a}_{n+1} = (1+\alpha)\boldsymbol{f}(\boldsymbol{q}_{n+1}, \dot{\boldsymbol{q}}_{n+1}) - \alpha\boldsymbol{f}(\boldsymbol{q}_n, \dot{\boldsymbol{q}}_n)\,, \tag{4.5}$$

see (4.3c) and (4.4). With parameters $\gamma = 0.5 - \alpha$, $\beta = (1-\alpha)^2/4$, the local truncation errors in (4.3a,b) are of size $\mathscr{O}(h^3)$ and the method is unconditionally stable for the linear test equation [45].

For the direct application of generalized-α methods to constrained systems (4.1), the time-discrete equilibrium conditions (4.4) are substituted by

$$\mathbf{M}(\boldsymbol{q}_{n+1})\ddot{\boldsymbol{q}}_{n+1} = \boldsymbol{f}(\boldsymbol{q}_{n+1}, \dot{\boldsymbol{q}}_{n+1}) - \mathbf{G}^\top(\boldsymbol{q}_{n+1})\,\boldsymbol{\lambda}_{n+1}\,, \tag{4.6a}$$

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q}_{n+1})\,, \tag{4.6b}$$

see [9] and the earlier work of Cardona and Géradin [31] and Negrut et al. [66] who applied HHT-α methods to constrained systems (4.1) with constant mass matrix $\mathbf{M}$. In each time step, the numerical solution is defined implicitly by linear update formulae (4.3) and nonlinear equilibrium conditions (4.6). Taking into account the linear equations (4.3), we may express $\boldsymbol{q}_{n+1}$, $\dot{\boldsymbol{q}}_{n+1}$, $\ddot{\boldsymbol{q}}_{n+1}$ and $\boldsymbol{a}_{n+1}$ in terms of the scaled increment

$$\boldsymbol{\Delta q}_n := \dot{\boldsymbol{q}}_n + h(0.5-\beta)\boldsymbol{a}_n + h\beta\boldsymbol{a}_{n+1} \tag{4.7a}$$

in the position update (4.3a) and get

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n) = \boldsymbol{q}_n + h\boldsymbol{\Delta q}_n\,, \tag{4.7b}$$

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_{n+1}(\boldsymbol{\Delta q}_n) = \frac{1}{\beta h}(\boldsymbol{\Delta q}_n - \dot{\boldsymbol{q}}_n - (0.5-\beta)h\boldsymbol{a}_n)\,, \tag{4.7c}$$

$$\dot{\boldsymbol{q}}_{n+1} = \dot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n) = \frac{\gamma}{\beta}\boldsymbol{\Delta q}_n + (1-\frac{\gamma}{\beta})\dot{\boldsymbol{q}}_n + h(1-\frac{\gamma}{2\beta})\boldsymbol{a}_n\,, \tag{4.7d}$$

$$\ddot{\boldsymbol{q}}_{n+1} = \ddot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n) = \frac{1-\alpha_m}{\beta(1-\alpha_f)}(\frac{\boldsymbol{\Delta q}_n - \dot{\boldsymbol{q}}_n}{h} - 0.5\boldsymbol{a}_n) + \frac{\boldsymbol{a}_n - \alpha_f\ddot{\boldsymbol{q}}_n}{1-\alpha_f}\,. \tag{4.7e}$$

In that way, the nonlinear system (4.3), (4.6) is condensed to $n_{\boldsymbol{q}} + n_{\boldsymbol{\lambda}}$ nonlinear equations

$$\mathbf{0} = \boldsymbol{r}_h^{n+1}(\boldsymbol{\Delta q}_n, h\boldsymbol{\lambda}_{n+1})\,, \tag{4.8a}$$

$$\mathbf{0} = \boldsymbol{g}_h^{n+1}(\boldsymbol{\Delta q}_n) \tag{4.8b}$$

in terms of $\boldsymbol{\Delta q}_n$ and $h\boldsymbol{\lambda}_{n+1}$. The nonlinear functions

$$\begin{aligned}\boldsymbol{r}_h^{n+1}(\boldsymbol{\Delta q}_n, h\boldsymbol{\lambda}_{n+1}) &:= \mathbf{M}(\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n))h\ddot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n)\\ &\quad - h\boldsymbol{f}(\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n), \dot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n)) + \mathbf{G}^\top(\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n))h\boldsymbol{\lambda}_{n+1}\,,\\ \boldsymbol{g}_h^{n+1}(\boldsymbol{\Delta q}_n) &:= \frac{1}{h}\boldsymbol{g}(\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n))\,.\end{aligned}$$

are defined by the constrained equilibrium conditions (4.6). They are scaled such that the Jacobian

$$\begin{pmatrix} \dfrac{\partial \boldsymbol{r}_h^{n+1}}{\partial \boldsymbol{\Delta q}_n} & \dfrac{\partial \boldsymbol{r}_h^{n+1}}{\partial (h\boldsymbol{\lambda}_{n+1})} \\ \dfrac{\partial \boldsymbol{g}_h^{n+1}}{\partial \boldsymbol{\Delta q}_n} & \dfrac{\partial \boldsymbol{g}_h^{n+1}}{\partial (h\boldsymbol{\lambda}_{n+1})} \end{pmatrix} = \begin{pmatrix} \dfrac{1-\alpha_m}{\beta(1-\alpha_f)}\mathbf{M}(\boldsymbol{q}_n) + \mathscr{O}(h) & \mathbf{G}^\top(\boldsymbol{q}_n) + \mathscr{O}(h) \\ \mathbf{G}(\boldsymbol{q}_n) + \mathscr{O}(h) & \mathbf{0} \end{pmatrix} \tag{4.9}$$

and its inverse remain bounded for $h \to 0$ for algorithmic parameters α_f, α_m, β according to [33] or [51]. Scaling techniques are mandatory for the direct time discretization of higher index DAEs [70], see also [47]. For the application to constrained mechanical systems, they have been studied again more recently in [23].

Note that the Jacobian in (4.9) has the characteristic 2×2 block structure that was considered in Lemma 2.1 above. For sufficiently small time step sizes $h > 0$, it is non-singular since the constraint matrix $\mathbf{G}(\boldsymbol{q}_n)$ was assumed to have full rank and the positive semi-definite mass matrix $\mathbf{M}(\boldsymbol{q}_n)$ is positive definite on the null space of $\mathbf{G}(\boldsymbol{q}_n)$.

Generalized-α and HHT-α methods for constrained systems have been used successfully in large scale practical applications [45, 66]. They may, however, suffer from a strange solution behaviour in transient phases after initialization and step size changes:

Example 4.1 (See [15, Example 1]) We consider two different initial configurations of the mathematical pendulum with equations of motion (2.6) and physical parameters $m = 1.0$, $l = 1.0$, $g_{\text{grav}} = 9.81$ (physical units are omitted). The consistent initial values x_0, y_0, $\dot{x}_0$, $\dot{y}_0$, λ_0 are defined such that the total mechanical energy $T + U = m(\dot{x}^2 + \dot{y}^2)/2 + mg_{\text{grav}}y$ at $t = t_0$ is given by $T_0 + U_0 = m/2 - mg_{\text{grav}}l$. In that way, all solution trajectories with initial values $x_0 \geq 0$, $\dot{x}_0 \geq 0$ coincide up to a phase shift to the one with initial values $x_0 = 0.0$, $y_0 = -1.0$, $\dot{x}_0 = 1.0$, $\dot{y}_0 = 0.0$, $\lambda_0 = 10.81$.

Figure 10 shows the global errors in the Lagrange multiplier λ for the generalized-α method (4.3), (4.6) with a damping ratio $\rho_\infty = 0.9$ and algorithmic parameters α_f, α_m, β and γ according to [33]. The left plot shows $\lambda(t_n) - \lambda_n$ for simulations that start at the equilibrium position $(x_0, y_0) = (0, -l)$. Comparing the simulation results for time step size $h = \bar{h} := 0.02$ (marked by "×") and the

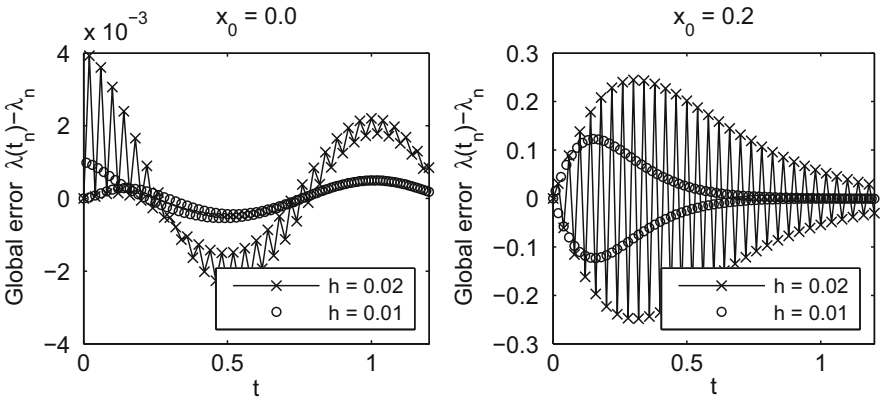


Fig. 10 Global error $\lambda(t_n) - \lambda_n$ of the generalized-α method (4.3), (4.6) for the equations of motion (2.6) of the mathematical pendulum with initial values $x_0 = 0$ (*left plot*) and $x_0 = 0.2$ (*right plot*)

ones for time step size $h = \bar{h}/2 = 0.01$ (marked by "∘"), we observe second order convergence since the maximum amplitudes are reduced from 4.0×10^{-3} to 1.0×10^{-3}, i.e., by a factor of $2^2 = 4$.

For simulations starting at $x_0 = 0.2$, the global errors are two orders of magnitude larger than before. The right plot of Fig. 10 shows a first order error term with maximum values that are reduced from 0.24 to 0.12 if the time step size is reduced by a factor of 2. This large oscillating error term is damped out after about 100 time steps. The detailed error analysis in [15, Remark 3a] shows that $|\lambda(t_n) - \lambda_n|$ is bounded by $C_1 n^2 \rho_\infty^n |\dot{y}_0| h + C_2 h^2$ with suitable constants $C_1, C_2 > 0$.

The test results in Fig. 10 are not sensitive to the initialization of (4.3) by starting values $(\boldsymbol{q}_0, \dot{\boldsymbol{q}}_0, \ddot{\boldsymbol{q}}_0, \boldsymbol{a}_0)$ with $\boldsymbol{q}_0 = \boldsymbol{q}(t_0)$, $\dot{\boldsymbol{q}}_0 = \dot{\boldsymbol{q}}(t_0)$, $\ddot{\boldsymbol{q}}_0 = \ddot{\boldsymbol{q}}(t_0)$ and $\boldsymbol{a}_0 = \ddot{\boldsymbol{q}}(t_0) + \mathscr{O}(h)$. In the numerical tests, we used starting values $\boldsymbol{a}_0 = \ddot{\boldsymbol{q}}(t_0 + (\alpha_m - \alpha_f)h) + \mathscr{O}(h^2)$ that are optimal in the sense that the local truncation error in (4.3c) is of size $\mathscr{O}(h^2)$ if $(\ddot{\boldsymbol{q}}_{n+\iota}, \boldsymbol{a}_{n+\iota})$ is substituted by $\big(\ddot{\boldsymbol{q}}(t_n + \iota h), \ddot{\boldsymbol{q}}(t_n + (\iota + \alpha_m - \alpha_f)h)\big)$, $(\iota = 0, 1)$, see, e.g., [55].

Example 4.1 illustrates that the direct application of generalized-α methods to constrained systems (4.1) may result in order reduction with a large transient oscillating error term that depends in a nontrivial way on the initial values $\boldsymbol{q}(t_0)$, $\dot{\boldsymbol{q}}(t_0)$. Modified starting values $\dot{\boldsymbol{q}}_0 = \dot{\boldsymbol{q}}(t_0) + \boldsymbol{\Delta}_0$ with a correction term $\boldsymbol{\Delta}_0$ of size $\mathscr{O}(h^2)$ have been proposed to eliminate this first order error term and to regain second order convergence [15].

Newmark type integrators like the HHT-α method and the generalized-α method are tailored to second order systems (4.1) and (4.2). They may, however, be extended to models with additional first order differential equations [28, 54]. In multibody dynamics, such coupled systems of first and second order differential equations are typical of systems with hydraulic components, see Sect. 3.2.

4.1.2 ODE and DAE Time Integration Methods for First Order Systems

As an alternative to time integration methods for the second order differential equations from structural dynamics we introduce velocity coordinates $\boldsymbol{v} := \dot{\boldsymbol{q}}$ and consider ODE and DAE time integration methods for first order systems that are applied to the linearly implicit DAE

$$\mathbf{0} = \mathbf{F}(t, \boldsymbol{x}, \dot{\boldsymbol{x}}) := \begin{pmatrix} \dot{\boldsymbol{q}} - \boldsymbol{v} \\ \mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} - \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}) + \mathbf{G}^\top(\boldsymbol{q})\boldsymbol{\lambda} \\ \boldsymbol{g}(\boldsymbol{q}) \end{pmatrix} \quad \text{with} \quad \boldsymbol{x} := \begin{pmatrix} \boldsymbol{q} \\ \boldsymbol{v} \\ \boldsymbol{\lambda} \end{pmatrix} \tag{4.10}$$

that is equivalent to the constrained system (4.1), see Remark 2.3. Implicit Runge–Kutta methods and implicit multi-step methods have originally been developed for the time integration of first order ODEs

$$\dot{\boldsymbol{x}} = \boldsymbol{\varphi}(t, \boldsymbol{x}) \tag{4.11}$$

but may (formally) be applied to DAE (4.10) as well:

Remark 4.1

(a) An implicit Runge–Kutta method uses s stage vectors

$$\mathbf{X}_{ni} = \boldsymbol{x}_n + h \sum_{j=1}^{s} a_{ij} \dot{\mathbf{X}}_{nj} \,, \ (i = 1, \ldots, s) \tag{4.12a}$$

to update the numerical solution $\boldsymbol{x}_n \approx \boldsymbol{x}(t_n)$ in time step $t_n \to t_{n+1} = t_n + h$ according to

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n + h \sum_{i=1}^{s} b_i \dot{\mathbf{X}}_{ni} \tag{4.12b}$$

with stage vectors $\dot{\mathbf{X}}_{ni} \approx \dot{\boldsymbol{x}}(t_n + c_i h)$ that are connected to $\mathbf{X}_{ni} \approx \boldsymbol{x}(t_n + c_i h)$ by

$$\dot{\mathbf{X}}_{ni} = \boldsymbol{\varphi}(t_n + c_i h, \mathbf{X}_{ni}) \,, \ (i = 1, \ldots, s) \tag{4.12c}$$

in the ODE case and by

$$\mathbf{0} = \mathbf{F}(t_n + c_i h, \mathbf{X}_{ni}, \dot{\mathbf{X}}_{ni}) \,, \ (i = 1, \ldots, s) \tag{4.12d}$$

in the application to DAE (4.10). The method is characterized by nodes c_i, weights b_i and Runge–Kutta parameters a_{ij}, $(i, j = 1, \ldots, s)$. The application of implicit Runge–Kutta methods to higher index DAEs was studied, e.g., in [69] and [47].

(b) For linear k-step methods with parameters α_j, β_j, $(j = 0, 1, \ldots, k)$, we have

$$\frac{1}{h} \sum_{j=0}^{k} \alpha_j \boldsymbol{x}_{n+1-j} = \sum_{j=0}^{k} \beta_j \dot{\boldsymbol{x}}_{n+1-j} \,. \tag{4.13a}$$

In time step $t_n \to t_{n+1} = t_n + h$, the vectors $\boldsymbol{x}_{n+1-j} \approx \boldsymbol{x}(t_n - (j-1)h)$, $\dot{\boldsymbol{x}}_{n+1-j} \approx \dot{\boldsymbol{x}}(t_n - (j-1)h)$, $(j = 1, \ldots, k)$ are assumed to be known and the numerical solution $\boldsymbol{x}_{n+1} \approx \boldsymbol{x}(t_{n+1})$ is defined such that

$$\dot{\boldsymbol{x}}_{n+1} = \boldsymbol{\varphi}(t_{n+1}, \boldsymbol{x}_{n+1}) \tag{4.13b}$$

for ODEs and

$$\mathbf{0} = \mathbf{F}(t_{n+1}, \boldsymbol{x}_{n+1}, \dot{\boldsymbol{x}}_{n+1}) \tag{4.13c}$$

in the application to DAEs (4.10). Considering k-step methods (4.13) with parameters $\beta_0 = 1$, $\beta_j = 0$, $(j = 1, \ldots, k)$, we may eliminate $\dot{\boldsymbol{x}}_{n+1}$ and get the k-step BDF methods

$$\mathbf{0} = \mathbf{F}(t_{n+1}, \boldsymbol{x}_{n+1}, \frac{1}{h}\sum_{j=0}^{k} \alpha_j \boldsymbol{x}_{n+1-j}) \,. \tag{4.14}$$

BDF are the most frequently used DAE time integration methods in technical simulation since they may be combined with very efficient step size and order control strategies [26].

For constrained mechanical systems (4.10), fixed step size BDF (4.14) define the update of position coordinates $\boldsymbol{q}_n$ explicitly in terms of $\boldsymbol{v}_{n+1}$:

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1}) = \frac{h}{\alpha_0}\boldsymbol{v}_{n+1} - \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0}\boldsymbol{q}_{n+1-j} \,.$$

Similar to generalized-α methods, we may use this expression to eliminate $\boldsymbol{q}_{n+1}$ in the BDF definition (4.14) and get a condensed system of $n_q + n_\lambda$ nonlinear equations

$$\mathbf{0} = \boldsymbol{r}_h^{n+1}(\boldsymbol{v}_{n+1}, h\boldsymbol{\lambda}_{n+1}) \,, \tag{4.15a}$$

$$\mathbf{0} = \boldsymbol{g}_h^{n+1}(\boldsymbol{v}_{n+1}) \tag{4.15b}$$

with

$$\begin{aligned} \boldsymbol{r}_h^{n+1}(\boldsymbol{v}_{n+1}, h\boldsymbol{\lambda}_{n+1}) &:= \mathbf{M}(\boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1})) \sum_{j=0}^{k} \alpha_j \boldsymbol{v}_{n+1-j} \\ &\quad - h\boldsymbol{f}(\boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1}), \boldsymbol{v}_{n+1}) + \mathbf{G}^\top(\boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1}))\, h\boldsymbol{\lambda}_{n+1} \,, \\ \boldsymbol{g}_h^{n+1}(\boldsymbol{v}_{n+1}) &:= \frac{1}{h}\boldsymbol{g}(\boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1})) \,, \end{aligned}$$

see (4.8). The Jacobian

$$\begin{pmatrix} \dfrac{\partial \boldsymbol{r}_h^{n+1}}{\partial \boldsymbol{v}_{n+1}} & \dfrac{\partial \boldsymbol{r}_h^{n+1}}{\partial (h\boldsymbol{\lambda}_{n+1})} \\ \dfrac{\partial \boldsymbol{g}_h^{n+1}}{\partial \boldsymbol{v}_{n+1}} & \dfrac{\partial \boldsymbol{g}_h^{n+1}}{\partial (h\boldsymbol{\lambda}_{n+1})} \end{pmatrix} = \begin{pmatrix} \alpha_0 \mathbf{M}\big(\boldsymbol{q}_{n+1}(\mathbf{0})\big) + \mathscr{O}(h) & \mathbf{G}^\top\big(\boldsymbol{q}_{n+1}(\mathbf{0})\big) + \mathscr{O}(h) \\ \mathbf{G}\big(\boldsymbol{q}_{n+1}(\mathbf{0})\big) + \mathscr{O}(h) & \mathbf{0} \end{pmatrix}$$

has the characteristic 2×2 block structure (2.18). The Jacobian and its inverse are bounded for $h \to 0$.

The direct application of implicit Runge–Kutta methods and BDF to the constrained system (4.10) may again result in order reduction [26, 47]. For variable time step sizes, the numerical solution may even fail to converge [68]:

Example 4.2 (See [6, Sect. 5]) The backward Euler method is both a one-stage implicit Runge–Kutta method (4.12) with parameters $a_{11} = b_1 = c_1 = 1$ and the one-step BDF (4.14). For equations of motion (4.10) with mass matrix $\mathbf{M} = \mathbf{I}$, force vector $\boldsymbol{f} = \mathbf{0}$ and time dependent constraints $\mathbf{0} = \boldsymbol{g}(t, \boldsymbol{q}(t)) := \mathbf{C}\boldsymbol{q}(t) - \boldsymbol{z}(t)$, it is defined by

$$\begin{aligned} \frac{\boldsymbol{q}_{n+1} - \boldsymbol{q}_n}{h_n} &= \boldsymbol{v}_{n+1}\,, \\ \frac{\boldsymbol{v}_{n+1} - \boldsymbol{v}_n}{h_n} &= -\mathbf{C}^\top \boldsymbol{\lambda}_{n+1}\,, \\ \mathbf{0} &= \mathbf{C}\boldsymbol{q}_{n+1} - \boldsymbol{z}(t_{n+1}) \end{aligned}$$

with h_n denoting the (variable) time step size of time step $t_n \to t_{n+1} = t_n + h_n$. Straightforward computations show that

$$\boldsymbol{\lambda}_{n+1} = -\frac{h_n + h_{n-1}}{2h_n}(\mathbf{C}\mathbf{C}^\top)^{-1}\ddot{\boldsymbol{z}}(t_{n+1}) + \mathscr{O}(h_n) + \mathscr{O}(\frac{h_{n-1}^2}{h_n})$$

if $\boldsymbol{z}$ is three times continuously differentiable. For $h_{n-1} \to 0$, $h_n \to 0$ and a fixed step size ratio $\sigma_n := h_n/h_{n-1} \neq 1$, the numerical solution $\boldsymbol{\lambda}_{n+1}$ does *not* converge to the analytical solution $\boldsymbol{\lambda}(t_{n+1}) = -(\mathbf{C}\mathbf{C}^\top)^{-1}\ddot{\boldsymbol{z}}(t_{n+1})$ that is obtained differentiating the constraints $\mathbf{0} = \mathbf{C}\boldsymbol{q}(t) - \boldsymbol{z}(t)$ twice and inserting $\ddot{\boldsymbol{q}} = -\mathbf{C}^\top \boldsymbol{\lambda}(t)$ afterwards.

The direct application of implicit ODE time integration methods to DAEs (4.1) and (4.10) is intuitive and may be extended straightforwardly to more complex model equations including, e.g., nonholonomic constraints or additional algebraic equations $\mathbf{0} = \boldsymbol{h}(\boldsymbol{q}, \boldsymbol{s})$ with non-singular Jacobian $\partial \boldsymbol{h}/\partial \boldsymbol{s}$, see (3.12). Special care is needed in the practical implementation of these methods to address ill-conditioning of iteration matrices [23, 47, 70] and reliable estimation of local errors in step size control algorithms [26, 50].

In Examples 4.1 and 4.2, we have verified by two trivial test problems that the direct time discretization of the constrained equations of motion by ODE methods results systematically in poor simulation results for the Lagrange multipliers $\boldsymbol{\lambda}$. For more realistic multibody system models from practical applications, these numerical problems may affect the result accuracy of position and velocity coordinates as well.

In [5], we discussed this numerical effect for test scenarios from railway dynamics. The dynamical behaviour of rail vehicles is strongly influenced by the contact and friction forces between wheel and rail. In a rigid body contact model, the permanent contact between wheel and rail is described by holonomic constraints [84]. In the constrained system (4.10), the friction forces are part of the force vector $\boldsymbol{f}$. They depend nonlinearly on the wheel-rail contact forces $-\mathbf{G}^\top(\boldsymbol{q})\ \boldsymbol{\lambda}$,

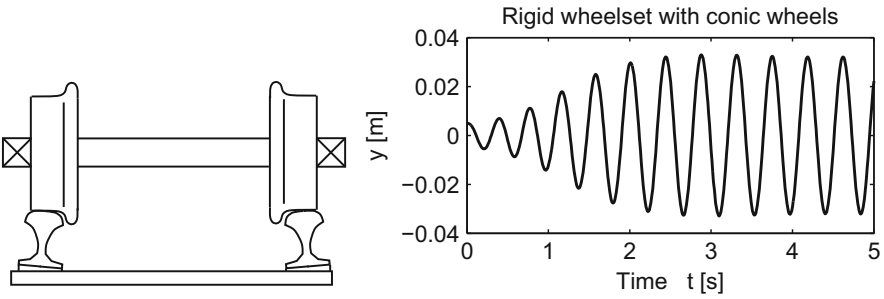


Fig. 11 Lateral displacement $y(t)$ of a rigid wheelset performing a hunting motion [6, Fig. 8]

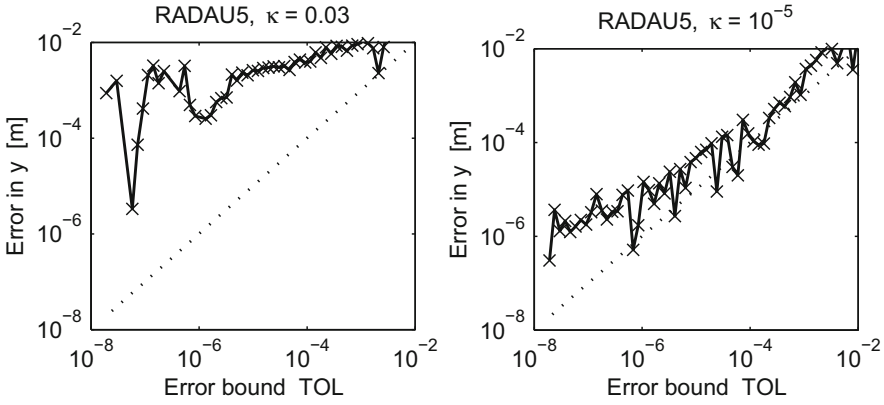


Fig. 12 Global error of the implicit Runge–Kutta solver RADAU5 being applied directly to DAE (4.10): hunting motion of a rigid wheelset [6, Fig. 9]

see [57, 84]. As a practical consequence, we get equations of motion (4.10) with $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}, \boldsymbol{\lambda})$.

We present numerical test results for a rigid wheelset following a straight track, see Fig. 11. These test results were published before in [6, Sect. 5]:

Example 4.3 We consider the multibody system model of a rigid wheelset with conic wheels moving with constant speed along a straight track. It is a well-known phenomenon from railway dynamics that the central position of the wheelset on the track gets unstable if the speed exceeds the so-called critical speed. Starting with a small initial lateral displacement $y(t_0)$ the wheelset oscillates in lateral direction (*hunting motion*), see Fig. 11.

The rigid wheelset has six degrees of freedom and is described by position coordinates $\boldsymbol{q}(t) \in \mathbb{R}^6$. The permanent contact between the two wheels and the rails is modelled by two holonomic constraints resulting in contact forces $-\mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\lambda}$ with $\boldsymbol{\lambda}(t) \in \mathbb{R}^2$.

Figure 12 shows numerical test results for the implicit Runge–Kutta solver RADAU5 that adjusts its (variable) time step size h_n automatically to meet user-

defined error bounds [50]. Practical experience in ODE applications shows that RADAU5 keeps the global error of the numerical solution usually well below these user-defined error tolerances TOL. But in the application to DAE (4.10), the relative errors remain even for very small error bounds in the size of $0.1 \ldots 1.0\,\%$. As a typical example, the left plot of Fig. 12 shows the error in the lateral displacement y for various values of TOL.

To analyse these unsatisfactory results, we modify one of the internal solver parameters. BDF (4.14) and implicit Runge–Kutta methods (4.12) define $\boldsymbol{x}_{n+1}$ solving a system of nonlinear equations. In the practical implementation this system is solved iteratively by Newton's method that is stopped if the residual is less than $\kappa \cdot$TOL. In this stopping criterion the user-defined error tolerance for time integration (TOL) is scaled by a constant $\kappa \leq 1$ that is a free control parameter of the solver. Default values are $\kappa = 0.33$ in the BDF solver DASSL, see [26], and $\kappa = 0.03$ in the implicit Runge–Kutta solver RADAU5, see [50].

The right plot of Fig. 12 shows that the error in time integration is reduced drastically and remains now roughly in the size of the error bounds TOL if κ is set to the very small value $\kappa = 10^{-5}$. The comparison of left and right plot in Fig. 12 illustrates that the direct application of RADAU5 to DAE (4.10) makes the solver very sensitive to (small) iteration errors in Newton's method. This practical observation coincides with the results of a detailed perturbation analysis for analytical and numerical solution [2].

4.2 Index Reduction and Projection

In the direct application of ODE time integration methods to DAE (4.10) the robustness of the solvers may be improved substantially by small values of κ that result in (very) small iteration errors in Newton's method, see the right plot of Fig. 12. On the other hand, values of κ that are less than 10^{-3} increase the number of Newton steps per time step substantially and may slow down the solver dramatically.

Instead of applying ODE time integration methods directly to (4.10) it proved to be much more advantageous to transform the equations of motion analytically before time integration. This *index reduction* is the key to the robust and efficient dynamical simulation of constrained systems. It results in several analytically equivalent DAE formulations of the constrained system that is originally given in its *index-3 formulation* [50]

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{4.16a}$$

$$\mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}) - \mathbf{G}^{\top}(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.16b}$$

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q}) \tag{4.16c}$$

with position coordinates $\boldsymbol{q}$, velocity coordinates $\boldsymbol{v}$ and Lagrange multipliers $\boldsymbol{\lambda}$, see (4.10) and Remark 2.3. Note that the index of (4.16) is three if the mass matrix $\mathbf{M}(\boldsymbol{q})$ is positive definite but may be less than three for rank-deficient mass matrices, see Example 2.5.

4.2.1 Index-2 Formulation

Substituting the constraints (4.16c) by the corresponding hidden constraints

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{q}(t)) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(\boldsymbol{q}(t))\,\dot{\boldsymbol{q}}(t) = \mathbf{G}(\boldsymbol{q}(t))\boldsymbol{v}(t) \tag{4.17}$$

at the level of velocity coordinates, see (2.8), we get the *index-2 formulation* [50]

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{4.18a}$$

$$\mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(\boldsymbol{q},\boldsymbol{v}) - \mathbf{G}^{\top}(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.18b}$$

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q})\boldsymbol{v} \tag{4.18c}$$

that is analytically equivalent to the original equations of motion (4.16) if the initial values $\boldsymbol{q}_0$ are consistent with the holonomic constraints (4.16c) since $\boldsymbol{g}(\boldsymbol{q}_0) = \mathbf{0}$ implies

$$\boldsymbol{g}(\boldsymbol{q}(t)) = \underbrace{\boldsymbol{g}(\boldsymbol{q}(t_0))}_{=\boldsymbol{g}(\boldsymbol{q}_0)=\mathbf{0}} + \int_{t_0}^{t} \underbrace{\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{q}(\tau))}_{=\mathbf{0},\ \text{see (4.17) and (4.18c)}}\,\mathrm{d}\tau = \mathbf{0} \tag{4.19}$$

for all $t \geq t_0$. Following step by step the index analysis in Remark 2.3, we see that DAEs (4.18) with positive definite mass matrices $\mathbf{M}(\boldsymbol{q})$ have differentiation index and perturbation index two.

Example 4.4 The perturbation analysis for index-2 systems (4.18) shows that the numerical solution is much less sensitive w.r.t. small constraint residuals than in the index-3 case [2]. This is nicely illustrated by numerical test results for the wheelset benchmark of Example 4.3. Applying the implicit Runge–Kutta solver RADAU5 to the index-2 formulation (4.18) of the equations of motion, we get much smaller errors than before, see Fig. 13. For default solver settings (left plot, $\kappa = 0.03$), the error remains roughly in the size of the user prescribed error tolerances TOL with some error saturation at the level of 10^{-6} for tolerances TOL $\leq 10^{-6}$. Further improvements are achieved by enforcing very small constraint residuals (right plot, $\kappa = 10^{-5}$) but these highly accurate simulation results require again much more computing time than the simulation with standard solver settings.

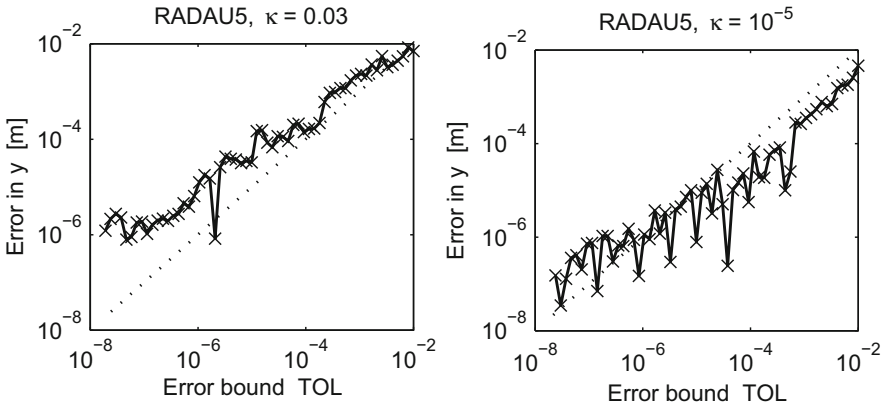


Fig. 13 Global error of the implicit Runge–Kutta solver RADAU5 being applied to the index-2 formulation (4.18) of the equations of motion: hunting motion of a rigid wheelset [5, Fig. 2.1]

The time discretization of index-2 DAEs by implicit Runge–Kutta methods and BDF is discussed in the monographs [26, 47, 50]. For generalized-α and HHT-α methods, the combination of index reduction and time discretization is studied, e.g., in [55, 56, 63], see also the recent analysis for configuration spaces with Lie group structure in [15]. The practical implementation of these implicit time integration methods for the index-2 formulation (4.18) follows the implementation scheme that was discussed in Sect. 4.1. In the systems of nonlinear equations (4.8) and (4.15), equations $\mathbf{0} = \boldsymbol{g}_h^{n+1}$ have to be substituted by

$$\mathbf{0} = \dot{\boldsymbol{g}}_h^{n+1}(\boldsymbol{\Delta q}_n) := \mathbf{G}(\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n))\dot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n)$$

for the generalized-α method and by

$$\mathbf{0} = \dot{\boldsymbol{g}}_h^{n+1}(\boldsymbol{v}_{n+1}) := \mathbf{G}(\boldsymbol{q}_{n+1}(\boldsymbol{v}_{n+1}))\boldsymbol{v}_{n+1}$$

in the BDF case.

Remark 4.2 For non-stiff constrained systems, the use of *half-explicit* Runge–Kutta methods for the index-2 formulation (4.18) of the equations of motion proves to be favourable since these methods avoid the solution of systems of nonlinear equations. The s stage half-explicit method has nodes c_i, weights b_i and Runge–Kutta parameters a_{ij}, ($i = 1, \ldots, s, j = 1, \ldots, i-1$). It updates the numerical solution $(\boldsymbol{q}_n, \boldsymbol{v}_n, \boldsymbol{\lambda}_n)$ in time step $t_n \to t_{n+1} = t_n + h$ using stage vectors

$$\mathbf{Q}_{ni} \approx \boldsymbol{q}(t_n + c_i h)\,,\ \mathbf{V}_{ni} \approx \boldsymbol{v}(t_n + c_i h)\,,\ \dot{\mathbf{V}}_{ni} \approx \dot{\boldsymbol{v}}(t_n + c_i h)\,,\ \boldsymbol{\Lambda}_{ni} \approx \boldsymbol{\lambda}(t_n + c_i h)\,,$$

($i = 1, \ldots, s$), that are initialized by

$$\mathbf{Q}_{n1} = \boldsymbol{q}_n\,, \quad \mathbf{V}_{n1} = \boldsymbol{v}_n\,, \quad \dot{\mathbf{V}}_{n1} = \dot{\boldsymbol{v}}_n\,, \quad \boldsymbol{\Lambda}_{n1} = \boldsymbol{\lambda}_n \tag{4.20a}$$

with $\dot{\boldsymbol{v}}_n$ satisfying the dynamical equations at $t = t_n$:

$$\mathbf{M}(\boldsymbol{q}_n)\dot{\boldsymbol{v}}_n = \boldsymbol{f}(\boldsymbol{q}_n, \boldsymbol{v}_n) - \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\lambda}_n \tag{4.20b}$$

The local error analysis shows that the method should start with an explicit stage

$$\mathbf{Q}_{n2} = \boldsymbol{q}_n + ha_{21}\mathbf{V}_{n1} = \boldsymbol{q}_n + ha_{21}\boldsymbol{v}_n \tag{4.20c}$$

to avoid order reduction [4, 10]. In the remaining $s-1$ stages, we may suppose that the stage vectors $\mathbf{Q}_{ni}$ and $(\mathbf{V}_{nj}, \dot{\mathbf{V}}_{nj})$, $(j = 1, \ldots, i-1)$, are known such that the stage vectors

$$\mathbf{V}_{ni} = \boldsymbol{v}_n + h\sum_{j=1}^{i-1} a_{ij}\dot{\mathbf{V}}_{nj}\,, \quad \mathbf{Q}_{n,i+1} = \boldsymbol{q}_n + h\sum_{j=1}^{i} a_{i+1,j}\mathbf{V}_{nj} \tag{4.20d}$$

may be computed explicitly. (For $i = s$ we use for simplicity $a_{s+1,j} := b_j$.)

According to [25], a half-explicit Runge–Kutta stage for index-2 systems (4.18) is defined by the dynamical equations (4.18b) at $t = t_n + c_i h$ and by the constraints (4.18c) that are evaluated at $t = t_n + c_{i+1}h$ using the stage vector $\mathbf{Q}_{n,i+1}$ from (4.20d):

$$\mathbf{M}(\mathbf{Q}_{ni})\dot{\mathbf{V}}_{ni} = \boldsymbol{f}(\mathbf{Q}_{ni}, \mathbf{V}_{ni}) - \mathbf{G}^\top(\mathbf{Q}_{ni})\,\boldsymbol{\Lambda}_{ni}\,,$$

$$\mathbf{0} = \mathbf{G}(\mathbf{Q}_{n,i+1})\mathbf{V}_{n,i+1} \quad \text{with} \quad \mathbf{V}_{n,i+1} = \boldsymbol{v}_n + h\sum_{j=1}^{i} a_{i+1,j}\dot{\mathbf{V}}_{nj}\,.$$

These equations are linear in the unknown stage vectors $\dot{\mathbf{V}}_{ni}$, $\boldsymbol{\Lambda}_{ni}$ and may be summarized to a system of $n_{\boldsymbol{q}} + n_{\boldsymbol{\lambda}}$ linear equations:

$$\begin{pmatrix} \mathbf{M}(\mathbf{Q}_{ni}) & \mathbf{G}^\top(\mathbf{Q}_{ni}) \\ \mathbf{G}(\mathbf{Q}_{n,i+1}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{V}}_{ni} \\ \boldsymbol{\Lambda}_{ni} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}(\mathbf{Q}_{ni}, \mathbf{V}_{ni}) \\ -\dfrac{1}{ha_{i+1,i}}\mathbf{G}(\mathbf{Q}_{n,i+1})\big(\boldsymbol{v}_n + h\displaystyle\sum_{j=1}^{i-1} a_{i+1,j}\dot{\mathbf{V}}_{nj}\big) \end{pmatrix}. \tag{4.20e}$$

Because of $\mathbf{Q}_{ni} = \boldsymbol{q}_n + \mathscr{O}(h)$, $\mathbf{Q}_{n,i+1} = \boldsymbol{q}_n + \mathscr{O}(h)$, these equations are uniquely solvable w.r.t. $\dot{\mathbf{V}}_{ni}$ and $\boldsymbol{\Lambda}_{ni}$ if the assumptions of Lemma 2.1 are satisfied with matrices $\mathbf{M} = \mathbf{M}(\boldsymbol{q}_n)$, $\mathbf{G} = \mathbf{G}(\boldsymbol{q}_n)$ and the time step size $h > 0$ is sufficiently small.

With (4.20d) and (4.20e) we compute stage by stage vectors $\mathbf{Q}_{n,i+1}$, $\mathbf{V}_{ni}$, $\dot{\mathbf{V}}_{ni}$ and $\boldsymbol{\Lambda}_{ni}$ for $i = 2, \ldots, s$. Finally, the numerical solution at $t = t_{n+1}$ is given by

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + h \sum_{i=1}^{s} b_i \mathbf{V}_{ni}\,, \quad \boldsymbol{v}_{n+1} = \boldsymbol{v}_n + h \sum_{i=1}^{s} b_i \dot{\mathbf{V}}_{ni}\,, \quad \boldsymbol{\lambda}_{n+1} = \sum_{i=1}^{s} d_i \boldsymbol{\Lambda}_{ni} \tag{4.20f}$$

with new algorithmic parameters d_i, $(i = 1, \ldots, s)$, being defined by order conditions and by a contractivity condition that has to be satisfied to guarantee zero-stability and convergence [4, 10], see also [50, Sect. VII.6].

The fifth order explicit Runge–Kutta method of Dormand and Prince [34] has $s = 6$ stages and may be extended to a half-explicit Runge–Kutta method (4.20) of order $p = 5$ with $\hat{s} = 7$ stages that was implemented in the solver HEDOP5 for non-stiff constrained systems [4]. Numerical tests for a wheel suspension benchmark problem [81] illustrate that half-explicit solvers like HEDOP5 are superior to the implicit BDF solver DASSL if the equations of motion (4.18) are non-stiff [4, 10]. On the other hand, implicit solvers are more flexible and may, e.g., be applied as well to systems with force vectors $\boldsymbol{f} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}, \boldsymbol{\lambda})$ that contain friction forces depending nonlinearly on the Lagrange multipliers $\boldsymbol{\lambda}$.

4.2.2 Index-1 Formulation

The index-2 formulation (4.18) was obtained substituting the holonomic constraints (4.16c) by their first time derivative. For index reduction, the second time derivative

$$\mathbf{0} = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \boldsymbol{g}(\boldsymbol{q}(t)) = \mathbf{G}(\boldsymbol{q}(t))\dot{\boldsymbol{v}}(t) + \boldsymbol{g}_{\boldsymbol{q}}(\boldsymbol{q}(t))\big(\boldsymbol{v}(t), \boldsymbol{v}(t)\big) \tag{4.21}$$

may be used as well. These hidden constraints at the level of acceleration coordinates have been used in Sect. 2.2 to prove the unique solvability of initial value problems for consistent initial values. They define the constraints in the *index-1 formulation* of the equations of motion:

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{4.22a}$$

$$\mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}) - \mathbf{G}^{\top}(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.22b}$$

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q})\dot{\boldsymbol{v}} + \boldsymbol{g}_{\boldsymbol{q}\boldsymbol{q}}(\boldsymbol{q})(\boldsymbol{v}, \boldsymbol{v})\,. \tag{4.22c}$$

Using similar arguments as in (4.19), we may verify that this index-1 formulation is equivalent to the original equations of motion (4.16) for any consistent initial values $\boldsymbol{q}_0$, $\boldsymbol{v}_0$ satisfying $\boldsymbol{g}(\boldsymbol{q}_0) = \mathbf{G}(\boldsymbol{q}_0)\boldsymbol{v}_0 = \mathbf{0}$.

The index-1 formulation is attractive from the numerical point of view since $\dot{\boldsymbol{v}}(t)$ and $\boldsymbol{\lambda}(t)$ may be eliminated from (4.22) solving a system of $n_{\boldsymbol{q}} + n_{\boldsymbol{\lambda}}$ linear equations,

see (2.17). Therefore, position and velocity coordinates may be obtained from the first order ODE

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{4.23a}$$

$$\dot{\boldsymbol{v}} = \boldsymbol{a}(\boldsymbol{q}, \boldsymbol{v}) \tag{4.23b}$$

with the right-hand side $\boldsymbol{a}(\boldsymbol{q}, \boldsymbol{v})$ being defined by

$$\begin{pmatrix} \mathbf{M}(\boldsymbol{q}) & \mathbf{G}^\top(\boldsymbol{q}) \\ \mathbf{G}(\boldsymbol{q}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{a} \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}(\boldsymbol{q}, \boldsymbol{v}) \\ -\boldsymbol{g}_{\boldsymbol{qq}}(\boldsymbol{q})(\boldsymbol{v}, \boldsymbol{v}) \end{pmatrix}. \tag{4.23c}$$

Initial value problems for (4.23) can be solved straightforwardly by any ODE time integration method including higher order explicit Runge–Kutta methods and predictor-corrector methods of Adams type.

Remark 4.3 Half-explicit Runge–Kutta methods for the index-1 formulation (4.22) of the equations of motion compute a numerical solution $(\boldsymbol{q}_n, \boldsymbol{v}_n)$ that is updated in time step $t_n \to t_{n+1} = t_n + h$ by s half-explicit stages. A half-explicit stage for the index-1 formulation (4.22) combines the explicit update

$$\mathbf{Q}_{ni} = \boldsymbol{q}_n + h \sum_{j=1}^{i-1} a_{ij} \mathbf{V}_{nj}\,, \qquad \mathbf{V}_{ni} = \boldsymbol{v}_n + h \sum_{j=1}^{i-1} a_{ij} \dot{\mathbf{V}}_{nj} \tag{4.24a}$$

with a system of $n_{\boldsymbol{q}} + n_{\boldsymbol{\lambda}}$ linear equations in terms of $\dot{\mathbf{V}}_{ni}$ and $\boldsymbol{\Lambda}_{ni}$:

$$\begin{pmatrix} \mathbf{M}(\mathbf{Q}_{ni}) & \mathbf{G}^\top(\mathbf{Q}_{ni}) \\ \mathbf{G}(\mathbf{Q}_{ni}) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \dot{\mathbf{V}}_{ni} \\ \boldsymbol{\Lambda}_{ni} \end{pmatrix} = \begin{pmatrix} \boldsymbol{f}(\mathbf{Q}_{ni}, \mathbf{V}_{ni}) \\ -\boldsymbol{g}_{\boldsymbol{qq}}(\mathbf{Q}_{ni})(\mathbf{V}_{ni}, \mathbf{V}_{ni}) \end{pmatrix}. \tag{4.24b}$$

With (4.24a) and (4.24b) we compute stage by stage vectors $\mathbf{Q}_{ni}$, $\mathbf{V}_{ni}$, $\dot{\mathbf{V}}_{ni}$ and $\boldsymbol{\Lambda}_{ni}$ for $i = 1, \ldots, s$. Finally, the numerical solution at $t = t_{n+1}$ is given by

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + h \sum_{i=1}^{s} b_i \mathbf{V}_{ni}\,, \qquad \boldsymbol{v}_{n+1} = \boldsymbol{v}_n + h \sum_{i=1}^{s} b_i \dot{\mathbf{V}}_{ni}\,. \tag{4.24c}$$

This solution strategy was implemented, e.g., in the half-explicit Runge–Kutta solver MDOP5 [80] that is based on the fifth order explicit Runge–Kutta method of Dormand and Prince [34]. For non-stiff problems, MDOP5 is as efficient as the half-explicit solver HEDOP5, see Remark 4.2. Numerical tests have shown that MDOP5 is slightly more efficient than HEDOP5 if the curvature term $\boldsymbol{g}_{\boldsymbol{qq}}(\boldsymbol{v}, \boldsymbol{v})$ may be evaluated with moderate numerical effort. On the other hand, the index-2 solver HEDOP5 is superior for problems with time consuming function evaluations $\boldsymbol{g}_{\boldsymbol{qq}}(\boldsymbol{v}, \boldsymbol{v})$ like the wheel suspension benchmark [81], see [4, 5].

4.2.3 Drift-Off Effect

Index reduction by differentiation does not only improve the robustness of implicit solvers substantially but offers additionally the chance to use explicit and half-explicit methods as well. The main drawback of index reduced formulations like (4.18) and (4.22) are large constraint residuals $\boldsymbol{g}(\boldsymbol{q}_n)$ in long-term simulations. This *drift-off effect* is illustrated by the numerical test results in Fig. 14 that show linearly growing constraint residuals of size 10^{-6} for the index-2 formulation (4.18) and quadratically growing constraint residuals of size 5.0×10^{-4} for the index-1 formulation (4.22).

Because of (4.19), the *analytical* solution of the index-2 formulation satisfies the original constraints $\boldsymbol{g}(\boldsymbol{q}) = \boldsymbol{0}$ exactly for all $t \geq t_0$. In the *numerical* solution the integrand $(\mathrm{d}\boldsymbol{g}/\mathrm{d}\, t)(\boldsymbol{q}(\tau))$ in (4.19) is still bounded by a small constant $\epsilon > 0$ but because of discretization and round-off errors it does *not* vanish identically. Therefore, the error in (4.16c) may increase linearly in time t:

$$\|\boldsymbol{g}(\boldsymbol{q}_n)\| \leq \|\boldsymbol{g}(\boldsymbol{q}_0)\| + \int_{t_0}^{t_n} \epsilon \, \mathrm{d}\, t = \epsilon \cdot (t_n - t_0)\,. \tag{4.25}$$

The numerical solution $\boldsymbol{q}_n$ *drifts* off the manifold $\mathfrak{M} = \{\, \boldsymbol{\eta} \,:\, \boldsymbol{g}(\boldsymbol{\eta}) = \boldsymbol{0} \,\}$ that is defined by the constraints (4.16c) on position level. The error bound ϵ summarizes discretization and round-off errors and the iteration errors of Newton's method.

For the index-1 formulation (4.22) a quadratic error growth

$$\|\boldsymbol{g}(\boldsymbol{q}_n)\| \leq \epsilon \cdot (t_n - t_0)^2$$

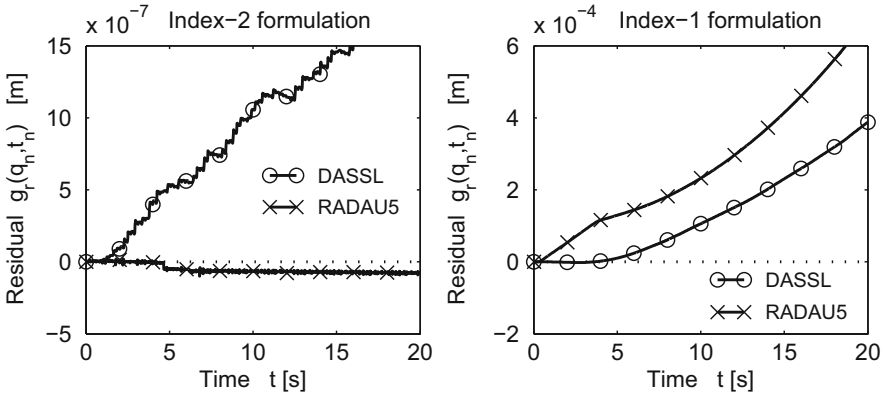


Fig. 14 Drift-off effect in the dynamical simulation of the rigid wheelset of Fig. 11 resulting in an increasing distance $g_r(\boldsymbol{q})$ between the right wheel and the rail, i.e., in an increasing error in the constraints $\boldsymbol{0} = \boldsymbol{g} = (g_l, g_r)^\top$ that are defined by the contact conditions for *left and right wheel* [6, Fig. 10]

has to be expected since the constraints (4.16c) on position level are substituted by their second derivatives (4.22c). Practical experience shows that ϵ depends on the solver and on the user-defined error tolerances TOL. In general, however, there is always a linear drift in the time integration of the index-2 formulation and a quadratic drift for the index-1 formulation.

4.2.4 Projection Techniques and Baumgarte Stabilization

An early attempt to avoid both the numerical problems for the index-3 formulation (4.16) and the drift-off effect in the index-2 and index-1 formulation goes back to the work of Baumgarte [21] who substituted the constraints (4.16c) by a linear combination of all three constraints (4.16c), (4.18c) and (4.22c). Because of the problems to select suitable coefficients for this linear combination (*Baumgarte coefficients*) the practical use of Baumgarte's approach is restricted to small scale models, see also the detailed analysis in [18].

A Baumgarte like method that substitutes the original constraints (4.16c) by a linear combination of (4.16c) and (4.18c) proved to be more favourable:

$$\dot{\boldsymbol{q}} = \boldsymbol{v}\,, \tag{4.26a}$$

$$\mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(\boldsymbol{q},\boldsymbol{v}) - \mathbf{G}^{\top}(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.26b}$$

$$\mathbf{0} = \alpha_0\,\boldsymbol{g}(\boldsymbol{q}) + \mathbf{G}(\boldsymbol{q})\boldsymbol{v}\,. \tag{4.26c}$$

This *index-2 Baumgarte approach* is used successfully for fixed step size computations in real-time applications. Here, the Baumgarte parameter should be set to $\alpha_0 = C/h$ with a suitable constant $C > 0$, see [13, 87].

In off-line simulation, it is state-of-the-art to avoid the drift-off effect by *projection* techniques [35, 62]. During the time integration of index reduced formulations like (4.18) and (4.22), the residual $\|\boldsymbol{g}(\boldsymbol{q}_n)\|$ in the constraints $\boldsymbol{g}(\boldsymbol{q}) = \mathbf{0}$ is monitored. If the residual exceeds at $t = t_n$ some user-defined small error bound $\varepsilon_g > 0$, then $\boldsymbol{q}_n$ is projected onto the manifold $\mathfrak{M} = \{\,\boldsymbol{\eta} \,:\, \boldsymbol{g}(\boldsymbol{\eta}) = \mathbf{0}\,\}$ resulting in projected position coordinates $\hat{\boldsymbol{q}}_n$, see Fig. 15. In a second stage, the velocity coordinates $\boldsymbol{v}_n$ are projected to the tangent space $T_{\boldsymbol{q}}\mathfrak{M}$ at $\boldsymbol{q} = \hat{\boldsymbol{q}}_n$. Finally, the position and velocity coordinates $(\boldsymbol{q}_n, \boldsymbol{v}_n)$ are substituted by their projections $(\hat{\boldsymbol{q}}_n, \hat{\boldsymbol{v}}_n)$ and the time integration is continued with the next time step.

For nonlinear constraints $\boldsymbol{g}(\boldsymbol{q}) = \mathbf{0}$, the projected position coordinates $\hat{\boldsymbol{q}}_n$ have to be computed iteratively. Following the approach of [62], we study the constrained minimization problem

$$\min\,\{\,\frac{1}{2}\|\boldsymbol{\eta} - \boldsymbol{q}_n\|^2_{\mathbf{M}(\boldsymbol{q}_n)} \,:\, \boldsymbol{g}(\boldsymbol{\eta}) = \mathbf{0}\,\} \tag{4.27}$$

with the semi-norm $\|\boldsymbol{\eta}\|_{\mathbf{M}(\boldsymbol{q}_n)} := (\boldsymbol{\eta}^{\top}\mathbf{M}(\boldsymbol{q}_n)\boldsymbol{\eta})^{1/2}$ that considers the mass distribution in the multibody system model. The constraints $\boldsymbol{g}(\boldsymbol{\eta}) = \mathbf{0}$ are coupled to the

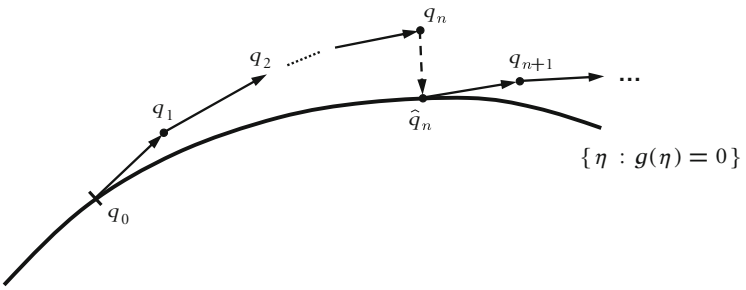


Fig. 15 Time integration with projection steps [6, Fig. 11]

objective function by Lagrange multipliers $\boldsymbol{\mu}$ resulting in

$$\mathscr{L}(\boldsymbol{\eta}, \boldsymbol{\mu}) := \frac{1}{2}(\boldsymbol{\eta} - \boldsymbol{q}_n)^\top \mathbf{M}(\boldsymbol{q}_n)(\boldsymbol{\eta} - \boldsymbol{q}_n) + \boldsymbol{\mu}^\top \boldsymbol{g}(\boldsymbol{\eta})\,.$$

The necessary conditions

$$\begin{aligned} \mathbf{0} &= \nabla_\eta \mathscr{L}(\boldsymbol{\eta}, \boldsymbol{\mu}) = \mathbf{M}(\boldsymbol{q}_n)(\boldsymbol{\eta} - \boldsymbol{q}_n) + \mathbf{G}^\top(\boldsymbol{\eta})\,\boldsymbol{\mu}\,, \\ \mathbf{0} &= \nabla_\mu \mathscr{L}(\boldsymbol{\eta}, \boldsymbol{\mu}) = \boldsymbol{g}(\boldsymbol{\eta}) \end{aligned}$$

for a local minimum of (4.27) motivate a projection step $\boldsymbol{q}_n \mapsto \hat{\boldsymbol{q}}_n$ with $\hat{\boldsymbol{q}}_n$ being defined by the nonlinear system

$$\begin{aligned} \mathbf{0} &= \mathbf{M}(\boldsymbol{q}_n)(\hat{\boldsymbol{q}}_n - \boldsymbol{q}_n) + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}\,, \\ \mathbf{0} &= \boldsymbol{g}(\hat{\boldsymbol{q}}_n) \end{aligned}$$

that may be solved iteratively by a simplified Newton method without re-evaluating the constraint matrix $\mathbf{G}$. The iteration matrix has the characteristic 2×2 block structure (2.18), see [62]:

$$\begin{pmatrix} \mathbf{M}(\boldsymbol{q}_n) & \mathbf{G}^\top(\boldsymbol{q}_n) \\ \mathbf{G}(\boldsymbol{q}_n) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{q}}_n^{(k+1)} - \hat{\boldsymbol{q}}_n^{(k)} \\ \boldsymbol{\mu}^{(k+1)} \end{pmatrix} = - \begin{pmatrix} \mathbf{M}(\boldsymbol{q}_n)(\hat{\boldsymbol{q}}_n^{(k)} - \boldsymbol{q}_n) \\ \boldsymbol{g}(\hat{\boldsymbol{q}}_n^{(k)}) \end{pmatrix}. \tag{4.28}$$

The method is initialized by $\hat{\boldsymbol{q}}_n^{(0)} := \boldsymbol{q}_n$ and needs typically only a few simplified Newton steps (4.28) to get an iterate $\hat{\boldsymbol{q}}_n = \hat{\boldsymbol{q}}_n^{(k)}$ satisfying $\|\boldsymbol{g}(\hat{\boldsymbol{q}}_n)\| \le \varepsilon_{\boldsymbol{g}}$, see [62].

The projection of $\boldsymbol{v}_n$ to the tangent space $T_{\boldsymbol{q}}\mathfrak{M}$ at $\boldsymbol{q} = \hat{\boldsymbol{q}}_n$ does not require the iterative solution of nonlinear equations because the hidden constraints $\mathbf{G}(\hat{\boldsymbol{q}}_n)\boldsymbol{v} = \mathbf{0}$ are linear in the velocity coordinates $\boldsymbol{v}$. The constrained minimization problem

$$\min \{ \frac{1}{2}\|\boldsymbol{\eta} - \boldsymbol{v}_n\|^2_{\mathbf{M}(\hat{\boldsymbol{q}}_n)} : \mathbf{G}(\hat{\boldsymbol{q}}_n)\boldsymbol{\eta} = \mathbf{0} \} \tag{4.29}$$

may be solved directly and defines the projected velocity coordinates $\hat{\boldsymbol{v}}_n$ by

$$\begin{pmatrix} \mathbf{M}(\hat{\boldsymbol{q}}_n) & \mathbf{G}^\top(\hat{\boldsymbol{q}}_n) \\ \mathbf{G}(\hat{\boldsymbol{q}}_n) & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{v}}_n - \boldsymbol{v}_n \\ \bar{\boldsymbol{\mu}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ -\mathbf{G}(\hat{\boldsymbol{q}}_n)\boldsymbol{v}_n \end{pmatrix}. \tag{4.30}$$

4.2.5 Stabilized Index-2 Formulation

In complex applications, the use of classical projection methods like (4.28) and (4.30) is restricted to Runge–Kutta, generalized-α and other one-step methods since the efficient implementation of projection steps in advanced BDF solvers with order and step size control is non-trivial.

Instead of implementing explicit projection steps in the solver, the index reduced formulations of the equations of motion are reformulated in a way that contains implicitly the projection onto the constraint manifold $\mathfrak{M} = \{\boldsymbol{q} : \boldsymbol{g}(\boldsymbol{q}) = \mathbf{0}\}$ and its tangent space $T_q\mathfrak{M}$. For the index-2 formulation (4.18), this approach goes back to the work of Gear et al. [44] who proposed to consider the (hidden) constraints (4.16c) and (4.18c) on position and velocity level simultaneously:

$$\dot{\boldsymbol{q}} = \boldsymbol{v} - \mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\mu}\,, \tag{4.31a}$$

$$\mathbf{M}(\boldsymbol{q})\dot{\boldsymbol{v}} = \boldsymbol{f}(\boldsymbol{q},\boldsymbol{v}) - \mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\lambda}\,, \tag{4.31b}$$

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q})\,, \tag{4.31c}$$

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q})\boldsymbol{v}\,. \tag{4.31d}$$

The increasing number of equations in this *stabilized index-2 formulation* [26] of the equations of motion is compensated by a correction term $-\mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\mu}$ with auxiliary variables $\boldsymbol{\mu}(t) \in \mathbb{R}^{n_\lambda}$. The correction term vanishes identically for the analytical solution since (4.31a,d) and the time derivative of (4.31c) imply

$$\mathbf{0} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{g}(\boldsymbol{q}(t)) = \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{q}}(\boldsymbol{q})\,\dot{\boldsymbol{q}} = \mathbf{G}(\boldsymbol{q})\big(\boldsymbol{v} - \mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\mu}\big) = -[\mathbf{G}\mathbf{G}^\top](\boldsymbol{q})\,\boldsymbol{\mu}$$

and the matrix product $[\mathbf{G}\mathbf{G}^\top](\boldsymbol{q})$ is non-singular for any full rank matrix $\mathbf{G}(\boldsymbol{q})$. Therefore, $\boldsymbol{\mu}(t) \equiv \mathbf{0}$. For the numerical solution, the correction term remains in the size of the user-defined error tolerances TOL.

Equation (4.31) form an index-2 DAE that may be solved robustly and efficiently by BDF [44] and implicit Runge–Kutta methods [50]. However, the error estimates in classical ODE solvers tend to overestimate the local errors in the algebraic components $\boldsymbol{\lambda}$, $\boldsymbol{\mu}$ of DAE (4.31), see [68]. Therefore the components $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ should not be considered in the automatic step size control of BDF solvers [70]. For implicit Runge–Kutta solvers the error estimates for $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ are scaled by the small factor h, see [50].

In implicit methods, the correction term $-\mathbf{G}^\top(\boldsymbol{q})\,\boldsymbol{\mu}$ may be evaluated efficiently taking into account the block structure of the iteration matrix in the corrector iteration [6, Sect. 5]. This implementation scheme goes back to the work of Führer [40] who considered furthermore a stabilized index-1 formulation in BDF time integration [41].

The stabilized index-2 formulation may be defined as well for the more complex model equations (3.12) with additional algebraic equations $\mathbf{0} = \boldsymbol{h}(\boldsymbol{q},\boldsymbol{s})$ but the structure of the correction term needs to be adapted carefully [3, 6]. This generalization of the classical approach of Gear, Gupta and Leimkuhler is closely related to the first index reduction step in the index reduction algorithm according to Kunkel and Mehrmann [58].

Stabilized index reduced formulations have also been investigated in the context of generalized-α and HHT-α methods [15, 55, 56, 63, 91]. Here, we consider the generalized-α method (4.3) for the stabilized index-2 formulation (4.31) of the equations of motion [15]. The constrained equilibrium conditions

$$\mathbf{M}(\boldsymbol{q}_{n+1})\ddot{\boldsymbol{q}}_{n+1} = \boldsymbol{f}(\boldsymbol{q}_{n+1},\dot{\boldsymbol{q}}_{n+1}) - \mathbf{G}^\top(\boldsymbol{q}_{n+1})\,\boldsymbol{\lambda}_{n+1}\,, \tag{4.32a}$$

$$\mathbf{0} = \boldsymbol{g}(\boldsymbol{q}_{n+1})\,, \tag{4.32b}$$

$$\mathbf{0} = \mathbf{G}(\boldsymbol{q}_{n+1})\boldsymbol{v}_{n+1} \tag{4.32c}$$

define a numerical solution $(\boldsymbol{q}_{n+1},\boldsymbol{v}_{n+1})$ that satisfies both the holonomic constraints at the level of position coordinates and the hidden constraints at the level of velocity coordinates. A correction term $-\mathbf{G}^\top(\boldsymbol{q}_n)\;\boldsymbol{\mu}_n$ is added to the update formula (4.3a) of the position coordinates, i.e., we get $\boldsymbol{q}_{n+1} = \boldsymbol{q}_n + h\boldsymbol{\Delta q}_n$ with the scaled increment

$$\boldsymbol{\Delta q}_n := \dot{\boldsymbol{q}}_n - \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n + h(0.5-\beta)\boldsymbol{a}_n + h\beta\boldsymbol{a}_{n+1}\,, \tag{4.33a}$$

see (4.7a). Using again the functions $\boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n)$, ... that were introduced in (4.7b–e), we may express $\boldsymbol{q}_{n+1}, \boldsymbol{a}_{n+1}, \dot{\boldsymbol{q}}_{n+1}$ and $\ddot{\boldsymbol{q}}_{n+1}$ in terms of $\boldsymbol{\Delta q}_n$ and $\boldsymbol{\mu}_n$:

$$\boldsymbol{q}_{n+1} = \boldsymbol{q}_{n+1}(\boldsymbol{\Delta q}_n)\,, \tag{4.33b}$$

$$\boldsymbol{a}_{n+1} = \boldsymbol{a}_{n+1}(\boldsymbol{\Delta q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n)\,, \tag{4.33c}$$

$$\dot{\boldsymbol{q}}_{n+1} = \dot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n)\,, \tag{4.33d}$$

$$\ddot{\boldsymbol{q}}_{n+1} = \ddot{\boldsymbol{q}}_{n+1}(\boldsymbol{\Delta q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n)\,. \tag{4.33e}$$

In each time step, the numerical solution is obtained solving the system of $n_{\boldsymbol{q}} + 2n_{\boldsymbol{\lambda}}$ nonlinear equations

$$\mathbf{0} = \boldsymbol{r}_h^{n+1}(\boldsymbol{\Delta q}_n, h\boldsymbol{\lambda}_{n+1}, \boldsymbol{\mu}_n)\,,$$

$$\mathbf{0} = \boldsymbol{g}_h^{n+1}(\boldsymbol{\Delta q}_n)\,,$$

$$\mathbf{0} = \dot{\boldsymbol{g}}_h^{n+1}(\boldsymbol{\Delta q}_n, \boldsymbol{\mu}_n)$$

with

$$\begin{aligned} \boldsymbol{r}_h^{n+1}(\Delta\boldsymbol{q}_n, h\boldsymbol{\lambda}_{n+1}, \boldsymbol{\mu}_n) &:= \mathbf{M}(\boldsymbol{q}_{n+1}(\Delta\boldsymbol{q}_n))\, h\ddot{\boldsymbol{q}}_{n+1}(\Delta\boldsymbol{q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n) \\ &\quad - h\boldsymbol{f}(\boldsymbol{q}_{n+1}(\Delta\boldsymbol{q}_n), \dot{\boldsymbol{q}}_{n+1}(\Delta\boldsymbol{q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n)) \\ &\quad + \mathbf{G}^\top(\boldsymbol{q}_{n+1}(\Delta\boldsymbol{q}_n))\, h\boldsymbol{\lambda}_{n+1}\,, \\ \boldsymbol{g}_h^{n+1}(\Delta\boldsymbol{q}_n) &:= \frac{1}{h}\boldsymbol{g}(\boldsymbol{q}_{n+1}(\Delta\boldsymbol{q}_n))\,, \\ \dot{\boldsymbol{g}}_h^{n+1}(\Delta\boldsymbol{q}_n) &:= \mathbf{G}(\boldsymbol{q}_{n+1}(\Delta\boldsymbol{q}_n))\,\dot{\boldsymbol{q}}_{n+1}(\Delta\boldsymbol{q}_n + \mathbf{G}^\top(\boldsymbol{q}_n)\,\boldsymbol{\mu}_n)\,. \end{aligned}$$

These equations are scaled again such that the Jacobian

$$\begin{pmatrix} \frac{1-\alpha_m}{\beta(1-\alpha_f)}\mathbf{M}(\boldsymbol{q}_n) + \mathscr{O}(h) & \mathbf{G}^\top(\boldsymbol{q}_n) + \mathscr{O}(h) & \frac{1-\alpha_m}{\beta(1-\alpha_f)}[\mathbf{M}\mathbf{G}^\top](\boldsymbol{q}_n) + \mathscr{O}(h) \\ \mathbf{G}(\boldsymbol{q}_n) + \mathscr{O}(h) & \mathbf{0} & \mathbf{0} \\ \frac{\gamma}{\beta}\mathbf{G}(\boldsymbol{q}_n) + \mathscr{O}(h) & \mathbf{0} & \frac{\gamma}{\beta}[\mathbf{G}\mathbf{G}^\top](\boldsymbol{q}_n) + \mathscr{O}(h) \end{pmatrix}$$

and its inverse remain bounded for $h \to 0$.

The generalized-α method (4.32), (4.33) converges with order $p = 2$ for all solution components if the starting values $\boldsymbol{q}_0$, $\dot{\boldsymbol{q}}_0$ and $\ddot{\boldsymbol{q}}_0$ are second order accurate and $\boldsymbol{a}_0 = \ddot{\boldsymbol{q}}(t_0 + (\alpha_m - \alpha_f)h) + \mathscr{O}(h^2)$, see [15]. In Fig. 16 this second order convergence result is illustrated for the application to the equations of motion of the mathematical pendulum, see Example 4.1. For the stabilized index-2 formulation, the global error $\lambda(t_n) - \lambda_n$ remains in the size of 2.0×10^{-3} for both initial

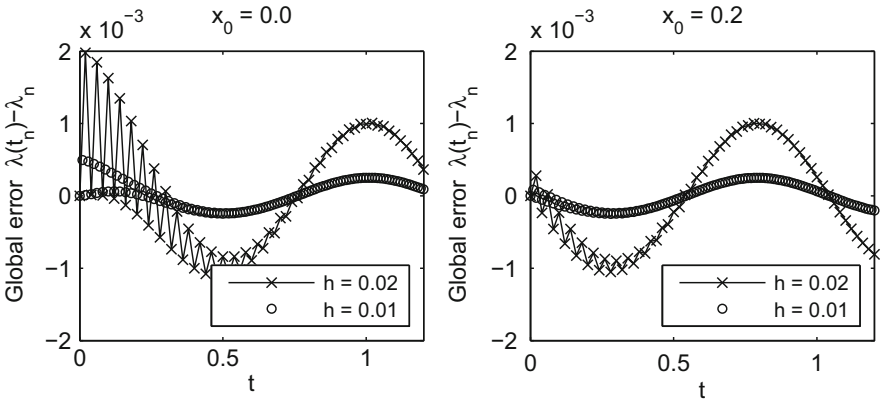


Fig. 16 Global error $\lambda(t_n) - \lambda_n$ of the generalized-α method (4.32), (4.33) for the stabilized index-2 formulation of the equations of motion for the mathematical pendulum with initial values $x_0 = 0$ (*left plot*) and $x_0 = 0.2$ (*right plot*)

configurations. A step size reduction by a factor of two results approximately in a reduction of global errors by the factor of $2^2 = 4$.

The stabilized index-2 formulation of the equations of motion may be extended straightforwardly to model equations with more complex structure including, e.g., nonholonomic constraints or additional differential or algebraic equations. In that sense it is considered to be the most flexible general purpose approach to time integration in multibody dynamics that combines robustness with numerical efficiency. In a practical implementation, the stabilized index-2 formulation is discretized by BDF, by implicit Runge–Kutta methods or by Newmark type methods resulting in nonlinear corrector equations that have to be solved in each time step.

5 Summary

Analysis and numerical solution of constrained mechanical systems have been an important subject of DAE theory for more than 25 years. The well-structured higher index model equations have inspired the development of very efficient index reduction and time integration methods. These DAE solution techniques offer much flexibility to multibody system dynamics w.r.t. model setup and choice of coordinates. Model equations with redundant coordinates resulting from a generic network approach for model setup or from kinematically closed loops in the multibody system model may be solved efficiently by a combination of index reduction techniques with time integration methods from nonlinear system dynamics (BDF, Runge–Kutta methods) or structural dynamics (generalized-α and HHT-α methods).

Half-explicit methods prove to be efficient in the non-stiff case but are typically restricted to the simulation of N-body systems. The model equations in multibody system dynamics may have a substantially more complex structure including nonlinear configuration spaces, additional differential and algebraic equations, rank-deficient mass matrices and redundant constraints. Often, they may be solved more efficiently by implicit methods. The combination of BDF, implicit Runge–Kutta methods or generalized-α methods with the stabilized index-2 formulation of the equations of motion is the method of choice in industrial multibody system simulation.

References

1. Arnold, V.I.: Mathematical Methods of Classical Mechanics. Graduate Texts in Mathematics, vol. 60, 2nd edn. Springer, New York, Berlin, Heidelberg (1989)
2. Arnold, M.: A perturbation analysis for the dynamical simulation of mechanical multibody systems. Appl. Numer. Math. **18**, 37–56 (1995)
3. Arnold, M.: Numerical problems in the dynamical simulation of wheel-rail systems. Z. Angew. Math. Mech. **76**(S3), 151–154 (1996)

4. Arnold, M.: Half-explicit Runge–Kutta methods with explicit stages for differential-algebraic systems of index 2. BIT Numer. Math. **38**, 415–438 (1998)
5. Arnold, M.: Zur Theorie und zur numerischen Lösung von Anfangswertproblemen für differentiell-algebraische Systeme von höherem Index. Fortschritt-Berichte VDI Reihe 20, vol. 264. VDI, Düsseldorf (1998)
6. Arnold, M.: Numerical methods for simulation in applied dynamics. In: Arnold, M., Schiehlen, W. (eds.) Simulation Techniques for Applied Dynamics. CISM Courses and Lectures, vol. 507, pp. 191–246. Springer, Wien, New York (2009)
7. Arnold, M.: A recursive multibody formalism for systems with small mass and inertia terms. Mech. Sci. **4**, 221–231 (2013)
8. Arnold, M.: Algorithmic aspects of singularly perturbed multibody system models. GACM Report. Summer **2013**, 10–16 (2013)
9. Arnold, M., Brüls, O.: Convergence of the generalized-α scheme for constrained mechanical systems. Multibody Sys. Dyn. **18**, 185–202 (2007)
10. Arnold, M., Murua, A.: Non-stiff integrators for differential-algebraic systems of index 2. Numer. Algorithms **19**, 25–41 (1998)
11. Arnold, M., Schiehlen, W. (eds.) Simulation Techniques for Applied Dynamics. CISM Courses and Lectures, vol. 507. Springer, Wien, New York (2009)
12. Arnold, M., Fuchs, A., Führer, C.: Efficient corrector iteration for DAE time integration in multibody dynamics. Comp. Meth. Appl. Mech. Eng. **195**, 6958–6973 (2006)
13. Arnold, M., Burgermeister, B., Eichberger, A.: Linearly implicit time integration methods in real-time applications: DAEs and stiff ODEs. Multibody Sys. Dyn. **17**, 99–117 (2007)
14. Arnold, M., Burgermeister, B., Führer, C., Hippmann, G., Rill, G.: Numerical methods in vehicle system dynamics: state of the art and current developments. Veh. Syst. Dyn. **49**, 1159–1207 (2011)
15. Arnold, M., Brüls, O., Cardona, A.: Error analysis of generalized-α Lie group time integration methods for constrained mechanical systems. Numer. Math. **129**, 149–179 (2015)
16. Arnold, M., Cardona, A., Brüls, O.: Order reduction in time integration caused by velocity projection. In: Proceedings of the 3rd Joint International Conference on Multibody System Dynamics and the 7th Asian Conference on Multibody Dynamics, June 30–July 3, 2014, BEXCO, Busan (2014). In revised version online available as Technical Report 02-2015, Martin Luther University Halle-Wittenberg, Institute of Mathematics (2015)
17. Arnold, M., Cardona, A., Brüls, O.: A Lie algebra approach to Lie group time integration of constrained systems. In: Betsch, P. (ed.) Structure-Preserving Integrators in Nonlinear Structural Dynamics and Flexible Multibody Dynamics, vol. 565. CISM Courses and Lectures, pp. 91–158. Springer International Publishing, Cham (2016)
18. Ascher, U.M., Chin, H., Reich, S.: Stabilization of DAEs and invariant manifolds. Numer. Math. **67**, 131–149 (1994)
19. Bae, D.S., Haug, E.J.: A recursive formulation for constrained mechanical system dynamics: part II. Closed loop systems. Mech. Struct. Mach. **15**, 481–506 (1987)
20. Bauchau, O.A.: Flexible Multibody Dynamics. Springer, Dordrecht, Heidelberg, London, New York (2011)
21. Baumgarte, J.: Stabilization of constraints and integrals of motion in dynamical systems. Comput. Methods Appl. Mech. Eng. **1**, 1–16 (1972)
22. Bottasso, C.L., Borri, M.: Integrating finite rotations. Comput. Methods Appl. Mech. Eng. **164**, 307–331 (1998)
23. Bottasso, C.L., Bauchau, O.A., Cardona, A.: Time-step-size-independent conditioning and sensitivity to perturbations in the numerical solution of index three differential algebraic equations. SIAM J. Sci. Comp. **29**, 397–414 (2007)
24. Brandl, H., Johanni, R., Otter, M.: A very efficient algorithm for the simulation of robots and similar multibody systems without inversion of the mass matrix. In: Kopacek, P., Troch, I., Desoyer, K. (eds.) Theory of Robots, pp. 95–100. Pergamon Press, Oxford (1988)
25. Brasey, V.: A half-explicit method of order 5 for solving constrained mechanical systems. Computing **48**, 191–201 (1992)

26. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial–Value Problems in Differential–Algebraic Equations, 2nd edn. SIAM, Philadelphia (1996)
27. Brüls, O., Cardona, A.: On the use of Lie group time integrators in multibody dynamics. J. Comput. Nonlinear Dyn. **5**, 031002 (2010)
28. Brüls, O., Golinval, J.C.: The generalized-α method in mechatronic applications. J. Appl. Math. Mech./Z. Angew. Math. Mech. **86**, 748–758 (2006)
29. Brüls, O., Arnold, M., Cardona, A.: Two Lie group formulations for dynamic multibody systems with large rotations. In: Proceedings of IDETC/MSNDC 2011, ASME 2011 International Design Engineering Technical Conferences, Washington (2011)
30. Campbell, S.L., Gear, C.W.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)
31. Cardona, A., Géradin, M.: Time integration of the equations of motion in mechanism analysis. Comput. Struct. **33**, 801–820 (1989)
32. Celledoni, E., Owren, B.: Lie group methods for rigid body dynamics and time integration on manifolds. Comput. Methods Appl. Mech. Eng. **192**, 421–438 (2003)
33. Chung, J., Hulbert, G.: A time integration algorithm for structural dynamics with improved numerical dissipation: the generalized-α method. ASME J. Appl. Mech. **60**, 371–375 (1993)
34. Dormand, J.R., Prince, P.J.: A family of embedded Runge–Kutta formulae. J. Comp. Appl. Math. **6**, 19–26 (1980)
35. Eich, E.: Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints. SIAM J. Numer. Anal. **30**, 1467–1482 (1993)
36. Eichberger, A.: Transputer–based multibody system dynamic simulation: part I. The residual algorithm – a modified inverse dynamic formulation. Mech. Struct. Mach. **22**, 211–237 (1994)
37. Eich-Soellner, E., Führer, C.: Numerical Methods in Multibody Dynamics. Teubner, Stuttgart (1998)
38. Featherstone, R.: The calculation of robot dynamics using articulated-body inertias. Int. J. Robot. Res. **2**, 13–30 (1983)
39. Frączek, J., Wojtyra, M.: On the unique solvability of a direct dynamics problem for mechanisms with redundant constraints and Coulomb friction in joints. Mech. Mach. Theory **46**, 312–334 (2011)
40. Führer, C.: Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen. Theorie, numerische Ansätze und Anwendungen. Ph.D. thesis, TU München, Mathematisches Institut und Institut für Informatik (1988)
41. Führer, C., Leimkuhler, B.J.: Numerical solution of differential-algebraic equations for constrained mechanical motion. Numer. Math. **59**, 55–69 (1991)
42. García de Jalón, J., Gutiérrez-López, M.D.: Multibody dynamics with redundant constraints and singular mass matrix: existence, uniqueness, and determination of solutions for accelerations and constraint forces. Multibody Sys. Dyn. **30**, 311–341 (2013)
43. Gear, C.W.: Maintaining solution invariants in the numerical solution of ODEs. SIAM J. Sci. Stat. Comput. **7**, 734–743 (1986)
44. Gear, C.W., Leimkuhler, B., Gupta, G.K.: Automatic integration of Euler–Lagrange equations with constraints. J. Comp. Appl. Math. **12&13**, 77–90 (1985)
45. Géradin, M.D., Cardona, A.: Flexible Multibody Dynamics: A Finite Element Approach. Wiley, Chichester (2001)
46. Golub, G.H., van Loan, Ch.F.: Matrix Computations, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
47. Hairer, E., Lubich, Ch., Roche, M.: The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods. Lecture Notes in Mathematics, vol. 1409. Springer, Berlin, Heidelberg, New York (1989)
48. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations. I. Nonstiff Problems, 2nd edn. Springer, Berlin, Heidelberg, New York (1993)
49. Hairer, E., Lubich, Ch., Wanner, G.: Geometric Numerical Integration. Structure–Preserving Algorithms for Ordinary Differential Equations, 2nd edn. Springer, Berlin, Heidelberg, New York (2006)

50. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations. II. Stiff and Differential-Algebraic Problems, 2nd edn. Springer, Berlin, Heidelberg, New York (1996)
51. Hilber, H.M., Hughes, T.J.R., Taylor, R.L.: Improved numerical dissipation for time integration algorithms in structural dynamics. Earthq. Eng. Struct. Dyn. **5**, 283–292 (1977)
52. Hoschek, M., Rentrop, P., Wagner, Y.: Network approach and differential-algebraic systems in technical applications. Surv. Math. Ind. **9**, 49–75 (1999)
53. Iserles, A., Munthe-Kaas, H.Z., Nørsett, S., Zanna, A.: Lie-group methods. Acta Numer. **9**, 215–365 (2000)
54. Jansen, K.E., Whiting, C.H., Hulbert, G.M.: A generalized-α method for integrating the filtered Navier-Stokes equations with a stabilized finite element method. Comput. Methods Appl. Mech. Eng. **190**, 305–319 (2000)
55. Jay, L.O., Negrut, D.: Extensions of the HHT-method to differential-algebraic equations in mechanics. Electron. Trans. Numer. Anal. **26**, 190–208 (2007)
56. Jay, L.O., Negrut, D.: A second order extension of the generalized-α method for constrained systems in mechanics. In: Bottasso, C. (ed.) Multibody Dynamics. Computational Methods and Applications. Computational Methods in Applied Sciences, vol. 12, pp. 143–158. Springer, Dordrecht (2008)
57. Kalker, J.J.: Three-Dimensional Elastic Bodies in Rolling Contact. Kluwer, Dordrecht, Boston, London (1990)
58. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations. Analysis and Numerical Solution. EMS Textbooks in Mathematics. European Mathematical Society, Zurich (2006)
59. Lötstedt, P.: Mechanical systems of rigid bodies subject to unilateral constraints. SIAM J. Appl. Math. **42**, 281–296 (1982)
60. Lötstedt, P., Petzold, L.R.: Numerical solution of nonlinear differential equations with algebraic constraints I: convergence results for backward differentiation formulas. Math. Comp. **46**, 491–516 (1986)
61. Lubich, Ch., Nowak, U., Pöhle, U., Engstler, Ch.: MEXX – Numerical software for the integration of constrained mechanical multibody systems. Technical Report SC 92–12, ZIB Berlin (1992)
62. Lubich, Ch., Engstler, Ch., Nowak, U., Pöhle, U.: Numerical integration of constrained mechanical systems using MEXX. Mech. Struct. Mach. **23**, 473–495 (1995)
63. Lunk, C., Simeon, B.: Solving constrained mechanical systems by the family of Newmark and α-methods. Z. Angew. Math. Mech. **86**, 772–784 (2006)
64. Möller, M., Glocker, C.: Rigid body dynamics with a scalable body, quaternions and perfect constraints. Multibody Sys. Dyn. **27**, 437–454 (2012)
65. Müller, A., Terze, Z.: The significance of the configuration space Lie group for the constraint satisfaction in numerical time integration of multibody systems. Mech. Mach. Theory **82**, 173–202 (2014)
66. Negrut, D., Rampalli, R., Ottarsson, G., Sajdak, A.: On the use of the HHT method in the context of index 3 differential algebraic equations of multi-body dynamics. In: Goicolea, J.M., Cuadrado, J., García Orden, J.C. (eds.) Proceedings of Multibody Dynamics 2005 (ECCOMAS Thematic Conference), Madrid (2005)
67. Orlandea, N.: Development and application of node–analogous sparsity–oriented methods for simulation of mechanical dynamic systems. Ph.D. thesis, University of Michigan (1973)
68. Petzold, L.R.: Differential/algebraic equations are not ODEs. SIAM J. Sci. Stat. Comput. **3**, 367–384 (1982)
69. Petzold, L.R.: Order results for implicit Runge–Kutta methods applied to differential/algebraic systems. SIAM J. Numer. Anal. **23**, 837–852 (1986)
70. Petzold, L.R., Lötstedt, P.: Numerical solution of nonlinear differential equations with algebraic constraints II: practical implications. SIAM J. Sci. Stat. Comput. **7**, 720–733 (1986)
71. Rheinboldt, W.C.: Differential-algebraic systems as differential equations on manifolds. Math. Comp. **43**, 473–482 (1984)
72. Rheinboldt, W.C.: On the existence and uniqueness of solutions of nonlinear semi-implicit differential-algebraic equations. Nonlinear Anal. Theory Methods Appl. **16**, 647–661 (1991)

73. Rulka, W.: Effiziente Simulation der Dynamik mechatronischer Systeme für industrielle Anwendungen. PhD Thesis, Vienna University of Technology, Department of Mechanical Engineering (1998)
74. Schiehlen, W.O. (ed.): Multibody Systems Handbook. Springer, Berlin, Heidelberg, New York (1990)
75. Schiehlen, W.: Multibody system dynamics: roots and perspectives. Multibody Sys. Dyn. **1**, 149–188 (1997)
76. Schiehlen, W., Eberhard, P.: Applied Dynamics. Springer, Dordrecht (2014)
77. Schwertassek, R., Wallrapp, O.: Dynamik Flexibler MehrKÖrperSysteme. Vieweg, Wiesbaden (1999)
78. Shabana, A.A.: Flexible multibody dynamics: Review of past and recent developments. Multibody Sys. Dyn. **1**, 189–222 (1997)
79. Shabana, A.A.: Dynamics of Multibody Systems, 2nd edn. Cambridge University Press, Cambridge (1998)
80. Simeon, B.: MBSPACK – Numerical integration software for constrained mechanical motion. Surv. Math. Ind. **5**, 169–202 (1995)
81. Simeon, B.: On the numerical solution of a wheel suspension benchmark problem. Comp. Appl. Math. **66**, 443–456 (1996)
82. Simeon, B.: Computational Flexible Multibody Dynamics: A Differential-Algebraic Approach. Differential-Algebraic Equations Forum. Springer, Berlin Heidelberg (2013)
83. Simeon, B.: On the history of differential-algebraic equations. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations, vol. III. Springer, Cham (2016)
84. Simeon, B., Führer, C., Rentrop, P.: Differential-algebraic equations in vehicle system dynamics. Surv. Math. Ind. **1**, 1–37 (1991)
85. Simo, J.C., Vu-Quoc, L.: On the dynamics in space of rods undergoing large motions – A geometrically exact approach. Comput. Methods Appl. Mech. Eng. **66**, 125–161 (1988)
86. Udwadia, F.E., Phohomsiri, P.: Explicit equations of motion for constrained mechanical systems with singular mass matrices and applications to multi-body dynamics. Proc. R. Soc. A **462**, 2097–2117 (2006)
87. Valášek, M., Šika, Z., Vaculín, O.: Multibody formalism for real-time application using natural coordinates and modified state space. Multibody Sys. Dyn. **17**, 209–227 (2007)
88. von Schwerin, R.: MultiBody System SIMulation – Numerical Methods, Algorithms, and Software. Lecture Notes in Computational Science and Engineering, vol. 7. Springer, Berlin, Heidelberg (1999)
89. Wehage, R.A., Haug, E.J.: Generalized coordinate partitioning for dimension reduction in analysis of constrained dynamic systems. J. Mech. Design **104**, 247–255 (1982)
90. Wehage, R.A., Shabana, A.A., Hwang, Y.L.: Projection methods in flexible multibody dynamics. part II: Dynamics and recursive projection methods. Int. J. Numer. Methods Eng. **35**, 1941–1966 (1992)
91. Yen, J., Petzold, L.R., Raha, S.: A time integration algorithm for flexible mechanism dynamics: The DAE α-method. Comput. Methods Appl. Mech. Eng. **158**, 341–355 (1998)

Model Order Reduction for Differential-Algebraic Equations: A Survey

Peter Benner and Tatjana Stykel

Abstract In this paper, we discuss the model order reduction problem for descriptor systems, that is, systems with dynamics described by differential-algebraic equations. We focus on linear descriptor systems as a broad variety of methods for these exist, while model order reduction for nonlinear descriptor systems has not received sufficient attention up to now. Model order reduction for linear state-space systems has been a topic of research for about 50 years at the time of writing, and by now can be considered as a mature field. The extension to linear descriptor systems usually requires extra treatment of the constraints imposed by the algebraic part of the system. For almost all methods, this causes some technical difficulties, and these have only been thoroughly addressed in the last decade. We will focus on these developments in particular for the popular methods related to balanced truncation and rational interpolation. We will review efforts in extending these approaches to descriptor systems, and also add the extension of the so-called *stochastic balanced truncation* method to descriptor systems which so far cannot be found in the literature.

Keywords Balanced truncation • Differential-algebraic equations • Interpolation-based approximation • Matrix equations • Matrix pencils • Model order reduction

Mathematics Subject Classification (2010) 15A22, 15A24, 34A09, 65D05, 65F30, 93C05

P. Benner (✉)
Computational Methods in Systems and Control Theory, Max Planck Institute for Dynamics of Complex Technical Systems, Sandtorstraße 1, 39106 Magdeburg, Germany
e-mail: benner@mpi-magdeburg.mpg.de

T. Stykel
Institut für Mathematik, Universität Augsburg, Universitätsstraße 14, 86159 Augsburg, Germany
e-mail: stykel@math.uni-augsburg.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7_3

1 Introduction

Consider a linear time-invariant descriptor system

$$E\dot{x}(t) = Ax(t) + Bu(t), \tag{1.1a}$$

$$y(t) = Cx(t) + Du(t), \tag{1.1b}$$

where $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, $C \in \mathbb{R}^{q\times n}$, $D \in \mathbb{R}^{q\times m}$, $x(t) \in \mathbb{R}^n$ is the generalized state space vector, $u(t) \in \mathbb{R}^m$ is the input, and $y(t) \in \mathbb{R}^q$ is the output. Here, (1.1a) represents a system of linear differential-algebraic equations (DAEs), while (1.1b) is an output equation, modeling observations or measurements of the system. Sometimes, the elements of $y(t)$ are also referred as the quantities-of-interest if the system is used in a (design) optimization context, and the output quantities in $y(t)$ are the subject of optimization.

Modeling by DAEs has become a ubiquitous tool in many engineering disciplines, in particular in structural dynamics and multi-body systems as well as in micro- and nanoelectronics, computational electromagnetics, and fluid mechanics, see, e.g., [28, 86, 93, 118], the DAE examples in [27, Part II], and the benchmarks provided at the Model Order Reduction Wiki [97]. In mechanics, algebraic constraints arise from holonomic or nonholonomic constraints, in circuit simulation and other network problems, among others, from Kirchhoff's laws, and in electromagnetics or fluid mechanics by the discretization of conservation laws like the preservation of mass in the incompressible Navier–Stokes equations. In these applications, the shear number of equations like in the modeling of semiconductor devices or the fine-grain spatial discretization of partial differential equations like the already mentioned Navier–Stokes or Maxwell's equations in electromagnetics, leads to descriptor systems with n in the thousands to millions or even larger than this. A single forward simulation of such a system is certainly feasible on modern computer architectures, but simulating a couple of hundreds of times in the context of design optimization, varying input signals, and control design, is often out of scope. In these situations, replacing the descriptor system (1.1) by a system with the same structure, but of much smaller size $r \ll n$ by approximating the input–output relation to a desired accuracy, is beneficial.

A *model order reduction* problem consists in approximating (1.1) by a reduced-order model

$$\begin{aligned} \tilde{E}\,\dot{\tilde{x}}(t) &= \tilde{A}\,\tilde{x}(t) + \tilde{B}\,u(t), \\ \tilde{y}(t) &= \tilde{C}\,\tilde{x}(t) + \tilde{D}\,u(t), \end{aligned} \tag{1.2}$$

where $\tilde{E}, \tilde{A} \in \mathbb{R}^{r\times r}$, $\tilde{B} \in \mathbb{R}^{r\times m}$, $\tilde{C} \in \mathbb{R}^{q\times r}$, $\tilde{D} \in \mathbb{R}^{q\times m}$ and $r \ll n$. Assume that the matrix pencil $\lambda E - A$ is *regular*, i.e., $\det(\lambda E - A) \neq 0$ for some $\lambda \in \mathbb{C}$. Applying the Laplace transform to system (1.1), it can be written in the frequency domain as

$$\hat{y}(s) = \boldsymbol{H}(s)\hat{u}(s) + C(sE - A)^{-1}Ex(0),$$

where $\hat{u}(s)$ and $\hat{y}(s)$ are the Laplace transforms of the input and output, respectively, and $\boldsymbol{H}(s) = C(sE - A)^{-1}B + D$ is a *transfer function* of (1.1). Then the model reduction problem can be formulated in the frequency domain as follows: given the transfer function $\boldsymbol{H}(s)$, find $\tilde{\boldsymbol{H}}(s) = \tilde{C}(s\tilde{E} - \tilde{A})^{-1}\tilde{B} + \tilde{D}$ of lower dimension that approximates $\boldsymbol{H}(s)$. The approximation quality can, for instance, be measured by the absolute error $\tilde{\boldsymbol{H}} - \boldsymbol{H}$ or by the relative error $\boldsymbol{H}^{-1}(\tilde{\boldsymbol{H}} - \boldsymbol{H})$, provided $\boldsymbol{H}^{-1}$ exists.

The structure of this survey is as follows: in the following section, we provide the relevant systems and control theoretic basics for linear descriptor systems. In Sect. 3, we review the most common methods for model order reduction of linear descriptor systems: balanced truncation and related methods in Sect. 3.1, and moment matching as well as other rational interpolation methods in Sect. 3.2. The computational bottleneck of many model reduction methods, in particular those related to balanced truncation, is the numerical solution of matrix equations (e.g., algebraic Lyapunov and Riccati equations). Therefore, we review the usually used methods and their adaptation to the DAE case in Sect. 4. Usually, descriptor systems have a certain block structure, often related to the differential and algebraic parts of the system. Exploiting these structures is mandatory for efficient methods for model order reduction and the associated matrix equations. This is discussed in Sect. 5, using some relevant example classes. In Sect. 6, we provide a brief outlook on topics not covered in depth in this survey and/or of current research interest.

Throughout the paper, $\mathbb{R}^{n\times m}$ and $\mathbb{C}^{n\times m}$ denote the spaces of $n \times m$ real and complex matrices, respectively. Furthermore, $\mathbb{C}_- = \{ s \in \mathbb{C} \ : \mathrm{Re}(s) < 0 \}$ and $\mathbb{C}_+ = \{ s \in \mathbb{C} \ : \mathrm{Re}(s) > 0 \}$ denote the open left and right half-planes, respectively, and $i = \sqrt{-1}$. The matrices A^T and A^* denote, respectively, the transpose and the conjugate transpose of $A \in \mathbb{C}^{n\times m}$, and $A^{-T} = (A^{-1})^T$. We use $\mathrm{rank}(A)$, $\mathrm{im}(A)$ and $\ker(A)$ for the rank, the image and the kernel of A, respectively. A matrix $A \in \mathbb{C}^{n\times n}$ is said to be positive semidefinite, if $v^*Av \geq 0$ for all $v \in \mathbb{C}^n$. Note that positive semidefiniteness of A does not require A to be Hermitian. For $A, B \in \mathbb{C}^{n\times n}$, we write $A \geq B$ if $A - B$ is positive semidefinite.

2 DAE Control Systems

In this section, we provide necessary notation and fundamental matrix and control theoretic concepts for DAE systems.

Any regular matrix pencil $\lambda E - A$ can be transformed into the *Weierstrass canonical form*

$$E = T_l \begin{bmatrix} I_{n_f} & 0 \\ 0 & E_\infty \end{bmatrix} T_r, \qquad A = T_l \begin{bmatrix} A_f & 0 \\ 0 & I_{n_\infty} \end{bmatrix} T_r, \tag{2.1}$$

where T_l and T_r are the left and right nonsingular transformation matrices, E_∞ is nilpotent with index of nilpotency ν, and $n_f + n_\infty = n$, e.g., [59]. The number ν is

called the *index* of $\lambda E - A$ and also of the DAE system (1.1). The eigenvalues of A_f are the finite eigenvalues of $\lambda E - A$, and $\lambda E_\infty - I$ has only eigenvalues at infinity. Thus, if E is singular, then $\lambda E - A$ has n_f finite and n_∞ infinite eigenvalues which together form a set of generalized eigenvalues.

The pencil $\lambda E - A$ is called *stable* if all its finite eigenvalues belong to the open left half-plane $\mathbb{C}_-$. In this case, the solution of system (1.1) with $u(t) \equiv 0$ tends to zero as $t \to \infty$, and, hence, the DAE system (1.1) is *asymptotically stable* .

We introduce now the *spectral projectors* onto the left and right deflating subspaces of the pencil $\lambda E - A$ corresponding to the finite eigenvalues along the left and right deflating subspaces corresponding to the eigenvalue at infinity as

$$P_l = T_l \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} T_l^{-1}, \qquad P_r = T_r^{-1} \begin{bmatrix} I_{n_f} & 0 \\ 0 & 0 \end{bmatrix} T_r. \tag{2.2}$$

Furthermore, the matrices

$$Q_l = I - P_l = T_l \begin{bmatrix} 0 & 0 \\ 0 & I_{n_\infty} \end{bmatrix} T_l^{-1}, \qquad Q_r = I - P_r = T_r^{-1} \begin{bmatrix} 0 & 0 \\ 0 & I_{n_\infty} \end{bmatrix} T_r \tag{2.3}$$

define the complementary projectors. All these projectors play an important role in model reduction of DAE systems.

Using the Weierstrass canonical form (2.1) and introducing

$$T_r x(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}, \qquad T_l^{-1} B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \qquad C T_r^{-1} = [\, C_1,\ C_2 \,], \tag{2.4}$$

we can decouple the DAE system (1.1) into a *slow subsystem*

$$\begin{aligned} \dot{x}_1(t) &= A_f x_1(t) + B_1 u(t), \\ y_1(t) &= C_1 x_1(t), \end{aligned} \tag{2.5}$$

and a *fast subsystem*

$$\begin{aligned} E_\infty \dot{x}_2(t) &= x_2(t) + B_2 u(t), \\ y_2(t) &= C_2 x_2(t) + D u(t). \end{aligned} \tag{2.6}$$

The output of (1.1) is then determined as $y(t) = y_1(t) + y_2(t)$.

Next, we introduce some algebraic properties of matrix triplets related to the DAE system (1.1). The equivalent definitions in terms of controllability and observability concepts relating to the dynamic behavior of the DAE system can be found in [36, 41] and the survey [37] contained in this volume. We restrict here to the definition of the algebraic properties as these are used in the rest of this paper.

Definition 2.1 Let the matrices Z_l and Z_r be of full rank such that $\operatorname{im}(Z_l) = \operatorname{im}(E^T)$ and $\operatorname{im}(Z_r) = \operatorname{im}(E)$. Then the matrix triplet $(E, A, B) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m}$ is called

1) *controllable in the behavioral sense* (*R-controllable*), if $\operatorname{rank}[\,\lambda E - A,\ B\,] = n$ for all $\lambda \in \mathbb{C}$;
2) *stabilizable in the behavioral sense* (*R-stabilizable*), if $\operatorname{rank}[\,\lambda E - A,\ B\,] = n$ for all $\lambda \in \mathbb{C} \setminus \mathbb{C}_-$;
3) *impulse controllable* (*I-controllable*), if $\operatorname{rank}[\,E,\ AZ_r,\ B\,] = n$;
4) *controllable at infinity* (*Inf-controllable*), if $\operatorname{rank}[\,E,\ B\,] = n$;
5) *strongly controllable* (*S-controllable*), if it is R-controllable and I-controllable;
6) *strongly stabilizable* (*S-stabilizable*), if it is R-stabilizable and I-controllable;
7) *completely controllable* (*C-controllable*), if it is R-controllable and Inf-controllable.

The matrix triplet $(E, A, C) \in \mathbb{R}^{n\times n} \times \mathbb{R}^{n\times n} \times \mathbb{R}^{q\times n}$ is called

8) *observable in the behavioral sense* (*R-observable*), if $\operatorname{rank}[\,\lambda E^T - A^T,\ C^T\,] = n$ for all $\lambda \in \mathbb{C}$;
9) *detectable in the behavioral sense* (*R-detectable*), if $\operatorname{rank}[\,\lambda E^T - A^T,\ C^T\,] = n$ for all $\lambda \in \mathbb{C} \setminus \mathbb{C}_-$;
10) *impulse observable* (*I-observable*), if $\operatorname{rank}[\,E^T,\ A^T Z_l,\ C^T\,] = n$;
11) *observable at infinity* (*Inf-observable*), if $\operatorname{rank}[\,E^T,\ C^T\,] = n$;
12) *strongly observable* (*S-observable*), if it is R-observable and I-observable;
13) *strongly detectable* (*S-detectable*), if it is R-detectable and I-observable;
14) *completely observable* (*C-observable*), if it is R-observable and Inf-observable.

In the following, we will not distinguish the algebraic and system-theoretic properties of the matrix triplets (E, A, B), (E, A, C) and the corresponding DAE system (1.1) and speak equivalently, e.g., of R-controllability of (E, A, B) and the DAE system (1.1).

In the frequency domain, the input–output behavior of the DAE system (1.1) is described by a transfer function $\boldsymbol{H}(s) = C(sE - A)^{-1}B + D$ which is a rational matrix-valued function. On the other side, for any rational matrix-valued function $\boldsymbol{H}(s)$, one can always find the matrices E, A, B, C and D such that

$$\boldsymbol{H}(s) = C(sE - A)^{-1}B + D,$$

e.g., [41]. Such a quintuple $\boldsymbol{H} = (E, A, B, C, D)$ is called a *realization* of $\boldsymbol{H}(s)$. If (E, A, B, C, D) is a realization of $\boldsymbol{H}(s)$, then for any nonsingular matrices W and T, (WET, WAT, WB, CT, D) is also a realization of $\boldsymbol{H}(s)$. This implies that $\boldsymbol{H}(s)$ has many different realizations. Moreover, there exist realizations of arbitrarily high order which is defined by the dimension of the matrices E and A. A realization $\boldsymbol{H} = (E, A, B, C, D)$ is called *minimal* if E and A have the smallest possible dimension. One can show that $\boldsymbol{H} = (E, A, B, C, D)$ is minimal if and only if system (1.1) is C-controllable, C-observable and $A\ker(E) \subseteq \operatorname{im}(E)$, see [147]. The latter condition means that the nilpotent matrix E_∞ in the Weierstrass canonical form (2.1) does not have any 1×1 Jordan blocks.

The transfer function $\boldsymbol{H}(s)$ is called *proper* if $H_\infty = \lim_{s\to\infty} \boldsymbol{H}(s)$ exists, and *improper*, otherwise. If $H_\infty = 0$, then $\boldsymbol{H}(s)$ is called *strictly proper*. Using (2.1) and (2.4), the transfer function $\boldsymbol{H}(s)$ can additively be decomposed as

$$\boldsymbol{H}(s) = \boldsymbol{H}_{sp}(s) + \boldsymbol{P}(s),$$

where

$$\boldsymbol{H}_{sp}(s) = C_1(sI - A_f)^{-1}B_1$$

is the *strictly proper part* of $\boldsymbol{H}(s)$, and

$$\boldsymbol{P}(s) = C_2(sE_\infty - I)^{-1}B_2 + D = \sum_{j=0}^{\nu-1} M_j s^j$$

with

$$M_j = -C_2 E_\infty^j B_2 + \delta_{0,j} D \tag{2.7}$$

is the *polynomial part* of $\boldsymbol{H}(s)$. Here, $\delta_{0,j}$ denotes the Kronecker delta. Note that $\boldsymbol{H}_{sp}(s)$ and $\boldsymbol{P}(s)$ are the transfer functions of the slow and fast subsystems (2.5) and (2.6), respectively. If the realization $\boldsymbol{H} = (E, A, B, C, D)$ is not minimal, then the degree of the polynomial $\boldsymbol{P}(s)$, denoted by $\deg(\boldsymbol{P})$, may be smaller than $\nu - 1$.

The transfer function $\boldsymbol{H}(s)$ can also be written as

$$\boldsymbol{H}(s) = \frac{\boldsymbol{N}(s)}{\boldsymbol{d}(s)},$$

where $\boldsymbol{N}(s)$ is a $q \times m$ matrix polynomial and $\boldsymbol{d}(s)$ is a scalar polynomial which is the least common denominator of the qm entries of $\boldsymbol{H}(s)$. The roots of the denominator $\boldsymbol{d}(s)$ are called the *finite poles* of $\boldsymbol{H}(s)$, and the roots of the numerator $\boldsymbol{N}(s)$ are called the *finite zeros* of $\boldsymbol{H}(s)$. The transfer function $\boldsymbol{H}(s)$ has a *pole* (*zero*) *at infinity* if $s = 0$ is a pole (zero) of $\boldsymbol{H}(1/s)$. If $\deg(\boldsymbol{N}) > \deg(\boldsymbol{d})$ or, equivalently, if $\boldsymbol{H}(s)$ is improper, then $\boldsymbol{H}(s)$ has a pole at infinity. If $\deg(\boldsymbol{N}) < \deg(\boldsymbol{d})$ or, equivalently, if $\boldsymbol{H}(s)$ is strictly proper, then $\boldsymbol{H}(s)$ has a zero at infinity. The poles of $\boldsymbol{H}(s)$ are generalized eigenvalues of the pencil $\lambda E - A$. The set of poles of $\boldsymbol{H}(s)$ coincides with the set of generalized eigenvalues of $\lambda E - A$ if and only if $\boldsymbol{H} = (E, A, B, C, D)$ is minimal. For the square transfer function $\boldsymbol{H}(s)$, the zeros of $\boldsymbol{H}(s)$ are generalized eigenvalues of the *system pencil*

$$\lambda \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A & B \\ C & D \end{bmatrix},$$

see [121]. If this pencil is regular, then $\boldsymbol{H}(s)$ is invertible and its inverse is given by

$$\boldsymbol{H}^{-1}(s) = [0,\ -I] \begin{bmatrix} sE - A & -B \\ -C & -D \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

This representation immediately follows from the relation

$$\begin{bmatrix} sE-A & 0 \\ 0 & \boldsymbol{H}(s) \end{bmatrix} = \begin{bmatrix} I & 0 \\ C(sE-A)^{-1} & I \end{bmatrix} \begin{bmatrix} sE-A & -B \\ -C & -D \end{bmatrix} \begin{bmatrix} I & -(sE-A)^{-1}B \\ 0 & -I \end{bmatrix}.$$

Note that if D is nonsingular, then $\boldsymbol{H}^{-1}(s)$ can also be realized as

$$\boldsymbol{H}^{-1}(s) = -D^{-1}C(sE-A+BD^{-1}C)BD^{-1}+D^{-1}.$$

An invertible transfer function $\boldsymbol{H}(s)$ is called (strictly) *minimum phase* if all its finite zeros have (negative) non-positive real part.

Let $s_1,\ldots,s_k$ be the pairwise different finite poles of $\boldsymbol{H}(s)$ of order ℓ_j, $j=1,\ldots,k$, then $\boldsymbol{H}(s)$ can be represented using a partial fraction expansion as

$$\boldsymbol{H}(s) = \sum_{j=1}^{k}\sum_{i=1}^{\ell_j} \frac{R_j^{(i)}}{(s-s_j)^i} + \sum_{j=0}^{\nu-1} M_j s^j, \tag{2.8}$$

where $R_j \equiv R_j^{(1)}$ is the *residue* of $\boldsymbol{H}$ at s_j.

Another useful representation of the transfer function $\boldsymbol{H}(s)$ is given by its power series expansion at $s_0 \in \mathbb{C}$ being not a pole of $\boldsymbol{H}$:

$$\boldsymbol{H}(s) = \sum_{j=0}^{\infty} M_j(s_0)(s-s_0)^j, \tag{2.9}$$

where the coefficients $M_j(s_0)$, also called *(shifted) moments*,[1] have the form

$$\begin{aligned} M_0(s_0) &= -C(A-s_0E)^{-1}B+D, \\ M_j(s_0) &= -C\big((A-s_0E)^{-1}E\big)^j(A-s_0E)^{-1}B, \quad j>0. \end{aligned}$$

For singular E, the Laurent expansion of $\boldsymbol{H}$ turns out to be beneficial as well:

$$\boldsymbol{H}(s) = \sum_{j=-\infty}^{\nu-1} M_j s^j, \tag{2.10}$$

where the coefficients M_j are the *Markov parameters* given by

$$\begin{aligned} M_j &= CT_r^{-1}\begin{bmatrix} A_f^{-j-1} & 0 \\ 0 & 0 \end{bmatrix} T_l^{-1}B = C_1 A_f^{-j-1}B_1, && j<0, \\ M_j &= CT_r^{-1}\begin{bmatrix} 0 & 0 \\ 0 & -E_\infty^j \end{bmatrix} T_l^{-1}B + \delta_{0,j}D = -C_2 E_\infty^j B_2 + \delta_{0,j}D, && j\geq 0. \end{aligned}$$

[1] Usually, the term *moments* is used to denote the coefficients of the Taylor series at $s_0 = 0$.

Thus, the Markov parameters corresponding to the nonnegative powers are the same as the coefficients M_j in the partial fraction expansion (2.8) and in (2.7), and, therefore, they determine the polynomial part of $\boldsymbol{H}(s)$.

In order to measure the approximation error of reduced-order models, we will employ classical system norms. Let $\mathcal{H}_\infty$ denote the space of matrix-valued functions that are analytic and bounded in the open right half-plane. The $\mathcal{H}_\infty$-norm of $\boldsymbol{H} \in \mathcal{H}_\infty$ is defined as

$$\|\boldsymbol{H}\|_{\mathcal{H}_\infty} = \sup_{\omega\in\mathbb{R}} \|\boldsymbol{H}(i\omega)\|_2,$$

where $\|\cdot\|_2$ denotes the spectral matrix norm. Furthermore, we consider the space $\mathcal{H}_2$ of matrix-valued functions that are analytic in the open right half-plane. The $\mathcal{H}_2$-norm of $\boldsymbol{H} \in \mathcal{H}_2$ is defined as

$$\|\boldsymbol{H}\|_{\mathcal{H}_2} = \left(\frac{1}{2\pi}\int_{-\infty}^{\infty} \|\boldsymbol{H}(i\omega)\|_F^2 d\omega\right)^{1/2},$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. We note that the rational matrix-valued functions given by the transfer functions corresponding to (1.1) are in $\mathcal{H}_\infty$ if the system is stable and proper, and in $\mathcal{H}_2$ if, in addition, it is strictly proper.

3 Model Order Reduction Techniques

Before describing different model reduction techniques, we would like to point out that most techniques are based on the *(Petrov–)Galerkin projection*. The basic idea can simply be described as follows, where we use (1.1) as a model problem. Assuming the dynamics of the system evolves in a low-dimensional subspace $\mathcal{T} \subset \mathbb{R}^n$ with basis matrix $T \in \mathbb{R}^{n\times r}$, we use the ansatz $x(t) \approx T\tilde{x}(t)$. Hence, $\mathcal{T}$ is considered as a *trial space*. Replacing $x(t)$ in the generalized state equation (the first equation in (1.1)), we obtain a residual

$$\tilde{r}(t) := ET\dot{\tilde{x}}(t) - AT\tilde{x}(t) - Bu(t).$$

In general, the residual is not zero. Therefore, we demand it to at least vanish on an r-dimensional *test space* $\mathcal{W} \subset \mathbb{R}^n$ with basis matrix $W \in \mathbb{R}^{n\times r}$, so that T and W are bi-orthogonal, i.e., $W^TT = I_r$. The requirement $W^T\tilde{r}(t) \equiv 0$ then leads to the reduced (generalized) state equation

$$W^TET\dot{\tilde{x}}(t) = W^TAT\tilde{x}(t) + W^TBu(t).$$

Applying the projection onto $\mathcal{T}$ also to the second equation in (1.1) leads to the reduced-order system

$$\left(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}\right) := \left(W^TET, W^TAT, W^TB, CT, D\right).$$

This process is called a *Petrov–Galerkin projection*, and TW^T defines an oblique projector onto $\mathcal{T}$. If one takes $\mathcal{W} = \mathcal{T}$, necessitating to choose an orthogonal basis matrix T, i.e., $T^T T = I_r$, we speak of a *Galerkin projection* and TT^T defines an orthogonal projector onto $\mathcal{T}$.

Note that the method "balanced truncation" described in Sect. 3.1 is, in general, a Petrov–Galerkin projection (turning into a Galerkin projection for symmetric systems with $E = E^T$, $A = A^T$ and $C = B^T$), while the interpolatory approaches in Sect. 3.2 can either be Galerkin or Petrov–Galerkin projection methods.

3.1 Balanced Truncation

Balanced truncation was initially introduced in the systems and control theory in the early 1980s [48, 63, 96] and has been continuously developed ever since. Due to new developments in Numerical Linear Algebra, it is now applicable to large-scale problems [13–15], and has already been used in many application areas including biochemical engineering [91], electrical circuit simulation [114, 116], mechanical systems [31, 113], computational fluid dynamics [21, 35, 71, 135] and power systems [53, 120].

A main idea of balanced truncation and its relatives is to transform a dynamical system to a balanced form defined in such a way that appropriately chosen controllability and observability Gramians are equal and diagonal. Then a reduced-order model is computed by truncating the states corresponding to the small diagonal elements of the Gramians. Depending on the choice of the Gramians, different balanced truncation techniques can be developed, see [15, 67] for surveys of balancing-related model reduction methods for standard state-space systems. In this section, we summarize the extensions of these methods to DAE systems.

3.1.1 Lyapunov Balanced Truncation

The most commonly used balanced truncation method is based on balancing the *controllability* and *observability Gramians* G_c and G_o which are defined for system (1.1) with $E = I$ as unique symmetric, positive semidefinite solutions of the continuous-time Lyapunov equations

$$AG_c + G_c A^T = -BB^T, \qquad A^T G_o + G_o A = -C^T C,$$

provided all eigenvalues of the matrix A have negative real part. These Gramians characterize the controllability and observability properties of the control system and quantify the input and output energy [96]. The square roots of the eigenvalues of the product $G_c G_o$ define the *Hankel singular values*, $\sigma_j = \sqrt{\lambda_j(G_c G_o)}$, which can be used to measure the importance of the state variables. We assume that σ_j are

ordered decreasingly. Finding a balancing transformation T_b such that

$$T_b G_c T_b^T = T_b^{-T} G_o T_b^{-1} = \operatorname{diag}(\sigma_1, \ldots, \sigma_n)$$

and truncating $n - r$ components of the transformed state vector $T_b x(t)$, which correspond to small $\sigma_j < \sigma_r$, yields an asymptotically stable reduced-order model [104]. Another important property of this method is the presence of the computable error estimates

$$\begin{aligned} &\|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty} \leq 2(\sigma_{r+1} + \ldots + \sigma_n), \\ &\| \tilde{y} - y \|_{\mathcal{L}_2} \leq \|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty} \|u\|_{\mathcal{L}_2} \leq 2(\sigma_{r+1} + \ldots + \sigma_n)\|u\|_{\mathcal{L}_2}, \end{aligned}$$

see [48, 63].

The Lyapunov-based balanced truncation approach was extended to DAEs in [16, 92, 103, 134]. A basic idea behind this extension is to decouple the DAE system (1.1) into the slow and fast subsystems (2.5) and (2.6), respectively, and reduce them separately. In the frequency domain, this corresponds to the separate approximation of the strictly proper part $\boldsymbol{H}_{sp}(s)$ and the polynomial part $\boldsymbol{P}(s)$ of the transfer function $\boldsymbol{H}(s) = \boldsymbol{H}_{sp}(s) + \boldsymbol{P}(s)$ resulting in an approximate system $\tilde{\boldsymbol{H}}(s) = \tilde{\boldsymbol{H}}_{sp}(s) + \tilde{\boldsymbol{P}}(s)$. It should, however, be noticed that if $\tilde{\boldsymbol{P}}(s) \neq \boldsymbol{P}(s)$ and $\deg(\boldsymbol{P}(s)) \geq 1$, then the error $\boldsymbol{H}(s) - \tilde{\boldsymbol{H}}(s)$ is unbounded. Also in the time domain, a naive reduction of the order of the fast subsystem (2.6) which, actually, describes the constraints in the model, may lead to an inaccurate approximation, see [92, 138]. These difficulties have been resolved in [134] by determining a minimal realization of $\boldsymbol{P}(s)$. This guarantees that $\boldsymbol{P}(s) = \tilde{\boldsymbol{P}}(s)$ and, hence, the error $\boldsymbol{H}(s) - \tilde{\boldsymbol{H}}(s)$ will be small if the error in the slow subsystem $\boldsymbol{H}_{sp}(s) - \tilde{\boldsymbol{H}}_{sp}(s)$ is small.

In practice, we do not need to compute the slow and fast subsystems explicitly. This is computationally expensive, especially for large-scale problems, and may be numerically ill-conditioned. Instead, we can define two pairs of controllability and observability Gramians in terms of the original data using the spectral projectors P_l, P_r and Q_l, Q_r given in (2.2) and (2.3), respectively. Assume that the DAE system (1.1) is asymptotically stable. Then the *proper controllability* and *observability Gramians* G_{pc} and G_{po} of (1.1) are defined as unique symmetric, positive semidefinite solutions of the projected continuous-time Lyapunov equations

$$E\,G_{pc}\,A^T + A\,G_{pc}\,E^T = -P_l B B^T P_l^T, \qquad G_{pc} = P_r G_{pc}\,P_r^T, \tag{3.1}$$

$$E^T G_{po}\,A + A^T G_{po}\,E = -P_r^T C^T C P_r, \qquad G_{po} = P_l^T G_{po}\,P_l, \tag{3.2}$$

respectively, whereas the *improper controllability* and *observability Gramians* G_{ic} and G_{io} of (1.1) are defined as unique symmetric, positive semidefinite solutions of

the projected discrete-time Lyapunov equations

$$A\,G_{ic}A^T - E\,G_{ic}E^T = Q_l B B^T Q_l^T, \qquad G_{ic} = Q_r G_{ic} Q_r^T, \tag{3.3}$$

$$A^T G_{io} A - E^T G_{io} E = Q_r^T C^T C Q_r, \qquad G_{io} = Q_l^T G_{io} Q_l, \tag{3.4}$$

respectively. The square roots of the largest n_f eigenvalues of $G_{pc}E^T G_{po}E$, denoted by σ_j, are called the *proper Hankel singular values* of (1.1), and the square roots of the largest n_∞ eigenvalues of $G_{ic}A^T G_{io}A$, denoted by θ_j, are called the *improper Hankel singular values*. System (1.1) is *balanced* if the Gramians satisfy

$$G_{pc} + G_{ic} = G_{po} + G_{io} = \mathrm{diag}(\sigma_1, \ldots \sigma_{n_f}, \theta_1, \ldots, \theta_{n_\infty}).$$

Thus, a reduced-order model (1.2) can be determined by truncating the states of the balanced system corresponding to the small proper Hankel singular values. In [134], it is shown that the states corresponding to the small eigenvalues of the proper controllability Gramian G_{pc} need the most energy to be reached. Also, the states corresponding to the small eigenvalues of the proper observability Gramian G_{po} contribute the least to the output energy

$$\mathbf{E}(y) = \int_0^\infty y(t)^T y(t)\, dt.$$

In balanced coordinates, the eigenvalues of G_{pc}, G_{po}, and the proper Hankel singular values coincide. Thus, the difficult-to-reach states coincide with those least involved in the output energy. Based on this energy interpretation of the proper Gramians, one can assert that these states are difficult to control and difficult to observe at the same time and can therefore be ignored in the system approximation. Furthermore, we can remove states which are not Inf-controllable and Inf-observable. Such states correspond to zero improper Hankel singular values.

Considering the Cholesky factorizations[2] of the Gramians

$$G_{pc} = Z_{pc}Z_{pc}^T, \qquad G_{po} = Z_{po}Z_{po}^T, \qquad G_{ic} = Z_{ic}Z_{ic}^T, \qquad G_{io} = Z_{io}Z_{io}^T,$$

and taking into account that the proper and improper Hankel singular values can be determined from the singular value decomposition of the matrices $Z_{po}^T E Z_{pc}$ and $Z_{io}^T A Z_{ic}$, respectively, we obtain the generalization of the square-root balanced truncation method [88, 143] for DAE systems shown in Algorithm 1. As in the

[2]It should be noted that by abuse of notation, these factors are neither necessarily upper triangular nor square, but we assume them to be of full rank. In particular, for non-minimal systems, these factors will in general be rectangular as then the Gramians will be rank deficient.

Algorithm 1 Lyapunov balanced truncation for DAE systems.

Input: an asymptotically stable system $\boldsymbol{H} = (E, A, B, C, D)$.

Output: a reduced-order asymptotically stable system $\tilde{\boldsymbol{H}} = (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

1: Compute the full rank Cholesky factors Z_{pc} and Z_{po} of the proper Gramians $G_{pc} = Z_{pc}Z_{pc}^T$ and $G_{po} = Z_{po}Z_{po}^T$ satisfying the projected Lyapunov equations (3.1) and (3.2), respectively.

2: Compute the full rank Cholesky factors Z_{ic} and Z_{io} of the improper Gramians $G_{ic} = Z_{ic}Z_{ic}^T$ and $G_{io} = Z_{io}Z_{io}^T$ satisfying the projected Lyapunov equations (3.3) and (3.4), respectively.

3: Compute the singular value decomposition $Z_{po}^T E Z_{pc} = [\,U_1,\, U_2\,]\mathrm{diag}(\Sigma_1, \Sigma_2)[\,V_1,\, V_2\,]^T$, where the matrices $[\,U_1, U_2\,]$ and $[\,V_1, V_2\,]$ have orthonormal columns, $\Sigma_1 = \mathrm{diag}(\sigma_1, \ldots, \sigma_{r_f})$ and $\Sigma_2 = \mathrm{diag}(\sigma_{r_f+1}, \ldots, \sigma_{n_f})$.

4: Compute the singular value decomposition $Z_{io}^T A Z_{ic} = U_3 \Theta V_3^T$, where U_3 and V_3 have orthonormal columns and Θ is nonsingular.

5: Compute the reduced-order system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (W^T E T, W^T A T, W^T B, C T, D)$ with $W = [\,Z_{po}U_1\Sigma_1^{-1/2},\, Z_{io}U_3\Theta^{-1/2}\,]$ and $T = [\,Z_{pc}V_1\Sigma_1^{-1/2},\, Z_{ic}V_3\Theta^{-1/2}\,]$.

standard state space case [48, 63], we have the error estimates

$$\|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty} \le 2(\sigma_{r_f+1} + \ldots + \sigma_{n_f}),$$
$$\|\,\tilde{y} - y\,\|_{\mathcal{L}_2} \le \|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty}\|u\|_{\mathcal{L}_2} \le 2(\sigma_{r_f+1} + \ldots + \sigma_{n_f})\|u\|_{\mathcal{L}_2}.$$

Moreover, for $\boldsymbol{P}(s) \neq D$, one can show that the index of the reduced-order model is equal to $\deg(\boldsymbol{P}) + 1$ and does not exceed the index of the original system (1.1). If $\boldsymbol{P}(s) = D$, then the reduced-order model is an ODE system.

Using the Weierstrass canonical form (2.1), one notices that the improper Gramians G_{ic} and G_{io} have usually low rank which can be estimated as

$$r_c = \mathrm{rank}(G_{ic}) \le \min(\nu m, n_\infty), \qquad r_o = \mathrm{rank}(G_{io}) \le \min(\nu q, n_\infty),$$

where ν is the index of (1.1). Furthermore, if the eigenvalues of the proper Gramians G_{pc} and G_{po} decay fast, then G_{pc} and G_{po} have low numerical rank. In this case, they can be well approximated by low-rank matrices $G_{pc} \approx \tilde{Z}_{pc}\tilde{Z}_{pc}^T$ and $G_{po} \approx \tilde{Z}_{po}\tilde{Z}_{po}^T$, where $\tilde{Z}_{pc} \in \mathbb{R}^{n\times n_c}$ and $\tilde{Z}_{po} \in \mathbb{R}^{n\times n_o}$ with $n_c, n_o \ll n$. Replacing the full rank factors Z_{pc} and Z_{po} in Algorithm 1 by the low-rank matrices $\tilde{Z}_{pc}$ and $\tilde{Z}_{po}$, respectively, reduces significantly the computational complexity and storage requirements for the balanced truncation method, making it applicable to large-scale problems. In fact, apart from solving the projected Lyapunov equations, only the singular value decomposition of the small matrices $\tilde{Z}_{po}^T E \tilde{Z}_{pc} \in \mathbb{R}^{n_o\times n_c}$ and $Z_{io}^T A Z_{ic} \in \mathbb{R}^{r_o\times r_c}$ needs to be computed. The computation of the (low-rank) Cholesky factors of the Gramians will be discussed in Sect. 4.1.

Remark 3.1 Unfortunately, in the literature [106, 142], one can often find the statement that the extension of balanced truncation from standard state-space systems to DAEs is as simple as replacing the identity matrix by E. In this case,

the Lyapunov equations take the form

$$AG_cE^T + EG_cA^T = -BB^T, \qquad A^TG_oE + E^TG_oA = -C^TC.$$

It should, however, be noted that for singular E, these equations may not be solvable even if the pencil $\lambda E - A$ is stable. Moreover, if the solutions exist, they are always non-unique. Hence, their use does not lead to a well-defined model reduction method.

In the Poor Man's truncated balanced reduction (PMTBR) method presented in [106], it was proposed to define the Gramians of system (1.1) as

$$X = \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega E - A)^{-1}BB^T(-i\omega E - A)^{-T}d\omega,$$
$$Y = \frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega E - A)^{-T}C^TC(i\omega E - A)^{-1}d\omega.$$

However, if E is singular, these integrals do not converge unless $B = P_lB$ and $C = CP_r$. Therefore, the correct definition should be

$$X = \frac{1}{2\pi}\int_{-\infty}^{\infty}(i\omega E - A)^{-1}P_lBB^TP_l^T(-i\omega E - A)^{-T}d\omega, \tag{3.5}$$
$$Y = \frac{1}{2\pi}\int_{-\infty}^{\infty}(-i\omega E - A)^{-T}P_r^TC^TCP_r(i\omega E - A)^{-1}d\omega.$$

It is worth noting that these matrices solve the projected Lyapunov Equations (3.1) and (3.2), respectively, which again justifies the above considerations.

3.1.2 Positive Real Balanced Truncation

Positive real balanced truncation was first developed for standard state-space systems in [70, 98] as a model reduction method preserving passivity. It was then extended to DAEs in [115].

The DAE system (1.1) is called *passive* if $m = q$ and

$$\int_0^t u(\tau)^Ty(\tau)\,d\tau \geq 0$$

for all $t > 0$ and all $u \in \mathcal{L}_2([0,t],\mathbb{R}^m)$ consistent with $x(0) = 0$. Physically, this property means that the system does not generate energy. It is of great importance especially for circuit equations. One can show that system (1.1) is passive if and only if its transfer function $\boldsymbol{H}(s)$ is *positive real*, i.e., $\boldsymbol{H}(s)$ is analytic in the open right half-plane $\mathbb{C}_+$ and $\boldsymbol{H}(s) + \boldsymbol{H}^*(s) \geq 0$ for all $s \in \mathbb{C}_+$, see [7]. Passivity of the DAE system (1.1) can also be characterized via the projected positive real Lur'e

equations

$$\begin{aligned} AXE^T + EXA^T &= -K_c K_c^T,\ X = P_r X P_r^T \geq 0, \\ EXC^T - P_l B &= -K_c J_c^T,\ M_0 + M_0^T = J_c J_c^T, \end{aligned} \tag{3.6}$$

and

$$\begin{aligned} A^T YE + E^T YA &= -K_o^T K_o,\ Y = P_l^T Y P_l \geq 0, \\ E^T YB - P_r^T C^T &= -K_o^T J_o,\ M_0 + M_0^T = J_o^T J_o, \end{aligned} \tag{3.7}$$

with M_0 as in (2.7) and unknowns $K_c, K_o^T \in \mathbb{R}^{n\times m}$, $J_c, J_o \in \mathbb{R}^{m\times m}$ and $X, Y \in \mathbb{R}^{n\times n}$. If system (1.1) is R-controllable, R-observable and passive, then the projected Lur'e Equations (3.6) are solvable. Conversely, the solvability of (3.6) together with the conditions $M_1 = M_1^T \geq 0$ and $M_j = 0$ for $j > 1$ implies that (1.1) is passive. A similar result holds also for the dual Lur'e Equations (3.7). Note that for some structured systems as they arise, for example, in modified nodal analysis (MNA) of electrical circuits, the existence of the solutions of the projected Lur'e equations can also be proved without R-controllability and R-observability conditions [114]. It should be emphasized that the solutions of (3.6) and (3.7) are not unique. There exist, however, unique extremal solutions satisfying

$$X_{\max} \geq X \geq X_{\min} \geq 0, \qquad Y_{\max} \geq Y \geq Y_{\min} \geq 0$$

for all symmetric solutions X and Y of (3.6) and (3.7), respectively. The minimal solutions $G_c^{PR} = X_{\min}$ and $G_o^{PR} = Y_{\min}$ are called, respectively, the *positive real controllability* and *observability Gramians* of system (1.1). Replacing the proper Gramians in the Lyapunov-based balanced truncation method by the positive real Gramians, we obtain the passivity-preserving model reduction method for DAE systems. In order to determine the positive real Gramians from the Lur'e Eqs. (3.6) and (3.7), we need first to calculate M_0. This matrix can be obtained from the polynomial part $\boldsymbol{P}(s)$ whose realization is given by $\boldsymbol{P} = (W_\infty^T E T_\infty, W_\infty^T A T_\infty, W_\infty^T B, C T_\infty, D)$ with $W_\infty = Z_{io} U_3 \Theta^{-1/2}$ and $T_\infty = Z_{ic} V_3 \Theta^{-1/2}$. Since $W_\infty^T A T_\infty = I$, we have

$$M_0 = D - C T_\infty W_\infty^T B = D - C Z_{ic} V_3 \Theta^{-1} U_3^T Z_{io}^T B.$$

The resulting positive real balanced truncation method is presented in Algorithm 2.

The values $\sigma_1^{PR} \geq \ldots \geq \sigma_{r_f}^{PR} > \sigma_{r_f+1}^{PR} \geq \ldots \geq \sigma_{n_f}^{PR}$ are called the *positive real characteristic values* of (1.1). Similar to the proper Hankel singular values, they can be used to estimate the approximation error. If $M_0 + M_0^T$ is nonsingular, we have the error bound

$$\|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty} \leq 2\|(M_0 + M_0^T)^{-1}\|_2 \|\boldsymbol{H} + M_0^T\|_{\mathcal{H}_\infty} \|\tilde{\boldsymbol{H}} + M_0^T\|_{\mathcal{H}_\infty} \sum_{j=r_f+1}^{n_f} \sigma_j^{PR}$$

that can be derived for DAE systems similarly to the standard state-space case [67].

Algorithm 2 Positive real balanced truncation for DAE systems.

Input: a passive system $\boldsymbol{H} = (E, A, B, C, D)$.

Output: a reduced-order passive system $\tilde{\boldsymbol{H}} = (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

1: Compute the full rank Cholesky factors Z_{ic} and Z_{io} of the improper Gramians $G_{ic} = Z_{ic}Z_{ic}^T$ and $G_{io} = Z_{io}Z_{io}^T$ satisfying the projected Lyapunov equations (3.3) and (3.4), respectively.
2: Compute the singular value decomposition $Z_{io}^T A Z_{ic} = U_3 \Theta V_3^T$ with nonsingular Θ.
3: Compute the matrix $M_0 = D - CZ_{ic}V_3\Theta^{-1}U_3^T Z_{io}^T B$.
4: Compute the Cholesky factors Z_c^{PR} and Z_o^{PR} of the positive real Gramians $G_c^{PR} = Z_c^{PR}(Z_c^{PR})^T$ and $G_o^{PR} = Z_o^{PR}(Z_o^{PR})^T$ that are the minimal solutions of the positive real projected Lur'e equations (3.6) and (3.7), respectively.
5: Compute $(Z_o^{PR})^T E Z_c^{PR} = [U_1, U_2]\mathrm{diag}(\Sigma_1^{PR}, \Sigma_2^{PR})[V_1, V_2]^T$ by singular value decomposition, where $\Sigma_1^{PR} = \mathrm{diag}(\sigma_1^{PR}, \ldots, \sigma_{r_f}^{PR})$ and $\Sigma_2^{PR} = \mathrm{diag}(\sigma_{r_f+1}^{PR}, \ldots, \sigma_{n_f}^{PR})$.
6: Compute the reduced-order system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (W^T E T, W^T A T, W^T B, CT, D)$ with $W = [Z_o^{PR} U_1 (\Sigma_1^{PR})^{-1/2}, Z_{io} U_3 \Theta^{-1/2}]$ and $T = [Z_c^{PR} V_1 (\Sigma_1^{PR})^{-1/2}, Z_{ic} V_3 \Theta^{-1/2}]$.

The positive real balanced truncation method requires solving the projected Lur'e equations. The numerical solution of standard Lur'e equations based on deflating subspaces of a certain even pencil has been considered in [107, 108]. However, so far no numerical method has been developed for projected Lur'e equations. In the case where $R_0 = M_0 + M_0^T$ is nonsingular, the projected Lur'e Equations (3.6) and (3.7) can be written as the projected positive real Riccati equations

$$AXE^T + EXA^T + (EXC^T - P_l B)R_0^{-1}(EXC^T - P_l B)^T = 0, \qquad X = P_r X P_r^T$$

and

$$A^T YE + E^T YA + (B^T YE - CP_r)^T R_0^{-1}(B^T YE - CP_r) = 0, \qquad Y = P_l^T Y P_l,$$

respectively. Such equations can be solved using Newton's method [23] briefly described in Sect. 4.2.

An alternative approach for passivity-preserving model reduction has been proposed in [145]. It relies on a combination of Lyapunov balancing and positive real balancing and involves solving only one Lyapunov equation and one Lur'e equation. However, there exists no error bound for this approach.

3.1.3 Bounded Real Balanced Truncation

If, instead of passivity, we aim to preserve contractivity, an important property in $\mathcal{L}_2$-gain constraint controller design, then bounded real balanced truncation [98, 100, 115] has to be used. The DAE system (1.1) is called *contractive* if

$$\int_0^t \|u(\tau)\|^2 - \|y(\tau)\|^2 \, d\tau \geq 0$$

for all $t > 0$ and all $u \in \mathcal{L}_2([0,t],\mathbb{R}^m)$ consistent with $x(0) = 0$. This condition implies that the $\mathcal{L}_2$-norm of the output is bounded by the $\mathcal{L}_2$-norm of the input. In the frequency domain, contractivity is equivalent to *bounded realness* of the transfer function $\boldsymbol{H}(s)$, meaning that $\boldsymbol{H}(s)$ is analytic in $\mathbb{C}_+$ and $I - \boldsymbol{H}(s)^*\boldsymbol{H}(s) \geq 0$, for all $s \in \mathbb{C}_+$. The latter condition yields that the bounded real transfer function $\boldsymbol{H}(s)$ is necessarily proper.

To verify contractivity, we use the projected bounded real Lur'e equations

$$\begin{aligned} AXE^T + EXA^T + P_l BB^T P_l^T &= -K_c K_c^T,\ X = P_r X P_r^T \geq 0, \\ EXC^T + P_l BM_0^T &= -K_c J_c^T,\ \ I - M_0 M_0^T = J_c J_c^T, \end{aligned} \tag{3.8}$$

and

$$\begin{aligned} A^T YE + E^T YA + P_r^T C^T CP_r &= -K_o^T K_o,\ Y = P_l^T Y P_l \geq 0, \\ E^T YB + P_r^T C^T M_0 &= -K_o^T J_o,\ \ I - M_0^T M_0 = J_o^T J_o. \end{aligned} \tag{3.9}$$

Similarly to the positive real case, one can show that these equations have the minimal solutions $G_c^{BR} = X_{\min}$ and $G_o^{BR} = Y_{\min}$ that are called the *bounded real controllability* and *observability Gramians*, respectively. They can be used to characterize the required supply energy and the available storage energy for contractive systems [115]. This immediately leads to the bounded real balanced truncation method presented in Algorithm 3.

One can show that the reduced-order system computed by Algorithm 3 is contractive and has the error bound

$$\|\tilde{\boldsymbol{H}} - \boldsymbol{H}\|_{\mathcal{H}_\infty} \leq 2 \sum_{j=r_f+1}^{n_f} \sigma_j^{BR}$$

with the *bounded real characteristic values* σ_j^{BR}.

Algorithm 3 Bounded real balanced truncation for DAE systems.

Input: a contractive system $\boldsymbol{H} = (E, A, B, C, D)$.

Output: a reduced-order contractive system $\tilde{\boldsymbol{H}} = (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

1: Compute the full rank Cholesky factors Z_{ic} and Z_{io} of the improper Gramians $G_{ic} = Z_{ic}Z_{ic}^T$ and $G_{io} = Z_{io}Z_{io}^T$ satisfying the projected Lyapunov equations (3.3) and (3.4), respectively.

2: Compute the singular value decomposition $Z_{io}^T E Z_{ic} = U_3 \Theta V_3^T$ with nonsingular Θ.

3: Compute the matrix $M_0 = D - CZ_{ic}V_3\Theta^{-1}U_3^T Z_{io}^T B$.

4: Compute the Cholesky factors Z_c^{BR} and Z_o^{BR} of the bounded real Gramians $G_c^{BR} = Z_c^{BR}(Z_c^{BR})^T$ and $G_o^{BR} = Z_o^{BR}(Z_o^{BR})^T$ that are the minimal solutions of the bounded real projected Lur'e equations (3.8) and (3.9), respectively.

5: Compute $(Z_o^{BR})^T E Z_c^{BR} = [\,U_1,\ U_2\,]\mathrm{diag}(\Sigma_1^{BR}, \Sigma_2^{BR})[\,V_1,\ V_2\,]^T$ by singular value decomposition, where $\Sigma_1^{BR} = \mathrm{diag}(\sigma_1^{BR}, \ldots, \sigma_{r_f}^{BR})$ and $\Sigma_2^{BR} = \mathrm{diag}(\sigma_{r_f+1}^{BR}, \ldots, \sigma_{n_f}^{BR})$.

6: Compute the reduced-order system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (W^T E T,\ W^T A T,\ W^T B,\ CT,\ M_0)$ with the projection matrices $W = Z_o^{BR} U_1 (\Sigma_1^{BR})^{-1/2}$ and $T = Z_c^{BR} V_1 (\Sigma_1^{BR})^{-1/2}$.

If $R_c = I - M_0 M_0^T$ is nonsingular, then $R_o = I - M_0^T M_0$ is also nonsingular and the projected Lur'e Equations (3.8) and (3.9) are equivalent to the projected bounded real Riccati equations

$$AXE^T + EXA^T + P_l BB^T P_l^T + (EXC^T + P_l BM_0^T)R_c^{-1}(EXC^T + P_l BM_0^T)^T = 0,$$
$$X - P_r X P_r^T = 0,$$

and

$$A^T YE + E^T YA + P_r^T C^T CP_r + (B^T YE + M_0^T CP_r)^T R_o^{-1}(B^T YE + M_0^T CP_r) = 0,$$
$$Y - P_l^T Y P_l = 0,$$

respectively. These equations can be solved using Newton's method described in Sect. 4.

Note that the bounded real systems are related to the positive real systems via a *Moebius transformation* defined as

$$\boldsymbol{H}_M(s) = \big(I - \boldsymbol{H}(s)\big)\big(I + \boldsymbol{H}(s)\big)^{-1}.$$

The transfer function $\boldsymbol{H}(s)$ is positive real if and only if the Moebius-transformed function $\boldsymbol{H}_M(s)$ is bounded real. For $\boldsymbol{H} = (E, A, B, C, D)$, a realization of $\boldsymbol{H}_M(s)$ is given by

$$\boldsymbol{H}_M = \big(E, A - B(I + D)^{-1}C, -\sqrt{2}B(I + D)^{-1}, \sqrt{2}(I + D)^{-1}C, (I - D)(I + D)^{-1}\big),$$

provided $I + D$ is invertible. This suggests another passivity-preserving balancing-related model reduction approach which consists of applying the bounded real balanced truncation method to $\boldsymbol{H}_M$ and computing the Moebius transformation $\tilde{\boldsymbol{H}}(s) = \big(I - \tilde{\boldsymbol{H}}_M(s)\big)\big(I + \tilde{\boldsymbol{H}}_M(s)\big)^{-1}$ of the obtained reduced-order model $\tilde{\boldsymbol{H}}_M$. This approach might be useful if the spectral projectors for the Moebius-transformed system are easier to compute than that for the original systems. Circuit equations belong, for example, to this class of problems [114].

3.1.4 Stochastic Balanced Truncation

Stochastic balanced truncation belongs to relative error model reduction methods attempting to minimize the relative error $\boldsymbol{H}^{-1}(\boldsymbol{H} - \tilde{\boldsymbol{H}})$ in an appropriate norm. It was first introduced for discrete-time and continuous-time standard state space systems in [45, 70] and studied further in [26, 65, 66, 146]. The stochastic balanced truncation method relies on an approximation of spectral factors of the *power spectrum* $\boldsymbol{\Phi}(s) = \boldsymbol{H}(s)\boldsymbol{H}^T(-s)$ and is known to preserve the right half-plane zeros of $\boldsymbol{H}(s)$. In this section, we present an extension of this method to DAEs.

Assume that system (1.1) is asymptotically stable and has a square proper and invertible transfer function $\boldsymbol{H}(s)$. Using the spectral projectors P_l and P_r, $\boldsymbol{H}(s)$ can then be written as $\boldsymbol{H}(s) = CP_r(sE-A)^{-1}P_lB + M_0$. Then the power spectrum can be written as

$$\begin{aligned}\boldsymbol{\Phi}(s) &= \boldsymbol{H}(s)\boldsymbol{H}^T(-s)\\ &= \begin{bmatrix} CP_r, & M_0B^TP_l^T \end{bmatrix} \begin{bmatrix} sE-A & -P_lBB^TP_l^T \\ 0 & -sE^T-A^T \end{bmatrix}^{-1} \begin{bmatrix} P_lBM_0^T \\ P_r^TC^T \end{bmatrix} + M_0M_0^T.\end{aligned}$$

Taking into account that the proper controllability Gramian G_{pc} solves the Lyapunov Equation (3.1), we obtain

$$\begin{bmatrix} sE-A & -P_lBB^TP_l^T \\ 0 & -sE^T-A^T \end{bmatrix} = \begin{bmatrix} I & -EG_{pc} \\ 0 & I \end{bmatrix} \begin{bmatrix} sE-A & 0 \\ 0 & -sE^T-A^T \end{bmatrix} \begin{bmatrix} I & -G_{pc}E^T \\ 0 & I \end{bmatrix}.$$

Therefore, introducing $B_0 = P_lBM_0^T + EG_{pc}C^T = P_lB_0$, we have

$$\begin{aligned}\boldsymbol{\Phi}(s) &= \begin{bmatrix} CP_r, & B_0^T \end{bmatrix} \begin{bmatrix} sE-A & 0 \\ 0 & -sE^T-A^T \end{bmatrix}^{-1} \begin{bmatrix} B_0 \\ P_r^TC^T \end{bmatrix} + M_0M_0^T\\ &= CP_r(sE-A)^{-1}B_0 + B_0^T(-sE-A)^{-T}P_r^TC^T + M_0M_0^T\\ &= \boldsymbol{Z}(s) + \boldsymbol{Z}^T(-s)\end{aligned}$$

with $\boldsymbol{Z}(s) = CP_r(sE-A)^{-1}B_0 + M_0M_0^T/2$. Since $\lambda E - A$ is stable and

$$\boldsymbol{Z}(i\omega) + \boldsymbol{Z}^*(i\omega) = \boldsymbol{H}(i\omega)\boldsymbol{H}^*(i\omega) \geq 0$$

for all $\omega \in \mathbb{R}$, it follows from [7, Theorem 2.7.2] that $\boldsymbol{Z}(s)$ is positive real. If $\boldsymbol{Z}$ is R-controllable and R-observable, then using the results from Sect. 3.1.2 we obtain that the corresponding positive real Lur'e equations

$$\begin{aligned} AXE^T + EXA^T &= -K_cK_c^T, & X &= P_rXP_r^T \geq 0,\\ EXC^T - B_0 &= -K_cJ_c^T, & M_0M_0^T &= J_cJ_c^T, \end{aligned} \tag{3.10}$$

and

$$\begin{aligned} A^TYE + E^TYA &= -K_o^TK_o, & Y &= P_l^TYP_l \geq 0,\\ E^TYB_0 - P_r^TC^T &= -K_o^TJ_o, & M_0M_0^T &= J_o^TJ_o \end{aligned} \tag{3.11}$$

are solvable. They have two extremal solutions satisfying

$$X_{\max} \geq X \geq X_{\min} \geq 0, \qquad Y_{\max} \geq Y \geq Y_{\min} \geq 0$$

for all symmetric solutions X and Y of (3.10) and (3.11), respectively. Moreover, one can also show that $X_{\max} = (E^TY_{\min}E)_r^-$, where $(M)_r^-$ denotes a *reflexive inverse*

of M with respect to P_r^T and P_r which is defined as the unique solution of the matrix equations

$$(M)_r^- M (M)_r^- = (M)_r^-, \qquad M(M)_r^- = P_r^T, \qquad (M)_r^- M = P_r.$$

Consider now $\boldsymbol{W}(s)$ being a square right spectral factor of the power spectrum $\boldsymbol{\Phi}(s) = \boldsymbol{H}(s)\boldsymbol{H}^T(-s) = \boldsymbol{W}^T(-s)\boldsymbol{W}(s)$. Its realization can be determined using the matrix Equations (3.1) and (3.11). We have

$$\begin{aligned}
\boldsymbol{\Phi}(s) &= \boldsymbol{Z}(s) + \boldsymbol{Z}^T(-s) \\
&= (CP_r - B_0^T Y E)(sE - A)^{-1} B_0 + B_0^T (-sE - A)^{-T} (CP_r - B_0^T Y E)^T + M_0 M_0^T \\
&\quad + B_0^T Y E (sE - A)^{-1} B_0 + B_0^T (-sE - A)^{-T} E^T Y B_0 \\
&= J_o^T K_o (sE - A)^{-1} B_0 + B_0^T (-sE - A)^{-T} K_o^T J_o + J_o^T J_o \\
&\quad + B_0^T (-sE - A)^{-T} K_o^T K_o (sE - A)^{-1} B_0 \\
&= \left(K_o(-sE - A)^{-1} B_0 + J_o\right)^T \left(K_o(sE - A)^{-1} B_0 + J_o\right),
\end{aligned}$$

and, hence, $\boldsymbol{W}(s) = K_o(sE - A)^{-1} B_0 + J_o$. Similarly to the standard state space case [110], we can show that for the minimal solution $Y_{\min}$ of (3.11), all finite eigenvalues of the pencil

$$\lambda \begin{bmatrix} E & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} A & B_0 \\ K_o & J_o \end{bmatrix}$$

have non-positive real part. Therefore, $\boldsymbol{W}(s)$ has no zeros in the open right half-plane meaning that $\boldsymbol{W}(s)$ is minimum phase. The matrices $G_c^S = G_{pc}$ and $G_o^S = Y_{\min}$ define the *stochastic controllability* and *observability Gramians* of system (1.1). A reduced-order model can then be computed by balancing these Gramians and truncating the states corresponding to small *stochastic characteristic values* σ_j^S defined as $\sigma_j^S = \sqrt{\lambda_j(G_c^S E^T G_o^S E)}$. The stochastic balanced truncation method is summarized in Algorithm 4.

Since $X = G_{pc}$ solves (3.10), we have $(E^T Y_{\min} E)_r^- = X_{\max} \geq G_{pc}$, and, hence, the eigenvalues of $G_{pc} E^T Y_{\min} E = G_c^S E^T G_o^S E$ do not exceed one. This implies that the stochastic characteristic values of (1.1) satisfy $0 \leq \sigma_j^S \leq 1$. Moreover, it follows from [65, Theorem 4.1] that $\boldsymbol{H}(s)$ has $k_z = \dim\left(\ker((E^T Y_{\min} E)_r^- - G_{pc})\right) - n_\infty$ infinite zeros and finite zeros in the closed right half-plane, and $\sigma_1^S = \ldots = \sigma_{k_z}^S = 1$. Similarly to [65, 66], one can show that if $r_f \geq k_z$ in Algorithm 4, then $\boldsymbol{H}(s)$ and $\tilde{\boldsymbol{H}}(s)$ have the same zeros in the closed right half-plane, and the relative error bound

$$\|\boldsymbol{H}^{-1}(\boldsymbol{H} - \tilde{\boldsymbol{H}})\|_{\mathcal{H}_\infty} \leq \prod_{j=r_f+1}^{n_f} \frac{1 + \sigma_j^S}{1 - \sigma_j^S} - 1$$

holds. Thus, if $\boldsymbol{H}(s)$ is minimum phase, then $\tilde{\boldsymbol{H}}(s)$ is also minimum phase.

Algorithm 4 Stochastic balanced truncation for DAE systems.

Input: an asymptotically stable system $\boldsymbol{H} = (E, A, B, C, D)$ with the proper and invertible transfer function.

Output: a reduced-order asymptotically stable system $\tilde{\boldsymbol{H}} = (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

1: Compute the full rank Cholesky factors Z_{ic} and Z_{io} of the improper Gramians $G_{ic} = Z_{ic}Z_{ic}^T$ and $G_{io} = Z_{io}Z_{io}^T$ satisfying the projected Lyapunov equations (3.3) and (3.4), respectively.
2: Compute the singular value decomposition $Z_{io}^T A Z_{ic} = U_3 \Theta V_3^T$ with nonsingular Θ.
3: Compute $M_0 = D - CZ_{ic}V_3\Theta^{-1}U_3^T Z_{io}^T B$.
4: Compute the Cholesky factors Z_c^S and Z_o^S of the stochastic controllability Gramian $G_c^S = Z_c^S(Z_c^S)^T = G_{pc}$ satisfying (3.1) and the stochastic observability Gramian $G_o^S = Z_o^S(Z_o^S)^T$ which is the minimal solution of the projected Lur'e equation (3.11).
5: Compute the singular value decomposition $(Z_o^S)^T E Z_c^S = [\,U_1,\, U_2\,]\mathrm{diag}(\Sigma_1^S, \Sigma_2^S)[\,V_1,\, V_2\,]^T$, where $\Sigma_1^S = \mathrm{diag}(\sigma_1^S, \ldots, \sigma_{\eta}^S)$ and $\Sigma_2^S = \mathrm{diag}(\sigma_{\eta+1}^S, \ldots, \sigma_{n_f}^S)$.
6: Compute the reduced-order system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (W^T E T,\, W^T A T,\, W^T B,\, CT,\, M_0)$ with the projection matrices $W = Z_o^S U_1 (\Sigma_1^S)^{-1/2}$ and $T = Z_c^S V_1 (\Sigma_1^S)^{-1/2}$.

If M_0 is nonsingular, then the projected Lur'e Equation (3.11) reduces to the projected Riccati equation

$$A^T Y E + E^T Y A + (B_0^T Y E - C P_r)^T (M_0 M_0^T)^{-1}(B_0^T Y E - C P_r) = 0, \quad Y = P_l^T Y P_l.$$

It has been shown in [159] that for standard state space systems with the invertible and strictly minimum phase transfer function $\boldsymbol{H}(s)$, the stochastic balanced truncation method is equivalent to a frequency-weighted balanced truncation approach with $\boldsymbol{H}^{-1}(s)$ as an output weight and I as an input weight. This approach is based on balancing the controllability Gramian of $\boldsymbol{H}$ against the observability Gramian of $\boldsymbol{H}^{-1}$. It can also be extended to the DAE system (1.1). If M_0 is nonsingular, then $\boldsymbol{H}^{-1}(s)$ can be realized as

$$\boldsymbol{H}^{-1} = (E,\ A - P_l B M_0^{-1} C P_r,\ P_l B M_0^{-1},\ -M_0^{-1} C P_r,\ M_0^{-1}).$$

The proper observability Gramian $\hat{G}_{po}$ of $\boldsymbol{H}^{-1}$ is defined as the solution of the projected Lyapunov equation

$$(A - P_l B M_0^{-1} C P_r)^T \hat{G}_{po} E + E^T \hat{G}_{po}(A - P_l B M_0^{-1} C P_r) = -P_r^T C^T (M_0 M_0^T)^{-1} C P_r,$$
$$\hat{G}_{po} = P_l^T \hat{G}_{po} P_l.$$

The stochastic characteristic values σ_j^S are related to the new characteristic values $\hat{\sigma}_j = \sqrt{\lambda_j(G_{pc}E^T\hat{G}_{po}E)}$ via $\sigma_j^S = \hat{\sigma}_j/\sqrt{(1+\hat{\sigma}_j^2)}$, see [158]. Thus, if (1.1) is asymptotically stable, $\boldsymbol{H}(s)$ is strictly minimum phase and M_0 is nonsingular, then the stochastic balanced truncation method involves solving two projected Lyapunov equations, and, hence, it is as expensive as Lyapunov-based balanced truncation.

3.1.5 LQG Balanced Truncation

Another balancing-related model reduction approach is linear-quadratic Gaussian (LQG) balanced truncation developed first for unstable standard state-space systems in [78]. An extension of this method to DAEs was presented in [95] and further developed in [21] for flow control problems. The LQG balanced truncation method is based on the generalized Riccati equations

$$\begin{aligned} AX^T + XA^T + BB^T - (XC^T + BD^T)(I + DD^T)^{-1}(CX^T + DB^T) = 0, \\ EX^T - XE^T = 0, \end{aligned} \tag{3.12}$$

and

$$\begin{aligned} A^TY + Y^TA + C^TC - (Y^TB + C^TD)(I + D^TD)^{-1}(B^TY + D^TC) = 0, \\ E^TY - Y^TE = 0, \end{aligned} \tag{3.13}$$

where the matrices $I + DD^T$ and $I + D^TD$ are assumed to be nonsingular. Note that these equations do not involve the spectral projectors. One can show that if the DAE system (1.1) is S-stabilizable and S-detectable, then Equations (3.12) and (3.13) have stabilizing solutions X and Y such that the pencils

$$\begin{aligned} &\lambda E - \big(A - (XC^T + BD^T)(I + DD^T)^{-1}C\big), \\ &\lambda E - \big(A - B(I + D^TD)^{-1}(B^TY + D^TC)\big) \end{aligned}$$

are both of index one and stable. The matrices $G_c^{LQG} = XE^T$ and $G_o^{LQG} = Y^TE$ are called the *LQG controllability* and *observability Gramians* of the DAE system (1.1). In contrast to X and Y, the Gramians G_c^{LQG} and G_o^{LQG} are symmetric, positive semidefinite and uniquely defined. The *LQG characteristic values* are defined as

$$\sigma_j^{LQG} = \sqrt{\lambda_j(G_c^{LQG}(E^+)^TG_o^{LQG}E^+)},$$

where E^+ denotes the Moore–Penrose pseudoinverse of E. Balancing the LQG Gramians and truncating the states corresponding to small LQG characteristic values provides the LQG balanced truncation model reduction method given in Algorithm 5.

For the LQG reduced-order system, there exists an error estimate in the gap metric [61] defined as follows. Let the DAE system (1.1) be S-stabilizable and S-detectable. Then its transfer function $\boldsymbol{H}(s)$ can be factored as $\boldsymbol{H}(s) = \boldsymbol{K}(s)\boldsymbol{M}^{-1}(s)$, where

$$\begin{aligned} \boldsymbol{K}(s) &= (C + DF)(sE - A - BF)^{-1}B(I + D^TD)^{-1/2} + D(I + D^TD)^{-1/2}, \\ \boldsymbol{M}(s) &= F(sE - A - BF)^{-1}B(I + D^TD)^{-1/2} + (I + D^TD)^{-1/2} \end{aligned}$$

Algorithm 5 LQG balanced truncation for DAE systems.

Input: $\boldsymbol{H} = (E, A, B, C, D)$

Output: a reduced-order model $\tilde{\boldsymbol{H}} = (\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$.

1: Compute the full rank matrices Z_r and Z_l such that $\operatorname{im}(Z_r) = \ker(E)$ and $\operatorname{im}(Z_l) = \ker(E^T)$.

2: Compute the Cholesky factors Z_c^{LQG} and Z_o^{LQG} such that $EX^T = EZ_c^{LQG}(Z_c^{LQG})^T E^T$ and $E^T Y = E^T Z_o^{LQG}(Z_o^{LQG})^T E$, where X and Y are the stabilizing solutions of the generalized Riccati equations (3.12) and (3.13), respectively.

3: Compute $(Z_o^{LQG})^T E Z_c^{LQG} = [U_1, U_2]\operatorname{diag}(\Sigma_1^{LQG}, \Sigma_2^{LQG})[V_1, V_2]^T$ by singular value decomposition with $\Sigma_1^{LQG} = \operatorname{diag}(\sigma_1^{LQG}, \ldots, \sigma_r^{LQG})$ and $\Sigma_2^{LQG} = \operatorname{diag}(\sigma_{r+1}^{LQG}, \ldots, \sigma_k^{LQG})$.

4: Compute the reduced-order system $(\tilde{E}, \tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) = (W^T E T, W^T A T, W^T B, CT, D)$ with the projection matrices $W = [Z_o^{LQG} U_1 (\Sigma_1^{LQG})^{-1/2}, Z_l]$ and $T = [Z_c^{LQG} V_1 (\Sigma_1^{LQG})^{-1/2}, Z_r]$.

with $F = -(I + D^T D)^{-1}(B^T Y + D^T C)$, are stable proper rational functions called the *right coprime factors* of $\boldsymbol{H}(s)$. Obviously, $\begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{K} \end{bmatrix} \in \mathcal{H}_\infty$, and we obtain the error estimate

$$\left\| \begin{bmatrix} \tilde{\boldsymbol{M}} \\ \tilde{\boldsymbol{K}} \end{bmatrix} - \begin{bmatrix} \boldsymbol{M} \\ \boldsymbol{K} \end{bmatrix} \right\|_{\mathcal{H}_\infty} \leq 2 \sum_{j=r+1}^{k} \frac{\sigma_j^{LQG}}{\sqrt{1 + \sigma_j^{LQG}}},$$

where $\tilde{\boldsymbol{H}}(s) = \tilde{\boldsymbol{K}}(s)\tilde{\boldsymbol{M}}^{-1}(s)$ is the right coprime factorization of $\tilde{\boldsymbol{H}}(s)$ and σ_j^{LQG} are the LQG characteristic values from Algorithm 5, see [95].

Projector-free generalized Riccati equations similar to (3.12) and (3.13) have also been studied in the context of linear-quadratic optimal control [79, 117, 154], spectral factorization problems [80, 81], and extensions of the positive real and bounded real lemmas to DAE systems [58, 149–151, 156]. Stability and the index-1 property of (1.1) can also be characterized via the projector-free generalized Lyapunov equations

$$\begin{aligned} AX^T + XA^T + BB^T &= 0, & EX^T - XE^T &= 0, \\ A^T Y + Y^T A + C^T C &= 0, & E^T Y - Y^T E &= 0, \end{aligned}$$

see [74, 141]. All these matrix equations provide an alternative way to define different types of Gramians for DAEs and also new balancing-related model reduction methods [112]. They might be advantageous if the spectral projectors are difficult to compute. It should, however, be noticed that currently existing numerical methods for such equations are restricted to small and medium-sized problems. Another disadvantage is that these new model reduction techniques would be limited, in most cases, to index one problems.

3.2 Interpolation-Based Approximation

Another family of methods for model reduction is based on (rational) interpolation. The unifying feature of the methods in this family is that the original transfer function $\mathbf{H}(s)$ is approximated by a rational matrix function $\tilde{\mathbf{H}}(s)$ of lower degree satisfying some interpolation conditions (that is, the original and the reduced-order transfer function coincide, e.g. $\mathbf{H}(s_0) = \tilde{\mathbf{H}}(s_0)$ at some predefined value s_0 such that $A - s_0E$ is nonsingular). Computationally, this is usually realized by certain Krylov subspace methods.

The classical approach is known under the name of *moment-matching* or *Padé(-type) approximation*. In these methods, the transfer functions of the original and the reduced-order systems are expanded into power series and the reduced-order system is then determined so that the first coefficients in the series expansions match. In this context, the coefficients of the power series expansion are called *moments*, which explains the term moment-matching. One speaks of Padé-approximation if the number of matching moments is maximized for a given degree of the approximating rational function.

Classically, the expansion of the transfer function in a power series about an expansion point s_0 as in (2.9) is used. Recall that the moments $M_j(s_0), j = 0, 1, 2, \ldots$ are given by

$$M_j(s_0) = -C\left((A - s_0E)^{-1}E\right)^j(A - s_0E)^{-1}B + \delta_{0,j}D.$$

Note that s_0 is necessarily chosen such that $A - s_0E$ is nonsingular, and hence s_0 is neither an eigenvalue of the matrix pencil $\lambda E - A$ nor a pole of the transfer function $\boldsymbol{H}(s)$. Thus, the approach described in the following can be applied regardless whether E is singular or not, so that no special adaptation to DAE systems is necessary.

Now consider the *block Krylov subspace*

$$\mathcal{K}_k(F, G) = \text{blockspan}\{G, FG, F^2G, \ldots, F^{k-1}G\}$$

generated by $F = (A - s_0E)^{-1}E$ and $G = -(A - s_0E)^{-1}B$ with an appropriately chosen expansion point s_0 which may be real or complex. From the definitions of A, B and E, it follows that $F \in \mathbb{K}^{n\times n}$ and $G \in \mathbb{K}^{n\times m}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ depending on whether s_0 is chosen in $\mathbb{R}$ or in $\mathbb{C}$. Considering $\mathcal{K}_k(F, G)$ columnwise, this leads to the observation that the number of column vectors in $[G, FG, F^2G, \ldots, F^{k-1}G]$ is given by $r = m \cdot k$, as there are k blocks $F^jG \in \mathbb{K}^{n\times m}$, $j = 0, \ldots, k - 1$. In the case when all r column vectors are linearly independent, the dimension of the Krylov subspace $\mathcal{K}_k(F, G)$ is r. Assume that a unitary basis for this block Krylov subspace is generated such that the column-space of the resulting unitary matrix $T \in \mathbb{K}^{n\times r}$ spans $\mathcal{K}_k(F, G)$. Applying the Galerkin projection $\Pi = TT^*$ to (1.1) yields a reduced system whose transfer function satisfies the

following Hermite interpolation conditions

$$\tilde{\boldsymbol{H}}^{(j)}(s_0) = \boldsymbol{H}^{(j)}(s_0), \qquad j = 0, 1, \ldots, k-1.$$

This means that transfer functions $\boldsymbol{H}$ and $\tilde{\boldsymbol{H}}$ and their first k derivatives coincide at s_0. Considering the power series expansion (2.9) of the original and the reduced-order transfer function, this is equivalent to saying that at least the first k moments $\tilde{M}_j(s_0)$ of the transfer function $\tilde{\boldsymbol{H}}(s)$ of the reduced system (1.2) are equal to the first k moments $M_j(s_0)$ of the transfer function $\boldsymbol{H}(s)$ of the original system (1.1) at the expansion point s_0, i.e.,

$$M_j(s_0) = \tilde{M}_j(s_0), \qquad j = 0, 1, \ldots, k-1.$$

If further the r columns of the unitary matrix W span the block Krylov subspace $\mathcal{K}_k(F, G)$ for $F = (A - s_0E)^{-T}E^T$ and $G = -(A - s_0E)^{-T}C^T$, applying the Petrov–Galerkin projection $\Pi = T(W^*T)^{-1}W^*$ to (1.1) yields a reduced system whose transfer function matches at least the first $2k$ moments of the transfer function $\boldsymbol{H}(s)$ of the original system.

Theoretically, the matrix T (and W) can be computed by explicitly forming the columns which span the corresponding Krylov subspace $\mathcal{K}_k(F, G)$ and using the Gram–Schmidt algorithm to generate unitary basis vectors for $\mathcal{K}_k(F, G)$. The forming of the moments (the Krylov subspace blocks F^jG) is numerically precarious and has to be avoided under all circumstances. Instead, it is recommended to use Krylov subspace methods to achieve an interpolation-based reduced-order model as described above. The unitary basis of a (block) Krylov subspace can be computed by employing a (block) Arnoldi or (block) Lanczos method, see e.g. [8, 55, 64].

In the case when an oblique projection is used, it is not necessary to compute two unitary bases as above. An alternative is then to use the nonsymmetric Lanczos process [64]. It computes bi-unitary bases for the above-mentioned Krylov subspaces and the reduced-order model as a by-product of the Lanczos process. An overview of the computational techniques for moment-matching and Padé approximation summarizing the work of a decade is given in [55] and the references therein.

The use of complex-valued expansion points will lead to a complex-valued reduced-order system (1.2). In some applications (in particular, if the original system is real-valued) this is undesired. In that case one can always use complex-conjugate pairs of expansion points as then the entire computations can be done in real arithmetic.

In general, the discussed model order reduction approaches are instances of rational interpolation. When the expansion point is chosen to be $s_0 = \infty$, the moments are called Markov parameters and the approximation problem is known as *partial realization*. Here, the singularity of E obviously makes a difference as then the Laurent expansion (2.10) is used. For singular E, using the reflexive inverse of E, a partial realization method for descriptor systems was derived in [24].

As the use of one single expansion point s_0 leads to good approximation only close to s_0, it might be desirable to use more than one expansion point. This leads to *multi-point moment-matching* methods, which can also be interpreted as rational Krylov methods, see, e.g., [8, 55].

Assume that ℓ expansion points s_i, $i = 1, 2, \ldots, \ell$ are considered. The column vectors of the matrix T are determined from the ℓ block Krylov subspaces $\mathcal{K}_{k_i}(F_i, G_i)$ generated by $F_i = (A - s_i E)^{-1}E$ and $G_i = -(A - s_i E)^{-1}B$ for $i = 1, 2, \ldots, \ell$. From each of these subspaces, the $m \cdot k_i$ column vectors are used to generate an $n \times r$ matrix

$$\hat{T} = \left[T_{[k_1]},\ T_{[k_2]},\ \ldots,\ T_{[k_\ell]} \right], \qquad r = m \sum_{i=1}^{\ell} k_i.$$

In order to obtain a unitary, full-rank matrix T, a rank-revealing QR decomposition can be used $\hat{T} = TR$, so that the numerical rank of $\hat{T}$ can be determined, $\hat{r} = \operatorname{rank}(\hat{T})$, and finally, $\hat{T}$ can be truncated to $T = [T(:, 1 : \hat{r})]$ (employing MATLAB® notation). The columns of T span the same subspace as the span of the union of the Krylov subspaces $\mathcal{K}_{k_i}(F_i, G_i)$, that is, $\operatorname{span}(T) = \cup_{i=1}^{\ell} \mathcal{K}_{k_i}(F_i, G_i)$. Then at least k_i moments are matched per expansion point s_i:

$$M_j(s_i) = \tilde{M}_j(s_i), \qquad j = 0, 1, \ldots, k_i - 1, \quad i = 1, 2, \ldots, \ell,$$

if the reduced system is generated by applying the Galerkin projection $\Pi = TT^*$. In this case, $\tilde{\boldsymbol{H}}$ fulfils the Hermite interpolation conditions

$$\tilde{\boldsymbol{H}}^{(j)}(s_i) = \boldsymbol{H}^{(j)}(s_i), \qquad j = 0, 1, \ldots, k_i - 1, \quad i = 1, 2, \ldots, \ell.$$

A Petrov–Galerkin projection can also be constructed following this idea. Then at least $2k_i$ moments are matched per expansion point s_i. It should be noted that at each s_i a different number of moments k_i is matched.

In contrast to balanced truncation, these (rational) interpolation methods do not necessarily preserve stability. Remedies have been suggested, see, e.g. [55].

The methods just described provide good approximation quality around the expansion points. They do not aim at a global approximation as measured by the $\mathcal{H}_2$- or $\mathcal{H}_\infty$-norm. In [68], an iterative procedure is presented which determines, upon convergence,[3] locally optimal expansion points with respect to the $\mathcal{H}_2$-norm approximation under the assumption that the order r of the reduced model is prescribed and such that only 0-th and 1-st order derivatives are matched. This is motivated by the necessary $\mathcal{H}_2$-norm optimality conditions for a stable, r-th order, rational interpolant $\tilde{\boldsymbol{H}}$ of $\boldsymbol{H}$. In order for $\tilde{\boldsymbol{H}}$ to be a local minimizer of the error measured in the $\mathcal{H}_2$-norm, it is necessarily a Hermite interpolant in the classical sense, i.e., interpolation of the function value and its first-order derivative at the

[3] For partial convergence results, see [52].

mirror images (with respect to the imaginary axis) of the poles of $\tilde{\boldsymbol{H}}$, see [94]. Also, for multi-input multi-output systems (that is, m and q in (1.1) are both larger than one), no full moment-matching is achieved, but only tangential interpolation

$$\boldsymbol{H}(s_j)b_j = \tilde{\boldsymbol{H}}(s_j)b_j, \quad c_j^*\boldsymbol{H}(s_j) = c_j^*\tilde{\boldsymbol{H}}(s_j), \quad c_j^*\boldsymbol{H}'(s_j)b_j = c_j^*\tilde{\boldsymbol{H}}'(s_j)b_j$$

for certain vectors b_j and c_j determined together with the optimal s_j by the iterative procedure. The $\mathcal{H}_2$-optimal approximation procedure was extended to DAE systems in [69]. Though the interpolation properties of the reduced-order transfer function are the same for ODE and DAE systems, one needs to take special care of behavior at infinity for DAE systems. In order for the error function $\boldsymbol{H} - \tilde{\boldsymbol{H}}$ to be an $\mathcal{H}_2$-function, it needs to be zero at infinity, which usually is not the case when only applying the necessary optimality conditions of the ODE case. In addition, it is necessary to "interpolate" at infinity. This requires some additional work and altering the realization of the reduced-order model without destroying the interpolation conditions in the mirror images of its poles. A procedure achieving this and requiring little extra effort is described in [69], but we refrain here from reproducing the technical details.

4 Solving Large Matrix Equations

In this section, we discuss the numerical solution of projected Lyapunov and Riccati matrix equations arising in balancing-related model reduction of DAE systems. We assume that the spectral projectors in these equations are given, though their computation may be a challenging task, especially for large-scale problems. Fortunately, for some structured problems, the spectral projectors can either be constructed explicitly or the DAE system (and also the matrix equations) can be modified such that the projectors are no longer required. This issue will be addressed in Sect. 5.

4.1 Projected Lyapunov Equations

We consider first the projected discrete-time Lyapunov equation

$$AXA^T - EXE^T = Q_lBB^TQ_l^T, \qquad X = Q_rXQ_r^T, \tag{4.1}$$

where $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ with $m \ll n$. If the pencil $\lambda E - A$ is stable, i.e., all its finite eigenvalues have negative real part, and it has index ν, then A is nonsingular and the solution of (4.1) can be represented as

$$X = \sum_{j=0}^{\nu-1}(A^{-1}E)^jA^{-1}Q_lBB^TQ_l^TA^{-T}((A^{-1}E)^T)^j = ZZ^T$$

Algorithm 6 Smith method for projected discrete-time Lyapunov equations.

Input: $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, the spectral projector Q_r, and a convergence tolerance $tol > 0$.
Output: a low-rank factor Z_k such that $X = Z_k Z_k^T$ is an approximate solution of (4.1).

$V_0 = Q_r A^{-1} B$;
$Z_0 = [\ ]$;
$k = 0$;
while $\|V_k\|_F > tol$ **do**
  $Z_{k+1} = [Z_k, V_k]$;
  $V_{k+1} = A^{-1} E V_k$;
  $k \leftarrow k + 1$;
end while

with $Z = Q_r[A^{-1}B, (A^{-1}E)A^{-1}B, \ldots, (A^{-1}E)^{\nu-1}A^{-1}B]$. If the index of $\lambda E - A$ is unknown a priori, then this low-rank factor can be computed using the generalized Smith iteration [137] which converges in a finite number of steps, see Algorithm 6.

We consider now the projected continuous-time Lyapunov equation

$$EXA^T + AXE^T = -P_l BB^T P_l^T, \qquad X = P_r X P_r^T, \tag{4.2}$$

where $\lambda E - A$ is assumed to be stable. We aim to determine the solution of this equation in the factored form $X = ZZ^T$, avoiding the computation of the solution matrix X. For problems of small and moderate size (up to a few thousands), this can be achieved using the generalized Schur–Hammarling method [133] which relies on computing the generalized Schur form of the pencil $\lambda E - A$. One can also employ the matrix sign function method which was initially developed for standard Lyapunov equations [18, 87, 119] and then extended to projected Lyapunov equations in [136]. This method is efficient, in particular, for large dense problems.

As mentioned in Sect. 3.1.1, to be able to apply the balanced truncation method to large-scale problems, we are rather interested in a low-rank approximation $X \approx \tilde{Z}\tilde{Z}^T$ with $\tilde{Z} \in \mathbb{R}^{n\times k}$ and $k \ll n$. The simplest way to compute such an approximation is based on the integral representation (3.5) for the solution of (4.2). Computing this integral by a quadrature rule

$$\begin{aligned} X \approx &\sum_{j=1}^{p} f_j (i\omega_j E - A)^{-1} P_l BB^T P_l^T (-i\omega_j E - A)^{-T} \\ &+ \sum_{j=1}^{p} f_j (-i\omega_j E - A)^{-1} P_l BB^T P_l^T (i\omega_j E - A)^{-T} \end{aligned}$$

with nonnegative nodes ω_j and positive weights f_j, we obtain the real low-rank factor

$$\tilde{Z} = [\mathrm{Re}(B_1), \mathrm{Im}(B_1), \ldots, \mathrm{Re}(B_p), \mathrm{Im}(B_p)] \in \mathbb{R}^{n\times 2pm}$$

with $B_j = \sqrt{2f_j}(i\omega_j E - A)^{-1} P_l B$. For the dual projected Lyapunov equation, the low-rank factor can be calculated analogously. Using these factors in Algorithm 1

can be viewed as an extension of the frequency domain POD approach [153] and the PMTBR method [106] to DAE systems.

4.1.1 Alternating Directions Implicit Method

A low-rank approximation to the solution of the projected Lyapunov Equation (4.2) can also be computed iteratively using a low-rank version of the alternating directions implicit method known as the LR-ADI method [89, 102, 137]. In recent years, several modifications concerning the efficient computation of Lyapunov residuals, adaptive choice of ADI shift parameters and handling the complex shifts have been proposed for Lyapunov equations with nonsingular E, which significantly improve the performance of the ADI iteration [30, 31, 33]. An extension of these results to the projected Lyapunov equation is straightforward [25, 137] and summarized in Algorithm 7.

One can see that this algorithm provides a real low-rank factor $Z_k \in \mathbb{R}^{n\times km}$ and the computational cost for the LR-ADI method is proportional to the cost of solving linear systems with the sparse matrix $E + \tau_k A$. The convergence rate of the ADI iteration is strongly influenced by the shift parameters $\tau_k \in \mathbb{C}_-$. Optimal parameters can be obtained by solving the minimax problem

$$\{\hat\tau_1,\dots,\hat\tau_p\} = \operatorname*{arg\,min}_{\{\tau_1,\dots,\tau_p\}\in\mathbb{C}_-} \max_{t\in \mathrm{Sp}(E,A)} \frac{|(1-\overline{\tau}_1 t)\cdots(1-\overline{\tau}_p t)|}{|(1+\tau_1 t)\cdots(1+\tau_p t)|},$$

Algorithm 7 LR-ADI method for projected continuous-time Lyapunov equations.

Input: $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, the spectral projector P_l, shifts $\tau_1,\dots,\tau_p \in \mathbb{C}_-$, a tolerance tol, and $k_{\max} \in \mathbb{N}$.
Output: a low-rank factor Z_k such that $X \approx Z_k Z_k^T$ solves (4.2) approximately.

$W_0 = P_l B$;
$Z_0 = [\ ]$;
$k = 1$;
while $(\|W_{k-1}^T W_{k-1}\|_F / \|W_0^T W_0\|_F > tol$ **and** $k < k_{\max})$ **do**
  $V_k = (E + \tau_k A)^{-1} W_{k-1}$;
  if $\tau_k \in \mathbb{R}$ **then**
    $W_k = W_{k-1} - 2\tau_k A\, V_k$;
    $Z_k = [Z_{k-1},\ \sqrt{-2\tau_k}\, V_k\,]$;
  else
    $\alpha_k = \sqrt{-2\mathrm{Re}(\tau_k)}$, $\beta_k = \mathrm{Re}(\tau_k)/\mathrm{Im}(\tau_k)$;
    $W_{k+1} = W_{k-1} - 4\mathrm{Re}(\tau_k) A\big(\mathrm{Re}(V_k) + \beta_k \mathrm{Im}(V_k)\big)$;
    $Z_k = [Z_{k-1},\ \alpha_k\big(\mathrm{Re}(V_k) + \beta_k \mathrm{Im}(V_k)\big),\ \alpha_k\sqrt{\beta_k^2 + 1}\,\mathrm{Im}(V_k)\,]$;
    $k \leftarrow k+1$;
  end if
  $k \leftarrow k+1$;
end while

where $\mathrm{Sp}(E, A)$ denotes the set of finite eigenvalues of the pencil $\lambda E - A$. Suboptimal ADI parameters can be determined from a set of largest and smallest in modulus approximate finite eigenvalues of $\lambda E - A$ computed by an Arnoldi or Lanczos procedure, or any other method to compute the extreme eigenvalues of a matrix pencil. Any other parameter selection technique developed for standard Lyapunov equations [33, 124, 148] can also be used for the projected Lyapunov equation.

4.1.2 Krylov Subspace Methods

Alternative iterative methods for Lyapunov equations are Krylov subspace methods [38, 75, 77, 122] which become competitive with the ADI iteration due to recent developments on extended and rational Krylov subspaces [46, 83, 128], see also [47] for a comparative analysis of the Krylov subspace and ADI methods. Employing the ADI iteration as a preconditioner in Krylov subspace methods has been considered in [38, 76]. An extension of these methods to projected Lyapunov equations can be found in [38, 140]. The approaches differ in the way the linear matrix equation is solved by either interpreting them as classical linear systems using their Kronecker product representation in $\mathbb{R}^{n^2}$, as is the case, e.g., for [38, 76, 77], or by directly working on the matrix equation and building the Krylov subspaces in $\mathbb{R}^n$ as done in [46, 75, 83, 122, 128, 140]. The latter approach appears to be more efficient (though also the first approach uses Krylov subspaces in $\mathbb{R}^{n^2}$ only implicitly), and we will therefore concentrate on this concept here.

In the Krylov subspace methods, an approximate solution to the projected Lyapunov Equation (4.2) is determined in the form $X \approx V\tilde{Y}\tilde{Y}^T V^T$, where columns of V span a certain Krylov subspace and $Y = \tilde{Y}\tilde{Y}^T$ solves the reduced Lyapunov equation

$$\tilde{A}\, Y + Y\tilde{A}^T = -\tilde{B}\tilde{B}^T,$$

where $\tilde{A} = V^T A^{-1} E V$ and $\tilde{B} = V^T A^{-1} P_l B$ or, alternatively, $\tilde{A} = V^T E^- A\, V$ and $\tilde{B} = V^T E^- B$. Here,

$$E^- = T_r^{-1} \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} T_l^{-1}$$

is a reflexive inverse of E with respect to the projectors P_l and P_r satisfying the matrix equations

$$E^- E E^- = E^-, \qquad E E^- = P_l, \qquad E^- E = P_r.$$

The projection subspace $\mathrm{im}(V)$ can be chosen as an extended block Krylov subspace

$$\mathcal{K}_k(A^{-1}E, A^{-1}P_l B) \cup \mathcal{K}_k(E^- A, E^- B).$$

Algorithm 8 Extended block Arnoldi method for projected Lyapunov equations.

Input: $A, E \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$, the spectral projector P_r, and $k \in \mathbb{N}$.
Output: a low-rank factor Z_k such that $X \approx Z_k Z_k^T$ solves (4.2) approximately.

$\hat{V}_1 = \operatorname{orth}([E^- B,\ P_r A^{-1} B])$ {orthogonalization of the columns of $[E^- B,\ P_r A^{-1} B]$};
$V_1 = \hat{V}_1,\ \ V_{1,1} = \hat{V}_1 [I_m,\ 0]^T,\ \ V_{1,2} = \hat{V}_1 [0,\ I_m]^T$;
for $j = 1, 2, \ldots, k$ **do**
  $V^{(j)} = [E^- A\, V_{j,1},\ A^{-1} E V_{j,2}]$;
  for $i = 1, 2, \ldots, j$ **do**
    $H_{i,j} = \hat{V}_i^T V^{(j)}$;
    $V^{(j)} = V^{(j)} - \hat{V}_i H_{i,j}$;
  end for
  $\hat{V}_{j+1} = \operatorname{orth}(V^{(j)})$ {orthogonalization of the columns of $V^{(j)}$};
  $V_{j+1} = [V_j, \hat{V}_{j+1}],\ \ V_{j+1,1} = \hat{V}_{j+1}[I_m, 0]^T,\ \ V_{j+1,2} = \hat{V}_{j+1}[0, I_m]^T$;
  $\Phi_j = V_j^T E^- A V_j,\ B_j = V_j^T E^- B$;
  solve the Lyapunov equation $\Phi_j Y_j + Y_j \Phi_j^T = -B_j B_j^T$ for $Y_j = \tilde{Y}_j \tilde{Y}_j^T$;
end for
$Z_k = V_k \tilde{Y}_k$.

The resulting numerical procedure based on a block Arnoldi method for computing an orthogonal basis of this subspace and solving the projected Lyapunov Equation (4.2) is given in Algorithm 8.

The iteration in this algorithm can be terminated as soon as the normalized residual defined by

$$\eta(Z_j) = \frac{\|E Z_j Z_j^T A^T + A Z_j Z_j^T E^T\|_F}{\|P_l B B^T P_l^T\|_F}$$

satisfies the condition $\eta(Z_j) \le tol$ with a tolerance tol. Since the computation of the residual is expensive for large-scale problems, it has been proposed in [140] to use the following stopping criterion:

$$\frac{\|E^-(E Z_j Z_j^T A^T + A Z_j Z_j^T E^T)(E^-)^T\|_F}{\|E^- B B^T (E^-)^T\|_F} = \frac{\sqrt{2}\|V_{j+1,1}^T E^- A V_j Y_j\|_F}{\|(E^- B)^T (E^- B)\|_F} \le tol,$$

where the matrix $V_{j+1,1}^T E^- A V_j$ can be obtained as a by-product of the iteration with no additional matrix-vector products with E^- and A and inner products with long vectors.

In the rational Krylov subspace method, the projection subspace $\operatorname{im}(V)$ is taken as the rational block Krylov subspace defined as

$$\mathcal{K}_k(E, A, B; s_1, \ldots, s_k) = \operatorname{blockspan}\Big\{(s_1 E - A)^{-1} P_l B,\ (s_2 E - A)^{-1} E (s_1 E - A)^{-1} P_l B,\ \ldots,\ (s_k E - A)^{-1} \prod_{j=1}^{k-1} E (s_j E - A)^{-1} P_l B\Big\}$$

for some shifts $s_1, \ldots, s_k$ which are not the eigenvalues of $\lambda E - A$. As in the LR-ADI method, these parameters should be chosen carefully to guarantee fast convergence [46, 47].

4.2 Projected Riccati Equations

We consider now the projected Riccati equation in the general form

$$EXF^T + FXE^T + EXQ^TQXE^T + P_lRR^TP_l^T = 0, \qquad X = P_rXP_r^T, \tag{4.3}$$

where the matrices $F \in \mathbb{R}^{n\times n}$, $R \in \mathbb{R}^{n\times m}$ and $Q \in \mathbb{R}^{q\times n}$ vary depending on the balanced truncation method:

$$F = A - P_lBJ_c^{-T}J_c^{-1}CP_r, \quad Q = J_c^{-1}C, \quad R = BJ_c^{-T}, \quad M_0 + M_0^T = J_cJ_c^T$$

in the positive real case and

$$\begin{aligned} &F = A + P_lBM_0J_c^{-T}J_c^{-1}CP_r, \quad Q = J_c^{-1}C, \quad R = BJ_o^{-1}, \\ &I - M_0M_0^T = J_cJ_c^T, \quad I - M_0^TM_0 = J_o^TJ_o \end{aligned}$$

in the bounded real case. In the stochastic balanced truncation method, where a dual Riccati equation has to be solved, E, P_r and P_l should be replaced by E^T, P_l^T and P_r^T, respectively, and

$$\begin{aligned} &F = (A - B_0M_0^{-T}M_0^{-1}CP_r)^T, \quad Q = M_0^{-1}B_0^T, \quad R = C^TM_0^{-T}, \\ &B_0 = P_lBM_0^T + EG_{pc}C^T. \end{aligned}$$

We assume that (4.3) has a unique stabilizing solution X_* such that the matrix pencil $\lambda E - (F + EX_*Q^TQP_r)$ is stable. Since the first equation in (4.3) is nonlinear, we can solve it by Newton's method presented in [25]. For this purpose, we define a Riccati operator

$$\mathcal{R}(X) = EXF^T + FXE^T + EXQ^TQXE^T + P_lRR^TP_l^T$$

and compute its Frechét derivative

$$\mathcal{R}'_X(N) = EN(F + EXQ^TQP_r)^T + (F + EXQ^TQP_r)NE^T.$$

Then Newton's method for the projected Riccati Eq. (4.3) is given by

$$N_j = -(\mathcal{R}'_{X_j})^{-1}(\mathcal{R}(X_j)), \qquad X_{j+1} = X_j + N_j. \tag{4.4}$$

Algorithm 9 Low-rank Newton method for projected Riccati equations.

Input: $E, F \in \mathbb{R}^{n\times n}$ such that $\lambda E - F$ is stable, $Q \in \mathbb{R}^{q\times n}$, $R \in \mathbb{R}^{n\times m}$, projectors P_r and P_l.
Output: an approximate low-rank factor of the stabilizing solution of (4.3).

1: Solve $EN_0F^T + FN_0E^T = -P_lRR^TP_l^T$, $N_0 = P_rN_0P_r^T$ for the low-rank factor $\tilde{N}_0$ such that $N_0 \approx \tilde{N}_0\tilde{N}_0^T$;
2: $\tilde{X}_1 = \tilde{N}_0$;
3: $F_0 = F$;
4: **for** $j = 1, 2, \ldots$ **do**
5: $K_j = E\tilde{N}_{j-1}\tilde{N}_{j-1}^TQ^T$;
6: $F_j = F_{j-1} + K_jQP_r$;
7: solve (4.5) for the low-rank factor $\tilde{N}_j$ such that $N_j \approx \tilde{N}_j\tilde{N}_j^T$;
8: $\tilde{X}_{j+1} = [\,\tilde{X}_j,\ \tilde{N}_j\,]$.
9: **end for**

It has been shown in [25] that this iteration converges quadratically towards X_* for any stabilizing initial guess X_0. If $\lambda E - F$ is stable, then we can take $X_0 = 0$. However, for unstable problems, the computation of a stabilizing X_0 might be challenging. For some methods to find an initial stabilizing feedback for descriptor systems, see [17].

Note that the first equation in (4.4) is equivalent to the projected Lyapunov equation

$$EN_jF_j^T + F_jN_jE^T = -P_lK_jK_j^TP_l^T, \qquad N_j = P_rN_jP_r^T \tag{4.5}$$

with $F_j = F + EX_jQ^TQP_r$ and $K_j = EN_{j-1}Q^T$. This equation can now be solved for a low-rank factor using the LR-ADI method discussed above. The resulting low-rank Newton method is summarized in Algorithm 9. It should be mentioned that taking the advantage of the special structure of $F_j = F + (EX_jQ^T)(QP_r)$, the inverse of $E + \tau_kF_j$ required in the LR-ADI iteration can be written using the Sherman–Morrison–Woodbury formula [64, Sect. 2.1.3] as

$$(E + \tau_kF_j)^{-1} = F_{jk}^{-1} - F_{jk}^{-1}(EX_jQ^T)\big(I_q + QP_rF_{jk}^{-1}(EX_jQ^T)\big)^{-1}QP_rF_{jk}^{-1},$$

with $F_{jk} = E + \tau_kF$. Thus, instead of solving the linear system with large and possibly dense $E + \tau_kF_j$, we can solve two large linear systems with sparse $E + \tau_kF$ and, additionally, one small system.

Substituting $N_j = X_{j+1} - X_j$ in (4.5), Newton's method can be reformulated as the Newton–Kleinman iteration, where the new approximation X_{j+1} to the solution of (4.3) is determined by solving the projected Lyapunov equation

$$EX_{j+1}F_j^T + F_jX_{j+1}E^T = -P_l(RR^T - EX_jQ^TQX_jE)P_l^T, \qquad X_{j+1} = P_rX_{j+1}P_r^T.$$

The low-rank version of the Newton–Kleinman iteration as well as a comparison of both the Newton-type techniques can be found in [25].

5 Structured DAE Systems

The main difficulty in the model reduction methods for DAE systems involving the spectral projectors is the determination of these projectors themselves. This is often a numerically ill-conditioned problem since it requires the computation of the deflating subspaces corresponding to the finite eigenvalues of $\lambda E - A$. Fortunately, for some structured problems, the projectors P_l and P_r can be determined employing the block structures of E and A. Of course, we should avoid forming them explicitly as they are usually $n \times n$ dense matrices. Since the projectors often inherit the block structures of E and A, projector-vector products can be computed block-wise, where multiplication with sparse matrices and solving sparse linear systems is involved [137]. Furthermore, some structured DAE systems can be transformed into the ODE form such that the computation of the projectors can even be completely avoided.

5.1 Semi-Explicit Systems of Index 1

First, we consider the semi-explicit DAE system

$$\begin{bmatrix} E_{11} & E_{12} \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t), \tag{5.1}$$

$$y(t) = C_1 x_1(t) + C_2 x_2(t) + Du(t). \tag{5.2}$$

Such systems arise in computational fluid dynamics [152] and power systems modeling [53, 120]. In the latter case, we have additionally $E_{12} = 0$. If the matrices E_{11} and $A_{22} - A_{21}E_{11}^{-1}E_{12}$ are both nonsingular, then (5.1) is of index 1, and the spectral projectors are given by

$$P_l = \begin{bmatrix} I & -(A_{12} - A_{11}E_{11}^{-1}E_{12})(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1} \\ 0 & 0 \end{bmatrix},$$

$$P_r = \begin{bmatrix} I + E_{11}^{-1}E_{12}(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{21} & E_{11}^{-1}E_{12}(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{22} \\ -(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{21} & I - (A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{22} \end{bmatrix},$$

see [137]. Furthermore, (5.1) can be rewritten as the ODE system

$$\begin{aligned} \hat{E}\dot{x}(t) &= \hat{A}x(t) + \hat{B}u(t), \\ y(t) &= \hat{C}x(t) + \hat{D}u(t), \end{aligned} \tag{5.3}$$

where $x(t) = x_1(t) + E_{11}^{-1}E_{12}x_2(t)$, $\hat{E} = E_{11}$, and

$$\begin{aligned}
\hat{A} &= A_{11} - (A_{12} - A_{11}E_{11}^{-1}E_{12})(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{21},\\
\hat{B} &= B_1 - (A_{12} - A_{11}E_{11}^{-1}E_{12})(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}B_2,\\
\hat{C} &= C_1 - (C_2 - C_1E_{11}^{-1}E_{12})(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}A_{21},\\
\hat{D} &= D - (C_2 - C_1E_{11}^{-1}E_{12})(A_{22} - A_{21}E_{11}^{-1}E_{12})^{-1}B_2.
\end{aligned}$$

We can now apply any model reduction method to system (5.3) with nonsingular $\hat{E}$, where the spectral projectors are no longer needed. In the LR-ADI method and the Krylov-based model reduction methods, one has to solve the shifted linear systems of the form $(\hat{E} + \tau\hat{A})z = f$. Their solutions can be obtained as $z = z_1 + E_{11}^{-1}E_{12}z_2$, where z_1 and z_2 solve the sparse linear system

$$\begin{bmatrix} E_{11} + \tau A_{11} & E_{12} + \tau A_{12} \\ \tau A_{21} & \tau A_{22} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}.$$

Another condition guaranteeing the index-1 property for (5.1) is nonsingularity of the matrices A_{22} and $E_{11} - E_{12}A_{22}^{-1}A_{21}$. In this case, the second equation in (5.1) gives

$$x_2(t) = -A_{22}^{-1}A_{21}x_1(t) - A_{22}^{-1}B_2u(t).$$

Substituting it in the first equation in (5.1) and in the output Equation (5.2), we obtain the ODE system

$$\begin{aligned}
\hat{E}_1\dot{x}_1(t) &= \hat{A}_1x_1(t) + \hat{B}_1u_1(t),\\
y(t) &= \hat{C}_1x_1(t) + \hat{D}_1u_1(t),
\end{aligned} \tag{5.4}$$

where

$$\begin{aligned}
\hat{E}_1 &= E_{11} - E_{12}A_{22}^{-1}A_{21}, & \hat{A}_1 &= A_{11} - A_{12}A_{22}^{-1}A_{21},\\
\hat{B}_1 &= [\,B_1 - A_{12}A_{22}^{-1}B_2,\ E_{12}A_{22}^{-1}B_2\,], & \hat{C}_1 &= C_1 - C_2A_{22}^{-1}A_{21},\\
\hat{D}_1 &= [\,D - C_2A_{22}^{-1}B_2,\ 0\,], & u_1(t) &= [\,u^T(t),\ \dot{u}^T(t)\,]^T
\end{aligned}$$

provided u is continuously differentiable. It should be emphasized that the matrices $\hat{E}_1$ and $\hat{A}_1$ will never be computed explicitly since they may be dense even if all matrices E_{ij} and A_{ij} are sparse. The solution of $(\hat{E}_1 + \tau\hat{A}_1)z = f$ can be obtained by solving the sparse linear system

$$\begin{bmatrix} E_{11} + \tau A_{11} & E_{12} + \tau A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} z \\ g \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}.$$

Note that if $E_{12} = 0$, then both systems (5.3) and (5.4) take the form

$$\begin{aligned} E_{11}\dot{x}_1(t) &= (A_{11} - A_{12}A_{22}^{-1}A_{21})x_1(t) + (B_1 - A_{12}A_{22}^{-1}B_2)u(t), \\ y(t) &= (C_1 - C_2A_{22}^{-1}A_{21})x_1(t) + (D - C_2A_{22}^{-1}B_2)u(t). \end{aligned}$$

Model reduction of such a system has been considered in [53, 120].

5.2 Magneto-Quasistatic Systems of Index 1

Magneto-quasistatic field systems arise in modeling of electromagnetic devices such as induction machines and transformers by neglecting the displacement currents. A spatial discretization of Maxwell's equations in magnetic vector potential formulation together with the circuit coupling equations using the finite integration technique or the finite element method yields the DAE system

$$\begin{bmatrix} M_{11} & 0 & 0 \\ 0 & 0 & 0 \\ X_1^T & X_2^T & 0 \end{bmatrix} \begin{bmatrix} \dot{a}_1(t) \\ \dot{a}_2(t) \\ j(t) \end{bmatrix} = \begin{bmatrix} -K_{11} & -K_{12} & X_1 \\ -K_{21} & -K_{22} & X_2 \\ 0 & 0 & -R \end{bmatrix} \begin{bmatrix} a_1(t) \\ a_2(t) \\ j(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ I \end{bmatrix} u(t), \tag{5.5}$$
$$y(t) = j(t),$$

where $[a_1^T, a_2^T]^T \in \mathbb{R}^{n_1+n_2}$ is a semidiscretized magnetic vector potential and $j(t) \in \mathbb{R}^m$ is a current vector, e.g., [126, 127]. The matrices M_{11}, K_{22} and R are symmetric, positive definite and X_2 is of full column rank. In this case, system (5.5) has index 1 [82]. Let the columns of Y form an orthonormal basis of the kernel of X_2^T and the columns of $Z = X_2(X_2^TX_2)^{-1/2}$ span the image of X_2. Multiplying the first equation in (5.5) with an orthogonal matrix

$$Q = \begin{bmatrix} I_{n_1} & 0 & 0 \\ 0 & Z^T & 0 \\ 0 & Y^T & 0 \\ 0 & 0 & I_m \end{bmatrix}$$

and introducing a vector $Qx(t) = \left[a_1^T(t),\ a_{21}^T(t),\ a_{22}^T(t),\ j^T(t)\right]^T$ partitioned according to Q, we obtain the ODE system

$$\begin{aligned} \hat{E}\dot{x}(t) &= \hat{A}x(t) + \hat{B}u(t), \\ y(t) &= \hat{C}x(t), \end{aligned} \tag{5.6}$$

where

$$\begin{aligned}
\hat{E} &= \begin{bmatrix} M_{11} + X_1 R^{-1} X_1^T & X_1 R^{-1} X_2^T Z \\ Z^T X_2 R^{-1} X_1^T & Z^T X_2 R^{-1} X_2^T Z \end{bmatrix}, \quad x(t) = \begin{bmatrix} a_1(t) \\ a_{21}(t) \end{bmatrix}, \\
\hat{A} &= -\begin{bmatrix} K_{11} & K_{12} Z \\ Z^T K_{21} & Z^T K_{22} Z \end{bmatrix} + \begin{bmatrix} K_{12} \\ Z^T K_{22} \end{bmatrix} Y \left(Y^T K_{22} Y \right)^{-1} Y^T [K_{21},\ K_{22} Z], \\
\hat{B} &= \begin{bmatrix} X_1 \\ Z^T X_2 \end{bmatrix} R^{-1}, \\
\hat{C} &= -(X_2^T X_2)^{-1} X_2^T \left(I - K_{22} Y (Y^T K_{22} Y)^{-1} Y^T \right) [K_{21},\ K_{22} Z].
\end{aligned} \tag{5.7}$$

In order to be able to apply the balanced truncation model reduction method to system (5.6), we need to solve linear systems of the form

$$(\hat{E} + \tau \hat{A}) z = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

Exploiting the block structure of the matrices $\hat{E}$ and $\hat{A}$ in (5.7), the solution of this system can be determined as $z = [z_1^T, (Z^T z_2)^T]^T$, where z_1 and z_2 solve the sparse linear system

$$\begin{bmatrix} M_{11} - \tau K_{11} & -\tau K_{12} & \tau X_1 \\ -\tau K_{21} & -\tau K_{22} & \tau X_2 \\ X_1^T & X_2^T & -\tau R \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ Z f_2 \\ 0 \end{bmatrix}.$$

Furthermore, the ADI shift parameters can be calculated by an Arnoldi procedure applied to the matrices $\hat{E}^{-1}\hat{A}$ and $\hat{A}^{-1}\hat{E}$. Again, the matrix-vector products $\hat{E}^{-1}\hat{A}v$ and $\hat{A}^{-1}\hat{E}v$ required in the Arnoldi procedure can be computed without the construction of the matrices $\hat{E}$, $\hat{A}$ and their inverses. A main difficulty here is the computation of the vector $z = Y(Y^T K_{22} Y)^{-1} Y^T w$. Fortunately, this vector can be determined by solving the sparse linear system

$$\begin{bmatrix} K_{22} & X_2 \\ X_2^T & 0 \end{bmatrix} \begin{bmatrix} z \\ g \end{bmatrix} = \begin{bmatrix} w \\ 0 \end{bmatrix},$$

see [82] for details. This shows that the computation of the large dense matrix Y can completely be avoided which reduces the computational complexity significantly.

5.3 Circuit Equations of Index 1 and 2

Linear RLC circuits consisting of linear resistors, inductors, capacitors and independent current and voltage sources can be described using modified nodal analysis

[72, 111]. Choosing currents through inductors and voltages of voltage sources as inputs, as well as voltages of current sources and currents through voltage sources as outputs, one obtains a DAE system of the form (1.1) described by

$$E = \begin{bmatrix} A_C \mathcal{C} A_C^T & 0 & 0 \\ 0 & \mathcal{L} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} -A_{\mathcal{R}} \mathcal{R}^{-1} A_{\mathcal{R}}^T & -A_{\mathcal{L}} & -A_{\mathcal{V}} \\ A_{\mathcal{L}}^T & 0 & 0 \\ A_{\mathcal{V}}^T & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -A_I & 0 \\ 0 & 0 \\ 0 & -I \end{bmatrix} = C^T,$$
$$D = 0, \quad x(t) = \begin{bmatrix} \eta(t) \\ j_{\mathcal{L}}(t) \\ j_{\mathcal{V}}(t) \end{bmatrix}, \quad u(t) = \begin{bmatrix} j_I(t) \\ v_{\mathcal{V}}(t) \end{bmatrix}, \quad y(t) = -\begin{bmatrix} v_I(t) \\ j_{\mathcal{V}}(t) \end{bmatrix}. \tag{5.8}$$

Here $\eta \in \mathbb{R}^{n_\eta}$ is a vector of node potentials, $j_{\mathcal{L}} \in \mathbb{R}^{n_{\mathcal{L}}}$, $j_I \in \mathbb{R}^{n_I}$ and $j_{\mathcal{V}} \in \mathbb{R}^{n_{\mathcal{V}}}$ are vectors of currents through inductors, current and voltage sources, respectively, and v_I and $v_{\mathcal{V}}$ are vectors of voltages of current and voltage sources, respectively. Furthermore, $A_{\mathcal{C}}$, $A_{\mathcal{L}}$, $A_{\mathcal{R}}$, $A_{\mathcal{V}}$ and A_I are the incidence matrices describing the topological structure of the circuit, and $\mathcal{C}$, $\mathcal{R}$ and $\mathcal{L}$ are the capacitance, resistance and inductance matrices. Under the assumptions that $A_{\mathcal{V}}$ has full column rank, $[A_{\mathcal{C}}, A_{\mathcal{L}}, A_{\mathcal{R}}, A_{\mathcal{V}}]$ has full row rank and $\mathcal{C}$, $\mathcal{R}$ and $\mathcal{L}$ are positive definite, system (1.1), (5.8) is of index at most 2 and passive [49, 109]. It has index 1 if, additionally, $[A_{\mathcal{C}}, A_{\mathcal{L}}, A_{\mathcal{R}}]$ has full row rank and $Z_{\mathcal{C}}^T A_{\mathcal{V}}$ has full column rank, where the columns of $Z_{\mathcal{C}}$ span $\ker(A_{\mathcal{C}}^T)$.

In model reduction of circuit equations, it is crucial to preserve passivity. This allows a back interpretation of the reduced-order model as an electrical circuit which has fewer electrical components than the original one [7, 109]. Passivity-preserving Krylov subspace based model reduction methods for structured circuit equations have been developed in [54, 56, 57, 84, 99], whereas balancing-related methods have been considered in [16, 105, 114, 116, 155]. Unfortunately, the application of the positive real balanced truncation method is currently restricted to small and medium-sized problems, since there exists no explicit representation for the spectral projectors required in the positive real Lur'e Equations (3.6) and (3.7). In contrast, for the Moebius-transformed system $\boldsymbol{H}_M = (E, A - BC, -\sqrt{2}B, \sqrt{2}C, I)$, the right and left spectral projectors are given by

$$P_r = \begin{bmatrix} H_5(H_4H_2 - I) & H_5H_4A_{\mathcal{L}}H_7 & 0 \\ 0 & H_7 & 0 \\ -A_{\mathcal{V}}^T(H_4H_2 - I) & -A_{\mathcal{V}}^T H_4 A_{\mathcal{L}} H_7 & 0 \end{bmatrix},$$
$$P_l = \begin{bmatrix} (H_2H_4 - I)H_6 & 0 & (H_2H_4 - I)A_{\mathcal{V}} \\ -H_8 A_{\mathcal{L}}^T H_4 H_6 & H_8 & -H_8 A_{\mathcal{L}}^T H_4 A_{\mathcal{V}} \\ 0 & 0 & 0 \end{bmatrix},$$

where

$$
\begin{aligned}
&H_1 = Z_{C\mathcal{R}I\mathcal{V}}^T A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^T Z_{C\mathcal{R}I\mathcal{V}},\\
&H_2 = A_{\mathcal{R}} \mathcal{R}^{-1} A_{\mathcal{R}}^T + A_I A_I^T + A_{\mathcal{V}} A_{\mathcal{V}}^T + A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^T Z_{C\mathcal{R}I\mathcal{V}} H_1^{-1} Z_{C\mathcal{R}I\mathcal{V}}^T A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^T,\\
&H_3 = Z_C^T H_2 Z_C, \qquad\qquad H_4 = Z_C H_3^{-1} Z_C^T,\\
&H_5 = Z_{C\mathcal{R}I\mathcal{V}} H_1^{-1} Z_{C\mathcal{R}I\mathcal{V}}^T A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^T - I, \quad H_6 = A_{\mathcal{L}} \mathcal{L}^{-1} A_{\mathcal{L}}^T Z_{C\mathcal{R}I\mathcal{V}} H_1^{-1} Z_{C\mathcal{R}I\mathcal{V}}^T - I,\\
&H_7 = I - \mathcal{L}^{-1} A_{\mathcal{L}}^T Z_{C\mathcal{R}I\mathcal{V}} H_1^{-1} Z_{C\mathcal{R}I\mathcal{V}}^T A_{\mathcal{L}}, \quad H_8 = I - A_{\mathcal{L}}^T Z_{C\mathcal{R}I\mathcal{V}} H_1^{-1} Z_{C\mathcal{R}I\mathcal{V}}^T A_{\mathcal{L}} \mathcal{L}^{-1},\\
&Z_C \text{ is a basis matrix for } \ker(A_C^T),\\
&Z_{C\mathcal{R}I\mathcal{V}} \text{ is a basis matrix for } \ker([A_C, A_{\mathcal{R}}, A_I, A_{\mathcal{V}}]^T),
\end{aligned}
$$

see [114, 139]. This allows us to compute the passive reduced-order model by applying the bounded real balanced truncation to $\boldsymbol{H}_M$ in the large-scale setting. Taking into account the block structure of the system matrices in (5.8), we can also determine the matrix $M_0 = \lim_{s\to\infty} \boldsymbol{H}_M(s)$ in the form

$$
M_0 = \begin{bmatrix} I - 2A_I^T Z H_0^{-1} Z^T A_I & 2A_I^T Z H_0^{-1} Z^T A_{\mathcal{V}} \\ -2A_{\mathcal{V}}^T Z H_0^{-1} Z^T A_I & -I + 2A_{\mathcal{V}}^T Z H_0^{-1} Z^T A_{\mathcal{V}} \end{bmatrix},
$$

where $H_0 = Z^T(A_{\mathcal{R}} \mathcal{R}^{-1} A_{\mathcal{R}}^T + A_I A_I^T + A_{\mathcal{V}} A_{\mathcal{V}}^T) Z$, $Z = Z_C Z'_{\mathcal{R}I\mathcal{V}-C}$ and $Z'_{\mathcal{R}I\mathcal{V}-C}$ is a basis matrix for $\operatorname{im}([A_{\mathcal{R}}, A_I, A_{\mathcal{V}}]^T Z_C)$. Having this matrix, we no longer need to compute the improper Gramians. Furthermore, if $\mathcal{C}$, $\mathcal{R}$ and $\mathcal{L}$ are symmetric, then $P_l = P_r^T$ and the bounded real Gramians G_c^{BR} and G_o^{BR} are related by

$$
G_c^{BR} = S_{\text{int}} G_o^{BR} S_{\text{int}}
$$

with a signature matrix $S_{\text{int}} = \operatorname{diag}(I_{n_\eta}, -I_{n_{\mathcal{L}}}, -I_{n_{\mathcal{V}}})$. In this case, only one Lur'e equation has to be solved which reduces the computational cost.

A further cost reduction can be achieved for RC and RL circuits. The underlying equations for such circuits are either symmetric or they can be transformed to symmetric systems for which passivity-preserving model reduction can be performed employing the Lyapunov balancing [116].

5.4 Stokes-Like Systems of Index 2

Another block structured DAE system arises in computational fluid dynamics, where the flow of an incompressible fluid is modeled by the Navier–Stokes equation. After a linearization along a stationary trajectory and discretization in space by the

finite element method, one gets the Stokes-like system

$$\begin{bmatrix} E_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{v}(t) \\ \dot{p}(t) \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix} \begin{bmatrix} v(t) \\ p(t) \end{bmatrix} + \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t), \tag{5.9}$$
$$y(t) = C_1 v(t) + C_2 p(t) + Du(t),$$

where $v(t)$ and $p(t)$ are the semidiscretized velocity and pressure vectors. Model reduction of such systems has been considered in [35, 69, 71, 135]. Note that unlike [71], we do not assume here that E_{11} is symmetric and $A_{21} = A_{12}^T$. If E_{11} and $A_{21}E_{11}^{-1}A_{12}$ are both nonsingular, then system (5.9) is of index 2, and the spectral projectors P_l and P_r have the form

$$P_l = \begin{bmatrix} \Pi_l & -\Pi_l A_{11}E_{11}^{-1}A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1} \\ 0 & 0 \end{bmatrix},$$
$$P_r = \begin{bmatrix} \Pi_r & 0 \\ -(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}A_{11}\Pi_r & 0 \end{bmatrix},$$

where

$$\Pi_l = I - A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1},$$
$$\Pi_r = I - E_{11}^{-1}A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21} = E_{11}^{-1}\Pi_l E_{11}.$$

Note that the conditions for A_{12} and A_{21} to be of full rank are, in general, not enough for the index-2 property. It has been shown in [71] that the velocity and pressure vectors can be determined as

$$v(t) = v_0(t) - E_{11}^{-1}A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}B_2u(t),$$
$$p(t) = -(A_{21}E_{11}^{-1}A_{12})^{-1}\big(A_{21}E_{11}^{-1}A_{11}v_0(t) + A_{21}E_{11}^{-1}B_{12}u(t) + B_2\dot{u}(t)\big),$$

where $B_{12} = B_1 - A_{11}E_{11}^{-1}A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}B_2$ and $v_0(t) = \Pi_r v_0(t)$ solves the DAE system

$$\begin{aligned} \hat{E}\dot{v}_0(t) &= \hat{A}v_0(t) + \hat{B}u(t), \\ y(t) &= \hat{C}v_0(t) + \hat{D}u(t) + \hat{D}_1\dot{u}(t), \end{aligned} \tag{5.10}$$

with

$$\hat{E} = \Pi_l E_{11}\Pi_r, \qquad \hat{A} = \Pi_l A_{11}\Pi_r, \qquad \hat{B} = \Pi_l B_{12}, \tag{5.11}$$
$$\hat{C} = C_1 - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}A_{11},$$
$$\hat{D} = D - C_1E_{11}^{-1}A_{12}(A_{21}E_{11}^{-1}A_{12})^{-1}B_2 - C_2(A_{21}E_{11}^{-1}A_{12})^{-1}A_{21}E_{11}^{-1}B_{12},$$
$$\hat{D}_1 = -C_2(A_{21}E_{11}^{-1}A_{12})^{-1}B_2.$$

Note that the matrices $\hat{E}$ and $\hat{A}$ in (5.11) have a common nontrivial kernel, and, hence, $\lambda\hat{E} - \hat{A}$ is singular for all $\lambda \in \mathbb{C}$. At first glance, this renders the application of balanced truncation and interpolatory-based model reduction methods to (5.10) impossible since there the inversion of $\hat{E} + \tau_k\hat{A}$ (or $\hat{A} - s_k\hat{E}$) is required. Fortunately, these matrices can be inverted on a subspace. Then the LR-ADI iteration for the projected Lyapunov equation

$$\hat{A}X\hat{E}^T + \hat{E}X\hat{A}^T = -\hat{B}\hat{B}^T$$

associated with (5.10) can be reformulated as

$$\begin{aligned} &\hat{W}_0 = B_{12}, \qquad Z_0 = [\ ],\\ &\hat{V}_k = (\hat{E} + \tau_k\hat{A})^-\hat{W}_{k-1},\\ &\hat{W}_k = \hat{W}_{k-1} - 2\mathrm{Re}(\tau_k)A_{11}\hat{V}_k,\\ &\hat{Z}_k = [\,\hat{Z}_{k-1},\ \sqrt{-2\mathrm{Re}(\tau_k)}\,\hat{V}_k\,], \end{aligned} \tag{5.12}$$

where $(\hat{E} + \tau_k\hat{A})^-$ is the reflexive inverse of $\hat{E} + \tau_k\hat{A}$ with respect to Π_l and Π_r. Taking into account the structure of $\hat{E}$ and $\hat{A}$, the matrices $\hat{V}_k = (\hat{E} + \tau_k\hat{A})^-\hat{W}_{k-1}$ can be computed by solving the linear matrix equation

$$\begin{bmatrix} E_{11} + \tau_k A_{11} & A_{12} \\ A_{21} & 0 \end{bmatrix} \begin{bmatrix} \hat{V}_k \\ V \end{bmatrix} = \begin{bmatrix} \hat{W}_{k-1} \\ 0 \end{bmatrix}$$

with sparse (if E_{11} and A_{ij} are sparse) coefficient matrix. The main advantage of the LR-ADI iteration (5.12) over those in Algorithm 7 is that the matrices $\hat{V}_k$, $\hat{W}_k$ and $\hat{Z}_k$ have smaller dimension than V_k, W_k and Z_k, respectively, and no multiplication with the projectors is required. For further details of this novel formulation of the ADI iteration and its specific implementation for Stokes-like equations, see [35], where also an extension of balanced truncation to unstable descriptor systems is considered. Further note that LQG balanced truncation for (Navier–)Stokes flow is discussed in [21].

5.5 *Mechanical Systems of Index 1 and 3*

Consider a second-order DAE system

$$\begin{aligned} \begin{bmatrix} M_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{p}(t) \\ \ddot{\eta}(t) \end{bmatrix} + \begin{bmatrix} \mathcal{D}_{11} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{p}(t) \\ \dot{\eta}(t) \end{bmatrix} + \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} p(t) \\ \eta(t) \end{bmatrix} &= \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} u(t),\\ C_1 p(t) + C_2\eta(t) &= y(t), \end{aligned} \tag{5.13}$$

where $p(t)$ is a displacement vector and $\eta(t)$ is a vector of electrical potentials. Such systems frequently arise in mechatronics, where micro-electromechanical devices are of great interest, e.g., [144]. Introducing $x(t) = [p^T(t), \dot{p}^T(t), \eta^T(t)]^T$, system (5.13) can be written as the first-order DAE system (1.1) with

$$E = \begin{bmatrix} I & 0 & 0 \\ 0 & M_{11} & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & I & 0 \\ -K_{11} & -\mathcal{D}_{11} & -K_{12} \\ -K_{21} & 0 & -K_{22} \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ B_1 \\ B_2 \end{bmatrix}, \quad C = [C_1, 0, C_2], \quad D = 0.$$

If M_{11} and K_{22} are both nonsingular, then this system (and also (5.13)) is of index 1. Similarly to the semi-explicit DAE system (5.1), (5.2), system (5.13) can be rewritten in the compact form

$$\begin{aligned} M_{11}\ddot{p}(t) + \mathcal{D}_{11}\dot{p}(t) + \hat{K}_{11}p(t) &= \hat{B}u(t), \\ y(t) &= \hat{C}p(t) + \hat{D}u(t), \end{aligned}$$

where $\hat{K}_{11} = K_{11} - K_{12}K_{22}^{-1}K_{21}$, $\hat{B} = B_1 - K_{12}K_{22}^{-1}B_2$, $\hat{C} = C_1 - C_2K_{22}^{-1}K_{21}$ and $\hat{D} = C_2K_{22}^{-1}B_2$. Applying the second-order balanced truncation method as proposed in [22, 31] or the second-order Krylov subspace methods [9, 125] requires the solution of the linear systems $(\tau^2 M_{11} \pm \tau\mathcal{D}_{11} + \hat{K}_{11})z = f$. Employing the structure of the involved matrices, the vector z can be determined by solving the sparse system

$$\begin{bmatrix} \tau^2 M_{11} \pm \tau\mathcal{D}_{11} + K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \begin{bmatrix} z \\ g \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

using a sparse LU factorization or Krylov subspace methods [123].

The dynamical behavior of linear multibody systems with holonomic constraints is described by the Euler–Lagrange equations

$$\begin{aligned} \begin{bmatrix} I & 0 & 0 \\ 0 & M & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{p}(t) \\ \dot{v}(t) \\ \dot{\lambda}_p(t) \end{bmatrix} &= \begin{bmatrix} 0 & I & 0 \\ -K & -\mathcal{D} & -G^T \\ G & 0 & 0 \end{bmatrix} \begin{bmatrix} p(t) \\ v(t) \\ \lambda_p(t) \end{bmatrix} + \begin{bmatrix} 0 \\ B \\ 0 \end{bmatrix} u(t), \\ y(t) &= C_p p(t) + C_v v(t), \end{aligned} \tag{5.14}$$

where $p(t)$ and $v(t)$ are the position and velocity vectors, $\lambda_p(t)$ is the Lagrange multiplier, M, $\mathcal{D}$ and K are the mass, stiffness and damping matrices, respectively, and G is a matrix of constraints. If M and $GM^{-1}G^T$ are both nonsingular, then system (5.14) is of index 3. Exploiting the block structure of the system matrices,

the spectral projectors P_l and P_r can be computed as

$$P_l = \begin{bmatrix} \Pi_r & 0 & \Pi_r M^{-1}\mathcal{D}G_1 \\ \Pi_r^T\mathcal{D}(I-\Pi_r) & \Pi_r^T & \Pi_r^T(K-\mathcal{D}\Pi_r M^{-1}\mathcal{D})G_1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$P_r = \begin{bmatrix} \Pi_r & 0 & 0 \\ \Pi_r M^{-1}\mathcal{D}(I-\Pi_r) & \Pi_r & 0 \\ -G_1^T(K\Pi_r + \mathcal{D}\Pi_r M^{-1}\mathcal{D}(I-\Pi_r)) & -G_1^T\mathcal{D}\Pi_r & 0 \end{bmatrix},$$

where $G_1 = M^{-1}G^T(GM^{-1}G^T)^{-1}$ and

$$\Pi_r = I - M^{-1}G^T(GM^{-1}G^T)^{-1}G = I - G_1G$$

is a projector onto the constraint manifold $\ker(G)$. Instead of using the spectral projectors P_l and P_r explicitly, one can reformulate the DAE system (5.14) in such a way that only the implicit projection is needed. This can be achieved by the Gear–Gupta–Leimkuhler formulation [60] given by

$$\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & M & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \dot p(t) \\ \dot v(t) \\ \dot\lambda_p(t) \\ \dot\lambda_v(t) \end{bmatrix} = \begin{bmatrix} 0 & I & 0 & -G^T \\ -K & -\mathcal{D} & -G^T & 0 \\ G & 0 & 0 & 0 \\ 0 & G & 0 & 0 \end{bmatrix}\begin{bmatrix} p(t) \\ v(t) \\ \lambda_p(t) \\ \lambda_v(t) \end{bmatrix} + \begin{bmatrix} 0 \\ B \\ 0 \\ 0 \end{bmatrix}u(t), \quad (5.15)$$

$$y(t) = C_p p(t) + C_v v(t)$$

which has index 2. In computational multibody dynamics, (5.15) is also known as stabilized index-2 formulation of the equations of motion, e.g., [39]. It can be obtained by differentiating the position-level constraint $Gp(t) = 0$ and adding the resulting velocity-level constraint equation $Gv(t) = 0$ to (5.14) by introducing an additional Lagrange multiplier λ_v. It was shown in [60] that if (p, v, λ_p) is a solution of (5.14), then $(p, v, \lambda_p, \lambda_v)$ with $\lambda_v = 0$ is a solution of (5.15). Conversely, if $(p, v, \lambda_p, \lambda_v)$ solves (5.15), then $\lambda_v = 0$ and (p, v, λ_p) satisfies (5.14). Observe that system (5.15) has the Stokes-like form (5.9) with

$$E_{11} = \begin{bmatrix} I & 0 \\ 0 & M \end{bmatrix},\ A_{11} = \begin{bmatrix} 0 & I \\ -K & -\mathcal{D} \end{bmatrix},\ A_{12} = \begin{bmatrix} 0 & -G^T \\ -G^T & 0 \end{bmatrix},\ A_{21} = \begin{bmatrix} G & 0 \\ 0 & G \end{bmatrix},$$
$$B_1 = \begin{bmatrix} 0 \\ B \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad C_1 = \begin{bmatrix} C_p, & C_v \end{bmatrix}, \quad C_2 = 0, \quad D = 0. \quad (5.16)$$

Therefore, all results of Sect. 5.4 can be applied to the constrained mechanical system (5.15). Exploiting the block structure of the matrices in (5.16), we obtain the second-order system

$$\begin{aligned} \hat{M}\ddot{p}(t) + \hat{\mathcal{D}}\dot{p}(t) + \hat{K}p(t) &= \hat{B}u(t), \\ y(t) &= \hat{C}_p p(t) + \hat{C}_v \dot{p}(t) \end{aligned} \tag{5.17}$$

for the position vector $p(t) = \Pi p(t)$, where

$$\begin{aligned} \hat{M} &= \Pi_l M \Pi, & \hat{\mathcal{D}} &= \Pi_l \mathcal{D} \Pi, & \hat{K} &= \Pi_l K \Pi, \\ \hat{B} &= \Pi_l B, & \hat{C}_p &= C_p \Pi, & \hat{C}_v &= C_v \Pi, \\ \Pi_l &= M \Pi_r M^{-1}, & \Pi &= I - G^T (G G^T)^{-1} G. \end{aligned}$$

Combining the balanced truncation technique from [71] with the second-order LR-ADI method presented in [22, 31], we can derive an efficient computational procedure for model reduction of system (5.17) which does not require forming the first-order system. This procedure involves solving projected linear systems

$$\Pi_l(\tau^2 M - \tau \mathcal{D} + K)\Pi z = \Pi_l f$$

whose solution $z = \Pi z$ can be determined from the saddle point linear system

$$\begin{bmatrix} \tau^2 M - \tau \mathcal{D} + K & G^T \\ G & 0 \end{bmatrix} \begin{bmatrix} z \\ g \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix}$$

without computing the projectors Π_l and Π.

6 Other Model Reduction Topics

In this section, we briefly discuss other works related to model reduction of DAE systems. This list is far from complete and rather provides a very short overview of recent developments in this active research area.

6.1 Model Reduction of Periodic Discrete-Time Descriptor Systems

The balanced truncation model reduction method can also be formulated for discrete-time DAEs. In this case, instead of projected continuous-time Lyapunov

Equations (3.1) and (3.2), one has to solve the projected discrete-time Lyapunov equations

$$AXA^T - EXE^T = -P_l BB^T P_l^T, \qquad X = P_r X P_r^T, \tag{6.1}$$

$$A^T YA - E^T YE = -P_r^T C^T C P_r, \qquad Y = P_l^T Y P_l, \tag{6.2}$$

introduced in [133].

Model reduction of periodic discrete-time descriptor systems

$$\begin{aligned} E_k x_{k+1} &= A_k x_k + B_k u_k, \\ y_k &= C_k x_k, \end{aligned} \tag{6.3}$$

where $E_k \in \mathbb{R}^{\mu_{k+1}\times n_{k+1}}$, $A_k \in \mathbb{R}^{\mu_{k+1}\times n_k}$, $B_k \in \mathbb{R}^{\mu_{k+1}\times m_k}$, $C_k \in \mathbb{R}^{q_k \times n_k}$ are periodic with a period $K \geq 1$, $\sum_{k=0}^{K-1} \mu_k = \sum_{k=0}^{K-1} n_k = n$, $\sum_{k=0}^{K-1} m_k = m$ and $\sum_{k=0}^{K-1} q_k = q$, has been considered in [32, 40]. The Gramians for such systems can be determined as solutions of periodic projected Lyapunov equations. Using a lifted representation [132] for the periodic descriptor system (6.3), these equations can be written in the form (3.3), (3.4) and (6.1), (6.2) with block structured matrices $E, A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$ and $C \in \mathbb{R}^{q\times n}$. The efficient solution of these lifted systems using methods from Sect. 4.1 adapted to exploit the block sparsity in the lifted system matrices is considered in [29, 32, 73].

6.2 Index-Aware Model Reduction for DAEs

In [2, 3], an index-aware model reduction approach was proposed for DAE systems which is based on splitting the DAE into an ODE system and a system of algebraic equations. It was shown in [2] that the index-1 DAE system (1.1) can be written in the form

$$\begin{aligned} \dot{x}_1(t) &= A_{11}x_1(t) + B_1 u(t), & y_1(t) &= C_1 x_1(t), \\ x_2(t) &= A_{21}x_1(t) + B_2 u(t), & y(t) &= y_1(t) + C_2 x_2(t) + Du(t), \end{aligned} \tag{6.4}$$

where

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} x(t), \quad \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} = \begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} E_1^{-1} A T_1, \quad \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} W_1^T \\ W_2^T \end{bmatrix} E_1^{-1} B,$$

$$E_1 = E - AT_2 W_2^T, \qquad [\,C_1,\ C_2\,] = C[\,T_1,\ T_2\,], \qquad [\,W_1,\ W_2\,]^T = [\,T_1,\ T_2\,]^{-1},$$

and the columns of the matrices T_1 and T_2 form the basis of $\operatorname{im}(E^T)$ and $\ker(E)$, respectively. Then the ODE system (6.4) is approximated by a reduced-order model

$$\dot{\tilde{x}}_1(t) = \tilde{A}_{11}\tilde{x}_1(t) + \tilde{B}_1 u(t), \qquad \tilde{y}_1(t) = \tilde{C}_1 \tilde{x}_1(t)$$

with $\tilde{A}_{11} = V^T A_{11} V$, $\tilde{B}_1 = V^T B_1$ and $\tilde{C}_1 = C_1 V$ using any projection-based model reduction method, and $\tilde{y}(t) = \tilde{y}_1(t) + C_2 A_{21} V \tilde{x}_1(t) + (C_2 B_2 + D)u(t)$ approximates the output $y(t)$. The transformation matrix $[T_1, \ T_2]$ can be determined from the sparse LUQ factorization [85] of E^T as a product of a permutation matrix and a sparse lower triangular matrix, and its inverse is computed by forward substitution, see [157] for a detailed discussion. The index-aware model reduction approach was also extended in [3] to DAEs of index 2. It should be noted that this approach does not require any special structure of the matrices E and A, but its efficiency strongly relies on sparsity of the matrix $A_{11} = W_1^T E_1^{-1} A T_1$. Even if $[T_1, \ T_2]$ is sparse, the multiplication with E_1^{-1} may result in a full matrix that makes this approach unfeasible for large-scale problems.

6.3 Parametric Model Reduction

In recent years, model reduction of parameterized systems has received a lot of attention, see [34] for an overview and numerous references. Here, we are only going to provide a brief sketch of some approaches, noting that a lot remains to be done to adapt some of them to descriptor systems, and to exploit special structures as in Sect. 5.

Consider a linear parametric DAE system

$$\begin{aligned} E(p)\dot{x}(t,p) &= A(p)x(t,p) + B(p)u(t), \\ y(t,p) &= C(p)x(t,p), \end{aligned} \tag{6.5}$$

where the system matrices and, hence, the state and the output depend on a parameter $p \in \mathbb{P} \subset \mathbb{R}^d$. Such systems appear frequently in control design and optimization problems, where parameters describe varying geometric configurations and material characteristics. When approximating the parametric system, it is important to preserve the parameter dependence in the reduced-order model. For parametric model reduction, different techniques have been developed over the years, that are, in some sense, extensions of traditional non-parametric model reduction approaches, see [34] for a survey of state-of-the-art parametric model reduction methods. In Krylov subspace based methods [11, 19, 42, 50, 90], the transfer function

$$\mathbf{H}(s,p) = C(p)(sE(p) - A(p))^{-1}B(p)$$

of (6.5) is approximated by

$$\tilde{\mathbf{H}}(s,p) = \tilde{C}(p)(s\tilde{E}(p) - \tilde{A}(p))^{-1}\tilde{B}(p)$$

of lower dimension that satisfies (tangential) interpolation conditions with respect to s and p. Another class of the parametric model reduction methods is based

on interpolation. For selected parameters $p_1, \ldots, p_k \in \mathbb{P}$, one computes first the reduced-order local models

$$\begin{aligned} \tilde{E}_j \dot{\tilde{x}}_j(t) &= \tilde{A}_j \tilde{x}_j(t) + \tilde{B}_j u(t), \\ \tilde{y}_j(t) &= \tilde{C}_j \tilde{x}_j(t), \end{aligned}$$

where $\tilde{E}_j = W_j^T E(p_j) T_j$, $\tilde{A}_j = W_j^T A(p_j) T_j$, $\tilde{B}_j = W_j^T B(p_j)$ and $\tilde{C}_j = C(p_j) T_j$. Then a parameter-dependent reduced-order model is constructed by using one of the following interpolation approaches:

1. *interpolation in the frequency domain* [10, 51, 129], where the reduced transfer function is obtained by interpolation of the reduced local transfer functions

$$\tilde{\boldsymbol{H}}(s, p) = \sum_{j=1}^{k} f_j(p) \tilde{C}_j (s\tilde{E}_j - \tilde{A}_j)^{-1} \tilde{B}_j;$$

2. *interpolation in the time domain* [5, 6, 43, 62, 101], where the reduced-order model is derived by interpolation of the reduced system matrices

$$\begin{aligned} \tilde{E}(p) &= \sum_{j=1}^{k} f_j(p) \tilde{E}_j, & \tilde{A}(p) &= \sum_{j=1}^{k} f_j(p) \tilde{A}_j, \\ \tilde{B}(p) &= \sum_{j=1}^{k} f_j(p) \tilde{B}_j, & \tilde{C}(p) &= \sum_{j=1}^{k} f_j(p) \tilde{C}_j; \end{aligned}$$

3. *interpolation of the projection subspaces* [4, 130], where the reduced-order model is determined by projection

$$\begin{aligned} \tilde{E}(p) &= W^T(p) E(p) T(p), & \tilde{A}(p) &= W^T(p) A(p) T(p), \\ \tilde{B}(p) &= W^T(p) B(p), & \tilde{C}(p) &= C(p) T(p), \end{aligned}$$

   and the projection matrices $W(p)$ and $T(p)$ are obtained by interpolation of $W_1, \ldots, W_k$ and $T_1, \ldots, T_k$, respectively, on the Grassmann manifolds.

For an extension of these methods to descriptor systems and a comparative analysis of them, with particular focus on their application to circuit equations, we refer to [131].

Model reduction of nonlinear parametric DAEs arising in circuit simulation using a reduced bases method was considered in [44].

7 Conclusions

We have surveyed model order reduction methods for linear descriptor systems, i.e., systems with input–output structure and dynamics described by systems of differential-algebraic equations. We have seen that most methods based on system-theoretic approaches such as balanced truncation and the related family of balancing-based methods as well as methods based on rational interpolation of the associated transfer function can be adapted to descriptor systems by using appropriate spectral projectors. As an extension of the available literature, we have extended the method of balanced stochastic truncation to descriptor systems. The presented approaches rely on the availability of the spectral projectors. Often, in applications, these can be formed explicitly without additional computation by a smart usage of the structure arising from the different applications. Moreover, the explicit formation of the spectral projectors can usually be avoided using clever implementations of the algorithms needed, e.g., to compute the factors of the system Gramians used in balancing-based methods. These Gramians are solutions of projected algebraic Lyapunov or Riccati equations. We have shown recent advances in the numerical methods to solve these projected matrix equations. Details of the projector-avoiding strategies have been discussed for various engineering problems leading to descriptor systems of index 1, 2, or 3, resulting in specialized implementations of the model order reduction methods.

Future work in this area will address extensions of the methods discussed to nonlinear systems. Such extensions of the system-theoretic methods for nonlinear systems described by ordinary differential equations have been surveyed recently in [12]. First attempts focusing on bilinear descriptor systems as discussed in [20] show that in particular the interpolatory approaches carry over directly when the underlying structure is carefully exploited. The extension of these results to more general classes of nonlinear descriptor systems will require further research efforts in the future.

Acknowledgements The first author acknowledges support by the collaborative project nanoCOPS: "Nanoelectronic COupled Problems Solutions" funded by the European Union in the FP7-ICT-2013-11 Program under Grant Agreement Number 619166.

The second author was supported by the Research Network KoSMos: *Model reduction based simulation of coupled PDAE systems* funded by the German Federal Ministry of Education and Science (BMBF), grant 05M13WAA, and by the project *Model reduction for elastic multibody systems with moving interactions* funded by the German Research Foundation (DFG), grant STY 58/1–2.

The authors gratefully acknowledge the careful proofreading by a reviewer and the editors which greatly enhanced the presentation of this survey.

References

1. Ahmad, M.I., Benner, P., Goyal, P.: Krylov subspace-based model reduction for a class of bilinear descriptor systems. J. Comput. Appl. Math. **315**, 303–318 (2017)
2. Alì, G., Banagaaya, N., Schilders, W., Tischendorf, C.: Index-aware model order reduction for linear index-2 DAEs with constant coefficients. SIAM J. Sci. Comput. **35**, A1487–A1510 (2013)
3. Alì, G., Banagaaya, N., Schilders, W., Tischendorf, C.: Index-aware model order reduction for differential-algebraic equations. Math. Comput. Model. Dyn. Syst. **20**(4), 345–373 (2014)
4. Amsallem, D., Farhat, C.: Interpolation method for adapting reduced-order models and application to aeroelasticity. AIAAJ **46**(7), 1803–1813 (2008)
5. Amsallem, D., Farhat, C.: An online method for interpolating linear reduced-order models. SIAM J. Sci. Comput. **33**(5), 2169–2198 (2011)
6. Amsallem, D., Cortial, J., Carlberg, K., Farhat, C.: A method for interpolating on manifolds structural dynamics reduced-order models. Int. J. Numer. Methods Eng. **80**(9), 1241–1258 (2009)
7. Anderson, B., Vongpanitlerd, S.: Network Analysis and Synthesis. Prentice Hall, Englewood Cliffs, NJ (1973)
8. Antoulas, A.: Approximation of Large-Scale Dynamical Systems. SIAM, Philadelphia, PA (2005)
9. Bai, Z., Su, Y.: Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method. SIAM J. Sci. Comput. **26**, 1692–1709 (2005)
10. Baur, U., Benner, P.: Modellreduktion für parametrisierte Systeme durch balanciertes Abschneiden und Interpolation. at-Automatisierungstechnik **57**(8), 411–419 (2009)
11. Baur, U., Beattie, C., Benner, P., Gugercin, S.: Interpolatory projection methods for parameterized model reduction. SIAM J. Sci. Comput. **33**(5), 2489–2518 (2011)
12. Baur, U., Benner, P., Feng, L.: Model order reduction for linear and nonlinear systems: a system-theoretic perspective. Arch. Comput. Meth. Eng. **21**(4), 331–358 (2014)
13. Benner, P.: Solving large-scale control problems. IEEE Contr. Syst. Mag. **24**(1), 44–59 (2004)
14. Benner, P.: Numerical linear algebra for model reduction in control and simulation. GAMM Mitteilungen **29**(2), 275–296 (2006)
15. Benner, P.: System-theoretic methods for model reduction of large-scale systems: simulation, control, and inverse problems. In: Troch, I., Breitenecker, F. (eds.) Proceedings of MathMod 2009 (Vienna, 11–13 February 2009), ARGESIM-Reports, vol. 35, pp. 126–145. Argesim, Wien (2009)
16. Benner, P.: Advances in balancing-related model reduction for circuit simulation. In: Roos, J., Costa, L. (eds.) Scientific Computing in Electrical Engineering SCEE 2008. Mathematics in Industry, vol. 14, pp. 469–482. Springer, Berlin, Heidelberg (2010)
17. Benner, P.: Partial stabilization of descriptor systems using spectral projectors. In: Van Dooren, P., Bhattacharyya, S.P., Chan, R.H., Olshevsky, V., Routray, A. (eds.) Numerical Linear Algebra in Signals, Systems and Control. Lecture Notes in Electrical Engineering, vol. 80, pp. 55–76. Springer, Netherlands (2011)
18. Benner, P., Quintana-Ortí, E.: Solving stable generalized Lyapunov equations with the matrix sign function. Numer. Algorim. **20**(1), 75–100 (1999)
19. Benner, P., Feng, L.: A robust algorithm for parametric model order reduction based on implicit moment matching. In: Quarteroni, A., Rozza, R. (eds.) Reduced Order Methods for Modeling and Computational Reduction, vol. 9, pp. 159–186. Springer, Berlin, Heidelberg (2014)
20. Benner, P., Goyal, P.: Multipoint interpolation of Volterra series and H_2-model reduction for a family of bilinear descriptor systems. Systems Control Lett. **96**, 1–11 (2016)
21. Benner, P., Heiland, J.: LQG-balanced truncation low-order controller for stabilization of laminar flows. In: King, R. (ed.) Active Flow and Combustion Control 2014. Notes on

Numerical Fluid Mechanics and Multidisciplinary Design, vol. 127, pp. 365–379. Springer International Publishing, Cham (2015)
22. Benner, P., Saak, J.: Efficient balancing based MOR for large scale second order systems. Math. Comput. Model. Dyn. Syst. **17**(2), 123–143 (2011)
23. Benner, P., Saak, J.: Numerical solution of large and sparse continuous time algebraic matrix Riccati and Lyapunov equations: a state of the art survey. GAMM Mitteilungen **36**(1), 32–52 (2013)
24. Benner, P., Sokolov, V.: Partial realization of descriptor systems. Syst. Control Lett. **55**(11), 929–938 (2006)
25. Benner, P., Stykel, T.: Numerical solution of projected algebraic Riccati equations. SIAM J. Numer. Anal. **52**(2), 581–600 (2014)
26. Benner, P., Quintana-Ortí, E., Quintana-Ortí, G.: Efficient numerical algorithms for balanced stochastic truncation. Int. J. Appl. Math. Comput. Sci. **11**(5), 1123–1150 (2001)
27. Benner, P., Mehrmann, V., Sorensen, D. (eds.): Dimension Reduction of Large-Scale Systems. Lecture Notes in Computational Science and Engineering, vol. 45. Springer, Berlin, Heidelberg (2005)
28. Benner, P., Hinze, M., ter Maten, E.J.W. (eds.): Model Reduction for Circuit Simulation. Lecture Notes in Electrical Engineering, vol. 74 Springer, Dodrecht (2011)
29. Benner, P., Hossain, M.S., Stykel, T.: Model reduction of periodic descriptor systems using balanced truncation. In: Benner, P., Hinze, M., ter Maten, E.J.W. (eds.) Model Reduction for Circuit Simulation. Lecture Notes in Electrical Engineering, vol. 74, pp. 187–200. Springer, Dodrecht (2011)
30. Benner, P., Kürschner, P., Saak, J.: Efficient handling of complex shift parameters in the low-rank Cholesky factor ADI method. Numer. Algoritm. **62**(2), 225–251 (2013)
31. Benner, P., Kürschner, P., Saak, J.: An improved numerical method for balanced truncation for symmetric second-order systems. Math. Comput. Model. Dyn. Syst. **19**(6), 593–615 (2013)
32. Benner, P., Hossain, M.S., Stykel, T.: Low-rank iterative methods for periodic projected Lyapunov equations and their application in model reduction of periodic descriptor systems. Numer. Algoritm. **67**(3), 669–690 (2014)
33. Benner, P., Kürschner, P., Saak, J.: Self-generating and efficient shift parameters in ADI methods for large Lyapunov and Sylvester equations. Electron. Trans. Numer. Anal. **43**, 142–162 (2014)
34. Benner, P., Gugercin, S., Willcox, K.: A survey of projection-based model reduction methods for parametric systems. SIAM Rev. **57**(4), 483–531 (2015)
35. Benner, P., Saak, J., Uddin, M.M.: Balancing based model reduction for structured index-2 unstable descriptor systems with application to flow control. Numer. Alg. Cont. Opt. **6**(1), 1–20 (2016)
36. Berger, T., Reis, T.: Controllability of linear differential-algebraic systems - 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)
37. Berger, T., Reis, T., Trenn, S.: Observability of linear differential-algebraic systems: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations IV. Differential-Algebraic Equations Forum, pp. 161–220. Springer, Heidelberg/New York/Dordrecht/London (2017)
38. Bollhöfer, M., Eppler, A.: Low-rank Cholesky factor Krylov subspace methods for generalized projected Lyapunov equations. In: Benner, P. (ed.) System Reduction for Nanoscale IC Design. Mathematics in Industry, vol. 20. Springer, Berlin, Heidelberg (to appear)
39. Brenan, K., Campbell, S., Petzold, L.: The Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. North-Holland Publishing Co., New York (1989)
40. Chu, E.W., Fan, H.Y., Lin, W.W.: Projected generalized discrete-time periodic Lyapunov equations and balanced realization of periodic descriptor systems. SIAM J. Matrix Anal. Appl. **29**(3), 982–1006 (2007)
41. Dai, L.: Singular Control Systems. Lecture Notes in Control and Information Sciences, vol. 118. Springer, Berlin, Heidelberg (1989)

42. Daniel, L., Siong, O., Chay, L., Lee, K., White, J.: A multiparameter moment-matching model-reduction approach for generating geometrically parameterized interconnect performance models. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **23**(5), 678–693 (2004)
43. Degroote, J., Vierendeels, J., Willcox, K.: Interpolation among reduced-order matrices to obtain parameterized models for design, optimization and probabilistic analysis. Int. J. Numer. Methods Fluids **63**, 207–230 (2010)
44. D'Elia, M., Dedé, L., Quarteroni, A.: Reduced basis method for parameterized differential algebraic equations. Bol. Soc. Esp. Math. Apl. **46**, 45–73 (2009)
45. Desai, U., Pal, D.: A transformation approach to stochastic model reduction. IEEE Trans. Autom. Control **AC-29**(12), 1097–1100 (1984)
46. Druskin, V., Simoncini, V.: Adaptive rational Krylov subspaces for large-scale dynamical systems. Syst. Control Lett. **60**, 546–560 (2011)
47. Druskin, V., Knizhnerman, L., Simoncini, V.: Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. SIAM J. Numer. Anal. **49**, 1875–1898 (2011)
48. Enns, D.: Model reduction with balanced realization: an error bound and a frequency weighted generalization. In: Proceedings of the 23rd IEEE Conference on Decision and Control (Las Vegas, 1984), pp. 127–132. IEEE, New York (1984)
49. Estévez Schwarz, D., Tischendorf, C.: Structural analysis for electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. **28**, 131–162 (2000)
50. Farle, O., Hill, V., Ingelström, P., Dyczij-Edlinger, R.: Multi-parameter polynomial order reduction of linear finite element models. Math. Comput. Model. Dyn. Syst. **14**, 421–434 (2008)
51. Ferranti, F., Antonini, G., Dhaene, T., Knockaert, L.: Passivity-preserving interpolation-based parameterized model order reduction of PEEC models based on scattered grids. Int. J. Numer. Model. **24**(5), 478–495 (2011)
52. Flagg, G.M., Beattie, C.A., Gugercin, S.: Convergence of the iterative rational krylov algorithm. Syst. Control Lett. **61**(6), 688–691 (2012)
53. Freitas, F., Rommes, J., Martins, N.: Gramian-based reduction method applied to large sparse power system descriptor models. IEEE Trans. Power Syst. **23**(3), 1258–1270 (2008)
54. Freund, R.: Krylov-subspace methods for reduced-order modeling in circuit simulation. J. Comput. Appl. Math. **123**(1–2), 395–421 (2000)
55. Freund, R.: Model reduction methods based on Krylov subspaces. Acta Numerica **12**, 267–319 (2003)
56. Freund, R.: The SPRIM algorithm for structure-preserving order reduction of general RCL circuits. Model reduction for circuit simulation. In: Benner, P., Hinze, M., ter Maten, E.J.W. (eds.) Model Reduction for Circuit Simulation. Lecture Notes in Electrical Engineering, vol. 74, pp. 25–52. Springer, Dodrecht (2011)
57. Freund, R., Feldmann, P.: The SyMPVL algorithm and its applications in interconnect simulation. In: Proceedings of the 1997 International Conference on Simulation of Semiconductor Processes and Devices, pp. 113–116. Ney York (1997)
58. Freund, R., Jarre, F.: An extension of the positive real lemma to descriptor systems. Optim. Methods Softw. **19**, 69–87 (2004)
59. Gantmacher, F.: Theory of Matrices. Chelsea Publishing Company, New York (1959)
60. Gear, C., Leimkuhler, B., Gupta, G.: Automatic integration of Euler-Lagrange equations with constraints. J. Comput. Appl. Math. **12–13**, 77–90 (1985)
61. Georgiou, T., Smith, M.: Optimal robustness in the gap metric. IEEE Trans. Autom. Control **35**(6), 673–686 (1990)
62. Geuss, M., Panzer, H., Lohmann, B.: On parametric model order reduction by matrix interpolation. In: Proceedings of the European Control Conference (Zürich, Switzerland, 17–19 July 2013), pp. 3433–3438 (2013)
63. Glover, K.: All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds. Int. J. Control **39**(6), 1115–1193 (1984)

64. Golub, G., Loan, C.V.: Matrix Computations, 3rd edn. The Johns Hopkins University Press, Baltimore, London (1996)
65. Green, M.: Balanced stochastic realizations. Linear Algebra Appl. **98**, 211–247 (1988)
66. Green, M.: A relative error bound for balanced stochastic truncation. IEEE Trans. Autom. Control **33**, 961–965 (1988)
67. Gugercin, S., Antoulas, A.: A survey of model reduction by balanced truncation and some new results. Int. J. Control **77**(8), 748–766 (2004)
68. Gugercin, S., Antoulas, A., Beattie, C.: $\mathcal{H}_2$ model reduction for large-scale linear dynamical systems. SIAM J. Matrix Anal. Appl. **30**(2), 609–638 (2008)
69. Gugercin, S., Stykel, T., Wyatt, S.: Model reduction of descriptor systems by interpolatory projection methods. SIAM J. Sci. Comput. **35**(5), B1010–B1033 (2013)
70. Harshavardhana, P., Jonckheere, E., Silverman, L.: Stochastic balancing and approximation - stability and minimality. IEEE Trans. Autom. Control **29**(8), 744–746 (1984)
71. Heinkenschloss, M., Sorensen, D., Sun, K.: Balanced truncation model reduction for a class of descriptor systems with application to the Oseen equations. SIAM J. Sci. Comput. **30**(2), 1038–1063 (2008)
72. Ho, C.W., Ruehli, A., Brennan, P.: The modified nodal approach to network analysis. IEEE Trans. Circuits Syst. **22**(6), 504–509 (1975)
73. Hossain, M.S., Benner, P.: Generalized inverses of periodic matrix pairs and model reduction for periodic control systems. In: Proceedings of the 1st International Conference on Electrical Engineering and Information and Communication Technology (ICEEICT), Dhaka, Bangladesh, pp. 1–6. IEEE Publications, Piscataway (2014)
74. Ishihara, J., Terra, M.: On the Lyapunov theorem for singular systems. IEEE Trans. Autom. Control **47**(11), 1926–1930 (2002)
75. Jaimoukha, I., Kasenally, E.: Krylov subspace methods for solving large Lyapunov equations. SIAM J. Numer. Anal. **31**(1), 227–251 (1994)
76. Jbilou, K.: ADI preconditioned Krylov methods for large Lyapunov matrix equations. Linear Algebra Appl. **432**(10), 2473–2485 (2010)
77. Jbilou, K., Riquet, A.: Projection methods for large Lyapunov matrix equations. Linear Algebra Appl. **415**(2–3), 344–358 (2006)
78. Jonckheere, E., Silverman, L.: A new set of invariants for linear systems with application to reduced order compensator. IEEE Trans. Autom. Control **28**(10), 953–964 (1983)
79. Katayama, T., Minamino, K.: Linear quadratic regulator and spectral factorization for continuous-time descriptor system. In: Proceedings of the 31st IEEE Conference on Decision and Control (Tuscon, 1992), pp. 967–972. IEEE, New York (1992)
80. Kawamoto, A., Katayama, T.: The semi-stabilizing solution of generalized algebraic Riccati equation for descriptor systems. Automatica **38**, 1651–1662 (2002)
81. Kawamoto, A., Takaba, K., Katayama, T.: On the generalized algebraic Riccati equation for continuous-time descriptor systems. Linear Algebra Appl. **296**, 1–14 (1999)
82. Kerler-Back, J., Stykel, T.: Model reduction for linear and nonlinear magneto-quasistatic equations. Int. J. Numer. Methods Eng. (to appear). doi:10.1002/nme.5507
83. Knizhnerman, L., Simoncini, V.: Convergence analysis of the extended Krylov subspace method for the Lyapunov equation. Numer. Math. **118**(3), 567–586 (2011)
84. Knockaert, L., De Zutter, D.: Laguerre-SVD reduced-order modeling. IEEE Trans. Microw. Theory Tech. **48**(9), 1469–1475 (2000)
85. Kowal, P.: Null space of a sparse matrix. MATLAB Central (2006). http://www.mathworks.fr/matlabcentral/fileexchange/11120
86. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations. Analysis and Numerical Solution. EMS Publishing House, Zürich, Switzerland (2006)
87. Larin, V., Aliev, F.: Construction of square root factor for solution of the Lyapunov matrix equation. Syst. Control Lett. **20**(2), 109–112 (1993)
88. Laub, A., Heath, M., Paige, C., Ward, R.: Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. IEEE Trans. Automat. Control **AC-32**(2), 115–122 (1987)

89. Li, J.R., White, J.: Low rank solution of Lyapunov equations. SIAM J. Matrix Anal. Appl. **24**(1), 260–280 (2002)
90. Li, Y., Bai, Z., Su, Y.: A two-directional Arnoldi process and its application to parametric model order reduction. J. Comput. Appl. Math. **226**, 10–21 (2009)
91. Liebermeister, W., Baur, U., Klipp, E.: Biochemical network models simplified by balanced truncation. FEBS J. **272**, 4034–4043 (2005)
92. Liu, W., Sreeram, V.: Model reduction of singular systems. Internat. J. Syst. Sci. **32**(10), 1205–1215 (2001)
93. Mehrmann, V., Stykel, T.: Descriptor systems: a general mathematical framework for modelling, simulation and control. at-Automatisierungstechnik **54**(8), 405–415 (2006)
94. Meier, III., L., Luenberger, D.: Approximation of linear constant systems. IEEE Trans. Autom. Control **AC-12**(10), 585–588 (1967)
95. Möckel, J., Reis, T., Stykel, T.: Linear-quadratic gaussian balancing for model reduction of differential-algebraic systems. Int. J. Control **84**(10), 1621–1643 (2011)
96. Moore, B.: Principal component analysis in linear systems: controllability, observability, and model reduction. IEEE Trans. Autom. Control **AC-26**(1), 17–32 (1981)
97. Model Order Reduction Wiki. http://www.modelreduction.org (visited 2015-11-02)
98. Ober, R.: Balanced parametrization of classes of linear systems. SIAM J. Control Optim. **29**(6), 1251–1287 (1991)
99. Odabasioglu, A., Celik, M., Pileggi, L.: PRIMA: Passive reduced-order interconnect macromodeling algorithm. IEEE Trans. Circuits Syst. **17**(8), 645–654 (1998)
100. Opdenacker, P., Jonckheere, E.: A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds. IEEE Trans. Circuits Syst. I **35**(2), 184–189 (1988)
101. Panzer, H., Mohring, J., Eid, R., Lohmann, B.: Parametric model order reduction by matrix interpolation. at-Automatisierungtechnik **58**(8), 475–484 (2010)
102. Penzl, T.: A cyclic low-rank Smith method for large sparse Lyapunov equations. SIAM J. Sci. Comput. **21**(4), 1401–1418 (1999/2000)
103. Perev, K., Shafai, B.: Balanced realization and model reduction of singular systems. Int. J. Syst. Sci. **25**(6), 1039–1052 (1994)
104. Pernebo, L., Silverman, L.: Model reduction via balanced state space representation. IEEE Trans. Autom. Control **AC-27**, 382–387 (1982)
105. Phillips, J., Daniel, L., Miguel Silveira, L.: Guaranteed passive balancing transformations for model order reduction. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **22**, 1027–1041 (2003)
106. Phillips, J., Miguel Silveira, L.: Poor Man's TBR: a simple model reduction scheme. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **24**(1), 43–55 (2005)
107. Poloni, F., Reis, T.: A structured doubling algorithm for the numerical solution of Lur'e equations. Preprint 748–2011, DFG Research Center MATHEON, Technische Universität Berlin (2011)
108. Poloni, F., Reis, T.: A deflation approach for large-scale Lur'e equations. SIAM J. Matrix Anal. Appl. **33**(4), 1339–1368 (2012)
109. Reis, T.: Circuit synthesis of passive descriptor systems - a modified nodal approach. Int. J. Circuit Theory Appl. **38**(1), 44–68 (2010)
110. Reis, T.: Lur'e equations and even matrix pencils. Linear Algebra Appl. **434**(1), 152–173 (2011)
111. Reis, T.: Mathematical modeling and analysis of nonlinear time-invariant RLC circuits. In: Benner, P., Findeisen, R., Flockerzi, D., Reichl, U., Sundmacher, K. (eds.) Large-Scale Networks in Engineering and Life Sciences. Modeling and Simulation in Science, Engineering and Technology, pp. 125–198. Birkhäuser, Basel (2014). Chapter 2
112. Reis, T., Rendel, O.: Projection-free balanced truncation for differential-algebraic systems. In: Workshop on Model Reduction of Complex Dynamical Systems MODRED 2013, Magdeburg, 13 December 2013
113. Reis, T., Stykel, T.: Balanced truncation model reduction of second-order systems. Math. Comput. Model. Dyn. Syst. **14**(5), 391–406 (2008)

114. Reis, T., Stykel, T.: PABTEC: Passivity-preserving balanced truncation for electrical circuits. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **29**(9), 1354–1367 (2010)
115. Reis, T., Stykel, T.: Positive real and bounded real balancing for model reduction of descriptor systems. Int. J. Control **83**(1), 74–88 (2010)
116. Reis, T., Stykel, T.: Lyapunov balancing for passivity-preserving model reduction of RC circuits. SIAM J. Appl. Dyn. Syst. **10**(1), 1–34 (2011)
117. Reis, T., Voigt, M.: Linear-quadratic infinite time horizon optimal control for differential-algebraic equations - a new algebraic criterion. In: Proceedings of the International Symposium on Mathematical Theory of Networks and Systems (MTNS 2012), Melbourne, Australia, 9–13 July 2012 (2012)
118. Riaza, R.: Differential-Algebraic Systems: Analytical Aspects and Circuit Applications. World Scientific Publishing Co. Pte. Ltd., Hackensack (2008)
119. Roberts, J.: Linear model reduction and solution of the algebraic Riccati equation by use of the sign function. Int. J. Control **32**(4), 677–687 (1980). Reprint of Technical Report TR-13, CUED/B-Control, Engineering Department, Cambridge University, 1971
120. Rommes, J., Martins, N.: Exploiting structure in large-scale electrical circuit and power system problems. Linear Algebra Appl. **431**(3–4), 318–333 (2009)
121. Rosenbrock, H.: The zeros of a system. Int. J. Control **18**(2), 297–299 (1973)
122. Saad, Y.: Numerical solution of large Lyapunov equations. In: Kaashoek, M., Schuppen, J.V., Ran, A. (eds.) Signal Processing, Scattering, Operator Theory, and Numerical Methods, pp. 503–511. Birkhäuser, Boston (1990)
123. Saad, Y.: Iterative Methods for Sparse Linear Systems. PWS Publishing Company, Boston (1996)
124. Sabino, J.: Solution of large-scale Lyapunov equations via the block modified Smith method. Ph.D. thesis, Rice University, Houston (2006)
125. Salimbahrami, B., Lohmann, B.: Order reduction of large scale second-order systems using Krylov subspace methods. Linear Algebra Appl. **415**, 385–405 (2006)
126. Schöps, S.: Multiscale modeling and multirate time-integration of field/circuit coupled problems. Ph.D. thesis, Bergische Universität Wuppertal (2011)
127. Schöps, S., De Gersem, H., Weiland, T.: Winding functions in transient magnetoquasistatic field-circuit coupled simulations. COMPEL **32**(6), 2063–2083 (2013)
128. Simoncini, V.: A new iterative method for solving large-scale Lyapunov matrix equations. SIAM J. Sci. Comput. **29**(3), 1268–1288 (2007)
129. Son, N.: Interpolation based parametric model order reduction. Ph.D. thesis, Universität Bremen, Germany (2012)
130. Son, N.: A real time procedure for affinely dependent parametric model order reduction using interpolation on Grassmann manifolds. Int. J. Numer. Methods Eng. **93**(8), 818–833 (2013)
131. Son, N., Stykel, T.: Model order reduction of parameterized circuit equations based on interpolation. Adv. Comput. Math. **41**(5), 1321–1342 (2015)
132. Steedhar, J., Van Dooren, P., Misra, P.: Minimal order time invariant representation of periodic descriptor systems. In: Proceedings of the American Control Conference, San Diego, California, June 1999, vol. 2, pp. 1309–1313 (1999)
133. Stykel, T.: Numerical solution and perturbation theory for generalized Lyapunov equations. Linear Algebra Appl. **349**, 155–185 (2002)
134. Stykel, T.: Gramian-based model reduction for descriptor systems. Math. Control Signals Syst. **16**, 297–319 (2004)
135. Stykel, T.: Balanced truncation model reduction for semidiscretized Stokes equation. Linear Algebra Appl. **415**, 262–289 (2006)
136. Stykel, T.: A modified matrix sign function method for projected Lyapunov equations. Syst. Control Lett. **56**, 695–701 (2007)
137. Stykel, T.: Low-rank iterative methods for projected generalized Lyapunov equations. Electron. Trans. Numer. Anal. **30**, 187–202 (2008)
138. Stykel, T.: Balancing-related model reduction of circuit equations using topological structure. In: Benner, P., Hinze, M., ter Maten, E.J.W. (eds.) Model Reduction for Circuit Simulation. Lecture Notes in Electrical Engineering, vol. 74, pp. 53–80. Springer, Dodrecht (2011)

139. Stykel, T., Reis, T.: The PABTEC algorithm for passivity-preserving model reduction of circuit equations. In: Proceedings of the 19th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2010, Budapest, Hungary, 5–9 July 2010), paper 363. ELTE, Budapest, Hungary (2010)
140. Stykel, T., Simoncini, V.: Krylov subspace methods for projected Lyapunov equations. Appl. Numer. Math. **62**, 35–50 (2012)
141. Takaba, K., Morihira, N., Katayama, T.: A generalized Lyapunov theorem for descriptor system. Syst. Control Lett. **24**, 49–51 (1995)
142. Tan, S., He, L.: Advanced Model Order Reduction Techniques in VLSI Design. Cambridge University Press, New York (2007)
143. Tombs, M., Postlethweite, I.: Truncated balanced realization of a stable non-minimal state-space system. Int. J. Control **46**(4), 1319–1330 (1987)
144. Uddin, M., Saak, J., Kranz, B., Benner, P.: Computation of a compact state space model for an adaptive spindle head configuration with piezo actuators using balanced truncation. Prod. Eng. Res. Dev. **6**(6), 577–586 (2012)
145. Unneland, K., Van Dooren, P., Egeland, O.: New schemes for positive real truncation. Model. Identif. Control **28**, 53–65 (2007)
146. Varga, A., Fasol, K.: A new square-root balancing-free stochastic truncation model reduction algorithm. In: Proceedings of 12th IFAC World Congress, Sydney, Australia, vol. 7, pp. 153–156 (1993)
147. Verghese, G., Lévy, B., Kailath, T.: A generalized state-space for singular systems. IEEE Trans. Autom. Control **AC-26**(4), 811–831 (1981)
148. Wachspress, E.: The ADI minimax problem for complex spectra. In: Kincaid, D., Hayes, L. (eds.) Iterative Methods for Large Linear Systems, pp. 251–271. Academic Press, Boston (1990)
149. Wang, H.S., Chang, F.R.: The generalized state-space description of positive realness and bounded realness. In: Proceedings of the 39th IEEE Midwest Symposium on Circuits and Systems, vol. 2, pp. 893–896. IEEE, New York (1996)
150. Wang, H.S., Yung, C.F., Chang, F.R.: Bounded real lemma and H_∞ control for descriptor systems. In: IEE Proceedings on Control Theory and Applications, vol. 145, pp. 316–322. IEE, Stevenage (1998)
151. Wang, H.S., Yung, C.F., Chang, F.R.: The positive real control problem and the generalized algebraic Riccati equation for descriptor systems. J. Chin. Inst. Eng. **24**(2), 203–220 (2001)
152. Willcox, K., Lassaux, G.: Model reduction of an actively controlled supersonic diffuser. In: Benner, P., Mehrmann, V., Sorensen, D. (eds.) Dimension Reduction of Large-Scale Systems. Lecture Notes in Computational Science and Engineering, vol. 45, pp. 357–361. Springer, Berlin, Heidelberg (2005). Chapter 20
153. Willcox, K., Peraire, J.: Balanced model reduction via the proper orthogonal decomposition. AIAA J. **40**(11), 2323–2330 (2002)
154. Xin, X.: Strong solutions and maximal solutions of generalized algebraic Riccati equations. In: Proceedings of the 47th IEEE Conference on Decision and Control, pp. 528–533. IEEE, Piscataway (2008)
155. Yan, B., Tan, S.D., McGaughy, B.: Second-order balanced truncation for passive-order reduction of RLCK circuits. IEEE Trans. Circuits Syst. II **55**(9), 942–946 (2008)
156. Zhang, L., Lam, J., Xu, S.: On positive realness of descriptor systems. IEEE Trans. Circuits Syst. **49**(3), 401–407 (2002)
157. Zhang, Z., Wong, N.: An efficient projector-based passivity test for descriptor systems. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **29**(8), 1203–1214 (2010)
158. Zhou, K.: Error bounds for frequency weighted balanced truncation and relative error model reduction. In: Proceedings of the IEEE Conference on Decision and Control (San Antonio, Texas, December 1993), pp. 3347–3352 (1993)
159. Zhou, K.: Frequency-weighted $\mathcal{L}_\infty$ norm and optimal Hankel norm model reduction. IEEE Trans. Autom. Control **40**, 1687–1699 (1995)

Observability of Linear Differential-Algebraic Systems: A Survey

Thomas Berger, Timo Reis, and Stephan Trenn

Abstract We investigate different concepts related to observability of linear constant coefficient differential-algebraic equations. Regularity, which, loosely speaking, guarantees existence and uniqueness of solutions for any inhomogeneity, is not required in this article. Concepts like impulse observability, observability at infinity, behavioral observability, strong and complete observability are described and defined in the time-domain. Special emphasis is placed on a normal form under output injection, state space and output space transformation. This normal form together with duality is exploited to derive Hautus-type criteria for observability. We also discuss geometric criteria, Kalman decompositions and detectability. Some new results on stabilization by output injection are proved.

Keywords Controllability • Differential-algebraic equations • Duality • Hautus Test • Kalman decomposition • Observability • Output injection • Wong sequences

Mathematics Subject Classification (2010) 93B07, 34A09, 93B10, 93B25, 93B27, 93B05, 93C05

1 Introduction

Observability is, roughly speaking, the property of a system that the state can be reconstructed from the knowledge of input and output. The precise concept however depends on the specific framework, as quite a number of different concepts of observability are present today.

T. Berger (✉) • T. Reis
Fachbereich Mathematik, Universität Hamburg, Bundesstraße 55, 20146 Hamburg, Germany
e-mail: thomas.berger@uni-hamburg.de; timo.reis@math.uni-hamburg.de

S. Trenn
Fachbereich Mathematik, Technische Universität Kaiserslautern, Postfach 3049, 67653 Kaiserslautern, Germany
e-mail: trenn@mathematik.uni-kl.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7_4

Like many crucial concepts in mathematical systems theory, observability goes back to Kalman [44–46], who introduced the notion of observability more than 50 years ago for finite-dimensional linear systems governed by ordinary differential equations (ODEs). Observability was defined via the property that the initial value of the state is uniquely determined by input and output trajectories. What is particularly nice about observability is the *duality principle*. An ODE system is observable if, and only if, a certain artificial system obtained by taking the transposes of the involved matrices is controllable.

The theory of observability was an essential ingredient for Luenberger's achievements on observer design [58–60], which is, on the other hand, an essential ingredient for the design of dynamic controllers. The idea behind controller design is amazingly simple: the observer reconstructs the state and this reconstructed state is fed back to the system.

A further milestone in mathematical systems theory was the *theory of behaviors* introduced by Willems [70, 84], where systems of differential equations of possibly higher order are considered. The novelty of this approach was to treat inputs, states, and outputs alike; in particular, the behavioral model allows for different choices of inputs and outputs. Nevertheless, or even maybe because of this, the behavioral approach provides a deep understanding of nearly all tasks of modern systems theory. Indeed, the essential systems theoretic concepts of controllability and observability are defined so that they coincide with the respective properties of ODE systems: behavioral controllability is defined via concatenability of trajectories [70, Definition 5.2.2], whereas observability uses a split of the dynamic variables into two kinds, namely *external* and *internal variables* [70, Definition 5.3.2]. For ODE systems, the external variables are inputs and outputs, whereas the internal variables are the states. Behavioral observability means that the external variables uniquely determine the internal variables. The behavioral approach reveals a certain lack of duality between controllability and observability: while controllable systems with additional equations of the form $0 = 0$ stay controllable in the behavioral sense, their dual may contain free variables and is not observable in general. This does not come as a surprise, especially in view of Willems' remark in [84]:

> ... controllability and observability are prima facie not dual concepts. Controllability is an intrinsic concept of the behavior of a dynamical system, while observability remains representation dependent.

The type of systems to be analyzed in the present article is "in between" ODE and behavioral systems: we consider linear constant coefficient descriptor systems given by differential-algebraic equations (DAEs) of the form

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t} Ex(t) &= Ax(t) + Bu(t) \\ y(t) &= Cx(t) + Du(t)\,, \end{aligned} \tag{1.1}$$

where $E, A \in \mathbb{R}^{l\times n}$, $B \in \mathbb{R}^{l\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$. A matrix pencil $sE - A \in \mathbb{R}[s]^{l\times n}$ is called *regular*, if $l = n$ and $\det(sE - A) \in \mathbb{R}[s] \setminus \{0\}$; otherwise it is called *singular*. In the present paper, we put special emphasis on the singular case.

We distinguish between *input* $u : \mathbb{R} \to \mathbb{R}^m$, *output* $y : \mathbb{R} \to \mathbb{R}^p$, and *(generalized) state* $x : \mathbb{R} \to \mathbb{R}^n$. One should keep in mind that in the singular case u might be constrained and some of the state variables may play the role of an input. Note that, strictly speaking, $x(t)$ is in general not a state in the sense that the free system (i.e., $u \equiv 0$) can be initialized with an arbitrary state $x(0) = x_0 \in \mathbb{R}^n$ [48, Sect. 2.2]. We will, however, speak of the state $x(t)$ for sake of brevity, especially since $x(t)$ contains the full information about the system at time t.

We recall that in DAE systems (1.1) the algebraic constraints may lead to consistency conditions on the input and cause non-existence of solutions to certain initial value problems. Furthermore, solutions may not be unique due to underdetermined parts. There is a vast amount of literature on the solution theory of DAEs; here we refer to the recent depiction of DAEs in a systems theoretic framework in [15], where also several application areas are mentioned and a comprehensive list of literature is given.

Though DAEs are a subclass of behavioral systems, the study of behavioral observability is not fully satisfactory in the DAE case: the reason is that there might be purely algebraic variables which do not exert influence on the output. An observability concept which also covers this effect is in particular indispensable for the *minimal realization* problem by differential-algebraic systems [32, Sect. 2.6]. This need has led to the notions of *impulse observability* and *observability at infinity* [3, 13, 25, 26, 31–33, 40, 43, 53, 76, 82]. However, a rigorous definition of these concepts is a delicate issue: in various publications, the theoretical claim that an inconsistent initial value causes Dirac impulses in the state was used to define impulse observability (which was actually the reason for the choice of the name) [31, 32, 40, 43]. In particular, this leads to the consideration of *distributional solutions*. However, this approach contains a grave paradox: the initial value is the evaluation of the state at initial time (which can always be chosen to be zero here because of time-invariance); Schwartz' celebrated theory of distributions [75] however does not allow for evaluations at certain time points. Loosely speaking, distributions are only defined by means of their average behavior along compactly supported, infinitely often differentiable functions. In the present article we also aim to circumvent this paradox by focusing on the smaller class of *piecewise-smooth distributions* as introduced in [76, 77]. This class indeed allows for evaluation at specific time points, and therefore it is apt to consider inconsistently initialized DAEs and rigorously define accordant observability concepts.

A survey article [15] on controllability of DAE systems appeared in the same series "Surveys on Differential-Algebraic Equations" within the "Differential-Algebraic Equations Forum". The present article on observability is the counterpart of that survey. The structure of the present paper is similar to [15]: we introduce different observability concepts using the solution behavior and thereafter we give

characterizations by means of properties of the involved matrices. We further analyze duality to the respective controllability concepts.

As in [15], many of our considerations utilize certain (normal) forms. Besides the Weierstrass and Kronecker canonical forms for matrix pencils (see [50, 83] and the classical book [36] by Gantmacher), we also use a form that we call "output injection (OI) normal form", which is a normal form under state space and output space transformation and output injection. Loosely speaking, the OI normal form is the transpose of the feedback canonical form derived by Loiseau, Özçaldiran, Malabre and Karcanias in [56].

The paper is organized as follows:

2 Weak and Distributional Solutions **p. 166**

The solution framework for the present article is introduced in this section. Besides weak solutions (which are basically solutions in a function setting), we consider distributional solutions of linear DAEs. The collection of solutions is called *behavior*. In particular we consider the behavior arising from *initial trajectory problems* which is, loosely speaking, the set of those solutions which satisfy the DAE only for times $t \geq 0$. The relation between the introduced behavior notions is discussed.

3 Observability Concepts **p. 170**

This section contains the definition of all observability notions which are treated in the present article, such as behavioral, impulse, strong and complete observability as well as observability at infinity. We further introduce corresponding concepts of relevant state (RS) observability. Loosely speaking, the latter concepts correspond to observability of the part of the state which is uniquely determined by input, output and initial values. The RS observability notions will later turn out to be weaker than the respective conventional observability notions and to be equivalent to them, if the system is regular. All the observability concepts are introduced by means of time-domain properties. That is, they are defined by means of the (distributional) behavior of the underlying system. We also present some basic properties.

4 Output Injection Normal Form **p. 180**

We introduce an "output injection (OI) normal form", which is a special form under output injection and coordinate transformation of state and output. We further show that all considered observability concepts from Sect. 3 are invariant under this type of transformation. This allows for an analysis of the observability concepts by means of a system being in this form. Since, in particular, the OI normal form consists of decoupled parts, this analysis leads to a test of the respective observability properties by means of certain "prototypes".

5 Duality of Observability and Controllability **p. 191**

It is well known from systems theory for ODEs that controllability and observability are dual in a certain sense. More precisely, an ODE system is observable if, and only if, the control system obtained by transposition is controllable. Here we analyze duality for the introduced observability concepts and behavioral, impulse, strong and complete controllability as well as controllability at infinity as considered in [15]. It turns out that there is a certain lack of duality. However, we show that

the aforementioned controllability concepts are dual to the respective relevant state observability notions.

6 Algebraic Criteria p. 195

Duality and the OI normal form enable us to give short proofs of equivalent criteria for the observability concepts which are in particular generalizations of the Hautus test. Most characterizations are well known and we discuss the relevant literature.

7 Geometric Criteria p. 200

We present some geometric viewpoints of DAE systems using so-called *restricted Wong sequences*. This leads to further equivalent criteria for the observability concepts from Sect. 3.

8 Kalman Decomposition p. 203

We consider different types of Kalman decompositions for DAE systems. We show that a combined Kalman decomposition for controllability and observability is possible as well as a refined pure observability decomposition.

9 Detectability and Stabilization by Output Injection p. 206

Finally, we introduce some notions related to detectability for DAE systems. Criteria of Hautus type and duality to stabilizability concepts from [15] are derived. We further prove some new results concerning the stabilization by output injection.

We close the introduction with the nomenclature used in this paper:

$\mathbb{Z}, \mathbb{N}, \mathbb{N}_0$		The set of integers, natural numbers, and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, resp.
$\ell(\alpha), \lvert\alpha\rvert$		Length $\ell(\alpha) = l$ and absolute value $\lvert\alpha\rvert = \sum_{i=1}^{l} \alpha_i$ of a multi-index $\alpha = (\alpha_1, \ldots, \alpha_l) \in \mathbb{N}^l$
$\mathbb{C}_+ (\mathbb{C}_-)$		Open set of complex numbers with positive (negative) real part, resp.
$\overline{\mathbb{C}_+}$		Closed set of complex numbers with non-negative real part
$\mathbb{R}[s]$		The ring of polynomials with coefficients in $\mathbb{R}$
$\mathbb{R}(s)$		The quotient field of $\mathbb{R}[s]$
$R^{n\times m}$		The set of $n \times m$ matrices with entries in a ring R
$\mathbf{Gl}_n(R)$		The group of invertible matrices in $R^{n\times n}$
$\sigma(M)$		The spectrum of $M \in \mathbb{R}^{n\times m}$
$\lVert x\rVert$	$=$	$\sqrt{x^\top x}$, the Euclidean norm of $x \in \mathbb{R}^n$
$M\mathscr{S}$	$=$	$\{ Mx \in \mathbb{R}^m \mid x \in \mathscr{S} \}$, the image of $\mathscr{S} \subseteq \mathbb{R}^n$ under $M \in \mathbb{R}^{m\times n}$
$M^{-1}\mathscr{S}$	$=$	$\{ x \in \mathbb{R}^m \mid Mx \in \mathscr{S} \}$, the pre-image of $\mathscr{S} \subseteq \mathbb{R}^m$ under M
$\mathscr{C}^\infty(\mathscr{T}; \mathbb{R}^n)$		The set of infinitely differentiable functions $f : \mathscr{T} \to \mathbb{R}^n$
$\mathscr{AC}(\mathbb{R}; \mathbb{R}^n)$		The set of locally absolutely continuous functions $f : \mathbb{R} \to \mathbb{R}^n$
$\mathscr{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^n)$		The set of locally Lebesgue integrable functions $f : \mathbb{R} \to \mathbb{R}^n$, where $\int_{K\cap\mathscr{T}} \lVert f(t)\rVert \, \mathrm{d}t < \infty$ for all compact $K \subseteq \mathbb{R}$
$\mathscr{D}'$		The set of distributions on $\mathbb{R}$

$\dot{f}$ $(f^{(i)})$	The (i-th) distributional derivative of $f \in \mathscr{D}'$, $i \in \mathbb{N}_0$
$f_{\mathscr{D}'}$	The distribution induced by the function $f \in \mathscr{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$
δ_t, δ	The Dirac impulse at $t \in \mathbb{R}$ and $\delta = \delta_0$
$f \overset{\text{a.e.}}{=} g$	Means that $f, g \in \mathscr{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^n)$ are equal "almost everywhere", i.e., $f(t) = g(t)$ for almost all $t \in \mathbb{R}$
$\operatorname{ess\,sup}_I \|f\|$	The essential supremum of the measurable function $f : \mathscr{T} \to \mathbb{R}^n$ over $I \subseteq \mathscr{T}$
f_I	The restriction of the function $f : \mathbb{R} \to \mathbb{R}^n$ to $I \subseteq \mathbb{R}$, i.e., $f_I(t) = f(t)$ for $t \in I$ and $f_I(t) = 0$ otherwise

We further use the following abbreviations in this article:

DAE	differential-algebraic equation,
ITP	initial trajectory problem, see p. 168,
ODE	ordinary differential equation,
OI	output injection, see p. 180,
RS	relevant state, see p. 174.

2 Weak and Distributional Solutions

We consider linear DAE systems of the form (1.1) with $E, A \in \mathbb{R}^{l\times n}$, $B \in \mathbb{R}^{l\times m}$, $C \in \mathbb{R}^{p\times n}$, $D \in \mathbb{R}^{p\times m}$. The set of these systems is denoted by $\Sigma_{l,n,m,p}$ and we write $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$.

A trajectory $(x, u, y) : \mathbb{R} \to \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p$ is said to be a *(weak) solution* of (1.1) if, and only if, it belongs to the *behavior* of (1.1):

$$\mathfrak{B}_{[E,A,B,C,D]} := \left\{ (x,u,y) \in \mathscr{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R}^{n+m+p}) \;\middle|\; \begin{array}{l} Ex \in \mathscr{AC}(\mathbb{R};\mathbb{R}^n) \text{ and } (x,u,y) \\ \text{fulfills (1.1) for almost all } t \in \mathbb{R} \end{array} \right\}.$$

Recall that $Ex \in \mathscr{AC}(\mathbb{R};\mathbb{R}^l)$ implies continuity of Ex (but x itself may be discontinuous). For studying inconsistent initial values and impulsive effects we will also consider distributional behaviors which are formally introduced in due course.

For the analysis of DAE systems in $\Sigma_{l,n,m,p}$ we assume that the states, inputs and outputs of the system are fixed a priori by the designer, i.e., the realization is given (but maybe not appropriate). This is different from other approaches based on the behavioral setting, see [28], where only the free variables in the system are viewed as inputs; this may require a reinterpretation of states as inputs and of inputs as states. In the present paper we will assume that such a reinterpretation of variables has already been done or is not feasible, and the given DAE system is fixed.

Next we consider solutions of (1.1) in the distributional sense. We primarily do formal and arithmetical calculations in the space of distributions; the latter is

usually denoted by $\mathscr{D}'$ because it is defined as a dual of a certain test function space $\mathscr{D}$. For a deeper introduction to the mathematical (in particular, analytical) background we refer to [74, Chap. 6]. Distributions are generalized functions and allow differentiation of arbitrary order. A key role is played by the *Dirac impulse* (also called the *δ distribution*) δ_t, which corresponds to evaluation of a test function at $t \in \mathbb{R}$.

The *distributional behavior* consists of the *distributional solutions*, i.e.,

$$\mathfrak{B}^{\mathscr{D}'}_{[E,A,B,C,D]} = \left\{ (x,u,y) \in (\mathscr{D}')^{n+m+p} \;\middle|\; \begin{array}{r} E\dot{x} = Ax + Bu \\ y = Cx + Du \end{array} \right\}.$$

Note that $\mathfrak{B}_{[E,A,B,C,D]}$ can be canonically embedded into $\mathfrak{B}^{\mathscr{D}'}_{[E,A,B,C,D]}$. We also consider a special subspace of the distributions which features further properties. To this end we utilize the distributional solution framework as introduced in [76, 77], namely the space of *piecewise-smooth distributions*

$$\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty} = \left\{ \sum_{i\in\mathbb{Z}} \left(\left(\alpha^i{}_{[t_i,t_{i+1}]}\right)_{\mathscr{D}'} + D_{t_i} \right) \;\middle|\; \begin{array}{l} \{\, t_i \in \mathbb{R} \mid i \in \mathbb{Z} \,\} \text{ is locally finite,} \\ \forall i \in \mathbb{Z}:\ t_i < t_{i+1} \ \wedge\ \alpha^i \in \mathscr{C}^\infty(\mathbb{R};\mathbb{R}) \\ \qquad \wedge\ D_{t_i} \in \operatorname{span}\left\{ \delta^{(k)}_{t_i} \;\middle|\; k \in \mathbb{N}_0 \right\} \end{array} \right\}.$$

We clearly have that $\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ is a subspace of $\mathscr{D}'$ which is invariant under differentiation, i.e., $\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty} = \mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$. Note that $\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ is not a (topologically) closed subspace of $\mathscr{D}'$. The behavior corresponding to $\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ is

$$\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]} = \mathfrak{B}^{\mathscr{D}'}_{[E,A,B,C,D]} \cap (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^{n+m+p}.$$

Note that $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]} \not\subseteq \mathfrak{B}_{[E,A,B,C,D]}$ and $\mathfrak{B}_{[E,A,B,C,D]} \not\subseteq \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]}$.

Any $D \in \mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ has a unique representation $D = f_{\mathscr{D}'} + \sum_{t\in T} D_t$, where $T \subseteq \mathbb{R}$ is locally finite and $f \in \mathscr{L}^1_{\mathrm{loc}}(\mathbb{R};\mathbb{R})$ is piecewise smooth. The *distributional restriction* to some interval $M \subseteq \mathbb{R}$ (cf. [77, Definition 8]) is given by

$$D_M = (f_M)_{\mathscr{D}'} + \sum_{t\in M\cap T} D_t \in \mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}.$$

Note that the restriction is not well-defined for general distributions [77, Theorem 2.2.2]. The class $\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ moreover allows to perform point evaluations in some sense. Namely, for $D \in \mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$ as above and $t_0 \in \mathbb{R}$, the expressions

$$D(t_0^+) := \lim_{t\searrow t_0} f(t), \qquad D(t_0^-) := \lim_{t\nearrow t_0} f(t)$$

are well-defined, since f is piecewise smooth. Furthermore, the *impulsive part of D at $t_0 \in \mathbb{R}$* is given by

$$D[t_0] := \begin{cases} 0, & \text{if } t_0 \notin T, \\ D_{t_0}, & \text{if } t_0 \in T. \end{cases} \tag{2.1}$$

An important property of DAEs is the fact that due to the algebraic constraints not all initial values $x_0 \in \mathbb{R}^n$ for $x(0^-)$ are possible (even in the above distributional solution framework). Indeed, we call $x_0 \in \mathbb{R}^n$ a *consistent initial value* if, and only if, there exists $(x, u, y) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}$ with $x(0^-) = x_0$. However, there are many reasons to consider also inconsistent initial values. The problem of inconsistent initial values may be formalized in the framework of *initial trajectory problems* (ITP) and its corresponding ITP-behavior

$$\mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]} = \left\{ (x,u,y) \in (\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty)^{n+m+p} \;\middle|\; \begin{aligned} (E\dot{x})_{[0,\infty)} &= (Ax+Bu)_{[0,\infty)} \\ y_{[0,\infty)} &= (Cx+Du)_{[0,\infty)} \end{aligned} \right\},$$

i.e., the DAE is supposed to hold only on the interval $[0, \infty)$ and there are no explicit constraints in the past.[1] Clearly, $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]} \subseteq \mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]}$, i.e., any "consistent" solution $(x, u, y) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}$ is also an ITP-solution, but it should be noted that in general

$$\left\{ (x,u,y)_{[0,\infty)} \;\middle|\; \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]} \right\} \neq \left\{ (x,u,y)_{[0,\infty)} \;\middle|\; \mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]} \right\},$$

because ITP-solutions may exhibit impulsive terms $x[0]$ induced by inconsistent initial values, which are not present in consistent solutions. In the ODE-case, $E = I$, this distinction vanishes, that is on $[0, \infty)$ the two behaviors $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[I,A,B,C,D]}$ and $\mathfrak{B}^{\mathrm{ITP}}_{[I,A,B,C,D]}$ are identical.

A different approach (motivated somewhat by the Laplace transform) handles inconsistent initial values by the consideration of the following behavior parametrized by the "initial value" $z_0 \in \mathbb{R}^l$

$$\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]} := \left\{ (x,u,y) \in (\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty)^{n+m+p} \;\middle|\; \begin{aligned} E\dot{x} &= Ax+Bu+\delta z_0 \\ y &= Cx+Du \end{aligned} \right\}.$$

[1] For singular DAEs it is however *not true* that all $x(0^-) \in \mathbb{R}^n$ are feasible for an ITP. For example, the overdetermined DAE $\dot{x} = 0$, $0 = x$ has no ITP solution with $x(0^-) \neq 0$, because then $x(0^+) = 0$ and $0 = \dot{x}[0] = (x(0^+) - x(0^-))\delta_0$ are conflicting.

Indeed, for ODE systems, the addition of δz_0 corresponds to an initialization $x(0^+) = z_0$ (under the assumption that $x(0^-) = 0$). Note that the behavior $\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]}$ can be seen as a variant of $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]}$ where an additional impulsive input δz_0 is present. We will need all of the above distributional solution spaces to define different notions of observability.

Before we begin the investigation of the different observability definitions and their characterizations, we provide a better understanding of the three different distributional solution spaces and their relationship with each other.

First, we highlight a fundamental property of general homogeneous DAEs $\mathscr{E}\dot{z} = \mathscr{A}z$ with $\mathscr{E}, \mathscr{A} \in \mathbb{R}^{r\times s}$ which follows easily from the definition of restriction in $\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}$:

$$\left\{ z_{(-\infty,0)} \;\middle|\; z \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^s,\ \mathscr{E}\dot{z} = \mathscr{A}z \right\} = \left\{ z_{(-\infty,0)} \;\middle|\; z \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^s, (\mathscr{E}\dot{z})_{(-\infty,0)} = (\mathscr{A}z)_{(-\infty,0)} \right\}, \tag{2.2}$$

in other words any solution given on $(-\infty, 0)$ can be extended to a global solution. This "causality" property is essential to prove the following result which allows to decouple inhomogeneous DAEs.

Lemma 2.1 *Let $\mathscr{E}, \mathscr{A} \in \mathbb{R}^{r\times s}$ and $f \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^r$. Then*

$$\left\{ z \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^s \;\middle|\; \mathscr{E}\dot{z} = \mathscr{A}z + f_{[0,\infty)} \right\} = \left\{ z \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^s \;\middle|\; \mathscr{E}\dot{z} = \mathscr{A}z \right\} + \left\{ z \in (\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty})^s \;\middle|\; z_{(-\infty,0)} = 0,\ \mathscr{E}\dot{z} = \mathscr{A}z + f_{[0,\infty)} \right\}.$$

Proof The subspace inclusion $\supseteq$ is clear. To show the converse let z be a solution of $\mathscr{E}\dot{z} = \mathscr{A}z + f_{[0,\infty)}$, then z satisfies $(\mathscr{E}\dot{z})_{(-\infty,0)} = (\mathscr{A}z + f_{[0,\infty)})_{(-\infty,0)} = (\mathscr{A}z)_{(-\infty,0)}$. By causality (2.2) we find a solution $\tilde{z}$ of $\mathscr{E}\dot{z} = \mathscr{A}z$ with $\tilde{z}_{(-\infty,0)} = z_{(-\infty,0)}$. Then $\hat{z} := z - \tilde{z}$ satisfies $\hat{z}_{(-\infty,0)} = 0$ and $\mathscr{E}\dot{\hat{z}} = \mathscr{A}z + f_{[0,\infty)} - \mathscr{A}\tilde{z} = \mathscr{A}\hat{z} + f_{[0,\infty)}$. This shows that $z = \tilde{z} + \hat{z}$ can be decomposed as claimed. □

Note that Lemma 2.1 is a generalization of the well-known property of linear ODEs that the influence from the initial value on the solution can be decoupled from the influence of the inhomogeneity. However, for DAEs the initial condition $z(0) = 0$ is not feasible for general inhomogeneous DAEs (with fixed inhomogeneity), that is why we restrict the influence of the inhomogeneity to the interval $[0, \infty)$, because then a zero initial value (in the past) is feasible.

We are now able to present the relationship between the ITP-behaviors (which allows for inconsistent initial values implicitly) and the δz_0- behavior which introduces an initial value explicitly.

Lemma 2.2 *For $z_0 \in \mathbb{R}^l$ define*

$$\left[\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]} \ominus \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}\right] := \left\{ (x,u,y) \in \mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]} \;\middle|\; (x,u,y)_{(-\infty,0)} = 0 \right\}.$$

Then

$$\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]} = \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]} + \left[\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]} \ominus \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}\right].$$

Furthermore, for all $x_0 \in \mathbb{R}^n$:

$$\left\{ (x,u,y)_{[0,\infty)} \;\middle|\; (x,u,y) \in \mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]} \wedge x(0^-) = x_0 \right\}$$
$$= \left\{ (x,u,y)_{[0,\infty)} \;\middle|\; (x,u,y) \in \left[\mathfrak{B}^{\delta E x_0}_{[E,A,B,C,D]} \ominus \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}\right] \right\},$$

i.e., the response on $[0,\infty)$ to the (potentially inconsistent) initial value x_0 within the ITP-framework is the same as the response of the DAE with the additional input $\delta E x_0$ and zero initial condition.

Proof The first equality follows directly from Lemma 2.1 with $z = (x,u,y)$ and $f_{[0,\infty)} = \delta z_0$, the second equality was shown in [78, Theorem 5.3]. □

Remark 2.1 Note that $\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]}$ is not a vector space for $z_0 \neq 0$. It might even be empty (for instance, consider $E = A = B = C = D = 0 \in \mathbb{R}$ and $z_0 = 1$). Lemma 2.2 shows that it is an affine linear space. More precisely, it is a shifted version of the distributional behavior $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,B,C,D]}$, where z_0 takes the role of an initial value in a certain sense. However, the following linearity property holds for any $z_0^1, z_0^2 \in \mathbb{R}^l$:

$$(x^1,u^1,y^1) \in \mathfrak{B}^{\delta z_0^1}_{[E,A,B,C,D]} \wedge (x^2,u^2,y^2) \in \mathfrak{B}^{\delta z_0^2}_{[E,A,B,C,D]}$$
$$\Rightarrow \quad (x^1+x^2, u^1+u^2, y^1+y^1) \in \mathfrak{B}^{\delta(z_0^1+z_0^2)}_{[E,A,B,C,D]}.$$

At this point it is not yet clear why we have introduced the solution set $\mathfrak{B}^{\delta z_0}_{[E,A,B,C,D]}$ but it will turn out that this is fruitful for defining some of the observability concepts.

3 Observability Concepts

Classically, observability is defined as the absence of indistinguishable states (see the textbook [79]) or, in a behavioral setting [70], as the absence of nontrivial solutions which generate a trivial output.

In contrast to the observability notions for systems given by ODEs, there are many conceptually different observability definitions for DAE systems (even in the regular case). We first present the most intuitive observability notions and will later present and discuss the remaining observability concepts.

3.1 Behavioral, Impulse and Strong Observability

For the definition of behavioral observability, we follow [70, Definition 5.3.2] and for impulse observability we are inspired by [76, Definition 5.2.1].

Definition 3.1 The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is called

(a) *behaviorally observable*

$$:\Longleftrightarrow \ \forall\, (x^1, u, y), (x^2, u, y) \in \mathfrak{B}_{[E,A,B,C,D]} : \quad x^1 \overset{\text{a.e.}}{=} x^2,$$

(b) *impulse observable*

$$:\Longleftrightarrow \ \forall\, (x^1, u, y), (x^2, u, y) \in \mathfrak{B}^{\text{ITP}}_{[E,A,B,C,D]} : \quad x^1[0] = x^2[0],$$

where $D[0]$ is the impulsive part of $D \in \mathscr{D}'_{\text{pw}\mathscr{C}^\infty}$ at $t = 0$, see (2.1),

(c) *strongly observable*

$$:\Longleftrightarrow \ \forall\, (x^1, u, y), (x^2, u, y) \in \mathfrak{B}^{\text{ITP}}_{[E,A,B,C,D]} : \quad (x^1)_{[0,\infty)} = (x^2)_{[0,\infty)}.$$

The intuition behind these observability notions is as follows: In general, a system is called observable if the knowledge of the external signals allows the reconstruction of the inner state. This idea is directly formalized by the behavioral observability definition. Note that the forthcoming observability characterization will yield that the system $[E, A, B, C, D]$ is behaviorally observable (defined for weak solutions) if, and only if, it is behaviorally observable in a distributional solution framework, i.e.,

$$\forall\, (x^1, u, y), (x^2, u, y) \in \mathfrak{B}^{\mathscr{D}'_{\text{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]} : \quad x^1 = x^2.$$

Most physical systems are turned on at some time (i.e., the system does not run for an infinitely long time) and it is well known that DAE systems (in contrast to ODE systems) exhibit new phenomena in response to inconsistent initial values. In particular, inconsistent initial values may lead to Dirac impulses in the solution and an important question is, whether these Dirac impulses in the state variable can uniquely be determined from the measurement of the external signals. This property is formalized by the impulse observability definition.

Example 3.1 Consider the DAE

$$\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \dot{x} = x, \quad y = Cx.$$

The only solution (also in a distributional solution framework) is $x \equiv \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$, in particular $\left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ is the only consistent initial value and the DAE is behaviorally observable. The ITP with initial value $x(0^-) = \left(\begin{smallmatrix} x_0^1 \\ x_0^2 \end{smallmatrix}\right)$ leads to the impulsive term $x[0] = \left(\begin{smallmatrix} 0 \\ x_0^1 \delta_0 \end{smallmatrix}\right)$. Hence, $C = [0, 1]$ makes the DAE impulse observable (because then $y[0] = x_0^1 \delta_0$ uniquely determines $x[0]$), while $C = [1, 0]$ makes the DAE not impulse observable (because the impulse in $x[0]$ is not visible in the output y).

The following result is an immediate consequence of Definition 3.1.

Proposition 3.2 *The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is strongly observable if, and only if, it is behaviorally and impulse observable.*

Linearity of the system (1.1) implies that $\mathfrak{B}_{[E,A,B,C,D]}$ and $\mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]}$ are vector spaces. As an immediate consequence, we can characterize the previously introduced notions by the following slightly simpler properties.

Lemma 3.3 (Distinction from Zero) *The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is*

(a) behaviorally observable

$$\iff \quad \forall (x, 0, 0) \in \mathfrak{B}_{[E,A,B,C,D]} : \quad x \overset{\text{a.e.}}{=} 0,$$

(b) impulse observable

$$\iff \quad \forall (x, 0, 0) \in \mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]} : \quad x[0] = 0,$$

(c) strongly observable

$$\iff \quad \forall (x, 0, 0) \in \mathfrak{B}^{\mathrm{ITP}}_{[E,A,B,C,D]} : \quad x_{[0,\infty)} = 0.$$

Corollary 3.4 *The DAE system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is behaviorally, impulse, or strongly observable if, and only if, the DAE system $[E, A, 0_{l\times 0}, C, 0_{p\times 0}]$ with corresponding DAE*

$$\tfrac{\mathrm{d}}{\mathrm{d}t} Ex = Ax, \quad y = Cx$$

has the respective property.

The above result justifies to restrict our attention in the following to the system class

$$\mathscr{O}_{l,n,p} := \left\{ [E, A, C] \;\middle|\; [E, A, 0_{l\times 0}, C, 0_{p\times 0}] \in \Sigma_{l,n,0,p} \right\}$$

with the corresponding behaviors

$$\mathfrak{B}_{[E,A,C]} := \mathfrak{B}_{[E,A,0_{l\times 0},C,0_{p\times 0}]}, \qquad \mathfrak{B}^{\mathrm{ITP}}_{[E,A,C]} := \mathfrak{B}^{\mathrm{ITP}}_{[E,A,0_{l\times 0},C,0_{p\times 0}]}$$

and the question whether a zero output implies a trivial state (behavioral observability) or an impulse free response to any inconsistent initial value (impulse observability). Analogously, we set

$$\mathfrak{B}^{\mathscr{D}'}_{[E,A,C]} := \mathfrak{B}^{\mathscr{D}'}_{[E,A,0_{l\times 0},C,0_{p\times 0}]} \qquad \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,C]} := \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,0_{l\times 0},C,0_{p\times 0}]},$$
$$\mathfrak{B}^{\delta_{z_0}}_{[E,A,C]} := \mathfrak{B}^{\delta_{z_0}}_{[E,A,0_{l\times 0},C,0_{p\times 0}]}.$$

Note that we allow $p = 0$, i.e., DAE systems without an output. At first glance this might look meaningless in the context of observability, however, the DAE $0 = x$ (for example) is behaviorally and impulse observable, although there is no output. This is also related to the fact that adding or removing zero output equations $y = 0$ does not change the observability properties.

3.2 Observability at Infinity and Complete Observability

Now we introduce two observability notions which will later on turn out to be stronger than impulse and strong observability, resp. To this end we seek a definition in terms of "observability of excitations" which is related to input observability as in [41]. The idea is that a Dirac impulse at time $t = 0$ is applied to the system's equations weighted by some constants represented by a vector $z_0 \in \mathbb{R}^l$.

Definition 3.2 The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is called

(a) observable at infinity

$$:\Longleftrightarrow\ \forall\, z_0^1, z_0^2 \in \mathbb{R}^l :$$
$$\Big[\ (x^1, u, y) \in \mathfrak{B}^{\delta_{z_0^1}}_{[E,A,B,C,D]}\ \wedge\ (x^2, u, y) \in \mathfrak{B}^{\delta_{z_0^2}}_{[E,A,B,C,D]}\ \wedge\ Ex^1 = Ex^2$$
$$\Rightarrow\quad z_0^1 = z_0^2\ \wedge\ x^1[0] = x^2[0]\ \Big],$$

(b) completely observable

$$:\Longleftrightarrow\ \forall\, z_0^1, z_0^2 \in \mathbb{R}^l :$$
$$\Big[\ (x^1, u, y) \in \mathfrak{B}^{\delta_{z_0^1}}_{[E,A,B,C,D]}\ \wedge\ (x^2, u, y) \in \mathfrak{B}^{\delta_{z_0^2}}_{[E,A,B,C,D]}$$
$$\Rightarrow\quad z_0^1 = z_0^2\ \wedge\ x^1[0] = x^2[0]\ \Big].$$

It is obvious that complete observability implies observability at infinity. The forthcoming observability characterizations will further yield that a system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is completely observable if, and only if, it is behaviorally observable and observable at infinity.

By using that for all $z_0^1, z_0^2 \in \mathbb{R}^l$ we have from Remark 2.1 that

$$\mathfrak{B}^{\delta z_0^1}_{[E,A,B,C,D]} + \mathfrak{B}^{\delta z_0^1}_{[E,A,B,C,D]} = \mathfrak{B}^{\delta(z_0^1+z_0^2)}_{[E,A,B,C,D]},$$

we can conclude that observability at infinity and complete observability can be characterized by the conditions from Definition 3.2 in which z_0^2, u and y are trivial (cf. Lemma 3.3).

Lemma 3.5 (Distinction from Zero II) *The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is*

(a) observable at infinity

$$\iff \forall\, z_0 \in \mathbb{R}^l : \left[(x, 0) \in \mathfrak{B}^{\delta z_0}_{[E,A,C]} \ \wedge\ Ex = 0 \ \Rightarrow\ z_0 = 0 \ \wedge\ x[0] = 0 \right],$$

(b) completely observable

$$\iff \forall\, z_0 \in \mathbb{R}^l : \left[(x, 0) \in \mathfrak{B}^{\delta z_0}_{[E,A,C]} \ \Rightarrow\ z_0 = 0 \ \wedge\ x[0] = 0 \right].$$

An immediate consequence is that we can again restrict out attention to systems in $\mathscr{O}_{[E,A,C]}$.

Example 3.2 Consider the DAE

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \dot{x} = x + \delta z_0, \quad y = Cx.$$

If $C = I_2$, then $y = 0$ implies $x = 0$ and thus $z_0 = 0$, i.e., the DAE is completely observable. If we choose $C = [0, 1]$, then $x_2 = y = 0$ implies $z_0 = \left(\begin{smallmatrix} z_1 \\ 0 \end{smallmatrix}\right)$ and a solution exists even for $z_1 \neq 0$. Therefore, the DAE is not completely observable. However, if additionally $Ex = 0$, then $x_1 = 0$ and thus $z_1 = 0$, so we have observability at infinity. If we choose $C = 0$, then $y = 0$ and $Ex = 0$ imply $x_1 = 0$, but for $z_0 = \left(\begin{smallmatrix} 0 \\ z_2 \end{smallmatrix}\right)$ with $z_2 \neq 0$ a solution is given by $x = \left(\begin{smallmatrix} 0 \\ -z_2\delta \end{smallmatrix}\right)$, whence the DAE is not observable at infinity.

3.3 Relevant State Observability

A classical result of control theory of linear time-invariant ODE systems is that controllability and observability are dual in a certain sense, see e.g. [79, Sect. 3.3]. We will see in Sect. 5 that for *regular* systems the concepts of behavioral, impulse,

strong and complete observability and observability at infinity, are indeed dual to the respective controllability concepts as introduced in [15]. The singular case however exhibits a certain lack of duality. To account for this we introduce the weaker concepts of *relevant state (RS) behavioral, impulse, strong and complete observability and RS observability at infinity*, which will prove to be dual to the respective controllability concepts in Sect. 5. These concepts refer, as their name suggests, to observability up to "a certain part of the state", i.e., state variables that are not uniquely determined by their past, input and output. The reason is that, from a physical point of view, these states only appear in the model because of "bad design" and the system should not be deemed unobservable because it contains free variables. The definitions are as follows.

Definition 3.3 The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is called

(a) *RS behaviorally observable*

$$:\Longleftrightarrow \quad \forall\, (x^1,u,y), (x^2,u,y) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]}\ \exists\, (x^3,u,y) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]} :$$
$$(x^3)_{(-\infty,0)} = (x^1)_{(-\infty,0)} \ \wedge\ (x^3)_{(0,\infty)} = (x^2)_{(0,\infty)},$$

(b) *RS impulse observable*

$$:\Longleftrightarrow \quad \forall\, x^1_0, x^2_0 \in \mathbb{R}^n :$$
$$\Big[(x^1,u,y) \in \mathfrak{B}^{\delta E x^1_0}_{[E,A,B,C,D]} \ \wedge\ (x^2,u,y) \in \mathfrak{B}^{\delta E x^2_0}_{[E,A,B,C,D]} \ \wedge\ Ex^1 = Ex^2$$
$$\Rightarrow \quad Ex^1_0 = Ex^2_0\Big],$$

(c) *RS strongly observable*

$$:\Longleftrightarrow \quad \forall\, x^1_0, x^2_0 \in \mathbb{R}^n :$$
$$\Big[(x^1,u,y) \in \mathfrak{B}^{\delta E x^1_0}_{[E,A,B,C,D]} \ \wedge\ (x^2,u,y) \in \mathfrak{B}^{\delta E x^2_0}_{[E,A,B,C,D]}$$
$$\Rightarrow \quad Ex^1_0 = Ex^2_0\Big],$$

(d) *RS observable at infinity*

$$:\Longleftrightarrow \quad \forall\, z^1_0, z^2_0 \in \mathbb{R}^l :$$
$$\Big[(x^1,u,y) \in \mathfrak{B}^{\delta z^1_0}_{[E,A,B,C,D]} \ \wedge\ (x^2,u,y) \in \mathfrak{B}^{\delta z^2_0}_{[E,A,B,C,D]} \ \wedge\ Ex^1 = Ex^2$$
$$\Rightarrow \quad z^1_0 = z^2_0\ \Big],$$

(e) *RS completely observable*

$$:\Longleftrightarrow \quad \forall\, z_0^1, z_0^2 \in \mathbb{R}^l :$$
$$\Big[(x^1, u, y) \in \mathfrak{B}^{\delta z_0^1}_{[E,A,B,C,D]} \ \wedge\ (x^2, u, y) \in \mathfrak{B}^{\delta z_0^2}_{[E,A,B,C,D]}$$
$$\Rightarrow \quad z_0^1 = z_0^2 \Big].$$

It is clear that RS strong (complete) observability implies RS impulse observability (RS observability at infinity). The forthcoming observability characterizations will further yield that a system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is RS strongly observable if, and only if, it is RS behaviorally observable and RS impulse observable; it is RS completely observable if, and only if, it is RS behaviorally observable and RS observable at infinity.

Remark 3.1 One may wonder why the definition of RS behavioral observability is given in terms of the distributional behavior $\mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,B,C,D]}$ instead of the behavior $\mathfrak{B}_{[E,A,B,C,D]}$. The reason is that the concatenation of two solutions will in general introduce a jump at $t = 0$. For ODEs any concatenation with a jump in the state variable cannot be a solution, but for DAEs this is not true in general. However, the presence of a jump makes it necessary to view the DAE in a distributional solution space; in particular, Dirac impulses at $t = 0$ may occur in the solution in response to the jump. Nevertheless, the definition of RS behavioral observability can also be given in terms of $\mathfrak{B}_{[E,A,B,C,D]}$ as follows

$$\forall (x^1, u, y), (x^2, u, y) \in \mathfrak{B}_{[E,A,B,C,D]}\ \boxed{\exists T > 0}\ \exists (x^3, u, y) \in \mathfrak{B}_{[E,A,B,C,D]} :$$
$$(x^3)_{(-\infty,0)} \overset{\text{a.e.}}{=} (x^1)_{(-\infty,0)} \ \wedge\ (x^3)_{(T,\infty)} \overset{\text{a.e.}}{=} (x^2)_{(T,\infty)},$$

i.e., the concatenation is not instantaneous. Despite the slight technicalities involved, we find the definition via instantaneous concatenability more appealing because it does not introduce the additional concatenation time $T > 0$.

We can conclude that RS behavioral, impulse, strong and complete observability and RS observability at infinity can be characterized by the conditions from Definition 3.3 in which x_0^2, z_0^2, x^2, u and y are trivial (cf. Lemma 3.3).

Lemma 3.6 (Distinction from Zero III) *The system* $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ *is*

(a) RS behaviorally observable

$$\Longleftrightarrow \quad \forall\, (x, 0) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,C]}\ \exists\, (\overline{x}, 0) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}\mathscr{C}^\infty}}_{[E,A,C]} :$$
$$x_{(-\infty,0)} = \overline{x}_{(-\infty,0)} \ \wedge\ \overline{x}_{(0,\infty)} = 0,$$

(b) RS impulse observable

$$\iff \quad \forall\, x_0 \in \mathbb{R}^n : \left[(x,0) \in \mathfrak{B}^{\delta Ex_0}_{[E,A,C]} \ \wedge\ Ex = 0 \quad \Rightarrow \quad Ex_0 = 0\right],$$

(c) RS strongly observable

$$\iff \quad \forall\, x_0 \in \mathbb{R}^n : \left[(x,0) \in \mathfrak{B}^{\delta Ex_0}_{[E,A,C]} \quad \Rightarrow \quad Ex_0 = 0\right],$$

(d) RS observable at infinity

$$\iff \quad \forall\, z_0 \in \mathbb{R}^l : \left[\ (x,0) \in \mathfrak{B}^{\delta z_0}_{[E,A,C]} \ \wedge\ Ex = 0 \quad \Rightarrow \quad z_0 = 0\right],$$

(e) RS completely observable

$$\iff \quad \forall\, z_0 \in \mathbb{R}^l : \left[\ (x,0) \in \mathfrak{B}^{\delta z_0}_{[E,A,C]} \quad \Rightarrow \quad z_0 = 0\right].$$

As a consequence from Lemmas 3.5 and 3.6 we can further state the following implications for the so far introduced observability notions.

Corollary 3.7 *The following implications hold true for any system in* $\Sigma_{l,n,m,p}$*:*

(i) *behaviorally observable* $\implies$ *RS behaviorally observable,*
(ii) *observable at infinity* $\implies$ *RS observable at infinity* $\implies$ *RS impulse observable,*
(iii) *completely observable* $\implies$ *RS completely observable* $\implies$ *RS strongly observable.*

Note that it is still not clear (however true) that impulse (strong) observability implies RS impulse (strong) observability. To show this we need the characterizations in terms of the output injection form derived in Sect. 4.

It will later turn out, see Corollary 4.8, that for regular systems the observability concepts from Sects. 3.1 and 3.2 are equivalent to the respective relevant state observability concepts from Definition 3.3. In view of this, Examples 3.1 and 3.2 provide some illustrative examples for the RS observability concepts.

3.4 Comparison of the Concepts with the Literature

We compare the relations of the observability concepts introduced in the present paper to existing notions in the literature in the following list of remarks.

(i) The observability concepts are not consistently treated in the literature. While some authors rely on intuitive extensions of the definition known for ODEs [29, 88], others insist on duality to the known controllability

concepts [31]. Furthermore, one has to pay attention if it is (tacitly) claimed that $[E^\top, C^\top] \in \mathbb{R}^{l\times(n+p)}$ or $[E^\top, A^\top, C^\top] \in \mathbb{R}^{l\times(2n+p)}$ have full rank. Some of the references introduce observability by means of certain rank criteria for the matrices E, A, C. The connection of the observability concepts to linear algebraic properties of E, A and C are highlighted in Sect. 6 (and are partly used to derive the following comparisons).

(ii) For *regular systems* the number of different observability concepts reduces to five by Corollary 4.8. We have the following relationships between the observability notions introduced here and the ones given in the literature:

Concept	Coincides with	Called [...] in
Behavioral obs.	–	Obs. in [29, 88]; R-obs. in [32]; jump obs. in [76]
Impulse obs.	[31, 32, 76]	Obs. at infinity in [3, 53, 82]
Strong obs.	[82]	–
Obs. at infinity	[13, 33]	Dual normalizability in [32]
Complete obs.	[25]	Obs. in [31, 32]

(iii) There is also a significant amount of literature dealing with observability for general DAEs; the relationship to the notions introduced here is as follows:

Concept	Coincides with	Called [...] in
Behavioral obs.	–	Obs. in [70]; right-hand side obs. in [40]; strong almost obs. in [66]
Impulse obs.	[25, 26, 40, 43]	Obs. at infinity [25, 26][2]
Strong obs.	[66]	Obs. in [40, 68]
Obs. at infinity	–	–
Complete obs.	–	Obs. in [35]; str. obs. in [68]; str. compl. obs. in [66]
RS behavioral obs.	–	–
RS impulse obs.	–	–
RS strong obs.	–	Obs. in [10, 66]; weakly obs. in [68]
RS obs. at infinity	–	–
RS complete obs.	–	Strong obs. in [10]; complete obs. in [66][3]

[2] In [25, 26] the notions of impulse observability and observability at infinity are both used for impulse observability.

[3] Note that although the notion of complete observability is used in [66], it is only introduced by a geometric condition and not by a time domain definition.

Observability concepts for general discrete time DAE systems have been introduced and investigated in [7–9].

(iv) Impulse observability and observability at infinity are usually defined by considering distributional solutions of (1.1) (similar to our definitions), see e.g. [31, 43], sometimes called impulsive modes, see [13, 40, 82]. For regular systems, impulse observability was introduced by Verghese et al. [82] (called observability at infinity in this work) as observability of the impulsive modes of the system, and later made more precise by Cobb [31], see also Armentano [3] (who also calls it observability at infinity) for a more geometric point of view. In [82] the authors also develop the notion of strong observability as impulse observability with, additionally, "observability in the sense of the regular theory".

The name "observability at infinity" comes from the claim that the system has no infinite unobservable modes: speaking in terms of rank criteria (see also Sect. 6) the system $[E, A, C] \in \mathscr{O}_{l,n,p}$ is said to have an unobservable mode at $\frac{\alpha}{\beta}$ if, and only if, $\operatorname{rk}[\alpha E^\top + \beta A^\top, C^\top] < \operatorname{rk}[E^\top, A^\top, C^\top]$ for some $\alpha, \beta \in \mathbb{C}$. If $\beta = 0$ and $\alpha \neq 0$, then the unobservable mode is infinite. Observability at infinity was introduced by Rosenbrock [73]—although he does not use this phrase—as the absence of infinite output decoupling zeros. Later, Cobb [31] compared the concepts of impulse observability and observability at infinity, see [31, Theorem 10]; the notions we use in the present paper go back to the distinction in this work.

(v) Observability concepts with a distributional solution setup have also been considered in [31, 66]. Distributional solutions for time-invariant DAEs have been considered by Cobb [30] and Geerts [37, 38] and for time-varying DAEs by Rabier and Rheinboldt [72], and by Kunkel and Mehrmann [52]. In the present paper we use the approach by Trenn [76, 77]. The latter framework is also the basis for several observability concepts for switched DAE systems [69].

(vi) Behavioral observability was first defined by Yip and Sincovec [88], although merely called observability, as the dual of R-controllability for regular DAEs. They define observability essentially as the state x being computable from the input u, the output y, and the system data E, A, C. This is equivalent to classical observability of the ODE part of the system. Furthermore, it is equivalent to trivial output implying trivial state and hence to behavioral observability. The same approach is followed in [29] and it is emphasized that this "obvious extension of observability is not the dual of complete controllability". We stress that it is not even the dual of R-controllability when it comes to general DAE systems; however, as it will be shown in Corollary 5.1, the dual of R-controllability is RS behavioral observability.

In the context of the behavioral approach, behavioral observability was introduced in [70], but it is different to RS behavioral observability. These concepts are suitable for generalizations in various directions, see e.g. [27, 42, 85]. Having found the behavior of the considered system one can take

over the definition of RS behavioral observability without the need for any further changes. From this point of view this appears to be the most natural of the observability concepts. However, this concept also seems to be the least regarded in the DAE literature.

(vii) The observability theory of DAE systems can also be treated with the theory of differential inclusions [4, 5] as shown by Frankowska [35]. However, Frankowska assumes observability at infinity in order to derive duality between controllability and observability as introduced in [35].

4 Output Injection Normal Form

In this section we recall the concept of output injection for DAE systems and show that it induces an equivalence relation on $\mathscr{O}_{l,n,p}$. Then we state a normal form under this equivalence relation, which we use to characterize the observability concepts introduced in Sect. 3.

4.1 Output Injection Equivalence and Normal Form

Output injection is usually understood as the addition of the output y of the system, weighted by some matrix $L \in \mathbb{R}^{l\times p}$, to the right-hand side of the systems equation. Since $y(t) = Cx(t)$, the resulting system has the form

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}Ex(t) &= (A + LC)x(t), \\ y(t) &= Cx(t). \end{aligned} \tag{4.1}$$

Output injection can be understood as an algebraic transformation (more precisely: a group operation) within the set $\mathscr{O}_{l,n,p}$:

$$\begin{bmatrix} E \\ A + LC \\ C \end{bmatrix} = \begin{bmatrix} I_l & 0 & 0 \\ 0 & I_l & L \\ 0 & 0 & I_p \end{bmatrix} \begin{bmatrix} E \\ A \\ C \end{bmatrix}.$$

Allowing also for state space and output space transformations leads to the following notion of output injection equivalence.

Definition 4.1 (Output Injection Equivalence) Two systems $[E_i, A_i, C_i] \in \mathscr{O}_{l,n,p}$, $i = 1, 2$, are called *output injection equivalent* (OI equivalent) if, and only if,

$$\begin{aligned} &\exists\, W \in \mathbf{Gl}_l(\mathbb{R}), T \in \mathbf{Gl}_n(\mathbb{R}), V \in \mathbf{Gl}_p(\mathbb{R}), L \in \mathbb{R}^{l\times p} : \\ &[E_1\,,\, A_1\,,\, C_1] = [WE_2T\,,\, WA_2T + LC_2T\,,\, VC_2T]; \end{aligned} \tag{4.2}$$

we write

$$[E_1\,,\,A_1\,,\,C_1] \overset{W,T,V,L}{\sim_{OI}} [E_2\,,\,A_2\,,\,C_2]\,. \tag{4.3}$$

OI equivalence seems to have been first considered by Morse [64] for linear ODE systems, and it has been termed a "nonphysically realizable transformation". For DAE systems, OI equivalence was first exploited by Karcanias [47] using the framework introduced by Morse.

Clearly, multiplying the first equation in the DAE (1.1) from the left with an invertible matrix W does not change the behaviors introduced in Sect. 2 at all and a coordinate transformation of the state via T and the output via V does not qualitatively change the behaviors. Provided that the output is zero, its addition to the state equation certainly does not change the behavior as well. This is made precise in the following.

Lemma 4.1 (Behavior and Output Injection) *If $[E_1, A_1, C_1], [E_2, A_2, C_2] \in \mathscr{O}_{l,n,p}$ are OI equivalent for $W \in \mathbf{Gl}_l(\mathbb{R})$, $T \in \mathbf{Gl}_n(\mathbb{R})$, $V \in \mathbf{Gl}_p(\mathbb{R})$, $L \in \mathbb{R}^{l\times p}$ as in (4.3), then we have*

(a) $(x,0) \in \mathfrak{B}_{[E_1,A_1,C_1]} \;\Leftrightarrow\; (Tx,0) \in \mathfrak{B}_{[E_2,A_2,C_2]}$.
(b) $(x,0) \in \mathfrak{B}^{\mathscr{D}'}_{[E_1,A_1,C_1]} \;\Leftrightarrow\; (Tx,0) \in \mathfrak{B}^{\mathscr{D}'}_{[E_2,A_2,C_2]}$.
(c) $(x,0) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E_1,A_1,C_1]} \;\Leftrightarrow\; (Tx,0) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E_2,A_2,C_2]}$.
(d) $(x,0) \in \mathfrak{B}^{\mathrm{ITP}}_{[E_1,A_1,C_1]} \;\Leftrightarrow\; (Tx,0) \in \mathfrak{B}^{\mathrm{ITP}}_{[E_2,A_2,C_2]}$.
(e) $\forall\, z_0 \in \mathbb{R}^l :\; (x,0) \in \mathfrak{B}^{\delta z_0}_{[E_1,A_1,C_1]} \;\Leftrightarrow\; (Tx,0) \in \mathfrak{B}^{\delta W^{-1} z_0}_{[E_2,A_2,C_2]}$.
In particular, $(x,0) \in \mathfrak{B}^{\delta z_0}_{[E_1,A_1,C_1]}$ *satisfies*

$$E_1 x = 0 \;\Leftrightarrow\; E_2(Tx) = 0.$$

Finally, due to Lemmas 3.3, 3.5 and 3.6 we can restrict our attention to the solutions which produce a zero output. In summary we have the following result.

Proposition 4.2 (Invariance Under Output Injection) *On the set $\mathscr{O}_{l,n,p}$, behavioral, impulse, strong and complete observability, observability at infinity and the corresponding relevant state RS concepts are all invariant under OI equivalence.*

Proposition 4.2 allows to analyze the observability concepts by means of a normal form under OI equivalence. In order to present such a normal form, we need to introduce the following notation: for $k \in \mathbb{N}$ let

$$N_k = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & & 1\;0 \end{bmatrix} \in \mathbb{R}^{k\times k}, \quad K_k = \begin{bmatrix} 1\;0 & & \\ & \ddots & \\ & & 1\;0 \end{bmatrix}, \; L_k = \begin{bmatrix} 0\;1 & & \\ & \ddots & \\ & & 0\;1 \end{bmatrix} \in \mathbb{R}^{(k-1)\times k}.$$

Further, let $e_i^{[k]} \in \mathbb{R}^k$ be the ith canonical unit vector, and, for some multi-index $\alpha = (\alpha_1, \ldots, \alpha_r) \in \mathbb{N}^r$, we define

$$\begin{aligned} N_\alpha &= \operatorname{diag}(N_{\alpha_1}, \ldots, N_{\alpha_r}) \in \mathbb{R}^{|\alpha| \times |\alpha|}, \\ K_\alpha &= \operatorname{diag}(K_{\alpha_1}, \ldots, K_{\alpha_r}) \in \mathbb{R}^{(|\alpha| - \ell(\alpha)) \times |\alpha|}, \\ L_\alpha &= \operatorname{diag}(L_{\alpha_1}, \ldots, L_{\alpha_r}) \in \mathbb{R}^{(|\alpha| - \ell(\alpha)) \times |\alpha|}, \\ E_\alpha &= \operatorname{diag}(e_{\alpha_1}^{[\alpha_1]}, \ldots, e_{\alpha_r}^{[\alpha_r]}) \in \mathbb{R}^{|\alpha| \times \ell(\alpha)}. \end{aligned}$$

We are now in a position to derive a normal form under OI equivalence for systems $[E, A, C] \in \mathscr{O}_{l,n,p}$. We stress that we use the terminus "normal form" in a colloquial way to distinguish it from the mathematical terminus "canonical form". Whether the following form is a normal or canonical form is clarified in Remark 4.3.

Theorem 4.3 (Normal Form Under OI Equivalence) *Let $[E, A, C] \in \mathscr{O}_{l,n,p}$. Then there exist $W \in \mathbf{Gl}_l(\mathbb{R})$, $T \in \mathbf{Gl}_n(\mathbb{R})$, $V \in \mathbf{Gl}_p(\mathbb{R})$, $L \in \mathbb{R}^{l \times p}$ such that*

$$[E, A, C] \overset{W,T,V,L}{\sim_{OI}} \left[\begin{bmatrix} I_{|\alpha|} & 0 & 0 & 0 & 0 & 0 \\ 0 & K_\beta^\top & 0 & 0 & 0 & 0 \\ 0 & 0 & L_\gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & K_\varepsilon & 0 & 0 \\ 0 & 0 & 0 & 0 & N_\kappa^\top & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{n_{\overline{\sigma}}} \end{bmatrix}, \begin{bmatrix} N_\alpha & 0 & 0 & 0 & 0 & 0 \\ 0 & L_\beta^\top & 0 & 0 & 0 & 0 \\ 0 & 0 & K_\gamma & 0 & 0 & 0 \\ 0 & 0 & 0 & L_\varepsilon & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{|\kappa|} & 0 \\ 0 & 0 & 0 & 0 & 0 & A_{\overline{\sigma}} \end{bmatrix}, \begin{bmatrix} E_\alpha^\top & 0 & 0 & 0\,0\,0 \\ 0 & 0 & E_\gamma^\top & 0\,0\,0 \\ 0 & 0 & 0 & 0\,0\,0 \end{bmatrix} \right], \tag{4.4}$$

for some multi-indices $\alpha, \beta, \gamma, \varepsilon, \kappa$ and a matrix $A_{\overline{\sigma}} \in \mathbb{R}^{n_{\overline{\sigma}} \times n_{\overline{\sigma}}}$.

Proof It is easy to see that $[E_1, A_1, C_1], [E_2, A_2, C_2] \in \mathscr{O}_{n,m,p}$ are OI equivalent if, and only if, $[E_1^\top, A_1^\top, C_1^\top]$ and $[E_1^\top, A_1^\top, C_1^\top]$ with corresponding DAEs

$$E_1^\top \dot{z} = A_1^\top z + C_1^\top u \quad \text{and} \quad E_2^\top \dot{z} = A_2^\top z + C_2^\top u$$

are feedback equivalent in the sense of [15, Definition 3.1]. Hence the transposed feedback normal form derived in [15, Theorem 3.3] is a normal form under OI equivalence. □

Remark 4.1 (Duality for DAEs) It should be noted that although we utilized a "duality" argument in the proof of Theorem 4.3, we have not really defined duality for DAEs or its corresponding behaviors yet. In fact, the proof of Theorem 4.3 just utilizes a normal form for matrix triples and is not related to certain solution concepts for DAEs. Further duality results for DAEs are presented in Sect. 5.

The interpretation of the OI normal form (4.4), in terms of solutions of DAEs is as follows: $(x, y) \in \mathfrak{B}_{[E,A,C]}$ if, and only if,

$$\begin{aligned}\left(x_{co}(\cdot)^\top, x_o(\cdot)^\top, x_{uo}(\cdot)^\top, x_u(\cdot)^\top, x_f(\cdot)^\top, x_{\overline{o}}(\cdot)^\top\right)^\top &:= Tx(\cdot),\\ \left(y_{co}(\cdot)^\top, y_{uo}(\cdot)^\top, y_{\overline{o}}(\cdot)^\top\right)^\top &:= Vy(\cdot),\end{aligned}$$

with

$$x_{co}(\cdot) = \begin{pmatrix} x_{co[1]}(\cdot) \\ \vdots \\ x_{co[\ell(\alpha)]}(\cdot) \end{pmatrix}, \quad y_{co}(\cdot) = \begin{pmatrix} y_{co[1]}(\cdot) \\ \vdots \\ y_{co[\ell(\alpha)]}(\cdot) \end{pmatrix}, \quad x_o(\cdot) = \begin{pmatrix} x_{o[1]}(\cdot) \\ \vdots \\ x_{o[\ell(\beta)]}(\cdot) \end{pmatrix},$$

$$x_{uo}(\cdot) = \begin{pmatrix} x_{uo[1]}(\cdot) \\ \vdots \\ x_{uo[\ell(\gamma)]}(\cdot) \end{pmatrix}, \quad y_{uo}(\cdot) = \begin{pmatrix} y_{uo[1]}(\cdot) \\ \vdots \\ y_{uo[\ell(\gamma)]}(\cdot) \end{pmatrix}, \quad x_u(\cdot) = \begin{pmatrix} x_{u[1]}(\cdot) \\ \vdots \\ x_{u[\ell(\epsilon)]}(\cdot) \end{pmatrix},$$

$$x_f(\cdot) = \begin{pmatrix} x_{f[1]}(\cdot) \\ \vdots \\ x_{f[\ell(\kappa)]}(\cdot) \end{pmatrix}$$

solves the decoupled DAEs

$$\frac{\mathrm{d}}{\mathrm{d}t} x_{co[i]} = N_{\alpha_i} x_{co[i]}, \quad y_{co[i]} = \left(e_{\alpha_i}^{[\alpha_i]}\right)^\top x_{co[i]}, \quad \text{for } i = 1, \ldots, \ell(\alpha), \tag{4.5a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} K_{\beta_i}^\top x_{o[i]} = L_{\beta_i}^\top x_{o[i]}, \quad \text{for } i = 1, \ldots, \ell(\beta), \tag{4.5b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} L_{\gamma_i} x_{uo[i]} = K_{\gamma_i} x_{uo[i]}, \quad y_{uo[i]} = \left(e_{\gamma_i}^{[\gamma_i]}\right)^\top x_{uo[i]}, \quad \text{for } i = 1, \ldots, \ell(\gamma), \tag{4.5c}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} K_{\varepsilon_i} x_{u[i]} = L_{\varepsilon_i} x_{u[i]}, \quad \text{for } i = 1, \ldots, \ell(\varepsilon), \tag{4.5d}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} N_{\kappa_i}^\top x_{f[i]} = x_{f[i]}, \quad \text{for } i = 1, \ldots, \ell(\kappa), \tag{4.5e}$$

$$\frac{\mathrm{d}}{\mathrm{d}t} x_{\overline{o}} = A_{\overline{o}} x_{\overline{o}}, \qquad y_{\overline{o}} = 0. \tag{4.5f}$$

An analogous interpretation holds for $(x,u) \in \mathfrak{B}^{\mathscr{D}'}_{[E,A,C]}$ and $(x,u) \in \mathfrak{B}^{\mathscr{D}'_{\mathrm{pw}}\mathscr{C}^\infty}_{[E,A,C]}$. For $(x,u) \in \mathfrak{B}^{\mathrm{ITP}}_{[E,A,C]}$ the equations in (4.5) have to be restricted to the interval $[0,\infty)$ and for $(x,u) \in \mathfrak{B}^{\delta z_0}_{[E,A,C]}$ an appropriate term $\delta\tilde{z}_0$ has to be added to the respective state space equations in (4.5).

Remark 4.2 (Regular Case) In general, the OI normal form (4.4) for a *regular* system $[E,A,C] \in \mathscr{O}_{n,n,p}$, that is a system with a regular pencil $sE-A$, is *not regular*. For example, the regular system

$$[E,A,C] = \left[\begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\1&1\end{bmatrix}, [0,1]\right]$$

has the nonregular OI normal form

$$\left[\begin{bmatrix}1&0\\0&0\end{bmatrix}, \begin{bmatrix}0&0\\1&0\end{bmatrix}, [0,1]\right]$$

which consists of a 2×1 β-block and a 0×1 γ-block. However, the OI normal form of a regular system cannot have underdetermined DAEs of the form (4.5d), i.e., $\ell(\varepsilon) = 0$, because these DAEs would correspond to underdetermined parts in the original coordinates as well (because the nonexisting output cannot "fix" this nonuniqueness).

Remark 4.3 (Canonical and Normal Form) To explain the difference between our notions of normal and canonical form, recall the definition of a canonical form: given a group G, a set $\mathscr{S}$, and a group action $\alpha : G \times \mathscr{S} \to \mathscr{S}$ which defines an equivalence relation $s \overset{\alpha}{\sim} s'$ if, and only if, there exists $U \in G$ such that $\alpha(U,s) = s'$. Then a map $\Gamma : \mathscr{S} \to \mathscr{S}$ is called a *canonical form for* α [22] if, and only if,

$$\forall\, s,s' \in \mathscr{S} \;:\; \Gamma(s) \overset{\alpha}{\sim} s \quad \wedge \quad \Big[s \overset{\alpha}{\sim} s' \Leftrightarrow \Gamma(s) = \Gamma(s')\Big].$$

Therefore, the set $\mathscr{S}$ is divided into disjoint orbits (i.e., equivalence classes) and the mapping Γ picks a unique representative in each equivalence class. In the setup of OI equivalence, the group is $G = \mathbf{Gl}_l(\mathbb{R}) \times \mathbf{Gl}_n(\mathbb{R}) \times \mathbf{Gl}_p(\mathbb{R}) \times \mathbb{R}^{l\times p}$, the considered set is $\mathscr{S} = \mathscr{O}_{l,n,p}$ and the group action

$$\alpha\big((W,T,V,L),[E,A,C]\big) = [WET, WAT + LCT, VCT]$$

corresponds to $\overset{W,T,V,L}{\sim}_{OI}$. However, Theorem 4.3 does not provide a mapping Γ because the matrix $A_{\overline{o}}$ is not uniquely specified. This means that the form (4.4) is not a unique representative within the equivalence class and hence it is not a canonical form.

However, the OI normal form (4.4) is very close to a canonical form in the following sense. By a further (in general complex-valued) transformation we may put $A_{\overline{\sigma}}$ into Jordan canonical form. If the entries of the multi-indices $\alpha, \beta, \gamma, \varepsilon, \kappa$ are in non-decreasing order and in the Jordan canonical form of $A_{\overline{\sigma}}$ the Jordan blocks are ordered non-decreasing in size and lexicographically with respect to the corresponding eigenvalues if the blocks have the same size, then the OI normal form (4.4) is a canonical form.

Summarizing, the form (4.4) is not a canonical form but can be transformed into a canonical form. We therefore call the form as it stands a *normal form*.

Remark 4.4 (Canonical Form Under Output Injection and State Feedback) A combination of the OI normal form with the feedback form from [56] (see also [15, Theorem 3.3]) leads to a canonical from of systems $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ under state space transformation, input space transformation, output space transformation, proportional output injection, proportional state feedback and transformation of the codomain of the state (i.e., left transformation of E, A, B) which was derived in [55]. However, this form is not suitable for either the analysis of controllability or observability, since it is necessary to apply state feedback and output injection simultaneously to obtain the canonical form; but controllability is not invariant under output injection and observability is not invariant under state feedback.

4.2 Characterization of Behavioral, Impulse and Strong Observability

Based on the OI normal form we will now present the characterization of behavioral, impulse and strong observability. To this end we first present the observability properties of each of the individual decoupled DAE systems in (4.5).

Lemma 4.4 *Consider the decoupled DAEs* (4.5) *resulting from the OI normal form. Then the DAEs*

(4.5a) *are always behaviorally, impulse and strongly observable.*
(4.5b) *are always behaviorally, impulse and strongly observable.*
(4.5c) *are always behaviorally observable; they are impulse and strongly observable if, and only if,* $|\gamma| = \ell(\gamma)$*, i.e.,* $\gamma_i = 1$ *for all* $i = 1, \ldots, \ell(\gamma)$.
(4.5d) *are neither behaviorally, impulse nor strongly observable.*
(4.5e) *are always behaviorally observable; they are impulse and strongly observable if, and only if,* $|\kappa| = \ell(\kappa)$.
(4.5f) *are never behaviorally and strongly observable and always impulse observable.*

Proof It suffices to consider behavioral and impulse observability, because the corresponding characterization for strong observability follows trivially from the combination of the characterizations of behavioral and impulse observability.

(4.5a): The solutions of the ODE with size $k \times k$

$$\dot{x} = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & 1 & 0 \end{bmatrix} x$$
$$y = [0, \ldots, 0, 1]x$$

satisfy $x = (y^{(k-1)}, y^{(k-2)}, \ldots, \dot{y}, y)^\top$, hence a zero output implies $x = 0$. For the corresponding ODE-ITP it is easy to see (cf. [76, Theorem 3.3]) that all solutions x exhibit no jumps and no impulses at $t = 0$, hence (irrespective of the actual output) it holds that $x[0] = 0$.

(4.5b): DAEs of size $k \times (k-1)$ of the form

$$\begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & 1 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & 0 & 1 \end{bmatrix} x \tag{4.6}$$

can be interpreted as DAEs of the form (4.5a) with size $(k-1) \times (k-1)$, where the last state variable x_{k-1} is equal to a zero output. Hence the same arguments as above show behavioral and impulse observability.

(4.5c): The solutions of the DAE with size $(k-1) \times k$

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} \dot{x} = \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} x$$
$$y = [0, \ldots, 0, 1]x \tag{4.7}$$

are given by $x \overset{\text{a.e.}}{=} (y^{(k-1)}, y^{(k-2)}, \ldots, \dot{y}, y)^\top$. Hence a zero output implies a zero state, which shows behavioral observability. If $k = 1$, then the DAE-ITP reduces to the output equation $y_{[0,\infty)} = x_{[0,\infty)}$ for the free (scalar) variable x, in particular, $y = 0$ implies $x[0] = 0$ and the DAE for $k = 1$ is impulse observable. If $k > 1$ we now have $(x_k)_{[0,\infty)} = y_{[0,\infty)}$ and $(x_{k-1})_{[0,\infty)} = (\dot{x}_k)_{[0,\infty)}$. In general $x_k(0^-) \neq 0 = y(0^+) = x_k(0^+)$, hence there will be a jump in x_k at $t = 0$ and consequently a Dirac impulse in x_{k-1}. Therefore, a zero output does not imply that $x[0] = 0$ and we do not have impulse observability.

(4.5d): The DAE of size $(k-1) \times k$

$$\begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} x \tag{4.8}$$

contains the free variable x_k (unrelated to the output), hence neither $x = 0$ nor $x[0] = 0$ holds true in general and the DAE cannot be behaviorally or impulse observable.

(4.5e): The solutions of DAEs with size $k \times k$

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \dot{x} = x \tag{4.9}$$

satisfy $x \overset{\text{a.e.}}{=} 0$, hence we have behavioral observability. If $k = 1$ then the corresponding ITP reads as $x_{[0,\infty)} = 0$; in particular $x[0] = 0$ and impulse observability follows. For $k > 0$ we have $(x_k)_{[0,\infty)} = 0$ and $(x_{k-1})_{[0,\infty)} = (\dot{x}_k)_{[0,\infty)}$. In general $x_k(0^-) \neq 0$ and hence there is a jump in x_k and consequently a Dirac impulse in x_{k-1}, i.e., $x[0] \neq 0$, which shows that the DAE is not impulse observable.

(4.5f): The ODE (4.5f) has nontrivial solutions and a zero output, hence it is not behaviorally observable. As already observed for (4.5a) an ODE-ITP does not exhibit jumps or impulses at the initial time, hence $x[0] = 0$ in any case and we have shown impulse observability.

□

4.3 Characterization of Observability at Infinity and Complete Observability

Here we analyze observability at infinity and complete observability by means of the OI normal form.

Lemma 4.5 *Consider the decoupled DAEs (4.5) resulting from the OI normal form. Then the DAEs*

(4.5a) are always completely observable and observable at infinity.
(4.5b) are always completely observable and observable at infinity.
(4.5c) are completely observable and observable at infinity if, and only if, $|\gamma| = \ell(\gamma)$, i.e., $\gamma_i = 1$ for all $i = 1, \ldots, \ell(\gamma)$.
(4.5d) are neither observable at infinity nor completely observable.
(4.5e) are neither observable at infinity nor completely observable.
(4.5f) are never completely observable and always observable at infinity.

Proof In the following we use that complete observability implies observability at infinity.

(4.5a) Any solution of the ODE with size $k \times k$

$$\dot{x} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} x + \delta z_0$$
$$y = [0, \ldots, 0, 1]x$$

satisfies $x_i = \dot{x}_{i+1}$ on the intervals $(-\infty, 0)$ and $(0, \infty)$ for $i = k-1, \ldots, 2, 1$. Hence $y = x_k = 0$ implies $x_i = 0$ on $(-\infty, 0)$ and $(0, \infty)$ for $i = k, k-1, \ldots, 1$. It is easy to see that $x[t] = 0$ for all $t \in \mathbb{R}$, hence $\delta z_0 = \dot{x}[0] = (x(0^+) - x(0^-))\delta = 0$, which implies $z_0 = 0$ and $x[0] = 0$.

(4.5b) Any solution of the DAE with size $k \times (k-1)$ of the form

$$\begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & & 1 \\ & & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & & 0 \\ & & 1 \end{bmatrix} x + \delta z_0$$

satisfies $x_{k-1} = 0$ on $(-\infty, 0)$ and $(0, \infty)$. From $x_i = \dot{x}_{i+1}$ on these two intervals for $i = k-2, \ldots, 2, 1$ it follows that $x = 0$ on $(-\infty, 0)$ and $(0, \infty)$. Hence $\dot{x}_1[0]$ does not contain a Dirac impulse (because x_1 does not have a jump at $t = 0$), and $\dot{x}_1[0] = \delta z_{0,1}$ implies $z_{0,1} = 0$ which in turn implies that $x_1[0] = 0$. Hence, inductively, for $i = 2, 3, \ldots, k-1$ we conclude analogously from $\dot{x}_i[0] = x_{i-1}[0] + \delta z_{0,i}$ that $z_{0,i} = 0$ and $x_i[0] = 0$. This gives $x[0] = 0$ and, finally, $0 = x_{k-1}[0] + \delta z_{0,k}$ implies $z_{0,k} = 0$, which shows that also $z_0 = 0$ is necessary for the existence of a solution.

(4.5c) Consider the DAE of size $(k-1) \times k$

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} \dot{x} = \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} x + \delta z_0$$
$$y = [0, \ldots, 0, 1]x.$$

If $k = 1$, then complete observability follows from $x = y$ (note that there is no z_0 in this case).

Now we consider the case $k \geq 2$: with $z_0 = e_1^{[k-1]} \in \mathbb{R}^{k-1} \setminus \{0\}$, a simple calculation shows that for $x = \delta e_1^{[k]}$ we have $(x, 0) \in \mathfrak{B}^{\delta z_0}_{[L_k, K_k, (e_1^{[k]})^\top]}$. In particular, we have $Ex = 0$ and $x[0] = \delta e_1^{[k]} \neq 0$, whence the system is not observable at infinity.

(4.5d) Consider the DAE of size $(k-1) \times k$

$$\begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} x + \delta z_0 \tag{4.10}$$

If $k = 1$, then $(\delta, 0) \in \mathfrak{B}^{\delta \cot 0}_{[K_0, L_0, 0_{0\times 1}]}$ and hence the system is not observable at infinity in this case. If $k \geq 2$, then for $z_0 = e_{k-1}^{[k-1]} \in \mathbb{R}^{k-1} \setminus \{0\}$ we have that $x = \delta e_k^{[k]}$ fulfills $(x, 0) \in \mathfrak{B}^{\delta z_0}_{[K_k, L_k, 0_{0\times k}]}$. In particular, we have $Ex = 0$ and $x[0] = \delta e_k^{[k]} \neq 0$. Hence, the DAE is not observable at infinity.

(4.5e) The DAE of size $k \times k$

$$\begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{bmatrix} \dot{x} = x + \delta z_0 \tag{4.11}$$

has the unique solution

$$x = x[0] = -\sum_{j=0}^{k-1} (N_k^\top)^j \delta^{(j)} z_0.$$

Hence it can never be observable at infinity.

(4.5f) The ODE of size $k \times k$

$$\dot{x} = Ax + \delta z_0$$
$$y = 0$$

has a solution x for any $z_0 \in \mathbb{R}^n$, hence it is never completely observable. The additional constraint $x = 0$ yields $\dot{x} = 0$ and hence $\delta z_0 = \dot{x} - Ax = 0$ which shows observability at infinity.

□

4.4 Characterization of Relevant State Observability

Finally we consider the observability notions from Sect. 4.4. First we focus on RS behavioral, impulse and strong observability.

Lemma 4.6 *Consider the decoupled DAEs (4.5) resulting from the OI normal form. Then the DAEs*

(4.5a) are always RS behaviorally, RS impulse and RS strongly observable.
(4.5b) are always RS behaviorally, RS impulse and RS strongly observable.
(4.5c) are always RS behaviorally observable; they are RS impulse and RS strongly observable if, and only if, $|\gamma| = \ell(\gamma)$, i.e., $\gamma_i = 1$ for all $i = 1, \ldots, \ell(\gamma)$.
(4.5d) are always RS behaviorally observable; they are RS impulse and RS strongly observable if, and only if, $|\varepsilon| = \ell(\varepsilon)$, i.e., $\varepsilon_i = 1$ for all $i = 1, \ldots, \ell(\varepsilon)$.
(4.5e) are always RS behaviorally observable; they are RS impulse and RS strongly observable if, and only if, $|\kappa| = \ell(\kappa)$, i.e., $\kappa_i = 1$ for all $i = 1, \ldots, \ell(\kappa)$.
(4.5f) are never RS behaviorally and RS strongly observable and always RS impulse observable.

Proof First we consider RS behavioral and RS impulse observability. The statements for RS behavioral observability in (4.5a)–(4.5c) and (4.5e) follow by a combination of Corollary 3.7 and Lemma 4.4. Since observability at infinity implies RS impulse observability by Corollary 3.7, it follows from Lemma 4.5 that (4.5a), (4.5b) and (4.5f) are RS impulse observable.

We prove the remaining statements for RS behavioral and impulse observability:

(4.5c): If $|\gamma| = \ell(\gamma)$, then the DAE (4.5c) is RS impulse observable by a combination of Corollary 3.7 and Lemma 4.5. If $|\gamma| > \ell(\gamma)$, then we can use the same counterexample as in the proof of Lemma 4.5 for (4.5c) by observing that $z_0 = e_1^{[k-1]} = Ee_2^{[k]}$. Hence, the system is not RS impulse observable.

(4.5d): DAEs with size $(k-1) \times k$ of the form (4.8) are RS behaviorally observable by the characterization in Remark 3.1 and the fact that, by [70, Theorem 5.2.10], for any two solutions x^1, x^2 we can find some $T > 0$ and some $(x^3, u, y) \in \mathfrak{B}_{[E,A,B,C,D]}$ with

$$(x^3)_{(-\infty,0)} \overset{\text{a.e.}}{=} (x^1)_{(-\infty,0)} \ \wedge \ (x^3)_{(T,\infty)} \overset{\text{a.e.}}{=} (x^2)_{(T,\infty)}.$$

If $|\varepsilon| = \ell(\varepsilon)$, then the DAE (4.10) is RS impulse observable by Lemma 3.6 (b) and the fact that $K_\varepsilon = 0 \in \mathbb{R}^{0\times|\varepsilon|}$. If $|\varepsilon| > \ell(\varepsilon)$, then we can use the same counterexample as in the proof of Lemma 4.5 for (4.5d) by observing that $z_0 = e_{k-1}^{[k-1]} = Ee_{k-1}^{[k]}$. Hence, the system is not RS impulse observable.

(4.5e): If $|\kappa| = \ell(\kappa)$, then we have RS impulse observability due to $N_\kappa^\top = 0$. If $|\kappa| > \ell(\kappa)$, then there is a DAE of the form (4.11) with $k \geq 2$. For $z_0 = N_k^\top e_2^{[k]} \in \mathbb{R}^k \setminus \{0\}$ the unique solution of (4.11) is $x = -\delta e_1^{[k]}$. Since $N_k^\top x = 0$ and $z_0 \neq 0$ the system is not RS impulse observable.

(4.5f): The ODE (4.5f) has nontrivial solutions that are uniquely determined by $x(0^+)$, whence it is not RS behaviorally observable.

The characterization of RS strong observability follows from analogous arguments. □

Now we prove the characterizations for RS complete observability and RS observability at infinity.

Lemma 4.7 *Consider the decoupled DAEs (4.5) resulting from the OI normal form. Then the DAEs*

(4.5a) are always RS completely observable and RS observable at infinity.
(4.5b) are always RS completely observable and RS observable at infinity.
(4.5c) are RS completely observable and RS observable at infinity if, and only if, $|\gamma| = \ell(\gamma)$, i.e., $\gamma_i = 1$ for all $i = 1, \ldots, \ell(\gamma)$.
(4.5d) are RS completely observable and RS observable at infinity if, and only if, $|\varepsilon| = \ell(\varepsilon)$, i.e., $\varepsilon_i = 1$ for all $i = 1, \ldots, \ell(\varepsilon)$.
(4.5e) are neither RS completely observable nor RS observable at infinity.
(4.5f) are never RS completely observable and always RS observable at infinity.

Proof The proof is analogous to the proof of Lemma 4.5 with the only difference that for DAEs (4.5d) in the case $|\varepsilon| = \ell(\varepsilon)$ the system is RS observable at infinity (and hence RS completely observable) since $K_\varepsilon, L_\varepsilon \in \mathbb{R}^{0\times|\varepsilon|}$ and hence there is no z_0 (the number of rows is zero). □

4.5 Summary of Observability Characterizations

The different observability characterizations derived in the previous subsections in terms of the OI normal form are summarized in Table 1.

We have separated the concepts into two groups of five concepts where the first group consists of the observability notions introduced in Sects. 3.1 and 3.2 and the second group consists of the corresponding relevant state observability notions introduced in Sect. 3.3.

Table 1 together with Lemma 4.1 allows for a characterization of the observability concepts in terms of the OI normal form.

In particular, for regular systems we can conclude the following simplifications from Remark 4.2 and Table 1.

Corollary 4.8 *Consider a* regular *system* $[E, A, C] \in \mathscr{O}_{n,n,p}$. *Then the following equivalences hold for the DAE system:*

(i)	*behaviorally observable*	$\Longleftrightarrow$	*RS behaviorally observable,*
(ii)	*impulse observable*	$\Longleftrightarrow$	*RS impulse observable,*
(iii)	*strongly observable*	$\Longleftrightarrow$	*RS strongly observable,*
(iv)	*observable at infinity*	$\Longleftrightarrow$	*RS observable at infinity,*
(v)	*completely observable*	$\Longleftrightarrow$	*RS completely observable.*

From Table 1 the dependencies between the different observability concepts can easily be concluded and are illustrated in Fig. 1.

5 Duality of Observability and Controllability

The intuitive definitions of behavioral and impulse observability given in Sect. 3.1 are not satisfying from a duality seeking point of view. Duality means that a system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ has a certain observability property if, and only if, the "formal dual" system

$$\begin{aligned} \tfrac{\mathrm{d}}{\mathrm{d}t}E^\top x(t) &= A^\top x(t) + C^\top u(t) \\ y(t) &= B^\top x(t) + D^\top u(t)\,, \end{aligned} \tag{5.1}$$

Table 1 Characterization of the observability concepts in terms of the OI normal form

	$[I_{\alpha_i}, N_{\alpha_i}, \left(e_{\alpha_i}^{[\alpha_i]}\right)^\top]$	$[K_{\beta_i}^\top, L_{\beta_i}^\top, 0_{0\times\beta_i-1}]$	$[L_{\gamma_i}, K_{\gamma_i}, \left(e_{\gamma_i}^{[\gamma_i]}\right)^\top]$	$[K_{\varepsilon_i}, L_{\varepsilon_i}, 0_{0\times\varepsilon_i}]$	$[N_{\kappa_i}^\top, I_{\kappa_i}, 0_{0\times\kappa_i}]$	$[I_{n_{\overline{0}}}, A_{\overline{0}}, 0_{q\times\overline{o}}]$
Behaviorally observable	✓	✓	✓	×	✓	×
Impulse observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	×	$\Leftrightarrow \kappa_i = 1$	✓
Strongly observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	×	$\Leftrightarrow \kappa_i = 1$	×
Observable at infinity	✓	✓	$\Leftrightarrow \gamma_i = 1$	×	×	✓
Completely observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	×	×	×
RS behaviorally observable	✓	✓	✓	✓	✓	×
RS impulse observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	$\Leftrightarrow \varepsilon_i = 1$	$\Leftrightarrow \kappa_i = 1$	✓
RS strongly observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	$\Leftrightarrow \varepsilon_i = 1$	$\Leftrightarrow \kappa_i = 1$	×
RS observable at infinity	✓	✓	$\Leftrightarrow \gamma_i = 1$	$\Leftrightarrow \varepsilon_i = 1$	×	✓
RS completely observable	✓	✓	$\Leftrightarrow \gamma_i = 1$	$\Leftrightarrow \varepsilon_i = 1$	×	×

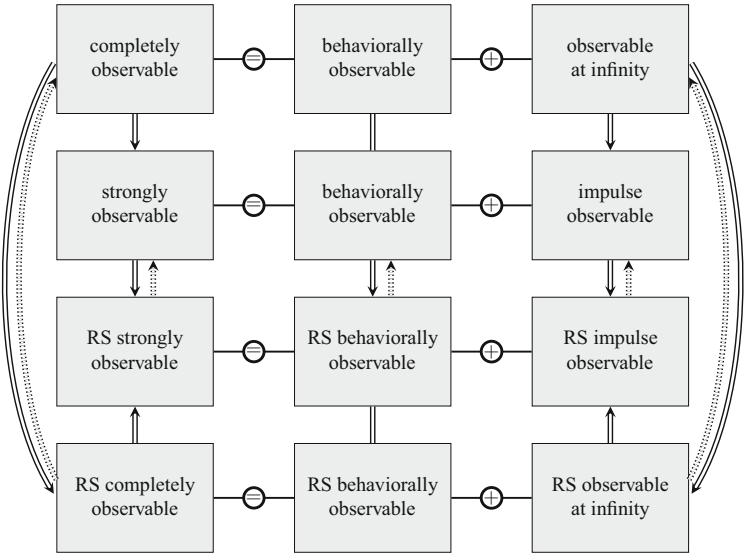


Fig. 1 Relationship between the different observability concepts. For each implication, the converse is false in general; *dotted* implications indicate the regular case

has the corresponding controllability property. Since the controllability properties of the dual system (5.1) do not depend on $B^\top$ and $D^\top$ it is sufficient to consider the class $\mathscr{C}_{l,n,m}$ of control systems governed by the equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t}Ex(t) = Ax(t) + Bu(t), \tag{5.2}$$

where $E, A \in \mathbb{R}^{l\times n}, B \in \mathbb{R}^{l\times m}$; we write $[E, A, B] \in \mathscr{C}_{l,n,m}$. Each controllability concept (see [14] and the survey [15]) is invariant under the addition of a zero row in $[E, A, B] \in \mathscr{C}_{l,n,m}$ or, equivalently, an equation $0 = 0$ in (5.2). However, if we consider the dual system $[E^\top, A^\top, B^\top] \in \mathscr{O}_{n,l,m}$, then $E^\top, A^\top, B^\top$ have a common zero column and hence there exists a free state in the system which is not visible at the output. This implies that the system is neither impulse nor behaviorally observable, although $[E, A, B]$ may be both impulse and behaviorally controllable as introduced in [15]. This means that these observability and controllability concepts or not dual.

As already pointed out in Sect. 3.3, it is not always reasonable to view a state as unobservable which actually does not appear in any of the systems equations; it only appears in the model because of "bad design". This viewpoint led us to the introduction of the relevant state observability concepts. It allows to provide duality results between the controllability concepts from [15] and the observability concepts from Sects. 3.1–3.3. The RS observability concepts cope with "design errors" as mentioned above by preserving the physical meaning of observability. The duality results will provide algebraic characterizations for the observability concepts.

Remark 5.1 The "design errors" mentioned above can be given an interpretation using the behavioral framework. If (5.2) contains an equation of the form $0 = 0$ or other redundant equations, then it is not minimal in the behavioral sense as introduced in [70, Definition 2.5.24], see also [16]. Minimality is equivalent to $\mathrm{rk}_{\mathbb{R}[s]}[sE - A, B] = l$, see [16] for further characterizations. If this condition is not satisfied, then $\mathrm{rk}_{\mathbb{R}[s]} \begin{bmatrix} sE^\top - A^\top \\ B^\top \end{bmatrix} < l$ and hence the equation

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} E^\top \\ 0 \end{bmatrix} x(t) = \begin{bmatrix} A^\top \\ B^\top \end{bmatrix} x(t)$$

does always have an underdetermined part and thus non-unique solutions independent of the properties of the original system $[E, A, B]$. This leads to the "lack of duality" between (for instance) behavioral controllability in the sense of [15] and behavioral observability, in the case of non-minimal systems. Note that if $sE - A$ is regular, then $[E, A, B]$ is always minimal.

If minimality is assumed, then it is easy to check that the controllability concepts from [15] are indeed dual to the observability concepts introduced in the present paper. This can also be deduced from a recent approach by Lomadze [57] to the definition of the dual of a behavioral system. When the definition of the dual system given in [57] is applied to DAE systems (1.1), then the dual is exactly the formal dual system (5.1).

Summarizing, this justifies to say that a lack of duality does not come from intrinsic system properties, but from a bad (i.e., not minimal) model of the underlying behavior.

Using the OI normal form (which is the "dual" of the feedback form derived in [15]), the characterizations summarized in Table 1 and the respective results in [15] lead to the following duality results between the RS observability and the controllability concepts.

Corollary 5.1 (Duality Between Observability and Controllability) *Let $[E, A, C] \in \mathscr{O}_{l,n,p}$ be given. Then we have the following equivalences:*

(a) $[E, A, C]$ *is RS behaviorally observable if, and only if,* $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ *is behaviorally controllable in the sense of [15],*

(b) $[E, A, C]$ *is RS impulse observable if, and only if,* $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ *is impulse controllable in the sense of [15],*

(c) $[E, A, C]$ *is RS strongly observable if, and only if,* $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ *is strongly controllable in the sense of [15],*

(d) $[E, A, C]$ *is RS observable at infinity if, and only if,* $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ *is controllable at infinity in the sense of [15],*

(e) $[E, A, C]$ *is RS completely observable if, and only if,* $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ *is completely controllable in the sense of [15].*

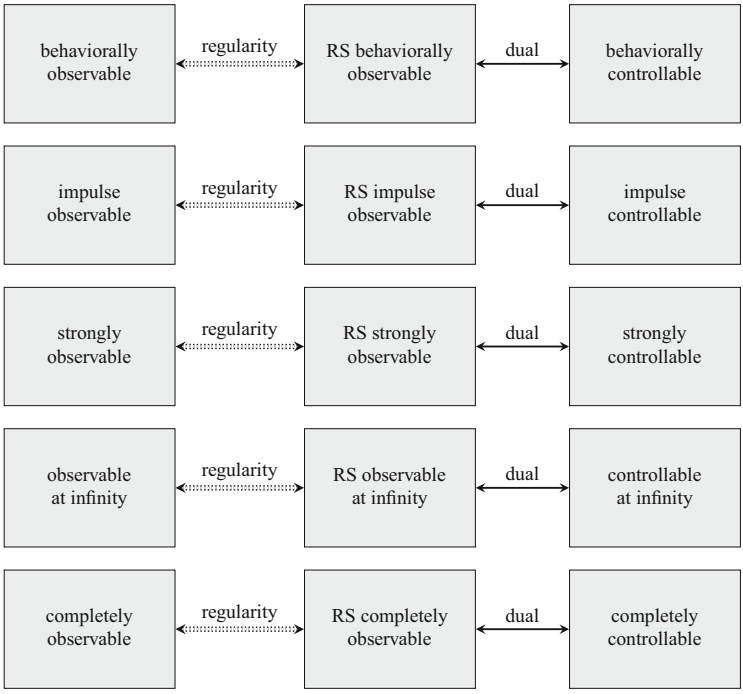


Fig. 2 Illustration of duality between observability and controllability

In particular, for regular DAE systems we have duality between the five remaining observability concepts and the corresponding controllability concepts. The duality properties are summarized in Fig. 2.

6 Algebraic Criteria

Using the duality results derived in Corollary 5.1, in this section we derive algebraic criteria for the observability concepts. These criteria are generalizations of the Hautus test (also called the Popov–Belevitch–Hautus test, since they were independently developed by Popov [71], Belevitch [12] and Hautus [39]) in terms of rank and kernel criteria on the involved matrices. Most of these conditions are known—we refer to the relevant literature.

Proposition 6.1 (Algebraic Criteria for Observability) *Let a system* $[E, A, C] \in \mathscr{O}_{l,n,p}$ *be given. Then we have the following:*

$[E, A, C]$ is	if, and only if,
Behaviorally observable	$\forall\, \lambda \in \mathbb{C}:\ \ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C = \{0\}$
Impulse observable	$\ker_{\mathbb{R}} E \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) \cap \ker_{\mathbb{R}} C = \{0\}$
Strongly observable	$\ker_{\mathbb{R}} E \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) \cap \ker_{\mathbb{R}} C = \{0\}$ $\wedge\ \forall\, \lambda \in \mathbb{C}:\ \ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C = \{0\}$
Observable at infinity	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} C = \{0\}$
Completely observable	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} C = \{0\}$ $\wedge\ \forall\, \lambda \in \mathbb{C}:\ \ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C = \{0\}$
RS behaviorally observable	$\forall\, \lambda \in \mathbb{C}:\ \dim \ker_{\mathbb{R}(s)} \begin{bmatrix} sE - A \\ C \end{bmatrix} = \dim \ker_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix}$.
RS impulse observable	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \ker_{\mathbb{R}} E \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) \cap \ker_{\mathbb{R}} C$
RS strongly observable	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \ker_{\mathbb{R}} E \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) \cap \ker_{\mathbb{R}} C$ $\wedge\ \forall\, \lambda \in \mathbb{C}: \ker_{\mathbb{C}} E \cap \ker_{\mathbb{C}} A \cap \ker_{\mathbb{C}} C = \ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C$
RS observable at infinity	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} C$
RS completely observable	$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} C$ $\wedge\ \forall\, \lambda \in \mathbb{C}: \ker_{\mathbb{C}} E \cap \ker_{\mathbb{C}} A \cap \ker_{\mathbb{C}} C = \ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C$

Proof Combining Corollary 5.1 and [15, Corollary 4.3] the criteria for RS behavioral, impulse, strong and complete observability and RS observability at infinity follow immediately. From the OI normal form (4.4) it can be concluded that

$$\ell(\varepsilon) = 0 \ \wedge\ \det A_{\overline{0}} \neq 0 \quad \Longleftrightarrow \quad \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}.$$

Therefore, invoking Table 1, behavioral observability is equivalent to RS behavioral observability together with the condition $\ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}$. Hence, the characterization of RS behavioral observability follows from observing that the conditions $\ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}$ and $\mathrm{rk}_{\mathbb{R}(s)} \left[\begin{smallmatrix} sE-A \\ C \end{smallmatrix}\right] = \mathrm{rk}_{\mathbb{C}} \left[\begin{smallmatrix} \lambda E-A \\ C \end{smallmatrix}\right]$ for all $\lambda \in \mathbb{C}$ are equivalent to $\ker_{\mathbb{C}}(\lambda E - A) \cap \ker_{\mathbb{C}} C = \{0\}$ for all $\lambda \in \mathbb{C}$.

Furthermore, it follows from the OI normal form that

$$\ell(\varepsilon) = 0 \quad \Longleftrightarrow \quad \ell(\varepsilon) = |\varepsilon| \ \wedge\ \ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}. \tag{6.1}$$

Therefore, invoking Table 1, impulse observability is equivalent to RS impulse observability together with the condition $\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}$, which yields the characterization in the statement of the corollary. The characterization of strong observability then follows from those of behavioral and impulse observability.

Likewise, Eq. (6.1) implies that observability at infinity is equivalent to RS observability at infinity together with the condition $\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} A \cap \ker_{\mathbb{R}} C = \{0\}$, which yields the characterization in the statement of the corollary. Finally, the characterization for complete observability then follows from those of behavioral observability and observability at infinity. □

In the following we consider further criteria for the observability concepts.

Remark 6.1 (RS Observability at Infinity) Proposition 6.1 immediately implies that RS observability at infinity is equivalent to

$$\ker_{\mathbb{R}} E \cap \ker_{\mathbb{R}} C \subseteq \ker_{\mathbb{R}} A.$$

In terms of a rank criterion, this is the same as

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix}. \tag{6.2}$$

Likewise, observability at infinity is equivalent to the rank condition

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = n. \tag{6.3}$$

As far as the authors are aware, the conditions (6.2) and (6.3) are new for general DAE systems. In the case of regular $sE - A \in \mathbb{R}[s]^{n\times n}$, condition (6.3) can be found for instance in [31].

Remark 6.2 (RS Impulse Observability) It follows from Proposition 6.1 that an equivalent characterization for RS impulse observability is that, for one (and hence any) matrix Z with $\operatorname{im}_{\mathbb{R}} Z = \ker_{\mathbb{R}} E^\top$, we have

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix}. \tag{6.4}$$

Likewise, impulse observability is equivalent to

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix} = n. \tag{6.5}$$

This was first derived in [43]. Furthermore, in [40, 43] it was shown that impulse observability is equivalent to

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E & A \\ 0 & C \\ 0 & E \end{bmatrix} = n + \operatorname{rk}_{\mathbb{R}} E, \tag{6.6}$$

which is in fact equivalent to (6.5). If the pencil $sE-A$ is regular, then condition (6.5) for impulse observability can also be inferred from [32, Theorem 2-3.4].

Remark 6.3 (RS Behavioral Observability) The algebraic criterion for RS behavioral observability in Proposition 6.1 is equivalent to the fact that the augmented matrix pencil

$$s\mathscr{E} - \mathscr{A} = s \begin{bmatrix} E \\ 0 \end{bmatrix} - \begin{bmatrix} A \\ C \end{bmatrix} \in \mathbb{R}[s]^{(l+p)\times n}$$

has no eigenvalues. Behavioral observability coincides with observability as defined in [70, Definition 5.3.2] for the larger class of linear differential behaviors, and the rank condition for behavioral observability in Proposition 6.1 has already been derived in [70, Theorem 5.3.3]; the condition has also been derived in [40] where this concept is called right-hand side observability. RS behavioral observability for systems with regular $sE - A$ is considered in [32, Theorem 2-3.2] (called R-observability in this work), where the condition

$$\forall\, \lambda \in \mathbb{C}: \quad \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n$$

is derived. This is, for regular $sE - A$, in fact equivalent to the criterion for RS behavioral observability in Proposition 6.1.

Remark 6.4 (RS Complete and Strong Observability) By Table 1, RS complete observability of $[E, A, C] \in \mathscr{O}_{l,n,p}$ is equivalent to $[E, A, C]$ being RS behaviorally observable and RS observable at infinity, whereas RS strong observability of $[E, A, C]$ is equivalent to $[E, A, C]$ being RS behaviorally observable and RS impulse observable.

The algebraic conditions for strong observability in Proposition 6.1 were first derived in [40] (called observability in that work). On the other hand, as far as the authors are aware, the algebraic criterion for RS complete observability is new for general DAE systems.

For regular systems, the conditions in Proposition 6.1 for complete observability are also derived in [32, Theorem 2-3.1].

The above considerations lead to the following alternative formulation of Proposition 6.1 in terms of rank criteria.

Corollary 6.2 *Let* $[E, A, C] \in \mathscr{O}_{l,n,p}$ *and* Z *be a matrix with* $\operatorname{im}_{\mathbb{R}} Z = \ker_{\mathbb{R}} E^\top$. *Then we have the following:*

$[E, A, C]$ is	if, and only if,
Behaviorally observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n$
Impulse observable	$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix} = n$
Strongly observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix} = n$
Observable at infinity	$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = n$
Completely observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = n$
RS behaviorally observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}(s)} \begin{bmatrix} sE - A \\ C \end{bmatrix}.$
RS impulse observable	$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$
RS strongly observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^\top A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$
RS observable at infinity	$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$
RS completely observable	$\forall\, \lambda \in \mathbb{C}:\ \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$

Remark 6.5 (Kalman Criterion for Regular Systems) For regular systems $[E, A, C] \in \mathscr{O}_{n,n,p}$ the usual Hautus and Kalman criteria for observability can be found in a summarized form e.g. in [32]. Other approaches to derive observability criteria rely on the expansion of $(sE-A)^{-1}$ as a power series in s at $s_0 = 0$, which is only feasible in the regular case. For instance, in [63] the numerator matrices of this expansion, i.e., the coefficients of the polynomial $\operatorname{adj}(sE - A)$, are used to derive a

rank criterion for complete observability. Then again, in [49] Kalman rank criteria for complete observability, behavioral observability (called R-observability in this work) and observability at infinity are derived in terms of the coefficients of the power series expansion of $(sE - A)^{-1}$. The advantage of these criteria, especially the last one, is that no transformation of the system needs to be performed as it is usually necessary in order to derive Kalman rank criteria for DAEs, see e.g. [32].

However, simple criteria can be obtained using only a left transformation of little impact: if $\alpha \in \mathbb{R}$ is chosen such that $\det(\alpha E - A) \neq 0$, then the system is completely observable if, and only if, [89, Corollary 2]

$$\operatorname{rk}_{\mathbb{R}} \begin{bmatrix} C \\ C(\alpha E - A)^{-1}E \\ \vdots \\ C\big((\alpha E - A)^{-1}E\big)^{n-1} \end{bmatrix} = n,$$

and it is impulse observable if, and only if, [89, Theorem 5]

$$\ker_{\mathbb{R}}(\alpha E - A)^{-1}E \cap \ker_{\mathbb{R}} C \cap \operatorname{im}_{\mathbb{R}}(\alpha E - A)^{-1}E = \mathbb{R}^n.$$

7 Geometric Criteria

In this section we derive geometric criteria for the observability concepts. Geometric theory plays a fundamental role in ODE system theory and was introduced independently by Wonham and Morse, and by Basile and Marro, see the famous books [11, 87] and also [79]. In [54], Lewis provided a survey of the to-date geometric theory of DAEs. As we will do here, he put special emphasis on the two fundamental sequences $(\mathscr{V}_i)_{i\in\mathbb{N}_0}$ and $(\mathscr{W}_i)_{i\in\mathbb{N}_0}$ of subspaces defined as follows:

$$\mathscr{V}_0 := \mathbb{R}^n, \qquad \mathscr{V}_{i+1} := A^{-1}(E\mathscr{V}_i) \cap \ker_{\mathbb{R}} C \subseteq \mathbb{R}^n, \qquad \mathscr{V}^* := \bigcap_{i\in\mathbb{N}_0} \mathscr{V}_i,$$

$$\mathscr{W}_0 := \{0\}, \qquad \mathscr{W}_{i+1} := E^{-1}(A\mathscr{W}_i) \cap \ker_{\mathbb{R}} C \subseteq \mathbb{R}^n, \qquad \mathscr{W}^* := \bigcup_{i\in\mathbb{N}_0} \mathscr{W}_i.$$

We will call the sequences $(\mathscr{V}_i)_{i\in\mathbb{N}}$ and $(\mathscr{W}_i)_{i\in\mathbb{N}}$ *restricted Wong sequences*. In [17, 18, 20] the Wong sequences for matrix pencils (i.e., $C = 0$) are investigated, the name chosen this way since Wong [86] was the first to use both sequences for the analysis of matrix pencils. In fact, the Wong sequences (with $C = 0$) can be traced back to Dieudonné [34], who focused on the first of the two Wong sequences. Bernhard [21] and Armentano [3] used the Wong sequences to carry out a geometric analysis of matrix pencils. They appear also in [1, 2, 51, 80]. The sequences $(\mathscr{V}_i)_{i\in\mathbb{N}}$

and $(\mathscr{W}_i)_{i\in\mathbb{N}}$ are no Wong sequences corresponding to any matrix pencils, that is why we call them *restricted Wong sequences* with respect to the system $[E, A, C] \in \mathscr{O}_{l,n,p}$.

For the investigation of observability of DAE systems, that is when $C \neq 0$, the restricted Wong sequences have been extensively studied by several authors, see e.g. [53, 61, 62, 65, 67, 81] for regular systems and [7–10, 23, 54, 55, 66, 68] for general DAE systems.

For regular systems Özçaldiran [65] (see also [68]) showed that $\mathscr{V}^*$ is the supremal (A, E)-invariant subspace contained in $\ker_{\mathbb{R}} C$ and $\mathscr{W}^*$ is the infimal restricted $(E, A; \ker_{\mathbb{R}} C)$-invariant subspace (which is also a subspace of $\ker_{\mathbb{R}} C$); note that by these invariance definitions, $\mathscr{W}^*$ is not the obvious dual to $\mathscr{V}^*$, but by the definition of the restricted Wong sequences this connection becomes more apparent. The aforementioned invariance concepts, which have also been used in [1, 6, 53, 62], are defined as follows.

Definition 7.1 ((A, E)- and $(E, A; \ker_{\mathbb{R}} C)$-Invariance [65]) Let $E, A \in \mathbb{R}^{l\times n}$. A subspace $\mathscr{V} \subseteq \mathbb{R}^n$ is called (A, E)-*invariant*, if

$$A\mathscr{V} \subseteq E\mathscr{V}.$$

For $C \in \mathbb{R}^{p\times n}$, a subspace $\mathscr{W} \subseteq \mathbb{R}^n$ is called *restricted* $(E, A; \ker_{\mathbb{R}} C)$-*invariant*, if

$$\mathscr{W} = \ker_{\mathbb{R}} C \cap E^{-1}(A\mathscr{W}).$$

It is easy to verify that the proofs given in [65, Lemmas 2.1 and 2.2] remain the same for general $E, A \in \mathbb{R}^{l\times n}$ and (in the notation of [65]) $K = \ker_{\mathbb{R}} C$ for $C \in \mathbb{R}^{p\times n}$ and $B = 0$; this is shown in [6] as well. For $\mathscr{V}^*$ this can be found in [1], see also [62]. We have the following proposition.

Proposition 7.1 (Restricted Wong Sequences as Invariant Subspaces) *Consider $[E, A, C] \in \mathscr{O}_{l,n,p}$ and the limits $\mathscr{V}^*$ and $\mathscr{W}^*$ of the restricted Wong sequences. Then the following statements hold true.*

(a) $\mathscr{V}^*$ is (A, E)-invariant with $\mathscr{V}^* \subseteq \ker_{\mathbb{R}} C$ and for any $\mathscr{V} \subseteq \ker_{\mathbb{R}} C$ which is (A, E)-invariant it holds that $\mathscr{V} \subseteq \mathscr{V}^*$;
(b) $\mathscr{W}^*$ is restricted $(E, A; \ker_{\mathbb{R}} C)$-invariant and for any $\mathscr{W} \subseteq \mathbb{R}^n$ which is restricted $(E, A; \ker_{\mathbb{R}} C)$-invariant it holds that $\mathscr{W}^* \subseteq \mathscr{W}$.

In the following we show how the observability concepts can be characterized in terms of the invariant subspaces $\mathscr{V}^*$ and $\mathscr{W}^*$ by using the OI normal form (4.4).

Theorem 7.2 (Geometric Criteria for Observability) *Consider $[E, A, C] \in \mathscr{O}_{l,n,p}$ and the limits $\mathscr{V}^*$ and $\mathscr{W}^*$ of the restricted Wong sequences. Then $[E, A, C]$ is*

(a) behaviorally observable if, and only if, $\mathscr{V}^* = \{0\}$;
(b) impulse observable if, and only if, $\mathscr{W}^* \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) = \{0\}$;
(c) strongly observable if, and only if, $(\mathscr{V}^* + \mathscr{W}^*) \cap A^{-1}(\mathrm{im}_{\mathbb{R}} E) = \{0\}$;
(d) observable at infinity if, and only if, $\mathscr{W}^* = \{0\}$;
(e) completely observable if, and only if, $\mathscr{V}^* + \mathscr{W}^* = \{0\}$;

(f) RS behaviorally observable if, and only if, $\mathscr{V}^* \subseteq \mathscr{W}^*$*;*
(g) RS impulse observable if, and only if, $A\mathscr{W}^* \cap \operatorname{im}_{\mathbb{R}} E = \{0\}$.
(h) RS strongly observable if, and only if, $(E\mathscr{V}^* + A\mathscr{W}^*) \cap \operatorname{im}_{\mathbb{R}} E = \{0\}$.
(i) RS observable at infinity if, and only if, $A\mathscr{W}^* = \{0\}$*;*
(j) RS completely observable if, and only if, $E\mathscr{V}^* + A\mathscr{W}^* = \{0\}$.

Proof We prove the assertions by deriving formulas for $\mathscr{V}^*$ and $\mathscr{W}^*$ in terms of the OI normal form (4.4) and then connect the geometric conditions to the observability concepts by Table 1. We proceed in several steps.

Step 1: Let $[E_1, A_1, C_1], [E_2, A_2, C_2] \in \mathscr{O}_{l,n,p}$ be such that for some $W \in \mathbf{Gl}_l(\mathbb{R})$, $T \in \mathbf{Gl}_n(\mathbb{R})$, $V \in \mathbf{Gl}_p(\mathbb{R})$ and $L \in \mathbb{R}^{l\times p}$ it holds that

$$[\,E_1\,,\,A_1\,,\,C_1\,] \overset{W,T,V,L}{\sim_{OI}} [\,E_2\,,\,A_2\,,\,C_2\,].$$

We show that the restricted Wong sequences $\mathscr{V}_i^1$, $\mathscr{W}_i^1$ of $[E_1, A_1, C_1]$ and the restricted Wong sequences $\mathscr{V}_i^2$, $\mathscr{W}_i^2$ of $[E_2, A_2, C_2]$ are related by

$$\forall\, i \in \mathbb{N}_0 :\ \mathscr{V}_i^1 = T^{-1}\mathscr{V}_i^2 \ \wedge\ \mathscr{W}_i^1 = T^{-1}\mathscr{W}_i^2.$$

We prove the statement by induction. It is clear that $\mathscr{V}_0^1 = T^{-1}\mathscr{V}_0^2$. Assuming that $\mathscr{V}_i^1 = T^{-1}\mathscr{V}_i^2$ for some $i \geq 0$ we find that, by (4.2),

$$\begin{aligned}
\mathscr{V}_{i+1}^1 &= \ker_{\mathbb{R}} C_1 \cap A_1^{-1}(E_1\mathscr{V}_i^1)\\
&= \left\{ x \in \mathbb{R}^n \;\middle|\; \exists\, y \in \mathscr{V}_i^1 :\ WA_2Tx = WE_2Ty \ \wedge\ VC_2Ty = 0 \right\}\\
&= \left\{ x \in \mathbb{R}^n \;\middle|\; \exists\, z \in \mathscr{V}_i^2 :\ A_2Tx = E_2z \ \wedge\ C_2z = 0 \right\}\\
&= T^{-1}\left(\ker_{\mathbb{R}} C_2 \cap A_2^{-1}(E_2\mathscr{V}_i^2)\right) = T^{-1}\mathscr{V}_{i+1}^2.
\end{aligned}$$

The statement about $\mathscr{W}_i^1$ and $\mathscr{W}_i^2$ can be proved analogously.

Step 2: By Step 1 we may without loss of generality assume that $[E, A, C]$ is given in OI normal form (4.4). We make the convention that if $\alpha \in \mathbb{N}^k$ is some multi-index, then $\alpha - 1 := (\alpha_1 - 1, \ldots, \alpha_k - 1)$. It follows that

$$\forall\, i \in \mathbb{N}_0 :\ \mathscr{V}_i = \bigcap_{j=0}^{i-1} \ker_{\mathbb{R}} E_\alpha^\top N_\alpha^j \times \operatorname{im}_{\mathbb{R}}(N_{\beta-1}^\top)^i \times \operatorname{im}_{\mathbb{R}}(N_\gamma^\top)^i \times \mathbb{R}^{|\varepsilon|} \times \operatorname{im}_{\mathbb{R}}(N_\kappa^\top)^i \times \mathbb{R}^{n\overline{\sigma}}, \tag{7.1}$$

which is immediate from observing that $L_\beta^\top x = K_\beta^\top y$ for some x, y of appropriate dimension yields $x = N_{\varepsilon-1}^\top y$, and $K_\gamma x = L_\gamma y$ with $E_\gamma^\top x = 0$ for some x, y yields $x = N_\gamma^\top y$. Note that in the case $\beta_j = 1$, i.e., we have a 1×0 block, we find that $N_{\beta_j-1}^\top$ is absent, so these relations are consistent.

On the other hand we find that

$$\forall\, i \in \mathbb{N}_0 :$$
$$\mathscr{W}_i = \{0\}^{|\alpha|} \times \{0\}^{|\beta|} \times \left(\ker_{\mathbb{R}} (N_\gamma^\top)^i \cap \ker_{\mathbb{R}} E_\gamma^\top\right) \times \ker_{\mathbb{R}} N_\varepsilon^i \times \ker_{\mathbb{R}} (N_\kappa^\top)^i \times \{0\}^{n_{\overline{\sigma}}}. \tag{7.2}$$

Step 3: From (7.1) and (7.2) it follows that

$$\mathscr{V}^* = \{0\}^{|\alpha|} \times \{0\}^{|\beta|-\ell(\beta)} \times \{0\}^{|\gamma|} \times \mathbb{R}^{|\varepsilon|} \times \{0\}^{|\kappa|} \times \mathbb{R}^{n_{\overline{\sigma}}},$$
$$\mathscr{W}^* = \{0\}^{|\alpha|} \times \{0\}^{|\beta|-\ell(\beta)} \times \ker_{\mathbb{R}} E_\gamma^\top \times \mathbb{R}^{|\varepsilon|} \times \mathbb{R}^{|\kappa|} \times \{0\}^{n_{\overline{\sigma}}}$$

and

$$E\mathscr{V}^* = \{0\}^{|\alpha|} \times \{0\}^{|\beta|} \times \{0\}^{|\gamma|-\ell(\gamma)} \times \mathbb{R}^{|\varepsilon|-\ell(\varepsilon)} \times \{0\}^{|\kappa|} \times \mathbb{R}^{n_{\overline{\sigma}}},$$
$$A\mathscr{W}^* = \{0\}^{|\alpha|} \times \{0\}^{|\beta|} \times K_\gamma(\ker_{\mathbb{R}} E_\gamma^\top) \times \mathbb{R}^{|\varepsilon|-\ell(\varepsilon)} \times \mathbb{R}^{|\kappa|} \times \{0\}^{n_{\overline{\sigma}}},$$
$$\operatorname{im}_{\mathbb{R}} E = \mathbb{R}^{|\alpha|} \times \operatorname{im}_{\mathbb{R}} K_\beta^\top \times \mathbb{R}^{|\gamma|-\ell(\gamma)} \times \mathbb{R}^{|\varepsilon|-\ell(\varepsilon)} \times \operatorname{im}_{\mathbb{R}} N_\kappa^\top \times \mathbb{R}^{n_{\overline{\sigma}}}.$$

The equivalences in (a)–(j) may now be inferred from Table 1. □

Under the additional assumption that $\operatorname{rk}[E^\top, A^\top, C^\top] = n$, the conditions for strong and complete observability as in Theorem 7.2 are derived in [10, 68] (which are called observability and strong observability in these works, resp.). The conditions for strong and complete observability are also derived in [66], as well as those for behavioral and RS strong observability; in [66] the observability concepts are defined within a distributional solution setup and other names are used than in the present work (cf. Sect. 3.4).

8 Kalman Decomposition

The famous decomposition of linear ODE control systems derived by Kalman [45] is one of the most important tools in the analysis of these systems. This decomposition was later been generalized to regular DAEs by Verghese et al. [82], see also [32]. A Kalman decomposition of general discrete-time DAE systems was provided by Banaszuk et al. [8] in a very nice way using the restricted/augmented Wong sequences (cf. Sect. 7 and [14]). They derived the following result.

Theorem 8.1 (Kalman Decomposition [8]) *For* $[E, A, B, C, 0] \in \Sigma_{l,n,m,p}$*, there exist* $S \in \mathbf{Gl}_l(\mathbb{R})$*,* $T \in \mathbf{Gl}_n(\mathbb{R})$ *such that*

$$[SET, SAT, SB, CT] = \left[\begin{bmatrix} E_{11} & E_{12} & E_{13} & E_{14} \\ 0 & E_{22} & 0 & E_{24} \\ 0 & 0 & E_{33} & E_{34} \\ 0 & 0 & 0 & E_{44} \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & A_{33} & A_{34} \\ 0 & 0 & 0 & A_{44} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \\ 0 \\ 0 \end{bmatrix}, [0, C_2, 0, C_4]\right], \tag{8.1}$$

where $E_{ij}, A_{ij} \in \mathbb{R}^{l_i \times n_j}$*,* $B_i \in \mathbb{R}^{l_i \times m}$*,* $C_j \in \mathbb{R}^{p \times n_j}$ *for* $i, j = 1, \ldots, 4$*, such that*

(i) $\left[\begin{bmatrix} E_{11} & E_{12} \\ 0 & E_{22} \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}\right] \in \mathscr{C}_{l_1+l_2, n_1+n_2, m}$ *is completely controllable and* $\operatorname{rk} \begin{bmatrix} E_{11} & E_{12} & B_1 \\ 0 & E_{22} & B_2 \end{bmatrix} = l_1 + l_2.$

(ii) $\left[\begin{bmatrix} E_{22} & E_{24} \\ 0 & E_{44} \end{bmatrix}, \begin{bmatrix} A_{22} & A_{24} \\ 0 & A_{44} \end{bmatrix}, [C_2, C_4]\right] \in \mathscr{O}_{l_2+l_4, n_2+n_4, p}$ *is completely observable.*

(iii) $\operatorname{rk}_{\mathbb{R}(s)} \begin{bmatrix} sE_{33} - A_{33} & sE_{34} - A_{34} \\ 0 & sE_{44} - A_{44} \end{bmatrix} = n_3 + n_4.$

(iv) $\operatorname{rk}_{\mathbb{R}(s)} \begin{bmatrix} sE_{11} - A_{11} & sE_{13} - A_{13} \\ 0 & sE_{33} - A_{33} \end{bmatrix} = l_1 + l_3.$

We would like to stress that several subtleties of the Kalman decomposition (8.1) are highlighted in [15, Remark 7.2] for a pure controllability decomposition and carry over to the general case.

Proposition 8.2 (Uniqueness of the Kalman Decomposition) *Let* $[E, A, B, C, 0] \in \Sigma_{l,n,m,p}$ *be given and assume that, for all* $i \in \{1, 2\}$*, the systems* $[E_i, A_i, B_i, C_i] = [S_i E T_i, S_i A T_i, S_i B, C T_i]$ *with*

$$sE_i - A_i = \begin{bmatrix} sE_{11,i} - A_{11,i} & sE_{12,i} - A_{12,i} & sE_{13,i} - A_{13,i} & sE_{14,i} - A_{14,i} \\ 0 & sE_{22,i} - A_{22,i} & 0 & sE_{24,i} - A_{24,i} \\ 0 & 0 & sE_{33,i} - A_{33,i} & sE_{34,i} - A_{34,i} \\ 0 & 0 & 0 & sE_{44,i} - A_{44,i} \end{bmatrix}, \quad B_i = \begin{bmatrix} B_{1,i} \\ B_{2,i} \\ 0 \\ 0 \end{bmatrix},$$

$$C_i = [0, C_{2,i}, 0, C_{4,i}]$$

where $E_{fg,i}, A_{fg,i} \in \mathbb{R}^{l_{f,i} \times n_{g,i}}$*,* $B_{f,i} \in \mathbb{R}^{l_{f,i} \times m}$*,* $C_g \in \mathbb{R}^{p \times n_{g,i}}$*,* $f, g = 1, \ldots, 4$*, satisfy the conditions (i)–(iv) in Theorem 8.1.*

Then $l_{j,1} = l_{j,2}$ *and* $n_{j,1} = n_{j,2}$ *for all* $j = 1, \ldots, 4$. *Moreover, for some* $W_{ij} \in \mathbb{R}^{l_{i,1} \times l_{j,1}}$, $T_{ij} \in \mathbb{R}^{n_{i,1} \times n_{j,1}}$ *such that* $\det W_{ii} \neq 0$ *and* $\det T_{ii} \neq 0$, $i, j = 1, \ldots, 4$, *we have*

$$W_2 W_1^{-1} = \begin{bmatrix} W_{11} & W_{12} & W_{13} & W_{14} \\ 0 & W_{22} & 0 & W_{24} \\ 0 & 0 & W_{33} & W_{34} \\ 0 & 0 & 0 & W_{44} \end{bmatrix}, \quad T_1^{-1} T_2 = \begin{bmatrix} T_{11} & T_{12} & T_{13} & T_{14} \\ 0 & T_{22} & 0 & T_{24} \\ 0 & 0 & T_{33} & T_{34} \\ 0 & 0 & 0 & T_{44} \end{bmatrix}.$$

Proof The result can be concluded from [15, Proposition 7.2] applied to $[E, A, B] \in \mathscr{C}_{l,n,m}$ and its dual $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ (invoking Corollary 5.1). □

Similar to [15, Corollary 7.3] several controllability, stabilizability, observability and detectability properties (and conditions for them) can be inferred for the subsystems appearing in the Kalman decomposition (8.1); we omit the details here.

The Kalman decomposition (8.1) is not satisfactory from a behavioral point of view: the trivial DAE $0 = x, y = 0$ given by $[0, I, 0, 0, 0]$ is behaviorally controllable and behaviorally observable, but in the decomposition (8.1) it is part of the uncontrollable and unobservable subsystem $[E_{33}, A_{33}, 0, 0]$. This is an unsatisfactory situation and is due to the fact that for DAE systems (both regular and singular) certain states can be inconsistent and it does not really make sense to label them controllable or uncontrollable (observable or unobservable, resp.). In the case of controllability decompositions this problem was treated in [19] and the following more detailed Kalman controllability decomposition was proved for $[E, A, B] \in \mathscr{C}_{l,n,m}$:

$$[SET, SAT, SB] = \left[\begin{bmatrix} E_{11} & E_{12} & E_{13} \\ 0 & E_{22} & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}, \begin{bmatrix} B_1 \\ 0 \\ 0 \end{bmatrix} \right],$$

where S and T are invertible matrices and the DAE system given by $[E_{11}, A_{11}, B_1]$ is completely controllable. Furthermore, E_{22} is invertible and the DAE $[E_{33}, A_{33}, 0]$ is such that it only has the trivial solution. Hence, we now have the decomposition into a completely controllable part, a classical uncontrollable part (given by an ODE) and an inconsistent part (which is behaviorally controllable but contains no completely controllable part). This decomposition seems to be more adequate for the analysis of DAE control systems as it takes into account the special DAE feature of possible inconsistent states which play a special role with respect to controllability. Using duality (see Sect. 5) we may derive the following analogous observability decomposition.

Theorem 8.3 (Kalman Observability Decomposition) *For* $[E, A, C] \in \mathscr{O}_{l,n,p}$ *there exist* $S \in \mathbf{Gl}_l(\mathbb{R})$ *and* $T \in \mathbf{Gl}_n(\mathbb{R})$ *such that*

$$[SET, SAT, CT] = \left[\begin{bmatrix} E_{11} & E_{12} & E_{13} \\ 0 & E_{22} & E_{23} \\ 0 & 0 & E_{33} \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{bmatrix}, [0, 0, C_3] \right], \tag{8.2}$$

where $E_{ij}, A_{ij} \in \mathbb{R}^{l_i \times n_j}$ *for* $i, j = 1, \ldots, 3$, $C_3 \in \mathbb{R}^{p \times n_3}$ *such that*

(i) $[E_{11}, A_{11}, 0] \in \mathscr{O}_{l_1,n_1,p}$ *with* $l_1 \leq n_1$ *and* $\operatorname{rk}_{\mathbb{C}}(\lambda E_{11} - A_{11}) = l_1$ *for all* $\lambda \in \mathbb{C}$,
(ii) $[E_{22}, A_{22}, 0] \in \mathscr{O}_{l_2,n_2,p}$ *with* $l_2 = n_2$ *and* E_{22} *is invertible,*
(iii) $[E_{33}, A_{33}, C_3] \in \mathscr{O}_{l_3,n_3,p}$ *is completely observable.*

Remark 8.1

(i) In the decomposition (8.2) we have an underdetermined and possibly inconsistent part $[E_{11}, A_{11}, 0]$, a classical unobservable part $[E_{22}, A_{22}, 0]$ and a completely observable part $[E_{33}, A_{33}, C_3]$. Note that furthermore
$\left[\begin{bmatrix} E_{11} & E_{13} \\ 0 & E_{33} \end{bmatrix}, \begin{bmatrix} A_{11} & A_{13} \\ 0 & A_{33} \end{bmatrix}, [0, C_3] \right]$ is RS behaviorally observable and
$\left[\begin{bmatrix} E_{22} & E_{23} \\ 0 & E_{33} \end{bmatrix}, \begin{bmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{bmatrix}, [0, C_3] \right]$ is observable at infinity.
(ii) Uniqueness of the Kalman observability decomposition (8.2) may be analyzed similar to Proposition 8.2 and [19, Theorem 3.5]; in particular the block sizes are unique.
(iii) Similar to [19, Theorem 3.3] it is possible to derive the decomposition (8.2) with the help of the restricted Wong sequences which were introduced in Sect. 7. In fact, the subspace decomposition leading to (8.2) is uniquely determined by the restricted Wong sequences. Also note that, especially in the singular case, the decomposition (8.2) bears several subtleties which can be analyzed similar to [19, Remark 3.2].
(iv) It is also possible to extend the pure observability decomposition (8.2) to a Kalman decomposition of the form (8.1) where additionally the classical (ODE) uncontrollable and unobservable parts are decomposed. However, due to the complexity of such a decomposition we omit it here.

9 Detectability and Stabilization by Output Injection

In this section we introduce detectability concepts for DAE systems. We characterize them in terms of the OI normal form and derive duality to the respective stabilizability concepts from [15]. This will enable us to infer algebraic criteria for the detectability concepts and to finally show that stabilization and index reduction can be achieved by output injection.

In general, detectability is a weaker version of observability in the sense that the state x is not exactly determined by the external signals but only asymptotically. In the following, we will use the simplified notation "$x(t) \to 0$ as $t \to \infty$" for $x \in \mathscr{L}^1_{\mathrm{loc}}(\mathbb{R}; \mathbb{R}^n)$ if, and only if,

$$\lim_{t\to\infty} \operatorname*{ess\,sup}_{\tau\in[t,\infty)} \|x(\tau)\| = 0.$$

Definition 9.1 The system $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ is called

(a) *behaviorally detectable*

$$:\Longleftrightarrow\ \forall\,(x^1,u,y),(x^2,u,y)\in\mathfrak{B}_{[E,A,B,C,D]}:\quad x^1(t)-x^2(t)\to 0 \text{ as } t\to\infty,$$

(b) *RS behaviorally detectable*

$$:\Longleftrightarrow\ \forall\,(x^1,u,y),(x^2,u,y)\in\mathfrak{B}_{[E,A,B,C,D]}\ \exists\,(x^3,u,y)\in\mathfrak{B}_{[E,A,B,C,D]}:$$
$$x^1_{(-\infty,0)} = x^3_{(-\infty,0)}\ \wedge\ x^2(t)-x^3(t)\to 0 \text{ as } t\to\infty,$$

(c) *strongly detectable*, if it is impulse observable and behaviorally detectable,
(d) *completely detectable*, if it is observable at infinity and behaviorally detectable,
(e) *RS strongly detectable*, if it is RS impulse observable and RS behaviorally detectable,
(f) *RS completely detectable*, if it is RS observable at infinity and RS behaviorally detectable.

The definitions of RS complete and strong detectability are motivated by the corresponding characterizations of RS complete and strong observability (see Fig. 1) in terms of RS observability at infinity, RS impulse observability and RS behavioral observability, where the latter is replaced by RS behavioral detectability. Similar as for the observability concepts, the detectability definitions can be simplified due to linearity.

Lemma 9.1 *The system* $[E, A, B, C, D] \in \Sigma_{l,n,m,p}$ *is*

(a) behaviorally detectable

$$\Longleftrightarrow\ \forall\,(x,0)\in\mathfrak{B}_{[E,A,C]}:\quad x(t)\to 0\ \textit{as}\ t\to\infty,$$

(b) RS behaviorally detectable

$$\Longleftrightarrow\ \forall\,(x,0)\in\mathfrak{B}_{[E,A,C]}\ \exists\,(\overline{x},0)\in\mathfrak{B}_{[E,A,C]}:$$
$$x_{(-\infty,0)} = \overline{x}_{(-\infty,0)}\ \wedge\ \overline{x}(t)\to 0\ \textit{as}\ t\to\infty.$$

Hence we may restrict our attention to systems in $\mathscr{O}_{l,n,p}$ and we can use the OI normal form (4.4) to obtain (similar to the observability characterizations given in Table 1) the following characterizations of the detectability concepts.

Corollary 9.2 (Detectability and OI Normal Form) *Let $[E, A, C] \in \mathscr{O}_{l,n,p}$ with OI normal form (4.4). Then $[E, A, C]$ is*

(a) behaviorally detectable if, and only if, $\ell(\varepsilon) = 0$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.
(b) RS behaviorally detectable if, and only if, $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.
(c) strongly detectable if, and only if, $\gamma = (1, \ldots, 1)$, $\ell(\varepsilon) = 0$, $\kappa = (1, \ldots, 1)$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.
(d) completely detectable if, and only if, $\gamma = (1, \ldots, 1)$, $\ell(\varepsilon) = 0$, $\ell(\kappa) = 0$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.
(e) RS strongly detectable if, and only if, $\gamma = (1, \ldots, 1)$, $\varepsilon = (1, \ldots, 1)$, $\kappa = (1, \ldots, 1)$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.
(f) RS completely detectable if, and only if, $\gamma = (1, \ldots, 1)$, $\varepsilon = (1, \ldots, 1)$, $\ell(\kappa) = 0$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$.

Using the OI normal form, the characterizations in Corollary 9.2 and the respective results for the feedback form derived in [15, Corollary 3.4], we are able to infer duality between detectability and stabilizability as follows.

Corollary 9.3 (Duality of Detectability and Stabilizability) *Let $[E, A, C] \in \mathscr{O}_{l,n,p}$ be given. Then we have the following equivalences:*

(a) $[E, A, C]$ is RS behaviorally detectable if, and only if, $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ is behaviorally stabilizable in the sense of [15].
(b) $[E, A, C]$ is RS strongly detectable if, and only if, $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ is strongly stabilizable in the sense of [15].
(c) $[E, A, C]$ is RS completely detectable if, and only if, $[E^\top, A^\top, C^\top] \in \mathscr{C}_{n,l,p}$ is completely stabilizable in the sense of [15].

In particular, for regular DAE systems, behavioral, strong and complete detectability are dual to behavioral, strong and complete stabilizability. The duality properties are illustrated in Fig. 3.

As a consequence of Corollaries 9.2 and 9.3 and [15, Corollary 4.3] we obtain the following algebraic criteria of Hautus type for the detectability concepts from Definition 9.1.

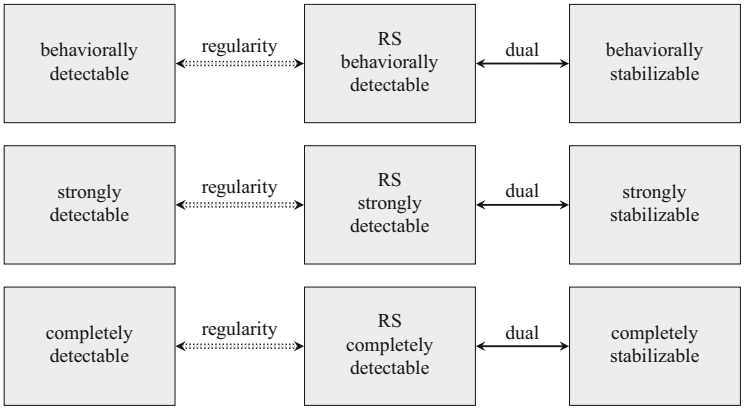


Fig. 3 Illustration of duality between detectability and stabilizability

Corollary 9.4 (Algebraic Criteria for Detectability) *Let* $[E, A, C] \in \mathscr{O}_{l,n,p}$ *and* Z *be a matrix with* $\operatorname{im}_{\mathbb{R}} Z = \ker_{\mathbb{R}} E^{\top}$. *Then we have the following:*

$[E, A, C]$ is	if, and only if,
Behaviorally detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = n$
RS behaviorally detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}(s)} \begin{bmatrix} sE - A \\ C \end{bmatrix}$
Strongly detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^{\top} A \\ C \end{bmatrix} = n$
Completely detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = n$
RS strongly detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ Z^{\top} A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$
RS completely detectable	$\forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}} \begin{bmatrix} \lambda E - A \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ C \end{bmatrix} = \operatorname{rk}_{\mathbb{R}} \begin{bmatrix} E \\ A \\ C \end{bmatrix}$

Remark 9.1 Behavioral detectability was investigated in [32] for regular systems, where it is called detectability. In this case, the algebraic criteria for RS behavioral detectability from Corollary 9.4 have been derived in [32, Theorem 3-1.3].

In the remainder of this section we consider stabilization and index reduction by output injection. As explained in Sect. 4, a system $[E, A, C] \in \mathscr{O}_{l,n,p}$ can, via output injection with some $L \in \mathbb{R}^{l\times p}$, be turned into a DAE of the form (4.1), that is a new system $[E, A + LC, C] \in \mathscr{O}_{l,n,p}$. It is our aim to choose L such that this new system is stable in a certain sense and its index is at most one. The *index* $\nu \in \mathbb{N}_0$ of a matrix pencil $sE - A \in \mathbb{R}[s]^{l\times n}$ is defined via its (quasi-)Kronecker form [17, 18, 36] as in [15, Definition 3.2]: if for some $S \in \mathbf{Gl}_l(\mathbb{R})$ and $T \in \mathbf{Gl}_n(\mathbb{R})$

$$S(sE-A)T = \begin{bmatrix} sI_r - J & 0 & 0 & 0 \\ 0 & sN_\alpha - I_{|\alpha|} & 0 & 0 \\ 0 & 0 & sK_\beta - L_\beta & 0 \\ 0 & 0 & 0 & sK_\gamma^\top - L_\gamma^\top \end{bmatrix}, \tag{9.1}$$

then $\nu = \max\{0, \alpha_1, \ldots, \alpha_{\ell(\alpha)}, \gamma_1, \ldots, \gamma_{\ell(\gamma)}\}$.

The index is independent of the choice of S, T and can be computed via the Wong sequences corresponding to $sE - A$ as shown in [17, 18].

The following result can now be inferred from Corollaries 5.1 and 9.3 and [15, Theorem 5.3].

Proposition 9.5 (Stabilization and Index Reduction) *For a system $[E, A, C] \in \mathscr{O}_{l,n,p}$ the following hold true:*

(a) $[E, A, C]$ is RS impulse observable if, and only if, there exists $L \in \mathbb{R}^{l\times p}$ such that the index of $sE^\top - (A + LC)^\top$ is at most one.

(b) $[E, A, C]$ is RS strongly detectable if, and only if, there exists $L \in \mathbb{R}^{l\times p}$ such that the index of $sE^\top - (A + LC)^\top$ is at most one and the pair $[E, A + LC]$ is behaviorally stabilizable in the sense of [15, Definition 5.1].

If we consider square systems $[E, A, C] \in \mathscr{O}_{n,n,p}$, then we may obtain an additional stabilization result via behavioral detectability which is false in general in the nonregular case. To this end, we call a system $[E, A, C] \in \mathscr{O}_{l,n,p}$ *behaviorally stable*, if $[E, A, 0]$ is behaviorally detectable. From Corollary 9.4 we obtain the characterization

$$[E, A, C] \text{ is behaviorally stable} \iff \forall\, \lambda \in \overline{\mathbb{C}_+} : \operatorname{rk}_{\mathbb{C}}(\lambda E - A) = n. \tag{9.2}$$

Furthermore, under the slightly stronger assumptions of impulse observability and strong detectability, resp., the results of Proposition 9.5 can be improved for square systems. It is then possible to show that the output injection leads to a system which is additionally regular. The equivalence of impulse observability and regularizability with index reduction by output injection (statement (a) of the following theorem) has been proved in [24], see also [25]. For completeness we provide a new proof using the OI normal form. To the best of our knowledge, statements (b) and (c) of the following theorem are new.

Theorem 9.6 (Stabilization and Index Reduction for Square Systems) *For a system* $[E, A, C] \in \mathscr{O}_{n,n,p}$ *the following hold true:*

(a) $[E, A, C]$ *is impulse observable if, and only if, there exists* $L \in \mathbb{R}^{n\times p}$ *such that* $sE - (A + LC)$ *is regular and its index is at most one.*

(b) $[E, A, C]$ *is behaviorally detectable if, and only if, there exists* $L \in \mathbb{R}^{n\times p}$ *such that* $sE - (A + LC)$ *is regular and* $[E, A + LC, C]$ *is behaviorally stable.*

(c) $[E, A, C]$ *is strongly detectable if, and only if, there exists* $L \in \mathbb{R}^{n\times p}$ *such that* $sE - (A + LC)$ *is regular, its index is at most one and* $[E, A + LC, C]$ *is behaviorally stable.*

Proof

(a) Without loss of generality, we may assume that $[E, A, C]$ is in OI normal form (4.4). First let $[E, A, C]$ be impulse observable, and hence it follows from Table 1 that $\gamma = (1, \ldots, 1)$, $\ell(\varepsilon) = 0$ and $\kappa = (1, \ldots, 1)$. Since E and A are square we may further deduce that $\ell(\beta) = \ell(\gamma)$, and therefore

$$E = \begin{bmatrix} I_{|\alpha|} & 0 & 0\,0 & 0 \\ 0 & K_\beta^\top & 0\,0 & 0 \\ 0 & 0 & 0\,0 & 0 \\ 0 & 0 & 0\,0 & I_{n_{\overline{\sigma}}} \end{bmatrix}, \quad A = \begin{bmatrix} N_\alpha & 0 & 0 & 0 & 0 \\ 0 & L_\beta^\top & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{|\kappa|} & 0 \\ 0 & 0 & 0 & 0 & A_{\overline{\sigma}} \end{bmatrix}, \quad C = \begin{bmatrix} E_\alpha^\top & 0 & 0 & 0\,0 \\ 0 & 0 & I_{|\gamma|} & 0\,0 \\ 0 & 0 & 0 & 0\,0 \end{bmatrix}. \tag{9.3}$$

It is easy to see that

$$s[K_\beta^\top, 0] - [L_\beta^\top, E_\beta] = S\left(s\begin{bmatrix} I_{|\beta-1|} & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} N_{\beta-1} & 0 \\ 0 & I_{\ell(\beta)} \end{bmatrix}\right) T$$

for some invertible matrices S, T, where $\beta - 1 = (\beta_1 - 1, \ldots, \beta_{\ell(\beta)} - 1)$. Therefore, the pencil $s[K_\beta^\top, 0] - [L_\beta^\top, E_\beta]$ is regular and has index at most one. Choosing

$$L = \begin{bmatrix} 0 & 0 & 0 \\ 0 & E_\beta & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we obtain that

$$sE - (A + LC) = \begin{bmatrix} sI_{|\alpha|} - N_\alpha & 0 & 0 & 0 & 0 \\ 0 & sK_\beta^\top - L_\beta^\top & -E_\beta & 0 & 0 \\ 0 & 0 & 0 & -I_{|\kappa|} & 0 \\ 0 & 0 & 0 & 0 & sI_{n_{\overline{\sigma}}} - A_{\overline{\sigma}} \end{bmatrix}$$

is regular and its index is at most one.

To show the opposite implication let $L \in \mathbb{R}^{n\times p}$ be such that $sE-(A+LC)$ is regular and its index is at most one. Then Proposition 9.5 implies that $[E,A,C]$ is RS impulse observable. To show impulse observability, by Table 1 it remains to show that $\ell(\varepsilon)=0$. Since an OI normal form of $[E,A,C]$ is also an OI normal form of $[E,A+LC,C]$, it follows from the regularity of $sE-(A+LC)$ and Remark 4.2 that $\ell(\varepsilon)=0$.

(b) Again, we assume that $[E,A,C]$ is in OI normal form (4.4). First let $[E,A,C]$ be behaviorally detectable, and hence it follows from Corollary 9.2 that $\ell(\varepsilon)=0$ and $\sigma(A_{\overline{\sigma}})\subseteq\mathbb{C}_-$. Since E and A are square we may further deduce that $\ell(\beta)=\ell(\gamma)$, and therefore

$$E=\begin{bmatrix} I_{|\alpha|} & 0 & 0 & 0 & 0\\ 0 & K_\beta^\top & 0 & 0 & 0\\ 0 & 0 & L_\gamma & 0 & 0\\ 0 & 0 & 0 & N_\kappa^\top & 0\\ 0 & 0 & 0 & 0 & I_{n_{\overline{\sigma}}}\end{bmatrix},\quad A=\begin{bmatrix} N_\alpha & 0 & 0 & 0 & 0\\ 0 & L_\beta^\top & 0 & 0 & 0\\ 0 & 0 & K_\gamma & 0 & 0\\ 0 & 0 & 0 & I_{|\kappa|} & 0\\ 0 & 0 & 0 & 0 & A_{\overline{\sigma}}\end{bmatrix},\quad C=\begin{bmatrix} E_\alpha^\top & 0 & 0 & 0\,0\\ 0 & 0 & E_\gamma^\top & 0\,0\\ 0 & 0 & 0 & 0\,0\end{bmatrix}.$$

By [79, Theorem 4.20] there exists $F_\alpha\in\mathbb{R}^{|\alpha|\times\ell(\alpha)}$ such that $\sigma(N_\alpha+F_\alpha E_\alpha^\top)\subseteq\mathbb{C}_-$. Furthermore, choosing

$$F_\beta=\operatorname{diag}(e_1^{[\beta_1]},\dots,e_1^{[\beta_{\ell(\beta)}]})$$

we find that, by the same argument as in the proof of [16, Theorem 3.5],

$$s\begin{bmatrix}K_\beta^\top & 0\\ 0 & L_\gamma\end{bmatrix}-\begin{bmatrix}L_\beta^\top & F_\beta E_\gamma^\top\\ 0 & K_\gamma\end{bmatrix}=S\left(s\begin{bmatrix}N_\beta^\top & 0\\ * & N_{\gamma-1}^\top\end{bmatrix}-\begin{bmatrix}I_{|\beta|} & 0\\ 0 & I_{|\gamma-1|}\end{bmatrix}\right)T$$

for some invertible matrices S,T, where $\gamma-1=(\gamma_1-1,\dots,\gamma_{\ell(\gamma)}-1)$. Therefore, the pencil $s\left[\begin{smallmatrix}K_\beta^\top & 0\\ 0 & L_\gamma\end{smallmatrix}\right]-\left[\begin{smallmatrix}L_\beta^\top & F_\beta E_\gamma^\top\\ 0 & K_\gamma\end{smallmatrix}\right]$ is regular and the system $\left[\left[\begin{smallmatrix}K_\beta^\top & 0\\ 0 & L_\gamma\end{smallmatrix}\right],\left[\begin{smallmatrix}L_\beta^\top & F_\beta E_\gamma^\top\\ 0 & K_\gamma\end{smallmatrix}\right],[0,E_\gamma^\top]\right]$ is behaviorally stable by (9.2). Choosing

$$L=\begin{bmatrix}F_\alpha & 0 & 0\\ 0 & F_\beta & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0\end{bmatrix},$$

we obtain that

$$sE-(A+LC) = \begin{bmatrix} sI_{|\alpha|}-(N_\alpha+F_\alpha E_\alpha^\top) & 0 & 0 & 0 & 0 \\ 0 & sK_\beta^\top - L_\beta^\top & -F_\beta E_\gamma^\top & 0 & 0 \\ 0 & 0 & sL_\gamma - K_\gamma & 0 & 0 \\ 0 & 0 & 0 & sN_\kappa^\top - I_{|\kappa|} & 0 \\ 0 & 0 & 0 & 0 & sI_{n_{\overline{\sigma}}} - A_{\overline{\sigma}} \end{bmatrix}$$

is regular and $[E, A+LC, C]$ is behaviorally stable by (9.2).

To show the opposite implication let $L \in \mathbb{R}^{n\times p}$ be such that $sE-(A+LC)$ is regular and $[E, A+LC, C]$ is behaviorally stable. Seeking a contradiction, assume that $[E, A, C]$ is not behaviorally detectable. Then it follows from Corollary 9.4 that there exist $\lambda \in \overline{\mathbb{C}_+}$ and $x \in \mathbb{C}^n \setminus \{0\}$ such that $\left[\begin{smallmatrix} \lambda E-A \\ C \end{smallmatrix}\right] x = 0$. This implies

$$(\lambda E-(A+LC))\,x = [I_n, -L]\begin{bmatrix} \lambda E-A \\ C \end{bmatrix} x = 0,$$

thus $\mathrm{rk}_{\mathbb{C}}(\lambda E-(A+LC)) < n$ which contradicts behavioral stability of $[E, A+LC, C]$ by (9.2).

(c) Again, we assume that $[E, A, C]$ is in OI normal form (4.4). First let $[E, A, C]$ be strongly detectable, and hence it follows from Corollary 9.2 that $\gamma = (1, \ldots, 1)$, $\ell(\varepsilon) = 0$, $\kappa = (1, \ldots, 1)$ and $\sigma(A_{\overline{\sigma}}) \subseteq \mathbb{C}_-$. Since E and A are square we may further deduce that $\ell(\beta) = \ell(\gamma)$ and (9.3) holds. Let $F_\alpha \in \mathbb{R}^{|\alpha|\times\ell(\alpha)}$ be such that $\sigma(N_\alpha + F_\alpha E_\alpha^\top) \subseteq \mathbb{C}_-$. Furthermore, let

$$a_j = [a_{j0}, \ldots, a_{j\beta_j-2}, 1]^\top \in \mathbb{R}^{\beta_j}$$

with the property that the polynomials

$$p_j(s) = s^{\beta_j} + a_{j\beta_j-1}s^{\beta_j-1} + \ldots + a_{j0} \in \mathbb{R}[s]$$

are Hurwitz for $j = 1, \ldots, \ell(\beta)$, and let

$$B_\beta = \mathrm{diag}\,(a_1, \ldots, a_{\ell(\beta)}) \in \mathbb{R}^{|\beta|\times\ell(\beta)}.$$

Consider the system

$$\tfrac{\mathrm{d}}{\mathrm{d}t}[K_\beta^\top, 0]\begin{pmatrix} z(t) \\ u(t) \end{pmatrix} = [L_\beta^\top, B_\beta]\begin{pmatrix} z(t) \\ u(t) \end{pmatrix}. \tag{9.4}$$

We see that the input u is uniquely determined by $u = -E_{\beta-1}^{\top} z$, where $\beta - 1 = (\beta_1 - 1, \ldots, \beta_{\ell(\beta)} - 1)$ and if $\beta_j = 1$ for some j, then the respective x-component does not exist and the equation simply reads $u_j = 0$. With $B_{\beta-1} = \operatorname{diag}(\tilde{a}_1, \ldots, \tilde{a}_{\ell(\beta)})$, where $\tilde{a}_j = [a_{j0}, \ldots, a_{j\beta_j-2}]^{\top}$, a permutation of rows in (9.4) and insertion of u gives

$$\begin{aligned} \dot{z}(t) &= (N_{\beta-1} - B_{\beta-1} E_{\beta-1}^{\top}) z(t), \\ u(t) &= E_{\beta-1}^{\top} z(t). \end{aligned}$$

It is now clear that the pencil $s[K_{\beta}^{\top}, 0] - [L_{\beta}^{\top}, B_{\beta}]$ in system (9.4) is regular and has index at most one. Furthermore, the characteristic polynomial of $N_{\beta-1} + B_{\beta-1} E_{\beta-1}^{\top}$ (which is a block diagonalization of companion matrices) is given by

$$\det\left(sI - (N_{\beta-1} + B_{\beta-1} E_{\beta-1}^{\top})\right) = \prod_{j=1}^{\ell(\beta)} p_j(s),$$

which is Hurwitz, since all $p_j(s)$ are Hurwitz. Therefore, $\left[\, [K_{\beta}^{\top},\, 0\,], [L_{\beta}^{\top},\, B_{\beta}], \right.$ $\left. [0, I_{|\gamma|}] \right]$ is also behaviorally stable. Choosing

$$L = \begin{bmatrix} F_{\alpha} & 0 & 0 \\ 0 & B_{\beta} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we obtain that

$$sE - (A + LC) = \begin{bmatrix} sI_{|\alpha|} - (N_{\alpha} + F_{\alpha} E_{\alpha}^{\top}) & 0 & 0 & 0 & 0 \\ 0 & sK_{\beta}^{\top} - L_{\beta}^{\top} & -B_{\beta} & 0 & 0 \\ 0 & 0 & 0 & -I_{|\kappa|} & 0 \\ 0 & 0 & 0 & 0 & sI_{n_{\overline{\sigma}}} - A_{\overline{\sigma}} \end{bmatrix}$$

is regular, its index is at most one and $[E, A + LC, C]$ is behaviorally stable by (9.2).

To show the opposite implication let $L \in \mathbb{R}^{n \times p}$ be such that $sE - (A + LC)$ is regular, its index is at most one and $[E, A + LC, C]$ is behaviorally stable. Then Proposition 9.5 implies that $[E, A, C]$ is RS strongly detectable. To show strong detectability, by Table 1 and Corollary 9.2 it remains to show that $\ell(\varepsilon) = 0$. As in (a), this follows from the regularity of $sE - (A + LC)$.

□

Note that in the proof of necessity in Theorem 9.6 (b) the regularity of $sE - (A + LC)$ has not been used explicitly, so one may wonder whether this property is necessary here. In fact, it is not: the regularity of $sE - (A + LC)$ is a direct consequence of the behavioral stability of $[E, A + LC, C]$ and the fact that E and $A + LC$ are square.

Remark 9.2

(i) It is a consequence of Theorem 9.6 that impulse observability or behavioral detectability in particular implies that the square system $[E, A, C] \in \mathscr{O}_{n,n,p}$ is regularizable by output injection, i.e., there exists $L \in \mathbb{R}^{n\times p}$ such that $sE-(A+LC)$ is regular. The dual of this concept is regularizability by state feedback and has been well investigated, see [16] and the references therein.

(ii) Another result on index reduction which is slightly different from both Proposition 9.5 (a) and Theorem 9.6 (a) was derived in [43, Theorem 5]. It is shown that $[E, A, C] \in \mathscr{O}_{l,n,p}$ is impulse observable if, and only if, there exists $L \in \mathbb{R}^{l\times p}$ such that

$$(A + LC)^{-1}(\mathrm{im}_{\mathbb{R}}\, E) \cap \ker_{\mathbb{R}} E = \{0\},$$

which is slightly stronger than to require that $sE - (A + LC)$ has index at most one; in fact, it is equivalent to the index being at most one and the absence of overdetermined γ-blocks in the quasi-Kronecker form (9.1).

(iii) Stabilization and index reduction by output injection for regular DAE systems have been investigated in [32]. In particular, under the additional assumption of regularity of $sE - A$, Theorem 9.6 (a) and (b) have been derived in [32, Theorem 3-2.1 and Corollary 3-3.2], resp.

Acknowledgements We thank the referees of this article for their valuable comments which very much helped to improve the manuscript.

References

1. Aplevich, J.D.: Minimal representations of implicit linear systems. Automatica **21**(3), 259–269 (1985). doi:10.1016/0005-1098(85)90059-7
2. Aplevich, J.D.: Implicit Linear Systems, vol. 152. Lecture Notes in Control and Information Sciences. Springer, Berlin (1991). doi:10.1007/BFb0044363
3. Armentano, V.A.: The pencil $(sE - A)$ and controllability-observability for generalized linear systems: a geometric approach. SIAM J. Control Optim. **24**, 616–638 (1986). doi:10.1137/0324037
4. Aubin, J.P., Cellina, A.: Differential Inclusions: Set-Valued Maps and Viability Theory, vol. 264. Grundlehren der mathematischen Wissenschaften. Springer, Berlin (1984). doi:10.1007/978-3-642-69512-4
5. Aubin, J.P., Frankowska, H.: Set Valued Analysis. Birkhäuser, Boston (1990). doi:10.1007/978-0-8176-4848-0

6. Banaszuk, A., Kocięcki, M., Przyłuski, K.M.: Implicit linear discrete-time systems. Math. Control Signals Syst. **3**(3), 271–297 (1990). doi:10.1007/BF02551372
7. Banaszuk, A., Kocięcki, M., Przyłuski, K.M.: Remarks on observability of implicit linear discrete-time systems. Automatica **26**(2), 421–423 (1990). doi:10.1016/0005-1098(90)90140-D
8. Banaszuk, A., Kocięcki, M., Lewis, F.L.: Kalman decomposition for implicit linear systems. IEEE Trans. Autom. Control **37**(10), 1509–1514 (1992). doi:10.1109/9.256370
9. Banaszuk, A., Kocięcki, M., Przyłuski, K.M.: On duality between observation and control for implicit linear discrete-time systems. IMA J. Math. Control Inf. **13**, 41–61 (1996). doi:10.1093/imamci/13.1.41
10. Baser, U., Özçaldiran, K.: On observability of singular systems. Circuits Systems Signal Process. **11**(3), 421–430 (1992). doi:10.1007/BF01190985
11. Basile, G., Marro, G.: Controlled and Conditioned Invariants in Linear System Theory. Prentice-Hall, Englewood Cliffs, NJ (1992)
12. Belevitch, V.: Classical Network Theory. Holden-Day, San Francisco (1968)
13. Bender, D.J., Laub, A.J.: Controllability and observability at infinity of multivariable linear second-order models. IEEE Trans. Autom. Control **AC-30**, 1234–1237 (1985). doi:10.1109/TAC.1985.1103869
14. Berger, T.: On differential-algebraic control systems. Ph.D. thesis, Institut für Mathematik, Technische Universität Ilmenau, Universitätsverlag Ilmenau, Ilmenau, Germany (2014). http://www.db-thueringen.de/servlets/DocumentServlet?id=22652
15. Berger, T., Reis, T.: Controllability of linear differential-algebraic systems - a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I, Differential-Algebraic Equations Forum, pp. 1–61. Springer, Berlin (2013). doi:10.1007/978-3-642-34928-7_1
16. Berger, T., Reis, T.: Regularization of linear time-invariant differential-algebraic systems. Syst. Control Lett. **78**, 40–46 (2015). doi:10.1016/j.sysconle.2015.01.013
17. Berger, T., Trenn, S.: The quasi-Kronecker form for matrix pencils. SIAM J. Matrix Anal. Appl. **33**(2), 336–368 (2012). doi:10.1137/110826278
18. Berger, T., Trenn, S.: Addition to "The quasi-Kronecker form for matrix pencils". SIAM J. Matrix Anal. Appl. **34**(1), 94–101 (2013). doi:10.1137/120883244
19. Berger, T., Trenn, S.: Kalman controllability decompositions for differential-algebraic systems. Syst. Control Lett. **71**, 54–61 (2014). doi:10.1016/j.sysconle.2014.06.004
20. Berger, T., Ilchmann, A., Trenn, S.: The quasi-Weierstraß form for regular matrix pencils. Linear Alg. Appl. **436**(10), 4052–4069 (2012). doi:10.1016/j.laa.2009.12.036
21. Bernhard, P.: On singular implicit linear dynamical systems. SIAM J. Control Optim. **20**(5), 612–633 (1982). doi:10.1137/0320046
22. Birkhoff, G., MacLane, S.: A Survey of Modern Algebra, 4th edn. Macmillan Publishing Co, New York (1977)
23. Bonilla, E.M., Malabre, M.: On the control of linear systems having internal variations. Automatica **39**, 1989–1996 (2003)
24. Bunse-Gerstner, A., Mehrmann, V., Nichols, N.K.: Regularization of descriptor systems by output feedback. IEEE Trans. Autom. Control **39**(8), 1742–1748 (1994). doi:10.1109/9.310065
25. Bunse-Gerstner, A., Byers, R., Mehrmann, V., Nichols, N.K.: Feedback design for regularizing descriptor systems. Linear Algebra Appl. **299**, 119–151 (1999). doi:10.1016/S0024-3795(99)00167-6
26. Byers, R., Geerts, A.H.W.T., Mehrmann, V.: Descriptor systems without controllability at infinity. SIAM J. Control Optim. **35**, 462–479 (1997). doi:10.1137/S0363012994269818
27. Campbell, S.L., Nichols, N.K., Terrell, W.J.: Duality, observability, and controllability for linear time-varying descriptor systems. Circuits Syst. Signal Process. **10**(4), 455–470 (1991). doi:10.1007/BF01194883
28. 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, vol. 23, pp. 17–36. SIAM, Philadelphia (2012)

29. Christodoulou, M.A., Paraskevopoulos, P.N.: Solvability, controllability, and observability of singular systems. J. Optim. Theorem Appl. **45**, 53–72 (1985). doi:10.1007/BF00940813
30. Cobb, J.D.: On the solution of linear differential equations with singular coefficients. J. Diff. Equ. **46**, 310–323 (1982). doi:10.1016/0022-0396(82)90097-3
31. Cobb, J.D.: Controllability, observability and duality in singular systems. IEEE Trans. Autom. Control **AC-29**, 1076–1082 (1984). doi:10.1109/TAC.1984.1103451
32. Dai, L.: Singular Control Systems, vol. 118. Lecture Notes in Control and Information Sciences. Springer, Berlin (1989). doi:10.1007/BFb0002475
33. Darouach, M., Boutayeb, M., Zasadzinski, M.: Kalman filtering for continuous descriptor systems. In: Proceedings of American Control Conference 1997, pp. 2108–2112. Albuquerque, NM (1997). doi:10.1109/ACC.1997.611062
34. Dieudonné, J.: Sur la réduction canonique des couples des matrices. Bull. de la Societé Mathématique de France **74**, 130–146 (1946). http://eudml.org/doc/86796
35. Frankowska, H.: On controllability and observability of implicit systems. Syst. Control Lett. **14**, 219–225 (1990). doi:10.1016/0167-6911(90)90016-N
36. Gantmacher, F.R.: The Theory of Matrices, vols. I & II. Chelsea, New York (1959)
37. Geerts, A.H.W.T.: Solvability conditions, consistency and weak consistency for linear differential-algebraic equations and time-invariant linear systems: the general case. Linear Alg. Appl. **181**, 111–130 (1993). doi:10.1016/0024-3795(93)90027-L
38. Geerts, A.H.W.T., Mehrmann, V.: Linear differential equations with constant coefficients: a distributional approach. Technical Report SFB 343 90-073, Bielefeld University, Germany (1990)
39. Hautus, M.L.J.: Controllability and observability condition for linear autonomous systems. Ned. Akademie. Wetenschappen, Proc. Ser. A **72**, 443–448 (1969)
40. Hou, M., Müller, P.C.: Causal observability of descriptor systems. IEEE Trans. Autom. Control **44**(1), 158–163 (1999). doi:10.1109/9.739111
41. Hou, M., Patton, R.: Input observability and input reconstruction. Automatica **34**(6), 789–794 (1988). doi:10.1016/S0005-1098(98)00021-1
42. Ilchmann, A., Mehrmann, V.: A behavioural approach to time-varying linear systems, Part 1: general theory. SIAM J. Control Optim. **44**(5), 1725–1747 (2005). doi:10.1137/S0363012904442239
43. Ishihara, J.Y., Terra, M.H.: Impulse controllability and observability of rectangular descriptor systems. IEEE Trans. Autom. Control **46**(6), 991–994 (2001). doi:10.1109/9.928613
44. Kalman, R.E.: On the general theory of control systems. In: Proceedings of the First International Congress on Automatic Control, Moscow 1960, pp. 481–493. Butterworth's, London (1961)
45. Kalman, R.E.: Canonical structure of linear dynamical systems. Proc. Nat. Acad. Sci. USA **48**(4), 596–600 (1962). http://www.ncbi.nlm.nih.gov/pmc/articles/PMC220821
46. Kalman, R.E.: Mathematical description of linear dynamical systems. SIAM J. Control Optim. **1**, 152–192 (1963). doi:10.1137/0301010
47. Karcanias, N.: Regular state-space realizations of singular system control problems. In: Proceedings of 26th IEEE Conference Decision Control, pp. 1144–1146. Los Angeles, CA (1987). doi:10.1109/CDC.1987.272588
48. Knobloch, H.W., Kwakernaak, H.: Lineare Kontrolltheorie. Springer, Berlin (1985)
49. Koumboulis, F.N., Mertzios, B.G.: On Kalman's controllability and observability criteria for singular systems. Circuits Syst. Signal Process. **18**(3), 269–290 (1999). doi:10.1007/BF01225698
50. Kronecker, L.: Algebraische Reduction der Schaaren bilinearer Formen. Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu, Berlin, pp. 1225–1237 (1890)
51. Kuijper, M.: First-Order Representations of Linear Systems. Birkhäuser, Boston (1994). doi:10.1007/978-1-4612-0259-2
52. Kunkel, P., Mehrmann, V.: Differential-algebraic equations. analysis and numerical solution. EMS Publishing House, Zürich, Switzerland (2006). doi:10.4171/017

53. Lewis, F.L.: A survey of linear singular systems. IEEE Proc. Circuits, Syst. Signal Process. **5**(1), 3–36 (1986). doi:10.1007/BF01600184
54. Lewis, F.L.: A tutorial on the geometric analysis of linear time-invariant implicit systems. Automatica **28**(1), 119–137 (1992). doi:10.1016/0005-1098(92)90012-5
55. Loiseau, J.J., Lebret, G.: A new canonical form for descriptor systems with outputs. In: Bensoussan, A., Lions, J.L. (eds.) Analysis and Optimization of Systems. Lecture Notes in Control and Information Sciences, vol. 144, pp. 371–380. Springer, Berlin (1990). doi:10.1007/BFb0120060
56. Loiseau, J.J., Özçaldiran, K., Malabre, M., Karcanias, N.: Feedback canonical forms of singular systems. Kybernetika **27**(4), 289–305 (1991). http://dml.cz/dmlcz/124568
57. Lomadze, V.: Duality in the behavioral systems theory. Automatica **49**(5), 1510–1514 (2013). doi:10.1016/j.automatica.2013.02.007
58. Luenberger, D.G.: Observing a state of a linear system. IEEE Trans. Mil. Electron. **MIL-8**, 74–80 (1964). doi:10.1109/TME.1964.4323124
59. Luenberger, D.G.: Observers for multivariable systems. IEEE Trans. Autom. Control **AC-11**(2), 190–197 (1966). doi:10.1109/TAC.1966.1098323
60. Luenberger, D.G.: An introduction to observers. IEEE Trans. Autom. Control **16**(6), 596–602 (1971). doi:10.1109/TAC.1971.1099826
61. Malabre, M.: More geometry about singular systems. IEEE Press, New York (1987). doi:10.1109/CDC.1987.272585
62. Malabre, M.: Generalized linear systems: geometric and structural approaches. Linear Algebra Appl. **122,123,124**, 591–621 (1989). doi:10.1016/0024-3795(89)90668-X
63. Mertzios, B.G., Christodoulou, M.A., Syrmos, B.L., Lewis, F.L.: Direct controllability and observability time domain conditions of singular systems. IEEE Trans. Autom. Control **33**(8), 788–791 (1988). doi:10.1109/9.1302
64. Morse, A.S.: Structural invariants of linear multivariable systems. SIAM J. Control Optim. **11**, 446–465 (1973). doi:10.1137/0311037
65. Özçaldiran, K.: A geometric characterization of the reachable and controllable subspaces of descriptor systems. IEEE Proc. Circuits Syst. Signal Process. **5**, 37–48 (1986). doi:10.1007/BF01600185
66. Özçaldiran, K., Haliloğlu, L.: Structural properties of singular systems. Kybernetika **29**(6), 518–546 (1993). http://dml.cz/dmlcz/125040
67. Özçaldiran, K., Lewis, F.L.: Generalized reachability subspaces for singular systems. SIAM J. Control Optim. **27**, 495–510 (1989). doi:10.1137/0327026
68. Özçaldiran, K., Fountain, D.W., Lewis, F.L.: Some generalized notions of observability. IEEE Trans. Autom. Control **37**(6), 856–860 (1992). doi:10.1109/9.256347
69. Petreczky, M., Tanwani, A., Trenn, S.: Observability of switched linear systems. In: Djemai, M., Defoort, M. (eds.) Hybrid Dynamical Systems, Lecture Notes in Control and Information Sciences, vol. 457, pp. 205–240. Springer, Berlin (2015). doi:10.1007/978-3-319-10795-0_8
70. Polderman, J.W., Willems, J.C.: Introduction to Mathematical Systems Theory. A Behavioral Approach. Springer, New York (1998). doi:10.1007/978-1-4757-2953-5
71. Popov, V.M.: Hyperstability of Control Systems. Springer, Berlin (1973). Translation based on a revised text prepared shortly after the publication of the Romanian ed. 1966
72. Rabier, P.J., Rheinboldt, W.C.: Classical and generalized solutions of time-dependent linear differential-algebraic equations. Linear Algebra Appl. **245**, 259–293 (1996). doi:10.1016/0024-3795(94)00243-6
73. Rosenbrock, H.H.: Structural properties of linear dynamical systems. Int. J. Control **20**, 191–202 (1974). doi:10.1080/00207177408932729
74. Rudin, W.: Functional Analysis. McGraw-Hill, New York (1973)
75. Schwartz, L.: Théorie des Distributions I,II. No. IX,X in Publications de l'institut de mathématique de l'Universite de Strasbourg. Hermann, Paris (1950, 1951)
76. Trenn, S.: Distributional differential algebraic equations. Ph.D. thesis, Institut für Mathematik, Technische Universität Ilmenau, Universitätsverlag Ilmenau, Ilmenau, Germany (2009). http://www.db-thueringen.de/servlets/DocumentServlet?id=13581

77. Trenn, S.: Regularity of distributional differential algebraic equations. Math. Control Signals Syst. **21**(3), 229–264 (2009). doi:10.1007/s00498-009-0045-4
78. Trenn, S.: Solution concepts for linear DAEs: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I, Differential-Algebraic Equations Forum, pp. 137–172. Springer, Berlin (2013). doi:10.1007/978-3-642-34928-7_4
79. Trentelman, H.L., Stoorvogel, A.A., Hautus, M.L.J.: Control Theory for Linear Systems. Communications and Control Engineering. Springer, London (2001). doi:10.1007/978-1-4471-0339-4
80. van der Schaft, A.J., Schumacher, J.M.H.: The complementary-slackness class of hybrid systems. Math. Control Signals Syst. **9**, 266–301 (1996). doi:10.1007/BF02551330
81. Verghese, G.C.: Further notes on singular systems. In: Proceedings of Joint American Control Conference (1981). Paper TA-4B
82. Verghese, G.C., Levy, B.C., Kailath, T.: A generalized state-space for singular systems. IEEE Trans. Autom. Control **AC-26**(4), 811–831 (1981)
83. Weierstrass, K.: Zur Theorie der bilinearen und quadratischen Formen. Berl. Monatsb. pp. 310–338 (1868)
84. Willems, J.C.: Paradigms and puzzles in the theory of dynamical systems. IEEE Trans. Autom. Control **AC-36**(3), 259–294 (1991). doi:10.1109/9.73561
85. Willems, J.C.: The behavioral approach to open and interconnected systems. IEEE Control Systems Magazine **27**(6), 46–99 (2007). doi:10.1109/MCS.2007.906923
86. Wong, K.T.: The eigenvalue problem $\lambda Tx + Sx$. J. Diff. Equ. **16**, 270–280 (1974). doi:10.1016/0022-0396(74)90014-X
87. Wonham, W.M.: Linear Multivariable Control: A Geometric Approach, 3rd edn. Springer, New York (1985)
88. Yip, E.L., Sincovec, R.F.: Solvability, controllability and observability of continuous descriptor systems. IEEE Trans. Autom. Control **AC-26**, 702–707 (1981). doi:10.1109/TAC.1981.1102699
89. Zhou, Z., Shayman, M.A., Tarn, T.J.: Singular systems: a new approach in the time domain. IEEE Trans. Autom. Control **32**(1), 42–50 (1987). doi:10.1109/TAC.1987.1104430

A Survey on Numerical Methods for the Simulation of Initial Value Problems with sDAEs

Michael Burger and Matthias Gerdts

Abstract This paper provides an overview on numerical aspects in the simulation of differential-algebraic equations (DAEs). Amongst others we discuss the basic construction principles of frequently used discretization schemes, such as BDF methods, Runge–Kutta methods, and ROW methods, as well as their adaption to DAEs. Moreover, topics like consistent initialization, stabilization, parametric sensitivity analysis, co-simulation techniques, aspects of real-time simulation, and contact problems are covered. Finally, some illustrative numerical examples are presented.

Keywords BDF methods • Consistent initialization • Contact problems • Co-simulation • Differential-algebraic equations • Real-time simulation • ROW methods • Runge–Kutta methods • Sensitivity analysis • Stabilization

Subject Classifications: 65L80, 65L05, 65L06

1 Introduction

Simulation is a well-established and indispensable tool in scientific research as well as in industrial development processes. Efficient tools are needed that are capable of simulating complex processes in, e.g., mechanical engineering, process engineering, or electrical engineering. Many of such processes (where appropriate after a spatial discretization of a partial differential equation) can be modeled as

M. Burger (✉)
Department Mathematical Methods in Dynamics and Durability MDF, Fraunhofer Institute for Industrial Mathematics ITWM, Fraunhofer-Platz 1, 67663 Kaiserslautern, Germany
e-mail: Michael.Burger@itwm.fraunhofer.de

M. Gerdts
Department of Aerospace Engineering, Institute of Mathematics and Applied Computing, Universität der Bundeswehr München, Werner-Heisenberg-Weg 39, 85577 Neubiberg, Germany
e-mail: matthias.gerdts@unibw.de

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7_5

differential-algebraic equations (DAEs), which are implicit differential equations that typically consist of ordinary differential equations as well as algebraic equations. Often, DAEs are formulated automatically by software packages such as MODELICA or SIMPACK. In its general form, the initial value problem for a DAE on the compact interval $I = [a, b]$ reads as

$$F(t, z(t), z'(t)) = 0, \qquad z(a) = z_a, \tag{1.1}$$

where $F : I \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is a given function and $z_a \in \mathbb{R}^n$ is an appropriate initial value at $t = a$. The task is to find a solution $z : I \longrightarrow \mathbb{R}^n$ of (1.1). Throughout it is assumed that F is sufficiently smooth, i.e., it possesses all the continuous partial derivatives up to a requested order.

Please note that (1.1) is not just an ordinary differential equation in implicit notation, since we permit the Jacobian of F with respect to z', i.e., $F'_{z'}$, to be *singular* along a solution. In such a situation, (1.1) cannot be solved directly for z'. Particular examples with singular Jacobian are semi-explicit DAEs of type

$$F(t, z, z') = \begin{pmatrix} M(t, x)x' - f(t, x, y) \\ g(t, x, y) \end{pmatrix}, \qquad z := (x, y)^\top, \tag{1.2}$$

with a non-singular matrix M and the so-called differential state vector x and the algebraic state vector y. Such systems occur, e.g., in process engineering and mechanical multi-body systems. More generally, quasi-linear DAEs of type

$$F(t, z, z') = Q(t, z)z' - f(t, z)$$

with a possibly singular matrix function Q frequently occur in electrical engineering.

The potential singularity of the Jacobian $F'_{z'}$ has implications with regard to theoretical properties (existence and uniqueness of solutions, smoothness properties, structural properties, ...) and with regard to the design of numerical methods (consistent initial values, order of convergence, stability, ...). A survey on the solution theory for linear DAEs can be found in the recent survey paper [141]. A comprehensive structural analysis of linear and nonlinear DAEs can be found in the monographs [90] and [92]. While explicit ordinary differential equations (ODEs) can be viewed as well-behaved systems, DAEs are inherently ill-conditioned and the degree of ill-conditioning increases with the so-called (perturbation) index, compare [75, Definition 1.1]. As such, DAEs require suitable techniques for its numerical treatment.

To this end, the paper aims to provide an overview on the numerical treatment of the initial value problem. The intention is to cover the main ideas without too many technical details, which, if required, can be found in full detail in a huge number of publications and excellent textbooks. Naturally not all developments can be covered, so we focus on a choice of methods and concepts that are relevant in industrial simulation environments for coupled systems of potentially large size.

These concepts enhance basic integration schemes by adding features like sensitivity analysis (needed, e.g., in optimization procedures), contact dynamics, real-time schemes, or co-simulation techniques. Still, the core challenges with DAEs, that is ill-conditioning, consistent initial values, index reduction, will be covered as well.

The outline of this paper is as follows. Section 2 introduces index concepts and summarizes stabilization techniques for certain classes of DAEs. Section 3 deals with the computation of the so-called consistent initial values for DAEs and their influence on parameters. Note in this respect that DAEs, in contrast to ODEs, do not permit solutions for arbitrary initial values and thus techniques are required to find suitable initial values. The basics of the most commonly used numerical discretization schemes are discussed in Sect. 4, amongst them are BDF methods, Runge–Kutta methods, and ROW methods. Co-simulation techniques for the interaction of different subsystems are presented in Sect. 5. Herein, the stability and convergence of the overall scheme are of particular importance. Section 6 discusses approaches for the simulation of time crucial systems in real-time. The influence of parameters on the (discrete and continuous) solution of an initial value problem is studied in Sect. 7. Hybrid systems and mechanical contact problems are discussed in Sect. 8.

Notation

We use the following notation. The derivative w.r.t. time of a function $z(t)$ is denoted by $z'(t)$. The partial derivative of a function f with respect to a variable x will be denoted by $f'_x = \partial f/\partial x$. As an abbreviation of a function of type $f(t, x(t))$ we use the notation $f[t]$.

2 Error Influence and Stabilization Techniques

DAEs are frequently characterized and classified according to its index. Various index definitions exist, for instance the differentiation index [62], the structural index [45], the strangeness index [90], the tractability index [92], and the perturbation index [75]. These index definitions are not equivalent for general DAEs (1.1), but they coincide for certain subclasses thereof, for instance semi-explicit DAEs in Hessenberg form. For our purposes we will focus on the differentiation index and the perturbation index only.

The *differentiation index* is one of the earliest index definitions for (1.1) and is based on a structural investigation of the DAE. It aims to identify the so-called underlying ordinary differential equation. To this end let the functions $F^{(j)} : [t_0, t_f] \times R^{(j+2)n} \longrightarrow \mathbb{R}^n$ for the variables $z, z', \ldots, z^{(j+1)} \in \mathbb{R}^n$ for $j = 0, 1, 2, \ldots$ be defined

by the recursion

$$F^{(0)}(t,z,z') := F(t,z,z'), \tag{2.1}$$

$$F^{(j)}(t,z,z',\ldots,z^{(j+1)}) := \frac{\partial F^{(j-1)}}{\partial t}(t,z,z',\ldots,z^{(j)}) \tag{2.2}$$

$$+\sum_{\ell=0}^{j}\frac{\partial F^{(j-1)}}{\partial z^{(\ell)}}(t,z,z',\ldots,z^{(j)})z^{(\ell+1)}, \quad j=1,2,\ldots. \tag{2.3}$$

Herein, F is supposed to be sufficiently smooth such that the functions $F^{(j)}$ are well defined.

The differentiation index is defined as follows:

Definition 2.1 (Differentiation Index, Compare [62]) The DAE (1.1) has *differentiation index* $d \in \mathbb{N}_0$, if d is the smallest number in $\mathbb{N}_0$ such that the so-called derivative array

$$F^{(j)}(t,z,z',\ldots,z^{(j+1)}) = 0, \qquad j=0,1,\ldots,d, \tag{2.4}$$

allows to deduce a relation of type $z' = f(t,z)$ by algebraic manipulations.

If such a relation exists, then the corresponding ordinary differential equation (ODE) $z'(t) = f(t,z(t))$ is called the *underlying ODE* of the DAE (1.1).

The definition leaves some space for interpretation as it is not entirely clear what is meant by "algebraic manipulations." However, for semi-explicit DAEs it provides a guideline to determine the differentiation index. Note that the special structure of semi-explicit DAEs is often exploited in the design of numerical schemes and stabilization techniques.

Definition 2.2 (Semi-Explicit DAE) A DAE of type

$$x'(t) = f(t,x(t),y(t)), \tag{2.5}$$

$$0 = g(t,x(t),y(t)), \tag{2.6}$$

is called *semi-explicit DAE*. Herein, $x(\cdot)$ is referred to as *differential variable* and $y(\cdot)$ is called *algebraic variable*. Correspondingly, (2.5) is called *differential equation* and (2.6) *algebraic equation*.

For semi-explicit DAEs the common approach is to differentiate the algebraic equation w.r.t. time and to substitute the occurring derivatives of x by the right-hand side of the differential equation. This procedure is repeated until the resulting equation can be solved for y'.

Example 2.1 (Semi-Explicit DAE with Differentiation Index One) Consider (2.5)–(2.6). Differentiation of the algebraic equation w.r.t. time yields

$$\begin{aligned} 0 &= g'_t[t] + g'_x[t]x'(t) + g'_y[t]y'(t) \\ &= g'_t[t] + g'_x[t]f[t] + g'_y[t]y'(t). \end{aligned}$$

Herein, we used the abbreviation $f[t]$ for $f(t, x(t), y(t))$ and likewise for the partial derivatives of g.

Now, if the Jacobian matrix $g'_y[t]$ is non-singular with a bounded inverse along a solution of the DAE, then the above equation can be solved for y' by the implicit function theorem and together with the differential equation (2.5) we obtain the underlying ODE

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)), \\ y'(t) &= -g'_y[t]^{-1} \left(g'_t[t] + g'_x[t]f[t] \right), \end{aligned}$$

and the differentiation index is $d = 1$.

In the above example, the situation becomes more involved, if the Jacobian matrix $g'_y[t]$ is singular. If it actually vanishes, then one can proceed as in the following example.

Example 2.2 (Semi-Explicit DAE with Differentiation Index Two) Consider (2.5)–(2.6). Suppose g does not depend on y and thus $g'_y[t] \equiv 0$. By differentiation of the algebraic equation we obtain

$$0 = g'_t[t] + g'_x[t]x'(t) = g'_t[t] + g'_x[t]f[t] =: g^{(1)}(t, x(t), y(t))$$

A further differentiation w.r.t. time yields

$$0 = (g^{(1)})'_t[t] + (g^{(1)})'_x[t]f[t] + (g^{(1)})'_y[t]y'(t)$$

with $(g^{(1)})'_y[t] = g'_x[t]f'_y[t]$. Now, if the matrix $g'_x[t]f'_y[t]$ is non-singular with a bounded inverse along a solution of the DAE, then the above equation can be solved for y' by the implicit function theorem and together with the differential equation (2.5) we obtain the underlying ODE

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)), \\ y'(t) &= -(g'_x[t]f'_y[t])^{-1} \left((g^{(1)})'_t[t] + (g^{(1)})'_x[t]f[t] \right), \end{aligned}$$

and the differentiation index is $d = 2$.

The procedure of the preceding examples works for semi-explicit Hessenberg DAEs, which are defined as follows:

Definition 2.3 (Hessenberg DAE)

(a) For a given $k \geq 2$ the DAE

$$\begin{aligned} x_1'(t) &= f_1(t, y(t), x_1(t), x_2(t), \ldots, x_{k-2}(t), x_{k-1}(t)), \\ x_2'(t) &= f_2(t, \quad x_1(t), x_2(t), \ldots, x_{k-2}(t), x_{k-1}(t)), \\ &\vdots \qquad\qquad\qquad \ddots \\ x_{k-1}'(t) &= f_{k-1}(t, \qquad\qquad\qquad x_{k-2}(t), x_{k-1}(t)), \\ 0 &= g(t, \qquad\qquad\qquad\qquad x_{k-1}(t)) \end{aligned} \tag{2.7}$$

is called *Hessenberg DAE of order k*, if the matrix

$$R(t) := g'_{x_{k-1}}[t] \cdot f'_{k-1,x_{k-2}}[t] \cdots f'_{2,x_1}[t] \cdot f'_{1,y}[t] \tag{2.8}$$

is non-singular for all $t \in [t_0, t_f]$ with a uniformly bounded inverse $\|R^{-1}(t)\| \leq C$ in $[t_0, t_f]$, where C is a constant independent of t.

(b) The DAE

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)), \\ 0 &= g(t, x(t), y(t)) \end{aligned} \tag{2.9}$$

is called *Hessenberg DAE of order* 1, if the matrix $g'_y[t]$ is non-singular with $\|g'_y[t]^{-1}\| \leq C$ for all $t \in [t_0, t_f]$ and some constant C independent of t.

Herein, y is called algebraic variable and $x = (x_1, \ldots, x_{k-1})^\top$ in (a) and x in (b), respectively, is called differential variable.

By repeated differentiation of the algebraic constraint $0 = g(t, x_{k-1}(t))$ w.r.t. to time and simultaneous substitution of the derivatives of the differential variable by the corresponding differential equations, it is straightforward to show that the differentiation index of a Hessenberg DAE of order k is equal to k, provided the functions g and f_j, $j = 1, \ldots, k-1$, are sufficiently smooth. In order to formalize this procedure, define

$$g^{(0)}(t, x_{k-1}(t)) := g(t, x_{k-1}(t)). \tag{2.10}$$

Differentiation of $g^{(0)}$ with respect to time and substitution of

$$x_{k-1}'(t) = f_{k-1}(t, x_{k-2}(t), x_{k-1}(t))$$

leads to the equation

$$\begin{aligned} 0 &= g'_t(t, x_{k-1}(t)) + g'_{x_{k-1}}(t, x_{k-1}(t)) \cdot f_{k-1}(t, x_{k-2}(t), x_{k-1}(t)) \\ &=: g^{(1)}(t, x_{k-2}(t), x_{k-1}(t)), \end{aligned}$$

which is satisfied implicitly as well. Recursive application of this differentiation and substitution process leads to the algebraic equations

$$0 = g^{(j)}(t, x_{k-1-j}(t), \ldots, x_{k-1}(t)), \qquad j = 1, 2, \ldots, k-2, \tag{2.11}$$

and

$$0 = g^{(k-1)}(t, y(t), x_1(t), \ldots, x_{k-1}(t)). \tag{2.12}$$

Since Eqs. (2.11)–(2.12) do not occur explicitly in the original system (2.7), these equations are called *hidden constraints* of the Hessenberg DAE. Note that the matrix R in (2.8) is given by $\partial g^{(k-1)}/\partial y$.

A practically important subclass of Hessenberg DAEs are mechanical multibody systems in descriptor form given by

$$\begin{aligned} q'(t) &= v(t), \\ M(t, q(t))v'(t) &= f(t, q(t), v(t)) - g'_q(t, q(t))^\top \lambda(t), \\ 0 &= g(t, q(t)), \end{aligned} \tag{2.13}$$

where $q(\cdot) \in \mathbb{R}^n$ denotes the vector of generalized positions, $v(\cdot) \in \mathbb{R}^n$ the vector of generalized velocities, and $\lambda(\cdot) \in \mathbb{R}^m$ are Lagrange multipliers. The mass matrix M is supposed to be symmetric and positive definite with a bounded inverse M^{-1} and thus, the second equation in (2.13) can be multiplied by $M(t, q(t))^{-1}$. The vector f denotes the generalized forces and torques. The term $g'_q(t, q)^\top \lambda$ can be interpreted as a force that keeps the system on the algebraic constraint $g(t, q) = 0$.

The constraint $g(t, q(t)) = 0$ is called *constraint on position level*. Differentiation with respect to time of this algebraic constraint yields the *constraint on velocity level*

$$g'_t(t, q(t)) + g'_q(t, q(t)) \cdot v(t) = 0$$

and the *constraint on acceleration level*

$$g''_{tt}(t, q(t)) + g''_{tq}(t, q(t)) \cdot v(t) + g'_q(t, q(t)) \cdot v'(t) + g''_{qq}(t, q(t))(v(t), v(t)) = 0.$$

Replacing v' by

$$v'(t) = M(t, q(t))^{-1} \left(f(q(t), v(t)) - g'_q(t, q(t))^\top \lambda(t) \right)$$

yields

$$\begin{aligned} 0 = {} & g''_{tt}(t, q(t)) + g''_{tq}(t, q(t)) \cdot v(t) \\ & + g'_q(t, q(t)) \cdot M(t, q(t))^{-1} \left(f(q(t), v(t)) - g'_q(t, q(t))^\top \lambda(t) \right) \\ & + g''_{qq}(t, q(t))(v(t), v(t)). \end{aligned}$$

If $g'_q(t, q))$ has full rank, then the matrix $g'_q(t, q) M(t, q)^{-1} g'_q(t, q)^\top$ is non-singular and the latter equation can be solved for the algebraic variable λ. Thus, the differentiation index is three.

Remark 2.1 Note that semi-explicit DAEs are more general than Hessenberg DAEs since no regularity assumptions are imposed in Definition 2.2. In fact, without additional regularity assumptions, the class of semi-explicit DAEs is essentially as large as the class of general DAEs (1.1), since the settings $z'(t) = y(t)$ and $F(t, y(t), z(t)) = 0$ transform the DAE (1.1) into a semi-explicit DAE (some care has to be taken with regard to the smoothness of solutions, though).

2.1 Error Influence and Perturbation Index

The differentiation index is based on a structural analysis of the DAE, but it does not indicate how perturbations influence the solution. In contrast, the perturbation index addresses the influence of perturbations on the solution and thus it is concerned with the *stability of DAEs*. Note that perturbations frequently occur, for instance they are introduced by numerical discretization schemes.

Definition 2.4 (Perturbation Index, See [75]) The DAE (1.1) has *perturbation index* $p \in \mathbb{N}$ along a solution z on $[t_0, t_f]$, if $p \in \mathbb{N}$ is the smallest number such that for all functions $\tilde{z}$ satisfying the perturbed DAE

$$F(t, \tilde{z}(t), \tilde{z}'(t)) = \delta(t), \tag{2.14}$$

there exists a constant S depending on F and $t_f - t_0$ with

$$\|z(t) - \tilde{z}(t)\| \leq S \left(\|z(t_0) - \tilde{z}(t_0)\| + \max_{t_0 \leq \tau \leq t} \|\delta(\tau)\| + \ldots + \max_{0 \leq \tau \leq t} \|\delta^{(p-1)}(\tau)\| \right) \tag{2.15}$$

for all $t \in [t_0, t_f]$, whenever the expression on the right is less than or equal to a given bound.

The *perturbation index* is $p = 0$, if the estimate

$$\|z(t) - \tilde{z}(t)\| \leq S\left(\|z(t_0) - \tilde{z}(t_0)\| + \max_{t_0 \leq \tau \leq t_f} \left\| \int_{t_0}^{\tau} \delta(s) ds \right\|\right) \tag{2.16}$$

holds. The DAE is said to be of *higher index*, if $p \geq 2$.

According to the definition of the perturbation index, higher index DAEs are ill-conditioned in the sense that small perturbations with high frequencies, i.e., with large derivatives, can have a considerable influence on the solution of a higher index DAE as it can be seen in (2.15). For some time it was believed that the difference between perturbation index and differentiation index is at most one, until it was shown in [34] that the difference between perturbation index and differentiation index can be arbitrarily large. However, for the subclass of Hessenberg DAEs as defined in Definition 2.3 both index concepts (and actually all other relevant index concepts) coincide.

The definition of the perturbation index shows that the degree of ill-conditioning increases with the perturbation index. Hence, in order to make a higher index DAE accessible to numerical methods it is advisable and common practice to reduce the perturbation index of a DAE. A straightforward idea is to replace the original DAE by a mathematically equivalent DAE with lower perturbation index. The index reduction process itself is nontrivial for general DAEs, since one has to ensure that it is actually the perturbation index, which is being reduced (and not some other index like the differentiation index).

For Hessenberg DAEs, however, the index reduction process is straightforward as perturbation index and differentiation index coincide. Consider a Hessenberg DAE of order k as in (2.7). Then, by replacing the algebraic constraint $0 = g(t, x_{k-1}(t))$ by one of the hidden constraints $g^{(j)}$, $j \in \{1, \ldots, k-1\}$, defined in (2.11) or (2.12) we obtain the Hessenberg DAE

$$\begin{array}{rcl}
x_1'(t) &=& f_1(t, y(t), x_1(t), x_2(t), \quad \ldots, \quad x_{k-2}(t), x_{k-1}(t)), \\
x_2'(t) &=& f_2(t, \quad x_1(t), x_2(t), \quad \ldots, \quad x_{k-2}(t), x_{k-1}(t)), \\
&\vdots& \qquad\qquad \ddots \\
x_{k-1}'(t) &=& f_{k-1}(t, \qquad\qquad\qquad\qquad x_{k-2}(t), x_{k-1}(t)), \\
0 &=& g^{(j)}(t, \qquad\qquad\qquad x_{k-1-j}(t), \quad \ldots, \quad x_{k-1}(t)),
\end{array} \tag{2.17}$$

where we use the setting $x_0 := y$ for notational convenience. The Hessenberg DAE in (2.17) has perturbation index $k - j$. Hence, this simple index reduction strategy actually reduces the perturbation index, and it leads to a mathematically equivalent DAE with the same solution as the original DAE, if the initial values $x(t_0)$ and $y(t_0)$ satisfy the algebraic constraints $g^{(\ell)}(t_0, x_{k-1-\ell}(t_0), \ldots, x_{k-1}(t_0)) = 0$ for all $\ell = 0, \ldots, k-1$.

On the other hand, the index reduced DAE (2.17) in general permits additional solutions for those initial values $x(t_0)$ and $y(t_0)$, which merely satisfy the algebraic constraints $g^{(\ell)}(t_0, x_{k-1-\ell}(t_0), \ldots, x_{k-1}(t_0)) = 0$ for all $\ell = j, \ldots, k-1$, but not the neglected algebraic constraints with index $\ell = 0, \ldots, j-1$. In the most extreme case $j = k-1$ (the reduced DAE has index-one) $x(t_0)$ can be chosen arbitrarily (assuming that the remaining algebraic constraint can be solved for $y(t_0)$ given the value of $x(t_0)$). The following theorem shows that the use of inconsistent initial values leads to a polynomial drift off the neglected algebraic constraints in time, compare [73, Sect. VII.2].

Theorem 2.1 *Consider the Hessenberg DAE of order k in (2.7) and the index reduced DAE in (2.17) with $j \in \{1, \ldots, k-1\}$. Let $x(t)$ and $y(t)$ be a solution of (2.17) such that the initial values $x(t_0)$ and $y(t_0)$ satisfy the algebraic constraints $g^{(\ell)}(t_0, x_{k-1-\ell}(t_0), \ldots, x_{k-1}(t_0)) = 0$ for all $\ell = j, \ldots, k-1$. Then for $\ell = 1, \ldots, j$ and $t \geq t_0$ we have*

$$g^{(j-\ell)}(t, x_{k-1-(j-\ell)}(t), \ldots, x_{k-1}(t)) = \sum_{\nu=0}^{\ell-1} \frac{1}{\nu!}(t-t_0)^\nu g^{(j-\ell+\nu)}[t_0]. \tag{2.18}$$

with $g^{(j-\ell+\nu)}[t_0] := g^{(j-\ell+\nu)}(t_0, x_{k-1-(j-\ell+\nu)}(t_0), \ldots, x_{k-1}(t_0))$.

Proof We use the abbreviation $g^{(\ell)}[t]$ for $g^{(\ell)}(t, x_{k-1-\ell}(t), x_{k-1}(t))$ for notational convenience. Observe that

$$g^{(j-\ell+1)}[t] = \frac{d}{dt} g^{(j-\ell)}[t], \qquad \ell = 1, \ldots, j,$$

and thus

$$g^{(j-\ell)}[t] = g^{(j-\ell)}[t_0] + \int_{t_0}^{t} g^{(j-\ell+1)}[\tau] d\tau.$$

We have $g^{(j)}[t] = 0$ and thus for $\ell = 1$:

$$g^{(j-1)}[t] = g^{(j-1)}[t_0] + \int_{t_0}^{t} g^{(j)}[\tau] d\tau = g^{(j-1)}[t_0].$$

This proves (2.18) for $\ell = 1$. Inductively we obtain

$$\begin{aligned} g^{(j-(\ell+1))}[t] &= g^{(j-(\ell+1))}[t_0] + \int_{t_0}^{t} g^{(j-\ell)}[\tau] d\tau \\ &= g^{(j-(\ell+1))}[t_0] + \int_{t_0}^{t} \sum_{\nu=0}^{\ell-1} \frac{1}{\nu!}(\tau-t_0)^\nu g^{(j-\ell+\nu)}[t_0] d\tau \end{aligned}$$

$$
\begin{aligned}
&= g^{(j-(\ell+1))}[t_0] + \sum_{\nu=0}^{\ell-1} \frac{1}{(\nu+1)!}(t-t_0)^{\nu+1} g^{(j-\ell+\nu)}[t_0] \\
&= g^{(j-(\ell+1))}[t_0] + \sum_{\nu=1}^{\ell} \frac{1}{\nu!}(t-t_0)^{\nu} g^{(j-\ell+\nu-1)}[t_0] \\
&= \sum_{\nu=0}^{\ell} \frac{1}{\nu!}(t-t_0)^{\nu} g^{(j-(\ell+1)+\nu)}[t_0],
\end{aligned}
$$

which proves the assertion. □

We investigate the practically relevant index-three case in more detail and consider the reduction to index one (i.e., $k = 3$ and $j = 2$). In this case Theorem 2.1 yields

$$g^{(0)}(t, x_2(t)) = g^{(0)}[t_0] + (t-t_0) g^{(1)}[t_0], \tag{2.19}$$

$$g^{(1)}(t, x_1(t), x_2(t)) = g^{(1)}[t_0]. \tag{2.20}$$

The drift-off property of the index reduced DAE causes difficulties for numerical discretization methods as the subsequent result shows, compare [73, Sect. VII.2].

Theorem 2.2 *Consider the DAE (2.7) with $k = 3$ and the index reduced problem (2.17) with $j = 2$. Let $z(t; t_m, z_m)$ denote the solution of the latter at time t with initial value z_m at t_m, where $z = (x_1, x_2, y)^\top$ denotes the vector of differential and algebraic states. Suppose the initial value z_0 at t_0 satisfies $g^{(0)}[t_0] = 0$ and $g^{(1)}[t_0] = 0$.*

Let a numerical method generate approximations $z_n = (x_{1,n}, x_{2,n}, y_n)^\top$ of $z(t_n; t_0, z_0)$ at time points $t_n = t_0 + nh$, $n \in \mathbb{N}_0$, with stepsize $h > 0$. Suppose the numerical method is of order $p \in \mathbb{N}$, i.e., the local error satisfies

$$\|z_{n+1} - z(t_{n+1}; t_n, z_n)\| = \mathcal{O}(h^{p+1}), \qquad n \in \mathbb{N}_0.$$

Then, for $n \in \mathbb{N}$ the algebraic constraint $g^{(0)} = g$ satisfies the estimate

$$\|g(t_n, x_{2,n})\| \le C h^p \left(L_0 (t_n - t_0) + \frac{L_1}{2} (t_n - t_0)^2 \right) \tag{2.21}$$

with constants C, L_0, and L_1.

Proof Since z_0 satisfies $g^{(0)}[t_0] = 0$ and $g^{(1)}[t_0] = 0$, the solution $z(t; t_0, z_0)$ satisfies these constraints for every t. For notational convenience we use the notion $g^{(0)}(t, z(t))$ instead of $g^{(0)}(t, x_2(t))$ and likewise for $g^{(1)}$. To this end, for a given t_n

we have

$$\begin{aligned}\|g^{(0)}(t_n, z_n)\| &= \|g^{(0)}(t_n, z_n) - g^{(0)}(t_n, z(t_n; t_0, z_0))\| \\ &= \| \sum_{m=0}^{n-1} \left(g^{(0)}(t_n, z(t_n; t_{m+1}, z_{m+1})) - g^{(0)}(t_n, z(t_n; t_m, z_m)) \right) \| \\ &\leq \sum_{m=0}^{n-1} \|g^{(0)}(t_n, z(t_n; t_{m+1}, z_{m+1})) - g^{(0)}(t_n, z(t_n; t_m, z_m))\|. \end{aligned} \tag{2.22}$$

Exploitation of (2.19)–(2.20) with t_0 replaced by t_m and t_{m+1}, respectively, yields

$$\begin{aligned}&\|g^{(0)}(t_n, z(t_n; t_{m+1}, z_{m+1})) - g^{(0)}(t_n, z(t_n; t_m, z_m))\| \\ &= \|g^{(0)}(t_{m+1}, z_{m+1}) + (t_n - t_{m+1}) g^{(1)}(t_{m+1}, z_{m+1}) \\ &\quad - g^{(0)}(t_m, z_m) - (t_n - t_m) g^{(1)}(t_m, z_m)\| \\ &= \|g^{(0)}(t_{m+1}, z_{m+1}) + (t_n - t_{m+1}) g^{(1)}(t_{m+1}, z_{m+1}) - g^{(0)}(t_{m+1}, z(t_{m+1}; t_m, z_m)) \\ &\quad + g^{(0)}(t_{m+1}, z(t_{m+1}; t_m, z_m)) - g^{(0)}(t_m, z_m) - (t_n - t_m) g^{(1)}(t_m, z_m)\| \\ &= \|g^{(0)}(t_{m+1}, z_{m+1}) + (t_n - t_{m+1}) g^{(1)}(t_{m+1}, z_{m+1}) - g^{(0)}(t_{m+1}, z(t_{m+1}; t_m, z_m)) \\ &\quad + g^{(0)}(t_m, z_m) + (t_{m+1} - t_m) g^{(1)}(t_m, z_m) - g^{(0)}(t_m, z_m) - (t_n - t_m) g^{(1)}(t_m, z_m)\| \\ &= \|g^{(0)}(t_{m+1}, z_{m+1}) - g^{(0)}(t_{m+1}, z(t_{m+1}; t_m, z_m)) \\ &\quad + (t_n - t_{m+1}) \left(g^{(1)}(t_{m+1}, z_{m+1}) - g^{(1)}(t_m, z_m) \right) \| \\ &\leq L_0 C h^{p+1} + (t_n - t_{m+1}) \|g^{(1)}(t_{m+1}, z_{m+1}) - g^{(1)}(t_{m+1}, z(t_{m+1}; t_m, z_m))\| \\ &\quad + (t_n - t_{m+1}) \|g^{(1)}(t_{m+1}, z(t_{m+1}; t_m, z_m)) - g^{(1)}(t_m, z_m)\| \\ &\leq C h^{p+1} \left(L_0 + L_1 (t_n - t_{m+1}) \right) + (t_n - t_{m+1}) \|g^{(1)}(t_m, z_m) - g^{(1)}(t_m, z_m)\| \\ &= C h^{p+1} \left(L_0 + L_1 (t_n - t_{m+1}) \right), \end{aligned}$$

where L_0 and L_1 are Lipschitz constants of $g^{(0)}$ and $g^{(1)}$. Together with (2.22) we thus proved the estimate

$$\begin{aligned}\|g^{(0)}(t_n, z_n)\| &\leq \sum_{m=0}^{n-1} C h^{p+1} \left(L_0 + L_1 (t_n - t_{m+1}) \right) \\ &\leq C h^p \left(L_0 (t_n - t_0) + \frac{L_1}{2} (t_n - t_0)^2 \right). \end{aligned}$$

□

The estimate (2.21) shows that the numerical solution may violate the algebraic constraint with a quadratic drift term in t_n for the setting in Theorem 2.2. This drift-off effect may lead to useless numerical simulation results, especially on long time

horizons. For DAEs with even higher index, the situation becomes worse as the degree of the polynomial drift term depends on the j in (2.17), i.e., on the number of differentiations used in the index reduction.

2.2 Stabilization Techniques

The basic index reduction approach in the previous section may lead to unsatisfactory numerical results. One possibility to avoid the drift-off on numerical level is to perform a projection step onto the neglected algebraic constraints after each successful integration step for the index reduced system, see [18, 47].

Another idea is to use stabilization techniques to stabilize the index reduced DAE itself. The common approaches are Baumgarte stabilization, Gear–Gupta–Leimkuhler stabilization, and the use of overdetermined DAEs.

2.2.1 Baumgarte Stabilization

The Baumgarte stabilization [22] was originally introduced for mechanical multibody systems (2.13). It can be extended to Hessenberg DAEs in a formal way. The idea is to replace the algebraic constraint in (2.7) by a linear combination of original and hidden algebraic constraints $g^{(\ell)}$, $\ell \in \{0, 1, \ldots, k-1\}$. With the setting $x_0 := y$, the resulting DAE reads as follows:

$$\begin{aligned}
x_1'(t) &= f_1(t, y(t), x_1(t), x_2(t), \ldots, x_{k-2}(t), x_{k-1}(t)),\\
x_2'(t) &= f_2(t, x_1(t), x_2(t), \ldots, x_{k-2}(t), x_{k-1}(t)),\\
&\vdots \qquad\qquad \ddots\\
x_{k-1}'(t) &= f_{k-1}(t, x_{k-2}(t), x_{k-1}(t)),\\
0 &= \sum_{\ell=0}^{k-1} \alpha_\ell g^{(\ell)}(t, x_{k-1-\ell}(t), \ldots, x_{k-1}(t)).
\end{aligned} \tag{2.23}$$

The DAE (2.23) has index one. The weights α_ℓ, $\ell = 0, 1, \ldots, k-1$, with $\alpha_{k-1} = 1$ have to be chosen such that the associated differential equation

$$0 = \sum_{\ell=0}^{k-1} \alpha_\ell \eta^{(\ell)}(t)$$

is asymptotically stable with $\|\eta^{(\ell)}(t)\| \longrightarrow 0$ for $\ell \in \{0, \ldots, k-2\}$ as $t \longrightarrow \infty$, compare [73, Sect. VII.2]. A proper choice of the weights is crucial since a balance between quick damping and low degree of stiffness has to be found.

The Baumgarte stabilization was used for real-time simulations in [14, 31], but on the index-two level and not on the index-one level.

2.2.2 Gear–Gupta–Leimkuhler Stabilization

The Gear–Gupta–Leimkuhler (GGL) stabilization [64] does not neglect algebraic constraints but couples them to the index reduced DAE using an additional multiplier. Consider the mechanical multibody system (2.13). The GGL stabilization reads as follows:

$$\begin{aligned} q'(t) &= v(t) - g_q'(t,q(t))^\top \mu(t), \\ M(t,q(t))v'(t) &= f(t,q(t),v(t)) - g_q'(t,q(t))^\top \lambda(t), \\ 0 &= g(t,q(t)), \\ 0 &= g_t'(t,q(t)) + g_q'(t,q(t)) \cdot v(t) \end{aligned} \tag{2.24}$$

The DAE is of Hessenberg type (if multiplied by M^{-1}) and it has index two, if M is symmetric and positive definite and g_q' has full rank. Differentiation of the first algebraic equation yields

$$0 = g_t'(t,q(t)) + g_q'(t,q(t)) \cdot \big(v(t) - g_q'(t,q(t))^\top \mu(t)\big) = -g_q'(t,q(t)) g_q'(t,q(t))^\top \mu(t).$$

Since g_q' is supposed to be of full rank, the matrix $g_q'[t]g_q'[t]^\top$ is non-singular and the equation implies $\mu \equiv 0$.

The idea of the GGL stabilization can be extended to Hessenberg DAEs. To this end consider (2.7) and the index reduced DAE (2.17) with $j \in \{1,\dots,k-1\}$ fixed. Define

$$G(t,x_1,\dots,x_{k-1}) := \begin{pmatrix} g^{(0)}(t,x_{k-1}) \\ g^{(1)}(t,x_{k-2},x_{k-1}) \\ \vdots \\ g^{(j-1)}(t,x_{k-j},\dots,x_{k-1}) \end{pmatrix}$$

and suppose the Jacobian

$$G'_{(x_1,\dots,x_{k-1})} = \begin{pmatrix} 0 \cdots 0 & 0 & \cdots & 0 & (g^{(0)})'_{x_{k-1}} \\ 0 \cdots 0 & \vdots & \iddots & (g^{(1)})'_{x_{k-2}} & (g^{(1)})'_{x_{k-1}} \\ 0 \cdots 0 & 0 & \iddots & \vdots & \vdots \\ 0 \cdots 0 & (g^{(j-1)})'_{x_{k-j}} & \cdots & (g^{(j-1)})'_{x_{k-2}} & (g^{(j-1)})'_{x_{k-1}} \end{pmatrix}$$

has full rank. A stabilized version of (2.17) is given by

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)) - G_x'(t, x(t))^\top \mu(t), \\ 0 &= G(t, x(t)), \\ 0 &= g^{(j)}(t, x_{k-j-1}(t), \ldots, x_{k-1}(t)), \end{aligned} \tag{2.25}$$

where μ is an additional algebraic variable, $x = (x_1, \ldots, x_{k-1})^\top$, and $f = (f_1, \ldots, f_{k-1})^\top$. The stabilized DAE has index $\max\{2, k-j\}$. Note that

$$\begin{aligned} G_t'[t] + G_x'[t] f[t] &= \begin{pmatrix} (g^{(0)})_t'[t] + (g^{(0)})_{x_{k-1}}'[t] f_{k-1}[t] \\ \vdots \\ (g^{(j-1)})_t'[t] + \sum_{\ell=1}^{j} (g^{(j-1)})_{x_{k-\ell}}'[t] f_{k-\ell}[t] \end{pmatrix} \\ &= \begin{pmatrix} g^{(1)}[t] \\ \vdots \\ g^{(j)}[t] \end{pmatrix} = 0. \end{aligned}$$

Moreover,

$$\begin{aligned} 0 &= \frac{d}{dt} G(t, x_1(t), \ldots, x_{k-1}(t)) \\ &= G_t'[t] + G_x'[t] \left(f[t] - G_x'[t]^\top \mu(t) \right) \\ &= G_t'[t] + G_x'[t] f[t] - G_x'[t] G_x'[t]^\top \mu(t) \\ &= -G_x'[t] G_x'[t]^\top \mu(t) \end{aligned}$$

and thus $\mu \equiv 0$ since G_x' was supposed to have full rank.

2.2.3 Stabilization by Over-Determination

The GGL stabilization approaches for the mechanical multibody system in (2.24) and the Hessenberg DAE in (2.25) are mathematically equivalent to the *overdetermined DAEs*

$$\begin{aligned} q'(t) &= v(t), \\ M(t, q(t)) v'(t) &= f(t, q(t), v(t)) - g_q'(t, q(t))^\top \lambda(t), \\ 0 &= g(t, q(t)), \\ 0 &= g_t'(t, q(t)) + g_q'(t, q(t)) \cdot v(t) \end{aligned}$$

and

$$\begin{aligned} x'(t) &= f(t, x(t), y(t)), \\ 0 &= G(t, x(t)), \\ 0 &= g^{(j)}(t, x_{k-j-1}(t), \ldots, x_{k-1}(t)), \end{aligned}$$

respectively, because the additional algebraic variable μ vanishes in either case. Hence, from an analytical point of view there is no difference between the respective systems. A different treatment is necessary from the numerical point of view, though. The GGL stabilized DAEs in (2.24) and (2.25) can be solved by standard discretization schemes, like BDF methods or methods of Runge–Kutta type, provided those are suitable for higher index DAEs. In contrast, the overdetermined DAEs require tailored numerical methods that are capable of dealing with overdetermined linear equations, which arise internally in each integration step. Typically, such overdetermined equations are solved in a least-squares sense, compare [56, 57] for details.

3 Consistent Initialization and Influence of Parameters

One of the crucial issues when dealing with DAEs is that a DAE in general only permits a solution for properly defined initial values, the so-called *consistent initial values*. The initial values not only have to satisfy those algebraic constraints that are explicitly present in the DAE, but hidden constraints have to be satisfied as well.

3.1 Consistent Initial Values

For the Hessenberg DAE (2.7) consistency is defined as follows.

Definition 3.1 (Consistent Initial Value for Hessenberg DAEs) The initial values $x(t_0) = (x_1(t_0), \ldots, x_{k-1}(t_0))^\top$ and $y(t_0)$ are *consistent with (2.7)*, if the equations

$$0 = g^{(j)}(t_0, x_{k-1-j}(t_0), \ldots, x_{k-1}(t_0)), \qquad j = 1, 2, \ldots, k-2, \tag{3.1}$$

$$0 = g^{(k-1)}(t_0, y(t_0), x_1(t_0), \ldots, x_{k-1}(t_0)) \tag{3.2}$$

hold.

Finding a consistent initial value for a Hessenberg DAE typically consists of two steps. Firstly, a suitable $x(t_0)$ subject to the constraints (3.1) has to be determined. Secondly, given $x(t_0)$ with (3.1), Eq. (3.2) can be solved for $y_0 = y(t_0)$ by Newton's method, if the matrix $R_0 = \partial g^{(k-1)}/\partial y$ is non-singular in a solution (assuming that

a solution exists). For mechanical multibody systems even a linear equation arises in the second step.

Example 3.1 Consider the mechanical multibody system (2.13). A consistent initial value (q_0, v_0, λ_0) at t_0 must satisfy

$$\begin{aligned} 0 &= g(t_0, q_0), \\ 0 &= g'_t(t_0, q_0) + g'_q(t_0, q_0) \cdot v_0, \\ 0 &= g''_{tt}(t_0, q_0) + g''_{tq}(t_0, q_0) \cdot v_0 + g'_q(t_0, q_0) \cdot v'_0 + g''_{qq}(t_0, q_0)(v_0, v_0) \end{aligned}$$

with

$$M(t_0, q_0)v'_0 = f(t_0, q_0, v_0) - g'_q(t_0, q_0)^\top \lambda_0.$$

The latter two equations yield a linear equation for v'_0 and λ_0:

$$\begin{pmatrix} M(t_0, q_0) & g'_q(t_0, q_0)^\top \\ g'_q(t_0, q_0) & 0 \end{pmatrix} \begin{pmatrix} v'_0 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} f(t_0, q_0, v_0) \\ -g''_{tt}(t_0, q_0) - g''_{tq}(t_0, q_0) \cdot v_0 - g''_{qq}(t_0, q_0)(v_0, v_0) \end{pmatrix}.$$

The matrix on the left-hand side is non-singular, if M is symmetric and positive definite and $g'_q(t_0, q_0)$ is of full rank.

Definition 3.2 (Consistent Initial Value for General DAEs, Compare [29, Sect. 5.3.4]) For a general DAE (1.1) with differentiation index d the initial value $z_0 = z(t_0)$ is said to be consistent at t_0, if the derivative array

$$F^{(j)}(t_0, z_0, z'_0, \ldots, z_0^{(j+1)}) = 0, \qquad j = 0, 1, \ldots, d, \tag{3.3}$$

in (2.4) has a solution $(z_0, z'_0, \ldots, z_0^{(d+1)})$.

Note that the system of nonlinear equations (3.3) in general has many solutions and additional conditions are required to obtain a particular consistent initial value, which might be relevant for a particular application. This can be achieved for instance by imposing additional constraints

$$G(t_0, z_0, z'_0) = 0, \tag{3.4}$$

which are known to hold for a specific application, compare [29, Sect. 5.3.4]. Of course, such additional constraints must not contradict the equations in (3.3).

If the user is not able to formulate relations in (3.4) such that the combined system of equations (3.3) and (3.4) returns a unique solution, then a least-squares

approach could be used to find a consistent initial value closest to a 'desired' initial value, compare [33]:

$$\begin{aligned} &\text{Minimize } \frac{1}{2}\|G(t_0, z_0, z_0')\|^2 + \frac{1}{2}\sum_{\ell=2}^{d+1} \|z_0^{(\ell)})\|^2 \\ &\text{w.r.t.} \quad (z_0, z_0', \ldots, z_0^{(d+1)})^\top \\ &\text{s.t.} \quad F^{(j)}(t_0, z_0, z_0', \ldots, z_0^{(j+1)}) = 0, \qquad j = 0, 1, \ldots, d. \end{aligned}$$

In practical computations the major challenge for higher index DAEs is to obtain analytical expressions or numerical approximations of the derivatives in $F^{(j)}$, $j = 1, \ldots, d$. For this purpose computer algebra packages like MAPLE, MATHEMATICA, or the symbolic toolbox of MATLAB can be used. Algorithmic differentiation tools are suitable as well, compare [72] for an overview. A potential issue is redundancy in the constraints (3.3) and the identification of the relevant equations in the derivative array. Approaches for the consistent initialization of general DAEs can be found in [1, 30, 35, 51, 71, 78, 93, 108]. A different approach is used in [127] in the context of shooting methods for parameter identification problems or optimal control problems. Herein, the algebraic constraints of the DAE are relaxed such that they are satisfied for any initial value. Then, the relaxation terms are driven to zero in the superordinate optimization problem in order to ensure consistency with the original DAE.

3.2 Dependence on Parameters

Initial values may depend on parameters that are present in the DAE. To this end the recomputation of consistent initial values for perturbed parameters becomes necessary or a parametric sensitivity analysis has to be performed, compare [66, 69]. Such issues frequently arise in the context of optimal control problems or parameter identification problems subject to DAEs, compare [68].

Example 3.2 Consider the equations of motion of a pendulum of mass m and length ℓ in the plane:

$$\begin{aligned} q_1'(t) &= v_1(t), \\ q_2'(t) &= v_2(t), \\ m v_1'(t) &= \quad -2q_1(t)\lambda(t), \\ m v_2'(t) &= -mg - 2q_2(t)\lambda(t), \\ 0 &= q_1(t)^2 + q_2(t)^2 - \ell^2. \end{aligned}$$

Herein, (q_1, q_2) denotes the pendulum's position, (v_1, v_2) its velocity, and λ the stress in the bar. A consistent initial value $(q_{1,0}, q_{2,0}, v_{1,0}, v_{2,0}, \lambda_0)$ has to satisfy the equations

$$0 = q_{1,0}^2 + q_{2,0}^2 - \ell^2, \tag{3.5}$$

$$0 = q_{1,0}v_{1,0} + q_{2,0}v_{2,0}, \tag{3.6}$$

$$0 = -\frac{\ell^2}{m}\lambda_0 + (v_{1,0}^2 + v_{2,0}^2) - q_{2,0}g. \tag{3.7}$$

Apparently, the algebraic component λ_0 depends on the parameter $p = (m, \ell, g)^\top$ according to

$$\lambda_0 = \lambda_0(p) = \frac{m}{\ell^2}\left(v_{1,0}^2 + v_{2,0}^2 - q_{2,0}g\right).$$

But in addition, the positions $q_{1,0}$ and $q_{2,0}$ depend on ℓ through the relation (3.5). So, if ℓ changes, then $q_{1,0}$ and/or $q_{2,0}$ have to change as well subject to (3.5) and (3.6). However, those equations in general do not uniquely define $q_{1,0}, q_{2,0}, v_{1,0}, v_{2,0}$ and the question arises, which set of values one should choose?

Firstly, we focus on the recomputation of an initial value for perturbed parameters. As the previous example shows, there is not a unique way to determine such a consistent initial value. A common approach is to use a projection technique, compare, e.g., [69] for a class of index-two DAEs, [68, Sect. 4.5.1] for Hessenberg DAEs, or [33] for general DAEs.

Consider the general parametric DAE

$$F(t, z(t), z'(t), p) = 0 \tag{3.8}$$

and the corresponding derivative array

$$F^{(j)}(t, z, z', \ldots, z^{(j+1)}, p) = 0, \qquad j = 0, 1, \ldots, d.$$

Remark 3.1 Please note that the differentiation index d of the general parametric DAE (3.8) may depend on p. For simplicity, we assume throughout that this is not the case (at least locally around a fixed nominal parameter).

Let $\tilde{p}$ be a given parameter. Suppose $\tilde{z}_0 = z_0(\tilde{p})$ with $\tilde{z}_0' = z_0'(\tilde{p}), \ldots, \tilde{z}_0^{(d+1)} = z_0^{(d+1)}(\tilde{p})$ is consistent. In order to find a consistent initial value for p, which is supposed to be close to $\tilde{p}$, solve the following parametric constrained least-squares problem:

LSQ(p): *Minimize*

$$\frac{1}{2}\|(\xi_0, \xi_0', \ldots, \xi_0^{(d+1)})^\top - (\tilde{z}_0, \tilde{z}_0', \ldots, \tilde{z}_0^{(d+1)})^\top\|^2$$

with respect to $(\xi_0, \xi_0', \ldots, \xi_0^{(d+1)})^\top$ *subject to the constraints*

$$F^{(j)}(t_0, \xi_0, \xi_0', \ldots, \xi_0^{(j+1)}, p) = 0, \qquad j = 0, 1, \ldots, d.$$

Remark 3.2 In case of a parametric Hessenberg DAE it would be sufficient to consider the hidden constraints up to order $k-2$ as the constraints in LSQ(p) and to compute a consistent algebraic component afterwards. Moreover, the quantities t_0 and $\tilde{z}_0^{(j)}$, $j = 0, \ldots, d+1$, could be considered as parameters of LSQ(p) as well, but here we are only interested in p.

The least-squares problem LSQ(p) is a parametric nonlinear optimization problem and allows for a sensitivity analysis in the spirit of [54] in order to investigate the sensitivity of a solution of LSQ(p) for p close to some nominal value $\hat{p}$. Let

$$L(\xi, \mu, p) = \frac{1}{2}\|\xi - \tilde{z}\|^2 + \mu^\top G(\xi, p) \tag{3.9}$$

with $\xi = (\xi_0, \xi_0', \ldots, \xi_0^{(d+1)})^\top$, $\tilde{z} = (\tilde{z}_0, \tilde{z}_0', \ldots, \tilde{z}_0^{(d+1)})^\top$, and

$$G(\xi, p) = \left(F^{(j)}(t_0, \xi_0, \xi_0', \ldots, \xi_0^{(j+1)}, p) \right)_{j=0,1,\ldots,d} \tag{3.10}$$

denote the Lagrange function of LSQ(p).

Theorem 3.1 (Sensitivity Theorem, Compare [54]) *Let G in (3.10) be twice continuously differentiable and $\hat{p}$ a nominal parameter. Let $\hat{\xi}$ be a local minimum of LSQ($\hat{p}$) with Lagrange multiplier $\hat{\mu}$ such that the following assumptions hold:*

(a) Linear independence constraint qualification: $G'_\xi(\hat{\xi}, \hat{p})$ has full rank.

(b) KKT conditions: $0 = \nabla_\xi L(\hat{\xi}, \hat{\mu}, \hat{p})$ with L from (3.9)

(c) Second-order sufficient condition:

$$L''_{\xi\xi}(\hat{\xi}, \hat{\mu}, \hat{p})(h, h) > 0 \qquad \forall h \neq 0 \,:\, G'_\xi(\hat{\xi}, \hat{p})h = 0.$$

Then there exist neighborhoods $B_\epsilon(\hat{p})$ and $B_\delta(\hat{\xi}, \hat{\mu})$, such that LSQ(p) has a unique local minimum

$$(\xi(p), \mu(p)) \in B_\delta(\hat{\xi}, \hat{\mu})$$

for each $p \in B_\epsilon(\hat{p})$. In addition, $(\xi(p), \mu(p))$ is continuously differentiable with respect to p with

$$\begin{pmatrix} L''_{\xi\xi}(\hat{\xi}, \hat{\mu}, \hat{p}) & G'_\xi(\hat{\xi}, \hat{p})^\top \\ G'_\xi(\hat{\xi}, \hat{p}) & 0 \end{pmatrix} \begin{pmatrix} \xi'(\hat{p}) \\ \mu'(\hat{p}) \end{pmatrix} = - \begin{pmatrix} L''_{\xi p}(\hat{\xi}, \hat{\mu}, \hat{p}) \\ G'_p(\hat{\xi}, \hat{p}) \end{pmatrix}. \tag{3.11}$$

The second equation in (3.11) reads

$$G'_{\xi}(\hat{\xi},\hat{p})\xi'(\hat{p}) + G'_{p}(\hat{\xi},\hat{p}) = 0$$

and in more detail using (3.10),

$$\sum_{\ell=0}^{j+1} \frac{\partial F^{(j)}}{\partial z^{(\ell)}}(t_0,\hat{\xi}_0,\ldots,\hat{\xi}_0^{(j+1)},\hat{p}) \cdot (\xi_0^{(\ell)})'(\hat{p}) + \frac{\partial F^{(j)}}{\partial p}(t_0,\hat{\xi}_0,\ldots,\hat{\xi}_0^{(j+1)},\hat{p}) = 0$$

for $j = 0, 1, \ldots, d$. Let us define $S_0^{(\ell)} := (\xi_0^{(\ell)})'(\hat{p})$ for $\ell = 0, \ldots, d+1$. Then, we obtain

$$\sum_{\ell=0}^{j+1} \frac{\partial F^{(j)}}{\partial z^{(\ell)}}(t_0,\hat{\xi}_0,\ldots,\hat{\xi}_0^{(j+1)},\hat{p}) \cdot S_0^{(\ell)} + \frac{\partial F^{(j)}}{\partial p}(t_0,\hat{\xi}_0,\ldots,\hat{\xi}_0^{(j+1)},\hat{p}) = 0 \qquad (3.12)$$

and in particular for $j = 0$,

$$\frac{\partial F}{\partial z}(t_0,\hat{\xi}_0,\hat{\xi}'_0,\hat{p}) \cdot S_0 + \frac{\partial F}{\partial z'}(t_0,\hat{\xi}_0,\hat{\xi}'_0,\hat{p}) \cdot S'_0 + \frac{\partial F}{\partial p}(t_0,\hat{\xi}_0,\hat{\xi}'_0,\hat{p}) = 0,$$

which is the linearization of (3.8) around $(t_0, \xi_0, \xi'_0, \hat{p})$ with respect to p. Taking into account the definition of the further components $F^{(j)}, j = 1, \ldots, d+1$, of the derivative array, compare (2.3), we recognize that (3.12) provides a linearization of (2.3) with respect to p. Hence, the settings

$$S(t_0) = S_0, \quad S'(t_0) = S'_0, \ldots, S^{(d+1)}(t_0) = S_0^{(d+1)}$$

provide consistent initial values for the sensitivity DAE

$$F'_z(t,z(t),z'(t),p)S(t) + F'_{z'}(t,z(t),z'(t),p)S'(t) + F'_p(t,z(t),z'(t),p) = 0,$$

where $S(t) := \partial z(t;p)/\partial p$ denotes the sensitivity of the solution of (3.8) with respect to the parameter p, compare Sect. 7. Herein, it is assumed that F is sufficiently smooth with respect to all arguments.

In summary, the benefits of the projection approach using LSQ(p) are twofold: Firstly, it allows to compute consistent initial values for the DAE itself. Secondly, the sensitivity analysis provides consistent initial values for the sensitivity DAE. Finally, the sensitivity analysis can be used to predict consistent initial values under perturbations through the Taylor expansion

$$\xi(p) = \xi(\hat{p}) + \xi'(\hat{p})(p-\hat{p}) + o(\|p-\hat{p}\|)$$

for $p \in B_\varepsilon(\hat{p})$.

4 Integration Methods

A vast number of numerical discretizations schemes exist for DAEs, most of them are originally designed for ODEs, such as BDF methods or Runge–Kutta methods. The methods for DAEs are typically (at least in part) implicit methods owing to the presence of algebraic equations. It is beyond the scope of the paper to provide a comprehensive overview on all the existing numerical discretization schemes for DAEs, since excellent textbooks with convergence results and many details are available, for instance [29, 73, 75–77, 90, 92, 134]. Our intention is to discuss the most commonly used methods, their construction principles, and some of their features. Efficient implementations use a bunch of additional ideas to improve the efficiency.

All methods work on a grid

$$\mathbb{G}_h = \{t_0 < t_1 < \ldots < t_{N-1} < t_N = t_f\}$$

with $N \in \mathbb{N}$ and step-sizes $h_k = t_{k+1} - t_k$, $k = 0, \ldots, N-1$. The maximum step-size is denoted by $h = \max_{k=0,\ldots,N-1} h_k$. The methods generate a grid function $z_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ with $z_h(t) \approx z(t)$ for $t \in \mathbb{G}_h$, where $z(t)$ denotes the solution of (1.1) with a consistent initial value z_0. The discretization schemes can be grouped into *one-step methods* with

$$z_h(t_{i+1}) = z_h(t_i) + h_i \Phi(t_i, z_h(t_i), h_i), \qquad i = 0, \ldots, N-1, \tag{4.1}$$

for a given consistent initial value $z_h(t_0) = z_0$ and *s-stage multi-step methods* with

$$z_h(t_{i+s}) = \Psi(t_i, \ldots, t_{i+s}, z_h(t_i), \ldots, z_h(t_{i+s-1}), h_i, \ldots, h_{i+s-1}), \quad i = 0, \ldots, N-s, \tag{4.2}$$

for given consistent initial values $z_h(t_0) = z_0, \ldots, z_h(t_{s-1}) = z_{s-1}$. Note that multi-step methods with $s > 1$ require an initialization procedure to compute $z_1, \ldots, z_{s-1}$. This can be realized by performing $s-1$ steps of a suitable one-step method or by using multi-step methods with $1, 2, \ldots, s-1$ stages successively for the first $s-1$ steps.

The aim is to construct convergent methods such that the *global error* $e_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n$ defined by

$$e_h := z_h - \Delta_h(z), \qquad e_h(t) = z_h(t) - \Delta_h(z)(t),\ t \in \mathbb{G}_h,$$

satisfies

$$\lim_{h \to 0} \|e_h\|_\infty = 0$$

or even exhibits the order of convergence $p \in \mathbb{N}$, i.e.

$$\|e_h\|_\infty = \mathcal{O}(h^p) \text{ as } h \to 0.$$

Herein, $\Delta_h : \{z : [t_0, t_f] \longrightarrow \mathbb{R}^n\} \longrightarrow \{z_h : \mathbb{G}_h \longrightarrow \mathbb{R}^n\}$ denotes the restriction operator onto the set of grid functions on $\mathbb{G}_h$ defined by $\Delta_h(z)(t) = z(t)$ for $t \in \mathbb{G}_h$.

A convergence proof for a specific discretization scheme typically resembles the reasoning

$$\text{consistency} \quad + \quad \text{stability} \qquad \Longrightarrow \qquad \text{convergence},$$

compare [131]. Herein, consistency is not to be confused with consistent initial values. Instead, consistency of a discretization method measures how well the exact solution satisfies the discretization scheme. Detailed definitions of consistency and stability and convergence proofs for various classes of DAEs (index-one, Hessenberg DAEs up to order 3, constant/variable step-sizes) can be found in the above-mentioned textbooks [29, 73, 75–77, 90, 92, 134]. As a rule, one cannot in general expect the same order of convergence for differential and algebraic variables for higher index DAEs.

4.1 BDF Methods

The Backward Differentiation Formulas (BDF) are implicit multi-step methods and were introduced in [40, 61]. A BDF method with $s \in \mathbb{N}$ stages is based on the construction of interpolating polynomials, compare Fig. 1. Suppose the method has produced approximations $z_h(t_{i+k})$, $k = 0, \ldots, s-1$, of z at the grid points t_{i+k}, $k = 0, \ldots, s-1$. The aim is to determine an approximation $z_h(t_{i+s})$ of $z(t_{i+s})$, where $i \in \{0, \ldots, N-s\}$.

To this end, let $P(t)$ be the interpolating polynomial of degree at most s with

$$P(t_{i+k}) = z_h(t_{i+k}), \qquad k = 0, \ldots, s.$$

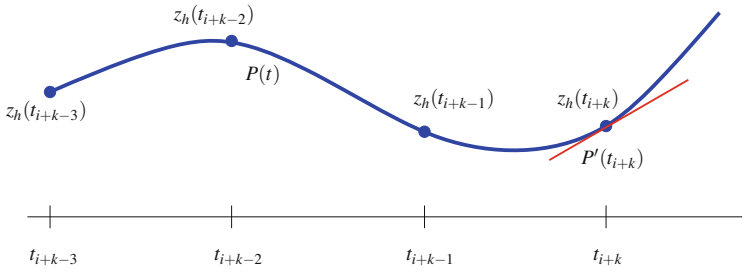


Fig. 1 Idea of BDF method: polynomial interpolation of approximations

The polynomial P can be expressed as

$$P(t) = \sum_{k=0}^{s} z_h(t_{i+k}) L_k(t), \qquad L_k(t) = \prod_{\ell=0, \ell \neq k}^{s} \frac{t - t_{i+\ell}}{t_{i+k} - t_{i+\ell}},$$

where the L_k's denote the Lagrange polynomials. Note that P interpolates the unknown vector $z_h(t_{i+s})$, which is determined implicitly by the postulation that P satisfies the DAE (1.1) at t_{i+s}, i.e.

$$F(t_{i+s}, z_h(t_{i+s}), P'(t_{i+s})) = 0. \tag{4.3}$$

The above representation of P yields

$$P'(t_{i+s}) = \sum_{k=0}^{s} z_h(t_{i+k}) L_k'(t_{i+s}) =: \frac{1}{h_{i+s-1}} \sum_{k=0}^{s} \alpha_k z_h(t_{i+k}),$$

with $\alpha_k = h_{i+s-1} L_k'(t_{i+s})$, $k = 0, \ldots, s$.

Example 4.1 The BDF methods with $s \leq 6$ and a constant step-size h read as follows, see [134, S. 335]:

$$s = 1 : hP'(t_{i+1}) = z_{i+1} - z_i \quad \text{(implicit Euler method)}$$

$$s = 2 : hP'(t_{i+2}) = \frac{1}{2}(3z_{i+2} - 4z_{i+1} + z_i)$$

$$s = 3 : hP'(t_{i+3}) = \frac{1}{6}(11z_{i+3} - 18z_{i+2} + 9z_{i+1} - 2z_i)$$

$$s = 4 : hP'(t_{i+4}) = \frac{1}{12}(25z_{i+4} - 48z_{i+3} + 36z_{i+2} - 16z_{i+1} + 3z_i)$$

$$s = 5 : hP'(t_{i+5}) = \frac{1}{60}(137z_{i+5} - 300z_{i+4} + 300z_{i+3} - 200z_{i+2} + 75z_{i+1} - 12z_i)$$

$$s = 6 : hP'(t_{i+6}) = \frac{1}{60}(147z_{i+6} - 360z_{i+5} + 450z_{i+4} - 400z_{i+3} + 225z_{i+2} - 72z_{i+1} + 10z_i).$$

Abbreviations: $z_{i+k} = z_h(t_{i+k})$, $k = 0, \ldots, 6$.

Introducing the expression for $P'(t_{i+s})$ into (4.3) yields the nonlinear equation

$$F\left(t_{i+s}, z_h(t_{i+s}), \frac{1}{h_{i+s-1}} \sum_{k=0}^{s} \alpha_k z_h(t_{i+k})\right) = 0 \tag{4.4}$$

for $z_h(t_{i+s})$. Suppose (4.4) possesses a solution $z_h(t_{i+s})$ and the matrix pencil

$$F'_z + \frac{\alpha_s}{h_{i+s-1}} F'_{z'} \tag{4.5}$$

is regular at this solution, i.e., there exists a step-size h_{i+s-1} such that the matrix (4.5) is non-singular. Then the implicit function theorem allows to solve Eq. (4.4) locally for $z_h(t_{i+s})$ and to express it in the form (4.2). In practice Newton's method or the simplified Newton method is used to solve Eq. (4.4) numerically, which requires the non-singularity of the matrix in (4.5) at the Newton iterates.

BDF methods are appealing amongst implicit methods since the effort per integration step amounts to solving just one nonlinear equation of dimension n, whereas a fully implicit s-stage Runge–Kutta method requires to solve a nonlinear equation of dimension $n \cdot s$. For numerical purposes only the BDF methods with $s \leq 6$ are relevant, since the methods for $s > 6$ are unstable. The maximal attainable order of convergence of an s-stage BDF method is s.

Convergence results assuming fixed step-sizes for BDF methods for certain subclasses of the general DAE (1.1), amongst them are index-one problems and Hessenberg DAEs, can be found in [27, 63, 99, 111]. Variable step-sizes may result in non-convergent components of the algebraic variables for index-three Hessenberg DAEs, compare [64]. This is another motivation to use an index reducing stabilization technique as in Sect. 2.2.

The famous code DASSL, [29, 109], is based on BDF methods, but adds several features like an automatic step-size selection strategy, a variable order selection strategy, a root finding strategy, and a parametric sensitivity module to the basic BDF method. Moreover, the re-use of Jacobians for one or more integration steps and numerically efficient divided difference schemes for the calculation of the interpolating polynomial P increase the efficiency of the code. The code ODASSL by Führer [56] and Führer and Leimkuhler [57] extends DASSL to overdetermined DAEs, which occur, e.g., for the GGL stabilization in Sect. 2.2. In these codes, the error tolerances for the algebraic variables of higher index DAEs have to be scaled by powers of $1/h$ compared to those of the differential states since otherwise the automatic step-size selection algorithm breaks down frequently, compare [110]. An enhanced version of DASSL is available in the package SUNDIALS, [80], which provides several methods (Runge–Kutta, Adams, BDF) for ODEs and DAEs in one software package.

4.2 Runge–Kutta Methods

A Runge–Kutta method with $s \in \mathbb{N}$ stages for (1.1) is a one-step method of type

$$z_h(t_{i+1}) = z_h(t_i) + h_i \Phi(t_i, z_h(t_i), h_i) \tag{4.6}$$

with the increment function

$$\Phi(t,z,h) := \sum_{j=1}^{s} b_j k_j(t,z,h) \tag{4.7}$$

and the stage derivatives $k_j(t,z,h)$, $j = 1,\dots,s$. The stage derivatives k_j are implicitly defined by the system of $n \cdot s$ nonlinear equations

$$F\left(t_i + c_1 h_i, z_{i+1}^{(1)}, k_1\right) = 0, \tag{4.8}$$

$$\vdots$$

$$F\left(t_i + c_s h_i, z_{i+1}^{(s)}, k_s\right) = 0, \tag{4.9}$$

where

$$z_{i+1}^{(\ell)} := z_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j}\, k_j, \qquad \ell = 1,\dots,s, \tag{4.10}$$

are approximations of z at the intermediate time points $t_i + c_\ell h$, $\ell = 1,\dots,s$. The coefficients in the Runge–Kutta method are collected in the Butcher array

$$\begin{array}{c|cccc} c_1 & a_{11} & a_{12} & \cdots & a_{1s} \\ c_2 & a_{21} & a_{22} & \cdots & a_{2s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & a_{s2} & \cdots & a_{ss} \\ \hline & b_1 & b_2 & \cdots & b_s \end{array}$$

Commonly used Runge–Kutta methods for DAEs are the RADAU IIA methods and the Lobatto IIIA and IIIC methods. These methods are stiffly accurate, i.e., they satisfy $c_s = 1$ and $a_{sj} = b_j$ for $j = 1,\dots,s$. This is a very desirable property for DAEs since it implies that (4.9) and $z_{i+1}^{(s)} = z_h(t_{i+1})$ hold at $t_{i+1} = t_i + c_s h_i$. Runge–Kutta methods, which are not stiffly accurate, can be used as well. However, for those it has to be enforced that the approximation $z_h(t_{i+1})$ satisfies the algebraic constraints of the DAE at t_{i+1}. This can be achieved by projecting the output of the Runge–Kutta method onto the algebraic constraints, compare [18].

Example 4.2 (RADAU IIA) The RADAU IIA methods with $s = 1, 2, 3$ are defined by the following Butcher arrays, compare [134, Beispiel 6.1.5]:

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} 1/3 & 5/12 & -1/12 \\ 1 & 3/4 & 1/4 \\ \hline & 3/4 & 1/4 \end{array}$$

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

The maximal attainable order of convergence is $2s - 1$.

Example 4.3 (Lobatto IIIA and Lobatto IIIC) The Lobatto IIIA methods with $s = 2, 3$ are defined by the following Butcher arrays, compare [134, Beispiel 6.1.6]:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & 0 & \\ 1/2 & 5/24 & 1/3 & -1/24 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

The Lobatto IIIC methods with $s = 2, 3$ are defined by the following Butcher arrays, compare [134, Beispiel 6.1.8]:

$$\begin{array}{c|cc} 0 & 1/2 & -1/2 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|ccc} 0 & 1/6 & -1/3 & 1/6 \\ 1/2 & 1/6 & 5/12 & -1/12 \\ 1 & 1/6 & 2/3 & 1/6 \\ \hline & 1/6 & 2/3 & 1/6 \end{array}$$

The maximal attainable order of convergence is $2s - 2$. A combined method of Lobatto IIIA and IIIC methods for mechanical multibody systems can be found in [124].

The main effort per integration step is to solve the system of nonlinear equations (4.8)–(4.9) for the unknown vector of stage derivatives $k = (k_1, \ldots, k_s)^\top$ by Newton's method or by a simplified version of it, where the Jacobian matrix is kept constant for a couple of iterations or integration steps. Another way to reduce the computational effort is to consider ROW methods or half-explicit Runge–Kutta methods as in Sect. 4.3.

Convergence results and order conditions for Runge–Kutta methods applied to DAEs can be found in, e.g., [28, 75, 84, 85].

4.3 Rosenbrock-Wanner (ROW) Methods

In this section, we introduce and discuss the so-called *Rosenbrock-Wanner(ROW) methods* for DAEs, cf. [121], where H.H. Rosenbrock introduced this method class. ROW methods are one-step methods, which are based on implicit Runge–Kutta methods. In literature, these methods are also called *Rosenbrock methods*, *linearly-implicit or semi-implicit Runge–Kutta methods*, cf. [73].

The motivation to introduce an additional class of integration methods is to avoid solving a fully nonlinear system of dimension $n \cdot s$ and to solve instead of that only linear systems. Thus, the key idea for the derivation of Rosenbrock-Wanner methods is to perform one Newton-step to solve Eqs. (4.8)–(4.9) for a Runge–Kutta method with $a_{ij} = 0$ for $i < j$ (diagonally implicit RK method, cf. [73]). We rewrite these

equations for such a method and an autonomous implicit DAE,

$$F(z, z') = 0, \tag{4.11}$$

as follows:

$$F\left(z_h(t_i) + h_i \left(\sum_{j=1}^{\ell-1} a_{\ell j}\, k_j + a_{\ell\ell}\, k_\ell\right), k_\ell\right) = 0, \quad \ell = 1, \ldots, s. \tag{4.12}$$

Due to the fact that we consider a diagonally implicit RK method, the above equations are decoupled and can be solved successively. Then, performing one Newton-step with starting value $k_\ell^{(0)}$ leads to

$$\left(F_z'\left(z_{i+1}^{(\ell-1)}, k_\ell^{(0)}\right) h_i \cdot a_{\ell\ell} + F_{z'}'\left(z_{i+1}^{(\ell-1)}, k_\ell^{(0)}\right)\right)\left(k_\ell - k_\ell^{(0)}\right) = -F\left(z_{i+1}^{(\ell-1)}, k_\ell^{(0)}\right), \tag{4.13}$$

for $\ell = 1, \ldots, s$.

We come to the general class of Rosenbrock-Wanner methods by proceeding with the following steps. First, we take as starting value $k_\ell^{(0)} = 0$ for $\ell = 1, \ldots, s$. Then, the Jacobians are evaluated at the fixed point $z_h(t_i)$ instead of $z_{i+1}^{(\ell-1)}$, which saves computational costs substantially. Moreover, linear combinations of the previous stages $k_j, j = 1, \ldots, \ell$ are introduced. And last but not least, the method is extended to general non-autonomous implicit DAEs as Eq. (1.1). We obtain the following class of Rosenbrock methods

$$F\left(t_i + c_\ell h_i, z_{i+1}^{(\ell-1)}, 0\right) + h_i J_z \sum_{j=1}^{\ell} \gamma_{\ell j} k_j + J_{z'} k_\ell + \gamma_\ell J_t = 0, \quad \ell = 1, \ldots, s, \tag{4.14}$$

with

$$J_z := F_z(t_i, z_h(t_i), 0), \tag{4.15}$$

$$J_{z'} := F_{z'}(t_i, z_h(t_i), 0), \tag{4.16}$$

$$J_t := F_t(t_i, z_h(t_i), 0). \tag{4.17}$$

The solution at the next time point t_{i+1} is computed exactly as in the case of Runge–Kutta methods:

$$z_h(t_{i+1}) = z_h(t_i) + h_i \Phi(t_i, z_h(t_i), h_i), \qquad \Phi(t, z, h) := \sum_{j=1}^{s} b_j k_j(t, z, h), \tag{4.18}$$

with the stage derivatives $k_j(t, z, h)$ defined by the *linear* system (4.14). An example for a ROW method is the linearly implicit Euler method ($s = 1$), for which the stage derivative is defined as follows

$$F(t_i, z_i, 0) + h_i J_z k_1 + J_{z'} k_1 = 0. \tag{4.19}$$

For semi-explicit DAEs of the form (2.5)–(2.6), a ROW method as defined above reads as

$$z_h^x(t_{i+1}) = z_h^x(t_i) + \sum_{j=1}^{s} b_j k_j^x(t_i, z_h^x(t_i), z_h^y(t_i), h_i) \tag{4.20}$$

$$z_h^y(t_{i+1}) = z_h^y(t_i) + \sum_{j=1}^{s} b_j k_j^y(t_i, z_h^x(t_i), z_h^y(t_i), h_i), \tag{4.21}$$

with

$$\begin{aligned} &\begin{pmatrix} f\left(t_i + c_\ell h_i, z_{i+1}^{x,(\ell-1)}, z_{i+1}^{y,(\ell-1)}\right) \\ g\left(t_i + c_\ell h_i, z_{i+1}^{x,(\ell-1)}, z_{i+1}^{y,(\ell-1)}\right) \end{pmatrix} + h_i \cdot \begin{pmatrix} f_{z^x} & f_{z^y} \\ g_{z^x} & g_{z^y} \end{pmatrix} \sum_{j=1}^{\ell} \gamma_{\ell j} \begin{pmatrix} k_j^x \\ k_j^y \end{pmatrix} \\ &\qquad + \begin{pmatrix} -k_\ell^x \\ 0 \end{pmatrix} + \gamma_\ell \begin{pmatrix} f_t \\ g_t \end{pmatrix} = 0, \end{aligned} \tag{4.22}$$

for $\ell = 1, \ldots, s$. Herein, we have set $z = ((z^x)^\top, (z^y)^\top)^\top = (x^\top, y^\top)^\top$ and

$$F(t, z, z') = \begin{pmatrix} f(t, x, y) - x' \\ g(t, x, y) \end{pmatrix}. \tag{4.23}$$

Up to now, we have considered ROW methods with exact Jacobian matrices $J_z = F_z, J_{z'} = F_{z'}$. There is an additional class of integration methods, which uses for J_z arbitrary matrices ('inexact Jacobians')—such methods are called *W-methods*, see [73, 146, 147].

We further remark that related integration methods can be derived, if other starting values are used for the stage derivatives, instead of $k_\ell^{(0)} = 0$ as it is done to derive Eq. (4.14), cf. [67, 68]—the methods derived there as well as ROW and W-methods can be seen to belong the common class of *linearized implicit Runge–Kutta methods*.

An introduction and more detailed discussion of ROW methods can be found in [73]; convergence results for general one-step methods (including ROW methods) applied to DAEs are available in [41]. Moreover, ROW methods for index-one DAEs in semi-explicit form are studied in [53, 117, 120, 135]; index-one problems and singularly perturbed problems are discussed in [23, 24, 74]. Analysis results and specific methods for the equations of motion of mechanical multibody systems, i.e.,

index-three DAEs in semi-explicit form are derived in [145, 146]; compare also the results in [14, 106, 123].

4.4 Half-Explicit Methods

In this section, we briefly discuss the so-called *half-explicit* Runge–Kutta methods, here for autonomous index-two DAEs in semi-explicit form. That is, we consider DAE systems of the form

$$x'(t) = f(x(t), y(t)), \tag{4.24}$$

$$0 = g(x(t)), \tag{4.25}$$

with initial values (x_0, y_0) that are assumed to be consistent. To derive the class of half-explicit Runge–Kutta methods, it is more convenient to use stages rather than the stage-derivatives k_ℓ as before. In particular, for the semi-explicit DAE (4.24), (4.25), we define stages for the differential and the algebraic variables as

$$X_{i\ell} := x_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j} k_j^x, \quad Y_{i\ell} := y_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j} k_j^y, \quad \ell = 1, \ldots, s. \tag{4.26}$$

Then, it holds

$$\begin{aligned} X_{i\ell} &= x_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j} k_j^x \\ &= x_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j} f\left(x_h(t_i) + \sum_{m=1}^{s} a_{jm} k_m^x,\ y_h(t_i) + \sum_{m=1}^{s} a_{jm} k_m^y \right) \\ &= x_h(t_i) + h \sum_{j=1}^{s} a_{\ell j} f(X_{ij}, Y_{ij}). \end{aligned} \tag{4.27}$$

Using this notation and the coefficients of an explicit Runge–Kutta scheme, half-explicit Runge–Kutta methods as firstly introduced in [75] are defined as follows

$$X_{i\ell} = x_h(t_i) + h_i \sum_{j=1}^{\ell-1} a_{\ell j} f(X_{nj}, Y_{nj}), \qquad \ell = 1, \ldots, s, \tag{4.28}$$

$$0 = g(X_{i\ell}), \tag{4.29}$$

$$x_h(t_{i+1}) = x_h(t_i) + h_i \sum_{\ell=1}^{s} b_\ell f(X_{i\ell}, Y_{i\ell}), \tag{4.30}$$

$$0 = g(x_h(t_{i+1})). \tag{4.31}$$

The algorithmic procedure is as follows: We start with $X_{i1} = x_h(t_i)$ assumed to be consistent. Then, taking Eq. (4.28) for X_{i2} and inserting into Eq. (4.29) lead to

$$0 = g(X_{i2}) = g\left(x_h(t_i) + a_{21} h_i f(X_{i1}, Y_{i1})\right), \tag{4.32}$$

this is a nonlinear equation that can be solved for Y_{i1}. Next, we calculate X_{i2} from Eq. (4.28) and, accordingly, Y_{i2}, etc. For methods with $c_s = 1$, one obtains an approximation for the algebraic variable at the next time-point by $y_h(t_{i+1}) = Y_{is}$. The key idea behind this kind of integration schemes is to apply an explicit Runge–Kutta scheme for the differential variable and to solve for the algebraic variable implicitly.

Convergence studies for this method class applied to index-two DAEs can be found in [26, 75]. In [7, 12, 105] the authors introduce a slight modification of the above stated scheme, which improves the method class concerning order conditions and computational efficiency. To be more precise, *partitioned half-explicit Runge–Kutta methods* for index-two DAEs in semi-explicit form are defined in the following way:

$$\begin{aligned}
X_{i1} &= x_h(t_i), \quad Y_{i1} = y_h(t_i),\\
X_{i\ell} &= x_h(t_i) + h_i \sum_{j=1}^{\ell-1} a_{\ell j} f(X_{ij}, Y_{ij}),\\
\bar{X}_{i\ell} &= x_h(t_i) + h_i \sum_{j=1}^{\ell} \bar{a}_{\ell j} f(X_{ij}, Y_{ij}),\\
0 &= g(\bar{X}_{i\ell}),\\
\ell &= 2, \ldots s+1,\\
x_h(t_{i+1}) &= X_{i,s+1}, \qquad y_h(t_{i+1}) = Y_{i,s+1}.
\end{aligned} \tag{4.33}$$

Results concerning the application of half-explicit methods to index-one DAE are available in [13]; the application to index-three DAEs is discussed in [107].

4.5 Examples

Some illustrative examples with DAEs are discussed. Example 4.4 addresses an index-three mechanical multibody system of a car on a bumpy road. A docking maneuver of a satellite to a tumbling target is investigated in Example 4.5. Herein, the use of quaternions leads to a formulation with an index-one DAE.

Example 4.4 We consider a vehicle simulation for the ILTIS on a bumpy road section. A detailed description of the mechanical multibody system is provided in [128]. The system was modeled by SIMPACK [81] and the simulation results were obtained using the code export feature of SIMPACK and the BDF method DASSL [29]. The mechanical multibody system consists of 11 rigid bodies with a total of 25 degrees of freedom (DOF) (chassis with 6 DOF, wheel suspension with 4 DOF in total, wheels with 12 DOF in total, steering rod with 1 DOF, camera with 2 DOF). The motion is restricted by 9 algebraic constraints. Figure 2 illustrates the test track with bumps and the resulting pitch and roll angles, and the vertical excitation of the chassis. The integration tolerance within DASSL is set to 10^{-4} for the differential states and to 10^{8} for the algebraic states (i.e., no error control was performed for the algebraic states).

Example 4.5 We consider a docking maneuver of a service satellite (S) to a tumbling object (T) on an orbit around the earth, compare [103]. Both objects are able to rotate freely in space and quaternions are used to parametrize their orientation. Note that, in contrast to Euler angles, quaternions lead to a continuous parametrization of the orientation without singularities.

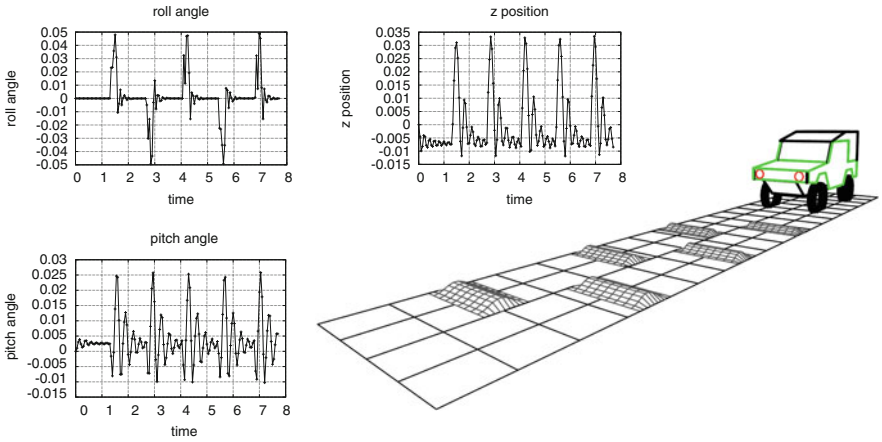


Fig. 2 Simulation results of the ILTIS on a bumpy road: roll angle, pitch angle, vertical excitation of chassis

The relative dynamics of S and T are approximately given by the *Clohessy-Wilshire-Equations*

$$\begin{aligned} x''(t) &= 2ny'(t) - 3n^2x(t) + a_x(t), \\ y''(t) &= -2nx'(t) + a_y(t), \\ z''(t) &= -n^2z(t) + a_z(t), \end{aligned}$$

where $(x, y, z)^\top$ is the relative position of S and T, $a = (a_x, a_y, a_z)^\top$ is a given control input (thrust) to S, $n = \sqrt{\mu/a_t^3}$, a_t is the semi-major axis of the orbit (assumed to be circular), and μ is the gravitational constant.

The direction cosine matrix using quaternions $q = (q_1, q_2, q_3, q_4)^\top$ is defined by

$$R(q)^\top = \begin{pmatrix} q_1^2 - q_2^2 - q_3^2 + q_4^2 & 2\,(q_1q_2 + q_3q_4) & 2\,(q_1q_3 - q_2q_4) \\ 2\,(q_1q_2 - q_3q_4) & -q_1^2 + q_2^2 - q_3^2 + q_4^2 & 2\,(q_2q_3 + q_1q_4) \\ 2\,(q_1q_3 + q_2q_4) & 2\,(q_2q_3 - q_1q_4) & -q_1^2 - q_2^2 + q_3^2 + q_4^2 \end{pmatrix}.$$

The matrix $R(q)$ represents the rotation matrix from rotated to non-rotated state. The orientation of S and T with respect to an unrotated reference coordinate system is described by quaternions $q^S = (q_1^S, q_2^S, q_3^S, q_4^S)^\top$ for S and $q^T = (q_1^T, q_2^T, q_3^T, q_4^T)^\top$ for T. With the angular velocities $\omega^S = (\omega_1^S, \omega_2^S, \omega_3^S)^\top$ and $\omega^T = (\omega_1^T, \omega_2^T, \omega_3^T)^\top$ the quaternions obey the differential equations

$$(q^\alpha)'(t) = \frac{1}{2} \begin{pmatrix} \omega^\alpha(t) \\ 0 \end{pmatrix} \otimes q^\alpha(t), \qquad \alpha \in \{S, T\}, \tag{4.34}$$

where the operator $\otimes$ is defined by

$$\begin{pmatrix} \omega \\ 0 \end{pmatrix} \otimes q = \begin{pmatrix} 0 & \omega_3 & -\omega_2 & \omega_1 \\ -\omega_3 & 0 & \omega_1 & \omega_2 \\ \omega_2 & -\omega_1 & 0 & \omega_3 \\ -\omega_1 & -\omega_2 & -\omega_3 & 0 \end{pmatrix} \begin{pmatrix} q_1 \\ q_2 \\ q_3 \\ q_4 \end{pmatrix}.$$

Assuming a constant mass distribution and body fixed coordinate systems that coincide with the principle axes, S and T obey the gyroscopic equations

$$\begin{aligned} (\omega_1^S)'(t) &= \frac{1}{J_{11}^S} \left(\omega_2^S(t)\omega_3^S(t) \left(J_{22}^S - J_{33}^S \right) + u_1(t) \right), \\ (\omega_2^S)'(t) &= \frac{1}{J_{22}^S} \left(\omega_1^S(t)\omega_3^S(t) \left(J_{33}^S - J_{11}^S \right) + u_2(t) \right), \\ (\omega_3^S)'(t) &= \frac{1}{J_{33}^S} \left(\omega_2^S(t)\omega_1^S(t) \left(J_{11}^S - J_{22}^S \right) + u_3(t) \right), \end{aligned}$$

$$(\omega_1^T)'(t) = \frac{1}{J_{11}^T}\left(\omega_2^T(t)\omega_3^T(t)\left(J_{22}^T - J_{33}^T\right)\right),$$
$$(\omega_2^T)'(t) = \frac{1}{J_{22}^T}\left(\omega_1^T(t)\omega_3^T(t)\left(J_{33}^T - J_{11}^T\right)\right),$$
$$(\omega_3^T)'(t) = \frac{1}{J_{33}^T}\left(\omega_2^T(t)\omega_1^T(t)\left(J_{11}^T - J_{22}^T\right)\right).$$

Herein $u = (u_1, u_2, u_3)^\top$ denotes a time-dependent torque input to S.

The quaternions are normalized to one by the algebraic constraints

$$0 = (q_1^\alpha)^2 + (q_2^\alpha)^2 + (q_3^\alpha)^2 + (q_4^\alpha)^2 - 1, \qquad \alpha \in \{S, T\},$$

which has to be obeyed since otherwise a drift-off would occur owing to numerical discretization errors. In order to incorporate these algebraic constraints, we treat $(q_4^S, q_4^T)^\top$ as algebraic variables and drop the differential equations for q_4^S and q_4^T in (4.34). In summary, we obtain an index-one DAE with differential state $(x, y, z, x', y', z', \omega_1^S, \omega_2^S, \omega_3^S, q_1^S, q_2^S, q_3^S, \omega_1^T, \omega_2^T, \omega_3^T, q_1^T, q_2^T, q_3^T)^\top \in \mathbb{R}^{18}$, algebraic state $(q_4^S, q_4^T)^\top \in \mathbb{R}^2$, and time-dependent control input $(a, u)^\top \in \mathbb{R}^6$ for S.

Figure 3 shows some snapshots of a docking maneuver on the time interval $[0, 667]$ with initial states

$$q^S(0) = (0, 0, 0, 1)^\top, \qquad q^T(0) = (-0.05, 0, 0, 0.99875)^\top,$$
$$\omega^S(0) = (0, 0, 0)^\top, \qquad \omega^T(0) = (0, 0.0349, 0.017453)^\top,$$
$$(x(0), y(0), z(0))^\top = (0, -100, 0)^\top, \qquad (x'(0), y'(0), z'(0))^\top = (0, 0, 0)^\top,$$

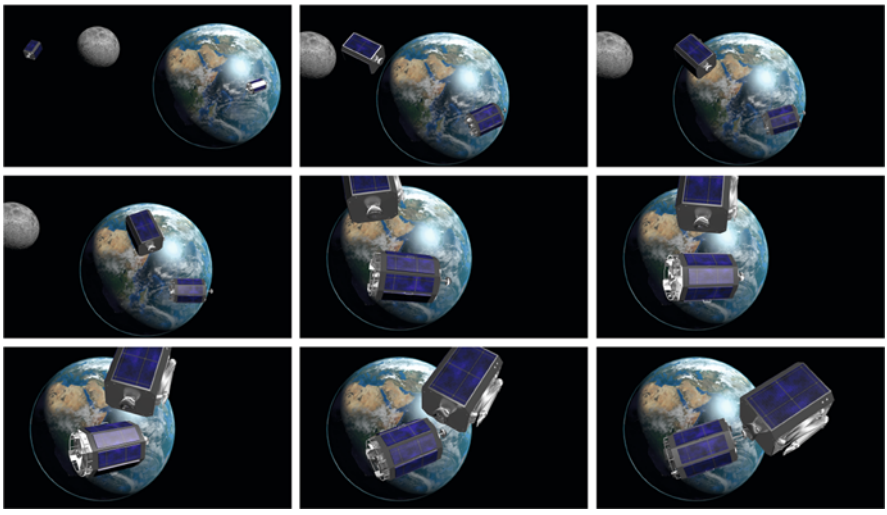


Fig. 3 Snapshots for the docking maneuver

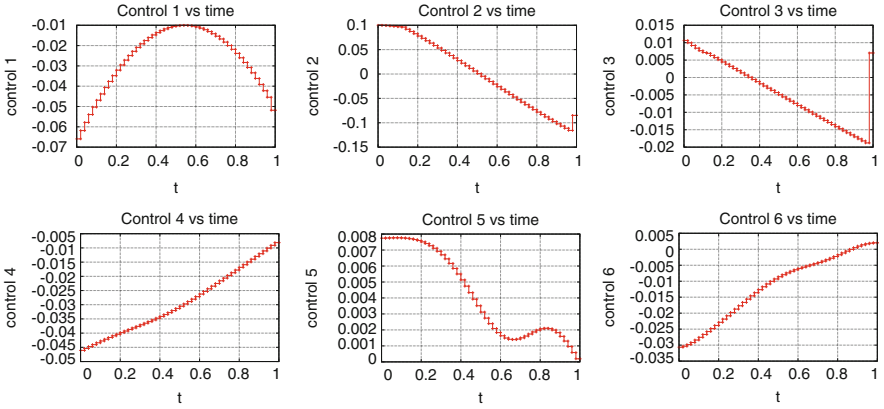


Fig. 4 Control input for the docking maneuver: ma_x, ma_y, ma_z with $m = 100$ (*top from left to right*), u_1, u_2, u_3 (*bottom from left to right*)

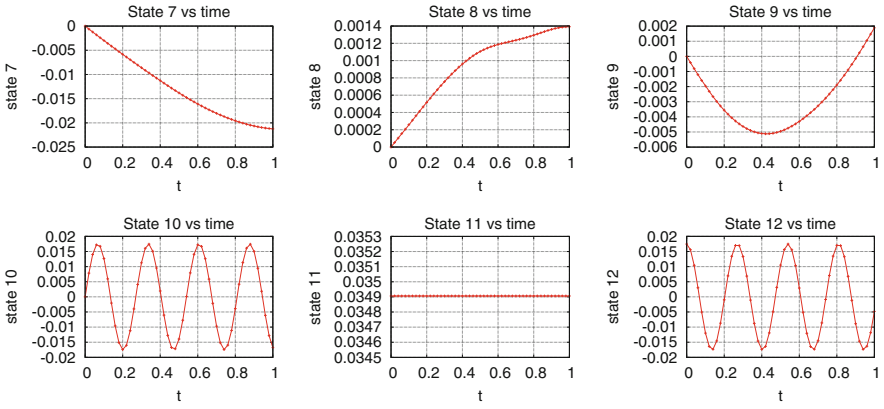


Fig. 5 Angular velocities of the service satellite S and the tumbling target T: $\omega_1^S, \omega_2^S, \omega_3^S$ (*top from left to right*), $\omega_1^T, \omega_2^T, \omega_3^T$ (*bottom from left to right*)

and parameters $a_t = 7071000$, $\mu = 398 \cdot 10^{12}$, $J_{11}^{\alpha} = 1000$, $J_{22}^{\alpha} = 2000$, $J_{33}^{\alpha} = 1000$, $\alpha \in \{S, T\}$. The integration tolerance within DASSL is set to 10^{-10} for the differential states and to 10^{-4} for the algebraic states. Figure 4 depicts the control inputs $m \cdot a = m \cdot (a_x, a_y, a_z)^\top$ with satellite mass $m = 100$ and $u = (u_1, u_2, u_3)^\top$. Finally, Fig. 5 shows the angular velocities ω^S and ω^T.

5 Co-simulation

In numerical system simulation, it is an essential task to simulate the dynamic interaction of different subsystems, possibly from different physical domains, modeled with different approaches, to be solved with different numerical solvers (multiphysical system models). Especially, in vehicle engineering, this becomes more and more important, because for a mathematical model of a modern passenger car or commercial vehicle, mechanical subsystems have to be coupled with flexible components, hydraulic subsystems, electronic and electric devices, and other control units. The mathematical models for all these subsystems are often given as DAE, but, typically, they substantially differ in their complexities, time constants, and scales; hence, it is not advisable to combine *all* model equations to *one entire* DAE and to solve it numerically with *one* integration scheme. In contrast, modern co-simulation strategies aim at using a specific numerical solver, i.e., DAE integration method, for each subsystem and to exchange only a limited number of coupling quantities at certain communication time points. Thus, it is important to analyze the behavior of such coupled simulation strategies, 'co-simulation', where the coupled subsystems are mathematically described as DAEs.

In addition to that, also the coupling may be described with an algebraic constraint equation; that is, DAE-related aspects and properties also arise here. Typical examples for such situations are network modeling approaches in general and, in particular, modeling of coupled electric circuits and coupled substructures of mechanical multibody systems, see [11].

Co-simulation techniques and their theoretical background are studied for a long time, see, for instance, the survey papers [83, 143]. In these days, a new interface standard has developed, the 'Functional Mock-Up Interface (FMI) for Model-Exchange and Co-Simulation', (https://www.fmi-standard.org/). This interface is supported from more and more commercial CAE-software tools and finds more and more interest in industry for application projects. Additionally, the development of that standard and its release has also stimulated new research activities concerning co-simulation.

A coupled system of $r \geq 2$ fully implicit DAEs initial value problems reads as

$$0 = F(t, z_i(t), z_i'(t), u_i(t)) = 0,\ t \in [t_0, t_f],\ z_i(t_0) = z_{i,0},\quad i = 1, \ldots r \tag{5.1}$$

with initial values assumed to be consistent and the *(subsystem-) outputs*

$$\xi_i(t) := \Xi_i(t, z_i(t), u_i(t)),$$

and the *(subsystem-) inputs* u_i that are given by *coupling conditions*

$$u_i(t) = h_i(\xi_1, \ldots, \xi_r),\ i = 1, 2, \ldots,\ \text{i.e.}, u = h(\xi),$$

where we have set $u := (u_1^\top, \ldots, u_r^\top)^\top$, $\xi := (\xi_1^\top, \ldots, \xi_r^\top)^\top$ and $h := (h_1^\top, \ldots, h_r^\top)^\top : \mathbb{R}^{n_y} \to \mathbb{R}^{n_u}$, with $n_y = n_{y_1} + \ldots + n_{y_r}$, $n_{u_1} + \ldots + n_{u_r} = n_u$. Moreover, we assume here

$$\frac{\partial h_i}{\partial \xi_i} = 0,$$

that is, the inputs of system i do not depend on his own output. If the subsystems DAEs are in semi-explicit form, Eq. (5.1) has to be replaced by

$$\begin{aligned} \dot{x}_i(t) &= f_i(t, x_i(t), y_i(t), u_i(t)), \\ 0 &= g_i(t, x_i(t), y_i(t), u_i(t)), \end{aligned}$$

with $t \in [t_0, t_f]$ and $(x_i(t_0), y_i(t_0)) = (x_{i,0}, y_{i,0})$ with consistent initial values. This representation is called *block-oriented*; it describes the subsystems as blocks with inputs and outputs that are coupled.

In principle, it is possible to set up one monolithic system including the coupling conditions and output equations as additional algebraic equations:

$$\begin{aligned} \dot{x}_i &= f_i(t, x_i(t), y_i(t), u_i(t)), \\ 0 &= g_i(t, x_i(t), y_i(t), u_i(t)), \\ 0 &= u_i(t) - h(\xi(t)), \\ 0 &= \xi_i(t) - \Xi_i(t, x_i(t), u_i(t)), \quad i = 1, \ldots, r. \end{aligned}$$

This entire system could be solved with one single integration scheme, which is, however, as indicated above typically not advisable. In contrast, in co-simulation strategies, also referred to as modular time-integration [125] or distributed time integration [11], the subsystem equations are solved separately on consecutive time-windows. Herein, the time integration of each subsystem within one time-window or macro step can be realized with a different step-size adapted to the subsystem (*multirate* approach), or even with different appropriate integration schemes (*multimethod* approach). During the integration process of one subsystem, the needed coupling quantities, i.e., inputs from other subsystems, are approximated—usually based on previous results. At the end of each macro step, coupling data is exchanged. To be more precise, for the considered time interval, we introduce a (macro) time grid $\mathbb{G} := \{T_0, \ldots, T_N\}$ with $t_0 = T_0 < T_1 < \ldots < T_N = t_f$. Then, the mentioned time-windows or macro steps are given by $[T_n, T_{n+1}]$, $n = 0, \ldots, N-1$ and each subsystem is integrated independently from the others in each macro step $T_n \to T_{n+1}$, only using a typically limited number of coupling quantities as information from the other subsystems. The macro time points T_n are also called communication points, since here, typically, coupling data is exchanged between the subsystems.

5.1 Jacobi, Gauss-Seidel, and Dynamic-Iteration Schemes

An overview on co-simulation schemes and strategies can be found, e.g., in [11, 104, 125]. There are, however, two main approaches how the above sketched co-simulation can be realized. The crucial differences are the strategy (order) how the subsystems are integrated within the macro steps and, accordingly, how coupling quantities are handled and approximated. The first possible approach is a completely *parallel* scheme and is called *Jacobi scheme* (or co-simulation/coupling of Jacobi-type). As the name indicates, the subsystems are integrated here in parallel and, thus, they have to use extrapolated input quantities during the current macro step, cf. Fig. 6. In contrast to this, the second approach is a *sequential* one, it is called *Gauss-Seidel scheme* (or co-simulation/coupling of Gauss-Seidel-type). For the special case of two coupled subsystems, $r = 2$, this looks as follows: one subsystem is integrated first on the current macro step using extrapolated input data yielding a (numerical) solution for this first system. Then, the second subsystem is integrated on the current macro step but, then, using already computed results from the first subsystem for the coupling quantities (since results from the first subsystem for the current macro step are available, in fact). The results from the first subsystem may be available on a fine micro time grid—within the macro step—or even as function of time, e.g., as dense output from the integration method; additionally, (polynomial) interpolation may also be used, cf. Fig. 6.

The sequential Gauss-Seidel scheme can be generalized straightforwardly to $r > 2$ coupled subsystems: The procedure is sequential, i.e., the subsystems are numerically integrated one after another and for the integration of the i-th subsystem results from the subsystems $1, \dots, i-1$ are available for the coupling quantities, whereas data from chronologically upcoming subsystems $i+1, \dots, r$ have to be extrapolated based on information from previous communication points.

The extra- and interpolation, respectively, are realized using data from previous communication points and, typically, polynomial extra- and interpolation approaches are taken. That is, in the macro step $T_n \to T_{n+1}$, the input of subsystem i is extrapolated using data from the communication points $T_{n-k}, \dots, T_n$,

$$\tilde{u}_i(t) = \Psi_i(t; u_i(T_{n-k}), \dots, u_i(T_n)) = \sum_{j=0}^{k} u_i(T_{n-j}) \prod_{l=0, l\neq j}^{k} \frac{t - T_{n-l}}{T_{n-j} - T_{n-l}},$$

$t \in [T_n, T_{n+1}]$ and with the extrapolation polynomial Ψ_i with degree $\leq k$; for interpolation, e.g., for Gauss-Seidel schemes, we have correspondingly

$$\tilde{u}_i(t) = \Psi_i(t; u_i(T_{n-k}), \dots, u_i(T_{n+1})) = \sum_{j=0}^{k+1} u_i(T_{n+1-j}) \prod_{l=0, l\neq j}^{k+1} \frac{t - T_{n+1-l}}{T_{n+1-j} - T_{n+1-l}}.$$

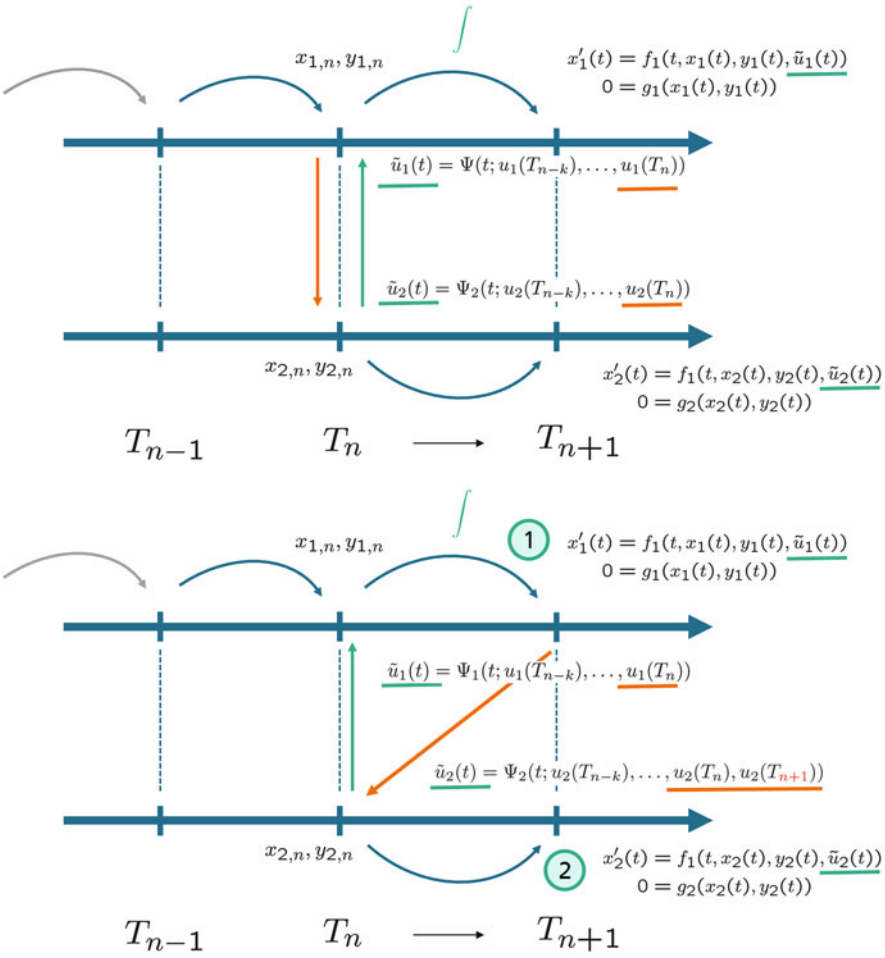


Fig. 6 Jacobi (*upper diagram*) and Gauss-Seidel (*lower diagram*) co-simulation schemes

The most simple extrapolation is that of zero-order, $k = 0$, leading to 'frozen' coupling quantities

$$u_i(t) = u_i(T_n), \quad t \in [T_n, T_{n+1}].$$

A third approach to establish a simulation of coupled systems are the so-called *dynamic iteration schemes*, [11, 20, 21], also referred to as waveform relaxation methods, [82, 94]. Here, the basic idea is to solve the subsystems iteratively on each macro step using coupling data information from previous iteration steps, in order to decrease simulation errors. How the subsystems are solved in each iteration step can be in a sequential fashion (Gauss-Seidel) or all in parallel (Jacobi or Picard),

cf. [11, 20]. The schemes defined above are contained in a corresponding dynamic iteration scheme by performing exactly one iteration step.

5.2 Stability and Convergence

First of all, we point out that there is a decisive difference between convergence and stability issues for coupled ODEs on the one hand and for coupled DAEs on the other hand. The stability problems that may appear for coupled ODEs with stiff coupling terms resemble the potential problems when applying an explicit integration method to stiff ODEs—thus, these difficulties can be avoided by using sufficiently small macro step-sizes $H_n = T_{n+1} - T_n$, cf. [9, 11, 104]. In the DAE-case, however, reducing the macro steps does not generally lead to an improvement; here, it is additionally essential that a certain contractivity condition is satisfied, see [9, 11, 21, 125].

5.2.1 The ODE-Case

For problems with coupled ODEs, convergence is studied, e.g., in [8, 10, 16, 17]. For coupled ODEs systems that are free of algebraic loops—this is guaranteed, for instance, provided that there is no direct feed-through, i.e., $\partial \Xi_i / \partial u_i = 0, \quad i = 1, \ldots, r$, for a precise definition see [10, 16]—we have the following global error estimation for a co-simulation with a Jacobi scheme with constant macro step-size $H > 0$ assumed to be sufficiently small,

$$\varepsilon^x \leq C \left(\sum_{i=1}^{r} \varepsilon_i^x + H^{k+1} \right), \tag{5.2}$$

where k denotes the order of the extrapolation and ε_i^x is the global error in subsystem i and ε^x is the overall global error, cf. [8, 10]. That is, the errors from the subsystems contribute to the global error, as well as the error from extra-(inter-)polation, $\mathcal{O}(H^{k+1})$. These results can be straightforwardly deduced following classical convergence analysis for ODE time integration schemes.

5.2.2 The DAE-Case

For detailed analysis and both convergence and stability results for coupled DAE systems, we refer the reader to [9, 11, 20, 125] and the literature cited therein. In the sequel we summarize and sketch some aspects from these research papers.

As already said, in the DAE case, the situation becomes more difficult. Following the lines of [20], we consider the following coupled DAE-IVP representation

$$x_i'(t) = f_i(x(t), y(t)), \tag{5.3}$$

$$0 = g_i(x(t), y(t)), \quad i = 1, \ldots, r, \tag{5.4}$$

with $x = (x_1^\top, \ldots, x_r^\top)^\top$, $y = (y_1^\top, \ldots, y_r^\top)^\top$ and initial conditions $(x_i(t_0), y_i(t_0)) = (x_{i,0}, y_{i,0})$, $i = 1, \ldots, r$. For the following considerations, we assume that the IVP(s) possess a unique global solution and that the right-hand side functions f_i, g_i are sufficiently often continuously differentiable and, moreover, that it holds

$$\frac{\partial g_i}{\partial y_i}$$

is non-singular for $i = 1, \ldots, r$ in a neighborhood of a solution (index-one condition for each subsystem). Notice that this representation differs from the previously stated block-oriented form. Equations (5.3)–(5.4) are, however, more convenient, in order to derive and to state the mentioned stability conditions, the coupling here is realized by the fact that all right-hand side functions f_i, g_i of each subsystem do depend on the entire differential and algebraic variables.

As before, we denote by a $\tilde{\cdot}$ quantities that are only available as extra- or interpolated quantity. Thus, establishing a co-simulation scheme of Jacobi-type yields for the i-th subsystem in macro step $T_n \to T_{n+1}$

$$\begin{aligned} x_{i,n}' = f_i(&\tilde{x}_{1,n}, \ldots, \tilde{x}_{i-1,n}, x_{i,n}, \tilde{x}_{i+1,n}, \ldots, \tilde{x}_{r,n},\\ &\tilde{y}_{1,n}, \ldots, \tilde{y}_{i-1,n}, y_{i,n}, \tilde{y}_{i+1,n}, \ldots, \tilde{y}_{r,n}),\\ 0 = g_i(&\tilde{x}_{1,n}, \ldots, \tilde{x}_{i-1,n}, x_{i,n}, \tilde{x}_{i+1,n}, \ldots, \tilde{x}_{r,n},\\ &\tilde{y}_{1,n}, \ldots, \tilde{y}_{i-1,n}, y_{i,n}, \tilde{y}_{i+1,n}, \ldots, \tilde{y}_{r,n}). \end{aligned}$$

Accordingly, for a Gauss-Seidel-type scheme, we obtain

$$\begin{aligned} x_{i,n}' = f_i(&x_{1,n}, \ldots, x_{i,n}, \tilde{x}_{i+1,n}, \ldots, \tilde{x}_{r,n},\\ &y_{1,n}, \ldots, y_{i,n}, \tilde{y}_{i+1,n}, \ldots, \tilde{y}_{r,n}),\\ 0 = g_i(&x_{1,n}, \ldots, x_{i,n}, \tilde{x}_{i+1,n}, \ldots, \tilde{x}_{r,n},\\ &y_{1,n}, \ldots, y_{i,n}, \tilde{y}_{i+1,n}, \ldots, \tilde{y}_{r,n}). \end{aligned}$$

With $g = (g_1, \ldots, g_r)^\top$, a sufficient (not generally necessary) *contractivity condition for stability* is derived and proven in [20]. The condition is given by

$$\alpha := \|g_y^{-1} g_{\tilde{y}}\| < 1,$$

with $\tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_r)$. For a detailed list of requirements and assumptions to be taken as well as for a proof and consequences, the reader is referred to [20, 21]. For the special case $r = 2$, the above condition leads to the following for the Jacobi-type scheme:

$$\alpha = \left\| \begin{pmatrix} 0 & g_{1,y_1}^{-1} g_{1,y_2} \\ g_{2,y_2}^{-1} g_{2,y_1} & 0 \end{pmatrix} \right\| < 1,$$

whereas, for the Gauss-Seidel-type scheme, we obtain

$$\alpha = \left\| \begin{pmatrix} g_{1,y_1} & 0 \\ g_{2,y_1} & g_{2,y_2} \end{pmatrix}^{-1} \begin{pmatrix} 0 & g_{1,y_2} \\ 0 & 0 \end{pmatrix} \right\| < 1.$$

An immediate consequence is that for a Jacobi-scheme of two coupled DAEs with no coupling in the algebraic equation, i.e., $g_{i,y_j} = 0$, for $i \neq j$, we have $\alpha = 0$.

As a further example, we discuss two mechanical multibody systems coupled via a kinematic constraint:

$$\begin{aligned} M_i(q_i) q_i'' &= \psi_i(q_i, q_i') - G_i^\top(q)\lambda, \quad i = 1, 2, \\ 0 &= \gamma(q), \end{aligned}$$

with $q = (q_1^\top, q_2^\top)^\top$ and $G(q) := \partial\gamma/\partial q$ and $G_i(q) := \partial\gamma/\partial q_i$, $i = 1, 2$. Performing an index reduction by twice differentiating the coupling constraint and setting $v_i := q_i'$, $a_i := v_i'$ as well as $x_i := (q_i^\top, v_i^\top)^\top$ and $y_1 = a_1^\top$, $y_2 := (a_2^\top, \lambda^\top)^\top$, $f_i := (v_i^\top, a_i^\top)^\top$, we are in the previously stated general framework:

$$\begin{aligned} x_1' &= f_1 & x_2' &= f_2 \\ 0 &= M_1 a_1 - \psi_1 + G_1^\top \lambda & 0 &= \begin{bmatrix} M_2 a_2 - \psi_2 + G_2^\top \lambda \\ G_1 a_1 + G_2 a_2 + \gamma^{(II)} \end{bmatrix} \\ &=: g_1(x_1, x_2, y_1, y_2) & &=: g_2(x_1, x_2, y_1, y_2) \end{aligned} .$$

Herein, $\gamma^{(II)}$ contains the remainder of the second derivative of γ without the term $G_1 a_1 + G_2 a_2$.

That is, the only coupling is via algebraic variables and in algebraic equations. If we set up a Jacobi-scheme, in macro step $T_n \to T_{n+1}$, in subsystem 1, we have to use extrapolated values from subsystem 2, i.e., y_2 is replaced by

$$\tilde{y}_2(t) = \Psi_1^y(t; y_2(T_{n-k}), \ldots, y_2(T_n))$$

and in subsystem 1, accordingly, $\tilde{y}_1(t) = \Psi_2^y(t; y_1(T_{n-k}), \ldots, y_1(T_n))$. The above contractivity condition in this case reads

$$\left\| \begin{pmatrix} 0 & 0 & M_1^{-1} G_1^\top \\ M_2^{-1} G_2^\top R_2^{-1} G_1 & 0 & 0 \\ -R_2^{-1} G_1 & 0 & 0 \end{pmatrix} \right\| < 1,$$

with $R_i := G_i M_i^{-1} G_i^\top$, $i = 1, 2$.

Analogously, we can consider a Gauss-Seidel-type scheme. Starting with subsystem 1, we have to extrapolate here y_2 from previous macro steps yielding x_1, y_1, which then can be evaluated during time-integration of subsystem 2. Stating the contractivity condition for this case and noticing that only the algebraic variable λ has to be extrapolated from previous time points, the relevant (λ-)part of the condition requires

$$\|R_2^{-1} R_1\| = \|(G_2 M_2 G_2^\top)^{-1} (G_1 M_1^{-1} G_1^\top)\| < 1. \tag{5.5}$$

We observe in both cases that mass and inertia properties of the coupled systems may strongly influence the stability of the co-simulation. In particular for the latter sequential Gauss-Seidel scheme, the order of integration has an essential impact on stability, i.e., the choice of system 1 and 2, respectively, should be taken such that the left-hand side of (5.5) is as small as possible.

This result has been developed and proven earlier in [11] for a more general framework, which is slightly different than our setup and for which the coupled mechanical systems are also a special case. In that paper, a method for stabilization (reducing α) is suggested. In [125], the authors also study stability and convergence of coupled DAE systems in a rather general framework and propose a strategy for stabilization as well.

For the specific application field of electric circuit simulation, the reader is referred to [20, 21] and the references therein. A specific consideration of coupled mechanical multibody systems is provided in [8, 9] and in [126], where the coupling of a multibody system and a flexible structure is investigated and an innovative coupling strategy is proposed. Lately, analysis results on coupled DAE systems solved with different co-simulation strategies and stabilization approaches are provided by the authors of [129, 130]. In [19], a multibody system model of a wheel-loader described as index-three DAE in a commercial software package is coupled with a particle code for soft-soil modeling, in order to establish a coupled digging simulation.

The general topic of coupled DAE system is additionally discussed in the early papers [82, 89, 94].

A multirate integrator for constrained dynamical systems is derived in [96], which is based on a discrete variational principle. The resulting integrator is symplectic and momentum preserving.

6 Real-Time Simulation

An important field in modern numerical system simulation is *real-time scenarios*. Here, a numerical model is coupled with the real world and both are interacting dynamically. A typical area, in which such couplings are employed, is interactive simulators ('human/man-in-the-loop'), such as driving simulators or flight simulators, see [58], but also interactively used software (simulators), e.g., for training purposes, cf. [98]. Apart from that, real-time couplings are used in tests for electronic control units (ECU tests) and devices ('hardware-in-the-loop'—HiL), see, e.g., [15, 122] and in the field of model based controllers ('model/software-in-the-loop'—MiL/SiL), see, e.g., [42, 43].

It is characteristic for all the mentioned fields that a numerical model *replaces* a part of the real world. In case of an automotive control unit test, the real control unit hardware is coupled with a numerical model of the rest of the considered vehicle; in case of an interactive driving simulator, the simulator hardware and, by that, the driver or the operator, respectively, is also coupled with a virtual vehicle. The benefits of such couplings are tremendous—tests and studies can be performed under fully accessible and reproducible conditions in the laboratory. Investigations and test runs with real cars and drivers can be reduced and partially avoided, which can save time, costs, and effort substantially. From the perspective of the numerical model, it receives from the real world environment signals as inputs (e.g., the steering-wheel angle from human driver in a simulator) and gives back its dynamical behavior as output (e.g., the car's reaction is transmitted to the simulator hardware, which, in turn, follows that motion making the driver feel as he would sit in a real car). It is crucial for a realistic realization of such a coupling that the simulation as well as the communication are sufficiently fast. That is, after delivering an input to the numerical model, the real world component expects a response after a fixed time ΔT—and the numerical model has to be simulated for that time span and has to feed back the response *on time*. Necessary for that is that the considered numerical simulation satisfies the *real-time condition*: the computation (or simulation) time ΔT_{comp} has to be smaller or equal than the simulated time ΔT.

Physical models are often described as differential equations (mechanical multibody systems that represent a vehicle model). Satisfying the real-time condition here means accordingly that the numerical time integration of the IVP

$$\begin{aligned} F(t, z(t), z'(t), u(t)) &= 0, \quad t \in [T_i; T_i + \Delta T] \\ z(T_i) &= z_{0,i}, \end{aligned}$$

is executed with a total computation time that is smaller or equal than ΔT. If a complete real-time simulation shall be run on a time horizon $[t_0; t_f]$ which is divided by an equidistant time-grid $\{T_0, \ldots, T_N\}$, $t_0 = T_0$, $t_f = T_N$, $T_{i+1} - T_i = \Delta T$, the real-time condition must be guaranteed for any subinterval of length ΔT. In fact, this is a coupling exactly as in classical co-simulation—with the decisive difference that one partner is not a numerical model, but a real world component and, thus,

the numerical model simulation must satisfy the real-time condition. Obviously, whether or not the real-time condition can be satisfied, strongly depends both on the numerical time integration method and the differential equation and its properties itself. In principle, any time integration method can be applied, provided that the resulting simulation satisfies the real-time condition.

The fulfillment of the real-time condition as stated above has, however, to be assured *deterministically* in each macro time step $T_i \rightarrow T_i + \Delta T$—at least in applications, where breaking this condition leads to a critical system shutdown (e.g., hardware simulators, HiL-tests). Whence, the chosen integration methods should not have indeterministic elements like step-size control or iterative inner methods (solution of nonlinear systems by Newton-like methods): varying iteration numbers lead to a varying computation time. Consequently, for real-time application, time integration methods with fixed time-steps and with a fixed number of possible iterations are preferred. Additionally, to save computation time, typically, low-order methods are in use, which is also caused by the fact that in the mentioned application situations, the coupled simulation needs not to be necessarily highly accurate, but stable.

6.1 Real-Time Integration of DAEs

For non-stiff ODE models, which have to be simulated under real-time conditions, even the simple explicit Euler scheme is frequently used. For stiff ODEs, the linearly implicit methods as discussed in Sect. 4.3 are evident, since for these method class, only linear systems have to be solved internally, which leads to an a priori known, fixed, and moderate computational effort, see [14, 15, 49, 118] and the references therein.

Since all typical and work-proven DAE time integration methods are at least partially implicit leading to the need of iterative computations, it is a common approach to avoid DAE models for real-time applications already in the modeling process (generally, for real-time applications, often specific modeling techniques are applied), whenever it is possible. However, this is often impossible in many application cases of practical relevance. For instance, the above-mentioned examples from the automotive area require a mechanical vehicle model, which is usually realized as mechanical multibody system model, whose underlying equations of motion are often a DAE as stated in Eq. (2.13). Thus, there is a need for DAE time integration schemes that are stable and highly efficient also for DAEs of realistic complexities.

Time integration methods for DAEs with a special focus on real-time applications and the fulfillment of the real-time condition are addressed, e.g., in [14, 15, 31, 32, 39, 44, 49, 50, 119]. In the sequel, we present a specific integration method for the MBS equations of motion (2.13) in its index-two formulation on velocity-level.

For the special case of the semi-explicit DAE describing a mechanical multibody system, compare (2.13), the following linearly implicit method can be applied, which is based on the linearly implicit Euler scheme. The first step is to reduce

the index from three to two by replacing the original algebraic equations by its first time-derivative,

$$G(q)v = 0,$$

which is linear in v. The numerical scheme proposed in [14, 31] consists in handling the time-step for the position coordinates explicitly and requiring that the algebraic equation on velocity level is satisfied, i.e.,

$$G(q_{i+1})v_{i+1} = 0.$$

In particular, this leads to the set of linear equations as follows

$$\begin{aligned} q_{i+1} &= q_i + h_i v_i, \\ \begin{pmatrix} M - hJ_v - h^2 J_v & G^\top(q_i) \\ G(q_{i+1}) & 0 \end{pmatrix} \begin{pmatrix} v_{i+1} - v_i \\ h\lambda_i \end{pmatrix} &= \begin{pmatrix} hf_i + h^2 J_q v_i \\ -G(q_{i+1})v_i \end{pmatrix}, \end{aligned}$$

where $J_{q/v} := \partial f/\partial(q/v)(q_i, v_i)$.

An important issue is naturally the *drift-off*, cf. Sect. 2, in the neglected algebraic constraints—here, in the above method for the index-two version of the MBS DAE, the error in the algebraic equation on position-level, i.e., $0 = g(q)$, may grow linearly in time; this effect is even more severe, since a low-order method is in use. Classical strategies to stabilize this drifting are projection approaches, cf., e.g., [73, 100], which are usually of adaptive and iterative character. The authors in [14, 31] propose and discuss a non-iterative projection strategy, which consists, in fact, in one special Newton-step for the KKT conditions related to the constrained optimization problem that is used for projection; thus, only one additional linear equation has to be solved in each time-step. The authors show that using this technique leads to a bound for the error on position level, which is independent of time. An alternative way to stabilize the drift-off effect without substantially increasing the computational effort is the Baumgarte stabilization, cf. Sect. 2 and [31, 48, 122].

7 Parametric Sensitivity Analysis and Adjoints

The parametric sensitivity analysis is concerned with parametric initial value problems subject to DAEs on the interval $[t_0, t_f]$ given by

$$F(t, z(t), z'(t), p) = 0, \tag{7.1}$$

$$z(t_0) = z_0(p), \tag{7.2}$$

where $p \in \mathbb{R}^m$ is a parameter vector and the mapping $z_0 : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is at least continuously differentiable. We assume that the initial value problem possesses a solution for every p in some neighborhood of a given nominal parameter $\hat{p}$ and denote the solution by $z(t; p)$. In order to quantify the influence of the parameter on the solution, we are interested in the so-called *sensitivities* (*sensitivity matrices*)

$$S(t) := \frac{\partial z}{\partial p}(t; \hat{p}) \qquad \text{for } t \in [t_0, t_f]. \tag{7.3}$$

Throughout we tacitly assume that the sensitivities actually exist.

In many applications, e.g., from optimal control or optimization problems involving DAEs, one is not directly interested in the sensitivities $S(\cdot)$ themselves but in the gradient of some function $g : \mathbb{R}^m \longrightarrow \mathbb{R}$ defined by

$$g(p) := \varphi(z(t_f; p), p), \tag{7.4}$$

where $\varphi : \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}$ is continuously differentiable. Of course, if the sensitivities $S(\cdot)$ are available, the gradient of g at $\hat{p}$ can easily be computed by the chain rule as

$$\nabla g(\hat{p}) = S(t_f)^\top \nabla_z \varphi(z(t_f; \hat{p}), \hat{p}) + \nabla_p \varphi(z(t_f; \hat{p}), \hat{p}). \tag{7.5}$$

However, often the explicit computation of S is costly and should be avoided. Then the question for alternative representations of the gradient $\nabla g(\hat{p})$ arises, which avoids the explicit computation of S. This alternative representation can be derived using an adjoint DAE. Both approaches are analytical in the sense that they provide the correct gradient, if round-off errors are not taken into account.

Remark 7.1 The computation of the gradient using S is often referred to as the *forward mode* and the computation using adjoints as the *backward* or *reverse mode* in the context of automatic differentiation, compare [72]. Using automatic differentiation is probably the most convenient way to compute the above gradient, since powerful tools are available, see the web-page familywww.autodiff.org.

The same kind of sensitivity investigations can be performed either for the problem (7.1)–(7.2) in continuous time or for discretizations thereof by means of one-step or multi-step methods.

7.1 Sensitivity Analysis in Discrete Time

7.1.1 The Forward Mode

Suppose a suitable discretization scheme of (7.1)–(7.2) is given, which provides approximations $z_h(t_i; p)$ at the grid points $t_i \in \mathbb{G}_h$ in dependence on the parameter

p. We are interested in the sensitivities

$$S_h(t_i) := \frac{\partial z_h}{\partial p}(t_i;\hat{p}) \in \mathbb{R}^{n\times m} \qquad \text{for } t_i \in \mathbb{G}_h$$

for a nominal parameter $\hat{p} \in \mathbb{R}^m$. As the computations are performed on a finite grid, these sensitivities can be obtained by differentiating the discretization scheme with respect to p. This procedure is called internal numerical differentiation (IND) and was introduced in [25].

To be more specific, let $\hat{p}$ be a given nominal parameter and consider the one-step method

$$z_h(t_0;\hat{p}) = z_0(\hat{p}), \tag{7.6}$$

$$z_h(t_{i+1};\hat{p}) = z_h(t_i;\hat{p}) + h_i\Phi(t_i, z_h(t_i;\hat{p}), h_i, \hat{p}), \quad i = 0, 1, \ldots, N-1. \tag{7.7}$$

Differentiating both equations with respect to p and evaluating the equations at $\hat{p}$ yields

$$S_h(t_0) = z_0'(\hat{p}), \tag{7.8}$$

$$S_h(t_{i+1}) = S_h(t_i) + h_i\left(\frac{\partial \Phi}{\partial z}[t_i]S_h(t_i) + \frac{\partial \Phi}{\partial p}[t_i]\right), \quad i = 0, 1, \ldots, N-1. \tag{7.9}$$

Herein, we used the abbreviation $[t_i]$ for $(t_i, z_h(t_i;\hat{p}), h_i, \hat{p})$. Evaluation of (7.8)–(7.9) yields the desired sensitivities $S_h(t_i)$ of $z_h(t_i;\hat{p})$ at the grid points, if the increment function Φ of the one-step method and the function z_0 are differentiable with respect to z and p, respectively. Note that the function z_0 can be realized by the projection method in LSQ(p) in Sect. 3.2 and sufficient conditions for its differentiability are provided by Theorem 3.1.

The computation of the partial derivatives of Φ is more involved. For a Runge–Kutta method Eqs. (4.7)–(4.9) (with an additional dependence on the parameter p) have to be differentiated with respect to z and p. Details can be found in [68, Sect. 5.3.2].

The same IND approach can be applied to multi-step methods. Differentiation of the scheme (4.2) and the consistent initial values

$$z_h(t_0;p) = z_0(p),\ z_h(t_1;p) = z_1(p),\ \ldots,\ z_h(t_{s-1};p) = z_{s-1}(p)$$

with respect to p and evaluation at $\hat{p}$ yields the formal scheme

$$S_h(t_\ell) = z_\ell'(\hat{p}), \qquad \ell = 0, \ldots, s-1,$$

$$S_h(t_{i+s}) = \sum_{\ell=0}^{s-1} \frac{\partial \Psi}{\partial z_{i+\ell}} \cdot S_h(t_{i+\ell}) + \frac{\partial \Psi}{\partial p}, \qquad i = 0, \ldots, N-s.$$

More specifically, for an s-stage BDF method the function ψ is implicitly given by (4.4) (with an additional dependence on the parameter p). Differentiation of (4.4) with respect to p yields

$$\left(\frac{\partial F}{\partial z}[t_{i+s}]+\frac{\alpha_s}{h_{i+s-1}}\frac{\partial F}{\partial z'}[t_{i+s}]\right)S_h(t_{i+s})+\sum_{\ell=0}^{s-1}\frac{\alpha_\ell}{h_{i+s-1}}\frac{\partial F}{\partial z'}[t_{i+s}]S_h(t_{i+\ell})+\frac{\partial F}{\partial p}[t_{i+s}]=0 \tag{7.10}$$

and, if the iteration matrix $M := F_z'[t_{i+s}] + \frac{\alpha_s}{h_{i+s-1}}F_{z'}'[t_{i+s}]$ is non-singular,

$$S_h(t_{i+s}) = -M^{-1}\cdot\left(\sum_{\ell=0}^{s-1}\frac{\alpha_\ell}{h_{i+s-1}}\frac{\partial F}{\partial z'}[t_{i+s}]S_h(t_{i+\ell})+\frac{\partial F}{\partial p}[t_{i+s}]\right).$$

Herein, we used the abbreviation $[t_{i+s}] = \left(t_{i+s}, z_h(t_{i+s}), \frac{1}{h_{i+s-1}}\sum_{k=0}^{s}\alpha_k z_h(t_{i+k})\right)$.

7.1.2 The Backward Mode and Adjoints

Consider the function g in (7.4) subject to a discretization scheme, i.e.

$$g_h(p) := \varphi(z_h(t_N;p),p). \tag{7.11}$$

We intend to compute the gradient of g_h at $\hat{p}$. Using the sensitivity $S_h(t_N)$ the gradient is given by

$$\nabla g_h(\hat{p}) = S_h(t_N)^\top \nabla_z\varphi(z_h(t_N;\hat{p}),\hat{p}) + \nabla_p\varphi(z_h(t_N;\hat{p}),\hat{p}).$$

Now we are interested in an alternative representation of the gradient without the sensitivity $S_h(t_N)$. To this end consider the one-step method in (7.6)–(7.7). Following [68, Sect. 5.3.2] define the auxiliary functional

$$g_h^a(p) := g_h(p) + \sum_{i=0}^{N-1}\lambda_h(t_{i+1})^\top\left(z_h(t_{i+1};p) - z_h(t_i;p) - h_i\Phi(t_i,z_h(t_i;p),h_i,p)\right)$$

with multipliers $\lambda_h(t_1),\ldots,\lambda_h(t_N)$ that will be specified later. Note that $g_h^a \equiv g_h$ for all discrete trajectories satisfying (7.6)–(7.7). The gradient of g_h^a at $\hat{p}$ computes to

$$\begin{aligned}\nabla g_h^a(\hat{p}) = {} & S_h(t_N)^\top\nabla_z\varphi(z_h(t_N;\hat{p}),\hat{p}) + \nabla_p\varphi(z_h(t_N;\hat{p}),\hat{p}) \\ & + \sum_{i=0}^{N-1}\left(S_h(t_{i+1}) - S_h(t_i) - h_i\frac{\partial\Phi}{\partial z}[t_i]S_h(t_i) - h_i\frac{\partial\Phi}{\partial p}[t_i]\right)^\top\lambda_h(t_{i+1})\end{aligned}$$

$$
\begin{aligned}
&= S_h(t_N)^\top \nabla_z \varphi(z_h(t_N;\hat{p}),\hat{p}) + \nabla_p \varphi(z_h(t_N;\hat{p}),\hat{p}) \\
&\quad + \sum_{i=1}^{N} S_h(t_i)^\top \lambda_h(t_i) - \sum_{i=0}^{N-1} \left(S_h(t_i) + h_i \frac{\partial \Phi}{\partial z}[t_i] S_h(t_i) \right)^\top \lambda_h(t_{i+1}) \\
&\quad - \sum_{i=0}^{N-1} h_i \frac{\partial \Phi}{\partial p}[t_i]^\top \lambda_h(t_{i+1}) \\
&= S_h(t_N)^\top \left(\lambda_h(t_N) + \nabla_z \varphi(z_h(t_N;\hat{p}),\hat{p}) \right) + \nabla_p \varphi(z_h(t_N;\hat{p}),\hat{p}) \\
&\quad + \sum_{i=1}^{N-1} S_h(t_i)^\top \left(\lambda_h(t_i) - \lambda_h(t_{i+1}) - h_i \frac{\partial \Phi}{\partial z}[t_i]^\top \lambda_h(t_{i+1}) \right) \\
&\quad - S_h(t_0)^\top \left(\lambda_h(t_1) + h_0 \frac{\partial \Phi}{\partial z}[t_0]^\top \lambda_h(t_1) \right) - \sum_{i=0}^{N-1} h_i \frac{\partial \Phi}{\partial p}[t_i]^\top \lambda_h(t_{i+1})
\end{aligned}
$$

In order to eliminate the sensitivities, we choose the multipliers λ_h such that they satisfy the *adjoint equations*

$$\lambda_h(t_N) = -\nabla_p \varphi(z_h(t_N;\hat{p}),\hat{p}), \tag{7.12}$$

$$\lambda_h(t_i) = \lambda_h(t_{i+1}) + h_i \frac{\partial \Phi}{\partial z}[t_i]^\top \lambda_h(t_{i+1}), \qquad i = 0,\ldots,N-1. \tag{7.13}$$

The adjoint equations have to be solved backwards in time starting at t_N. With this choice the gradient of g_h^a reduces to

$$\nabla g_h^a(\hat{p}) = \nabla_p \varphi(z_h(t_N;\hat{p}),\hat{p}) - S_h(t_0)^\top \lambda_h(t_0) - \sum_{i=0}^{N-1} h_i \frac{\partial \Phi}{\partial p}[t_i]^\top \lambda_h(t_{i+1})$$

width $S_h(t_0) = z_0'(\hat{p})$. Since g_h and g_h^a coincide for all discrete trajectories satisfying (7.6)–(7.7), the following theorem holds, see [68, Theorems 5.3.2, 5.3.3] for a proof:

Theorem 7.1 *We have*

$$\nabla g_h(\hat{p}) = \nabla g_h^a(\hat{p}) = \nabla_p \varphi(z_h(t_N;\hat{p}),\hat{p}) - S_h(t_0)^\top \lambda_h(t_0) - \sum_{i=0}^{N-1} h_i \frac{\partial \Phi}{\partial p}[t_i]^\top \lambda_h(t_{i+1}),$$

where $\lambda_h(\cdot)$ satisfies the adjoint Eqs. (7.12)–(7.13). Moreover, the combined discretization scheme (7.7) and (7.13) for z_h and λ_h is symplectic.

Remark 7.2 Computing the gradient of g_h via the adjoint approach is more efficient than using the sensitivities, because the adjoint equations do not depend on the dimension of p, whereas the sensitivity equations (7.8)–(7.9) are matrix difference equations for the $n \times m$-matrices $S_h(\cdot)$.

7.2 Sensitivity Analysis in Continuous Time

7.2.1 The Forward Mode

The IND approach is based on the differentiation of the discretization scheme. Applying the same idea to the parametric DAE in continuous time (7.1)–(7.2) yields the *sensitivity DAE*

$$\frac{\partial F}{\partial z}[t] \cdot S(t) + \frac{\partial F}{\partial z'}[t] \cdot S'(t) + \frac{\partial F}{\partial p}[t] = 0, \qquad t \in [t_0, t_f], \tag{7.14}$$

$$S(t_0) = z_0'(\hat{p}) \tag{7.15}$$

for the sensitivities $S(t)$ in (7.3). We used the abbreviation $[t] = (t, z(t;\hat{p}), z'(t;\hat{p}), \hat{p})$ and assumed that

$$S'(t) = \frac{\partial^2 z}{\partial p \partial t}(t;\hat{p}).$$

Note that the derivative $z_0'(\hat{p})$ can be obtained by a sensitivity analysis of the least-squares problem LSQ(p) in Sect. 3.2. Moreover, the sensitivity analysis in Theorem 3.1 provides a consistent initial value for the sensitivity DAE (7.14)–(7.15).

Now, the initial value problems for z and S in (7.1)–(7.2) and (7.14)–(7.15) can be solved simultaneously using some suitable one-step or multi-step method. Since efficient implementations often use approximate Jacobians, automatic step-size algorithms, or order selection strategies, the resulting numerical solutions $z_h(\cdot;\hat{p})$ and $S_h(\cdot)$ satisfy $S_h(\cdot) \approx \partial z_h / \partial p(\cdot;\hat{p})$ only up to some tolerance. As a result, the gradient of g in (7.5) will be accurate only in the range of a given integration tolerance. The forward approach using sensitivities is investigated in more detail, e.g., in [29, 37, 79, 87, 101] and a comparison is provided in [52].

A connection to the IND approach arises if the same discretization scheme and the same step-sizes for both DAEs are used. For the BDF method we obtain

$$F\left(t_{i+s}, z_h(t_{i+s}), \frac{1}{h_{i+s-1}} \sum_{\ell=0}^{s} \alpha_\ell z_h(t_{i+\ell}), \hat{p}\right) = 0$$

for $i = 0, \ldots, N-s$. Application of the same BDF method with the same step-sizes to the sensitivity DAE (7.14) yields

$$\frac{\partial F}{\partial z}[t_{i+s}] \cdot S_h(t_{i+s}) + \sum_{\ell=0}^{s} \frac{\alpha_\ell}{h_{i+s-1}} \frac{\partial F}{\partial z'}[t_{i+s}] \cdot S_h(t_{i+\ell}) + \frac{\partial F}{\partial p}[t_{i+s}] = 0,$$

for $i = 0, \ldots, N-s$. The latter coincides with the IND approach in (7.10). Hence, the discrete and continuous forward modes commute under discretization with the same method and the same step-sizes. The same is true for the Runge–Kutta method (4.7)–(4.10) applied to (7.1), i.e.,

$$z_h(t_{i+1}) = z_h(t_i) + h_i \sum_{j=1}^{s} b_j k_j(t_i, z_h(t_i), h_i, \hat{p}), \tag{7.16}$$

where $k_j(t_i, z_h(t_i), h_i, \hat{p})$, $j = 1, \ldots, s$, are implicitly defined by

$$F\left(t_i + c_\ell h_i, z_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j}\, k_j, k_\ell, \hat{p}\right) = 0, \qquad \ell = 1, \ldots, s. \tag{7.17}$$

Application of the same Runge–Kutta method with the same step-sizes to the sensitivity DAE (7.14) yields

$$S_h(t_{i+1}) = S_h(t_i) + h_i \sum_{j=1}^{s} b_j K_j,$$

where K_j, $j = 1, \ldots, s$, are implicitly given by the system of linear equations

$$\frac{\partial F}{\partial z}[t_i + c_\ell h_i]\left(S_h(t_i) + h_i \sum_{j=1}^{s} a_{\ell j} K_j\right) + \frac{\partial F}{\partial z'}[t_i + c_\ell h_i] \cdot K_\ell + \frac{\partial F}{\partial p}[t_i + c_\ell h_i] = 0$$

for $\ell = 1, \ldots, s$. With

$$K_j = \frac{\partial k_j}{\partial z}[t_i] S_h(t_i) + \frac{\partial k_j}{\partial p}[t_i], \qquad j = 1, \ldots, s,$$

the latter coincides with the IND approach for (7.16)–(7.17).

7.2.2 The Backward Mode and Adjoints

Consider the function g in (7.4), i.e., $g(p) = \varphi(z(t_f; p), p)$. Using the sensitivity $S(t_f)$ the gradient is given by

$$\nabla g(\hat{p}) = S(t_f)^\top \nabla_z \varphi(z(t_f; \hat{p}), \hat{p}) + \nabla_p \varphi(z(t_f; \hat{p}), \hat{p}).$$

As in the discrete case we are interested in an alternative representation of the gradient without the sensitivity $S(t_f)$. To this end we define the auxiliary functional

$$g^a(p) := g(p) + \int_{t_0}^{t_f} \lambda(t)^\top F(t, z(t; p), z'(t; p), p) dt,$$

where λ is a suitable function to be defined later. Differentiation with respect to p, evaluation at $\hat{p}$, and integration by parts yield

$$\begin{aligned}
\nabla g^a(\hat{p}) &= S(t_f)^\top \nabla_z \varphi(z(t_f; \hat{p}), \hat{p}) + \nabla_p \varphi(z(t_f; \hat{p}), \hat{p}) \\
&\quad + \int_{t_0}^{t_f} \left(F'_z[t] \cdot S(t) + F'_{z'}[t] \cdot S'(t) + F'_p[t] \right)^\top \lambda(t) dt \\
&= S(t_f)^\top \left(F'_{z'}[t_f]^\top \lambda(t_f) + \nabla_z \varphi(z(t_f; \hat{p}), \hat{p}) \right) - S(t_0)^\top F'_{z'}[t_0]^\top \lambda(t_0) \\
&\quad + \nabla_p \varphi(z(t_f; \hat{p}), \hat{p}) + \int_{t_0}^{t_f} F'_p[t]^\top \lambda(t) dt \\
&\quad + \int_{t_0}^{t_f} S(t)^\top \left(F'_z[t]^\top \lambda(t) - \frac{d}{dt} \left(F'_{z'}[t]^\top \lambda(t) \right) \right) dt.
\end{aligned}$$

Since we like to avoid the sensitivities $S(t)$ and $S(t_f)$ we define the *adjoint DAE*

$$F'_{z'}[t_f]^\top \lambda(t_f) + \nabla_z \varphi(z(t_f; \hat{p}), \hat{p}) = 0, \tag{7.18}$$

$$F'_z[t]^\top \lambda(t) - \frac{d}{dt} \left(F'_{z'}[t]^\top \lambda(t) \right) = 0. \tag{7.19}$$

Please note that this derivation is a formal derivation only and it is not clear whether the adjoint DAE (7.18)–(7.19) actually possesses a solution. In fact, it may not have a solution in general. The existence and stability of solutions of the adjoint DAE subject to structural assumptions were investigated in [36]. Details can be found in [68, Sect. 5.3.3] as well.

If the adjoint DAE possesses a solution, then the gradient of g^a is represented by

$$\nabla g^a(\hat{p}) = -S(t_0)^\top F'_{z'}[t_0]^\top \lambda(t_0) + \nabla_p \varphi(z(t_f;\hat{p}),\hat{p}) + \int_{t_0}^{t_f} F'_p[t]^\top \lambda(t)dt$$

with $S(t_0) = z'_0(\hat{p})$ and as in the discrete case it coincides with $\nabla g(\hat{p})$, compare [68, Sect. 5.3.3].

Remark 7.3 Solving the DAE (7.1)–(7.2) and (7.18)–(7.19) simultaneously by some suitable one-step or multi-step method in general does not commute with the discrete adjoint approach.

7.3 Example

Example 7.1 is concerned with a trolley moving on a surface, which leads to an index-three DAE. Herein, a parametric sensitivity analysis is performed and the sensitivity of the states w.r.t. to some parameters is computed using the forward mode.

Example 7.1 Consider the motion of a trolley of mass m_1 on a one-dimensional surface described by the function $h(x)$, which is supposed to be at least twice continuously differentiable, see Fig. 7.

Let a load of mass m_2 be attached to the trolley's center of gravity with a massless rod of length $\ell > 0$. The equations of motion are given by the following index-

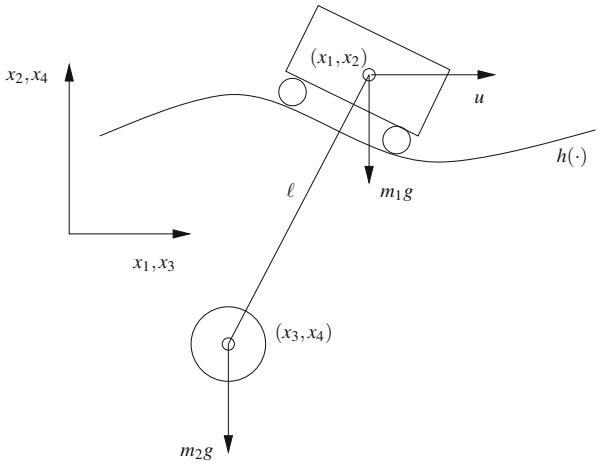


Fig. 7 Configuration of the trolley

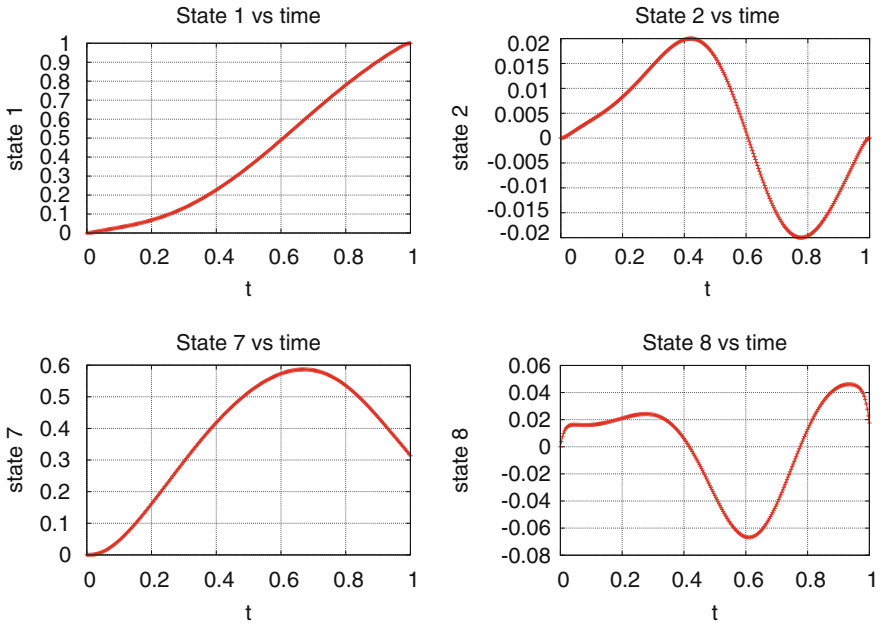


Fig. 8 Positions of trolley (*top*) and velocities of load (*bottom*) (normalized time interval $[0, 1]$)

three DAE:

$$
\begin{aligned}
m_1 x_1''(t) &= u(t) - 2\lambda_1(t)(x_1(t) - x_3(t)) + \lambda_2(t)h'(x_1(t)),\\
m_1 x_2''(t) &= -m_1 g - 2\lambda_1(t)(x_2(t) - x_4(t)) - \lambda_2(t),\\
m_2 x_3''(t) &= 2\lambda_1(t)(x_1(t) - x_3(t)),\\
m_2 x_4''(t) &= -m_2 g + 2\lambda_1(t)(x_2(t) - x_4(t)),\\
0 &= (x_1(t) - x_3(t))^2 + (x_2(t) - x_4(t))^2 - \ell^2,\\
0 &= x_2(t) - h(x_1(t)).
\end{aligned}
$$

Herein, (x_1, x_2) denotes the trolley's center of gravity, (x_3, x_4) the load's position, λ_1, λ_2 the algebraic variables, and $u(t)$ a given control input.

Figures 8 shows the results of a simulation using the software OCPID-DAE1, see http://www.optimal-control.de, on the interval $[0, 2.79]$ (scaled to the normalized interval $[0, 1]$) with $m_1 = 0.3$, $m_2 = 0.5$, $\ell = 0.75$, $g = 9.81$, and $h(x) = 0.02\sin(2\pi x)$. Figure 9 shows the control input u and the algebraic variables λ_1 and λ_2. The computations were performed for the GGL-stabilized system.
Figure 10 shows the sensitivities of some states w.r.t. to m_1.

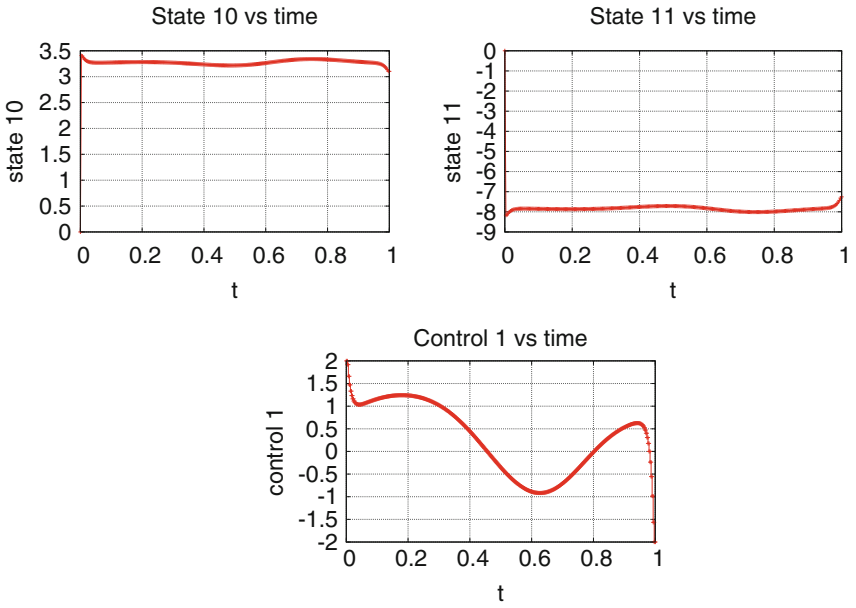


Fig. 9 Algebraic variables (λ_1, λ_2) (*top*) and control input u (*bottom*) (normalized time interval $[0, 1]$)

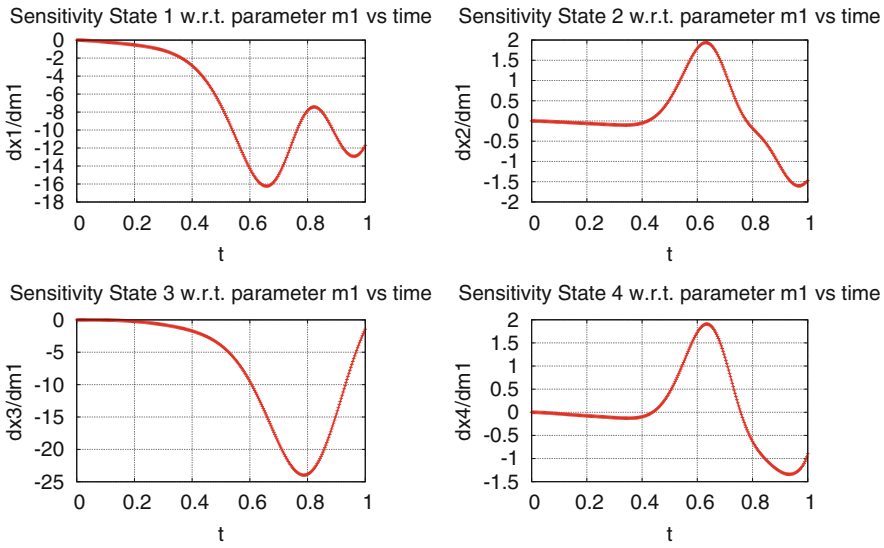


Fig. 10 Sensitivities of positions of trolley (*top*) and load (*bottom*) w.r.t. to m_1 (normalized time interval $[0, 1]$)

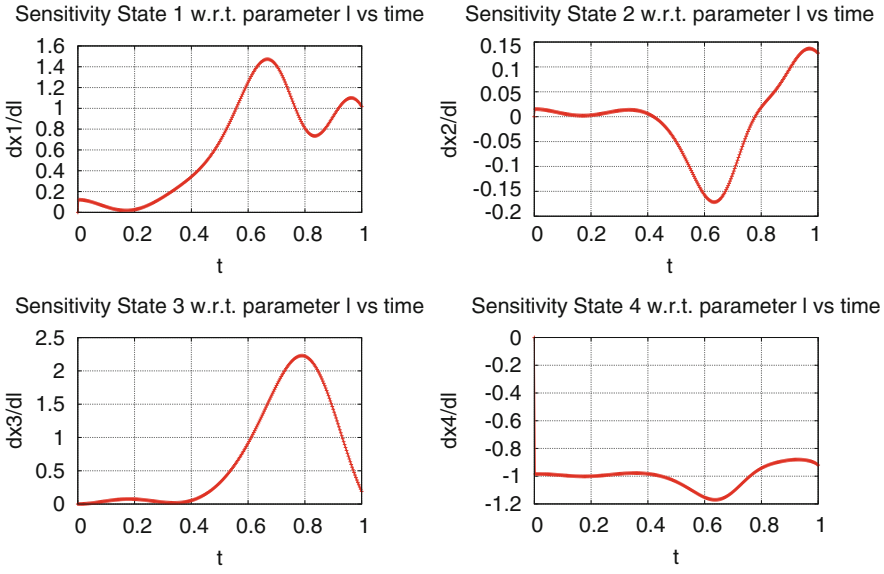


Fig. 11 Sensitivities of positions of trolley (*top*) and load (*bottom*) w.r.t. to ℓ (normalized time interval [0, 1])

Figure 11 shows the sensitivities of some states w.r.t. to ℓ.

8 Switched Systems and Contact Problems

Many applications lead to DAE models with piecewise defined dynamics. Herein, the different DAE models are only valid in defined regions of the state space. Those regions are separated and bounded by manifolds, which are typically implicitly defined by state-dependent switching functions. A transition from one region (i.e., one DAE) to another (with another DAE) occurs, if the switching function changes its sign, i.e., the switching function indicates a switch in the dynamic system. Moreover, a transition from one region to another may come along with a discontinuity of some state components. For instance, contact and friction forces acting between two or more colliding rigid bodies typically lead to discontinuities in the velocity components of the state of a mechanical multibody system.

More general classes of switched DAEs and the existence and stability of solutions are discussed in [97]. The controllability of switched DAEs is investigated in [91]. Hybrid optimal control problems and necessary conditions can be found in [59, 133, 136].

8.1 Hybrid Systems and Switching Functions

It is convenient to view the dynamic process as a hybrid system, compare [114, 142]. To this end, the status of the system is characterized by a finite set of modes $M = \{1, \ldots, P\}$. In mode $m \in M$, the state evolves according to the DAE

$$\begin{aligned} x'(t) &= f^m(x(t), y(t)), \\ 0 &= g^m(x(t), y(t)). \end{aligned}$$

The system remains in mode m as long as the trajectory $z(t) = (x(t), y(t))^\top$ stays within the set

$$\mathscr{S}^m := \{x \in X \mid s^m(x) \geq 0\},$$

where for each $m \in M$, $s^m : \mathbb{R}^n \longrightarrow \mathbb{R}$ is called *switching function of mode m*. For simplicity we exclude vector-valued switching functions in order to avoid situations with multiple active switching functions, which are difficult to resolve. $Z = X \times Y \subseteq \mathbb{R}^n \times \mathbb{R}^m$ defines the space of possible differential and algebraic states.

A transition from mode m to another mode $\tilde{m}$ becomes possible only in the event that x is about to cross the boundary of $\mathscr{S}^m$ at some time point $\hat{t}$, i.e., if $s^m(x(\hat{t}^-)) = 0$ and $s^m(x(t)) < 0$ for some $t > \hat{t}$ provided the process would be continued with the dynamics of mode m. Herein, $x(\hat{t}^\pm)$ denote the left- and right-sided limits of x at $\hat{t}$, respectively. The time point $\hat{t}$ in the above situation is called *switching point*.

In case of a transition from mode m to $\tilde{m}$ at time $\hat{t}$, the following jump condition applies to the differential state:

$$x(\hat{t}^+) = x(\hat{t}^-) + d^{m \to \tilde{m}}(x(\hat{t}^-)). \tag{8.1}$$

Herein, $d^{m \to \tilde{m}} : X \longrightarrow X$ denotes the jump function for a transition from mode m to mode $\tilde{m}$. The transition from mode m to mode $\tilde{m}$ is possible only if the state $x(\hat{t}^-)$ belongs to some set $X^{m \to \tilde{m}} \subseteq X$. Moreover, $x(\hat{t}^+)$ is supposed to be consistent with the DAE.

The following assumption provides a sufficient condition for a proper crossing of the switching manifold $\{x \in X \mid s^m(x) = 0\}$ in mode m.

Assumption 8.1 *Let the condition*

$$x'(\hat{t}^-)^\top \nabla s^m(x(\hat{t}^-)) < 0$$

be satisfied whenever the system is in mode $m \in M$ and $\hat{t}$ is a point with $s^m(x(\hat{t}^-))=0$.

In the case $x'(\hat{t}^-)^\top \nabla s^m(x(\hat{t}^-)) = 0$, the trajectory is tangential to the manifold $\mathscr{S}^m$ and it may or may not cross the manifold or it may even stay on the manifold. These cases are difficult to handle in general and bifurcation and non-uniqueness

issues may occur. Even if Assumption 8.1 holds, infinitely many switches (Zeno phenomenon) may occur with $\lim_{i\to\infty}(\hat{t}_{i+1} - \hat{t}_i) = 0$, where the $\hat{t}_i$'s denote the switching times. The continuation of the trajectory beyond such an accumulation point (the Zeno point) is nontrivial in general. Often, the trajectory is continued such that it stays on the switching manifold.

The simulation of a hybrid system subject to Assumption 8.1 can be performed as follows:

Algorithm 1 (Hybrid System Simulation Using Switching Functions)

(0) Init: Choose a consistent initial value $z_h(t_0) = (x_h(t_0), y_h(t_0))^\top$ at $t = t_0$, an initial mode $m_0 \in M$ with $s^{m_0}(x_h(t_0)) > 0$, a final time $t_f > t_0$, and set $k = 0$.

(1) Stop the integration, if $t_k = t_f$.

(2) Perform one step of a numerical integration scheme with a suitable step-size h to the DAE

$$\begin{aligned} x'(t) &= f^{m_k}(x(t), y(t)), \\ 0 &= g^{m_k}(x(t), y(t)), \end{aligned}$$

and compute the approximation $z_h(t_{k+1}) = (x_h(t_{k+1}), y_h(t_{k+1}))^\top$ at time $t_{k+1} = \min\{t_f, t_k + h\}$.

(3) If $s^{m_k}(x_h(t_{k+1})) > 0$, set $m_{k+1} = m_k$, $k \leftarrow k + 1$, and go to (1). Otherwise go to (4).

(4) If $s^{m_k}(x_h(t_{k+1})) = 0$, find $\tilde{m}$ with $x_h(t_{k+1}) \in X^{m_k \to \tilde{m}}$, update the state by

$$x_h(t_{k+1}) = x_h(t_{k+1}) + d^{m_k \to \tilde{m}}(x_h(t_{k+1})),$$

compute a corresponding consistent initial value $y_h(t_{k+1})$, update the mode by $m_{k+1} = \tilde{m}$, set $k \leftarrow k + 1$, and go to (1). Otherwise go to (5).

(5) If $s^{m_k}(x_h(t_{k+1})) < 0$, determine a step-size $\tau \in [0, 1]$ such that $s^{m_k}(x_h(t_k + \tau h)) = 0$ using the integration scheme in (2), set $t_{k+1} = t_k + \tau h$ and $z_h(t_{k+1}) = (x_h(t_{k+1}), y_h(t_{k+1}))^\top$, and go to (4).

In order to determine the step-size τ in step (5) of Algorithm 1, a root of the function

$$r(\tau) := s^{m_k}(x_h(t_k + \tau h))$$

has to be found. If $r(0) = s^{m_k}(x_h(t_k)) > 0$ and $r(1) = s^{m_k}(x_h(t_k + h)) < 0$, a root can be located by the bisection method with a linear rate of convergence. Such a root τ is guaranteed to exist in $[0, 1]$ by the intermediate value theorem, if the mapping $\zeta : [0, 1] \longrightarrow \mathbb{R}^n$ with $\zeta(\tau) := x_h(t_k + \tau h)$ is continuous. If ζ is continuously differentiable, then we may apply Newton's method in order to find a root of r and

hope for an at least super-linear convergence rate. The Newton iteration reads

$$\tau_{\ell+1} = \tau_\ell - \frac{r(\tau_\ell)}{r'(\tau_\ell)}, \qquad \ell = 0, 1, 2, \ldots,$$

where

$$r'(\tau) = \zeta'(\tau)^\top \nabla s^{m_k}(\zeta(\tau)).$$

Note that the iteration is well defined, if Assumption 8.1 holds. In fact, the weaker condition $\zeta'(\hat{\tau})^\top \nabla s^{m_k}(\zeta(\hat{\tau})) \neq 0$ in a root $\hat{t}$ of r would be sufficient for a locally super-linear convergence of Newton's method. Often, interpolation techniques are used to compute $\zeta(\tau)$ and $\zeta'(\tau)$ approximately in order to avoid the frequent evaluation of the discretization scheme, see [29, Sect. 5.3.3] for further details.

Example 8.1 Let $x = (x_1, x_2, x_3, x_4)^\top$ and $y = (y_1, y_2)^\top$ be the differential and algebraic states of a pendulum of mass 1 and length 1 with a wall described by the switching function $s^1(x) = x_2 + \frac{1}{2}$ for mode 1 and the set

$$\mathscr{S}^1 = \{(x_1, x_2, x_3, x_4)^\top \mid s^1(x_2) \geq 0\}.$$

In mode 1 (free mode) the pendulum moves according to the DAE (GGL-stabilization)

$$\begin{aligned}
x_1'(t) &= x_3(t) - 2x_1(t)y_2(t),\\
x_2'(t) &= x_4(t) - 2x_2(t)y_2(t),\\
x_3'(t) &= \quad - 2x_1(t)y_1(t),\\
x_4'(t) &= -g - 2x_2(t)y(t),\\
0 &= x_1(t)^2 + x_2(t)^2 - 1,\\
0 &= x_1(t)x_3(t) + x_2(t)x_4(t).
\end{aligned}$$

If the position $(x_1, x_2)^\top$ hits the boundary of $\mathscr{S}^1$ at some $\hat{t}$, i.e., if $x_2(\hat{t}) = -\frac{1}{2}$, the jump condition

$$\begin{pmatrix} x_3(\hat{t}^+) \\ x_4(\hat{t}^+) \end{pmatrix} = \begin{pmatrix} x_3(\hat{t}^-) \\ x_4(\hat{t}^-) \end{pmatrix} - (1+\varepsilon) \begin{pmatrix} x_3(\hat{t}^-) \\ x_4(\hat{t}^-) \end{pmatrix} = -\varepsilon \begin{pmatrix} x_3(\hat{t}^-) \\ x_4(\hat{t}^-) \end{pmatrix}$$

applies and the mode remains unchanged. Herein, $\varepsilon \in [0, 1]$ is the elasticity constant.

Figure 12 shows the results of a simulation with the code DASRT of the contact problem using switching functions in the time interval $[0, 6]$ with $\varepsilon = 0.9$, initial state $x(0) = (1, 0, 0, 0)^\top$, $y(0) = (0, 0)^\top$, and error tolerance 10^{-10} for the differential states.

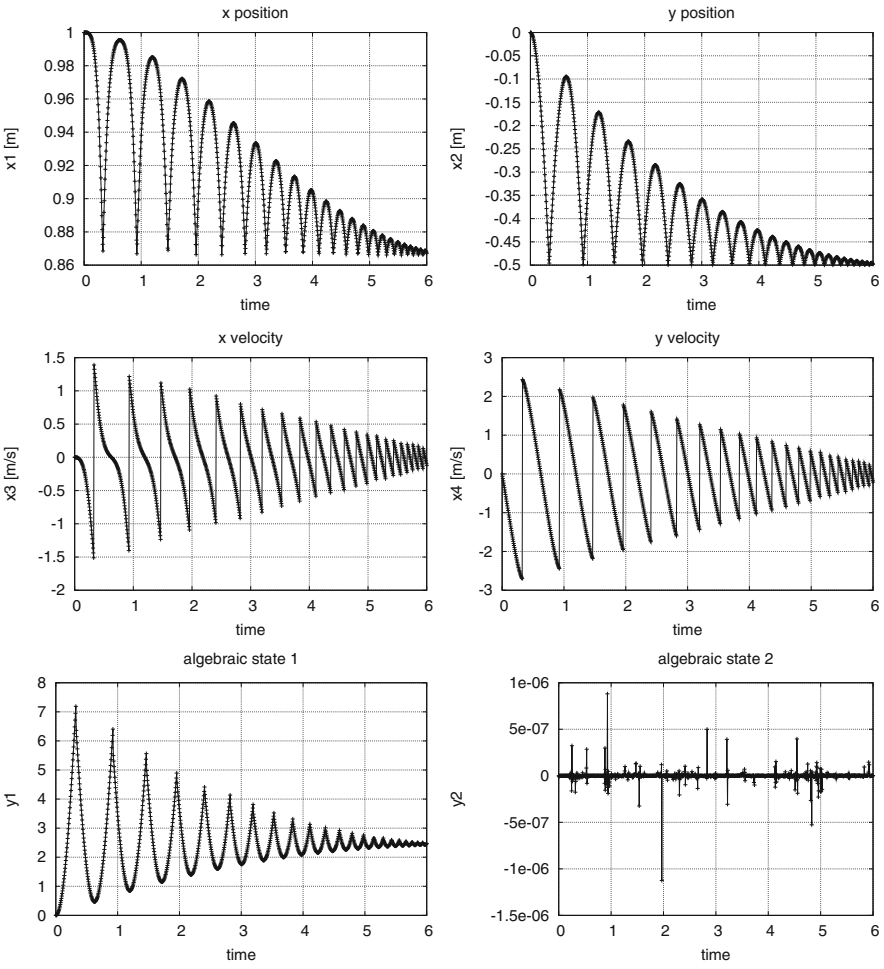


Fig. 12 Numerical simulation of pendulum with contact surface and switching functions

The results illustrate the Zeno phenomenon since the sequence of contact points accumulates. A natural continuation beyond the accumulation point is the constant solution with the pendulum being at rest on the switching manifold.

8.2 Parametric Sensitivity Analysis for Switched Systems

We add a parameter vector $p \in \mathbb{R}^q$ to the problem setting in Sect. 8.1 and investigate the sensitivity of a solution of the hybrid system with respect to the parameter vector.

To this end, let the state evolve according to the parameter-dependent DAE

$$\begin{aligned} x'(t) &= f^m(x(t), y(t), p), \\ 0 &= g^m(x(t), y(t), p) \end{aligned}$$

in mode $m \in M$ within the set

$$\mathscr{S}^m(p) := \{x \in X \mid s^m(x, p) \geq 0\}, \qquad s^m : \mathbb{R}^n \times \mathbb{R}^q \longrightarrow \mathbb{R}.$$

In case of a transition from mode m to $\tilde{m}$ at time $\hat{t}$, the following jump condition applies to the differential state:

$$x(\hat{t}^+) = x(\hat{t}^-) + d^{m \to \tilde{m}}(x(\hat{t}^-), p). \tag{8.2}$$

Herein, $d^{m \to \tilde{m}} : X \times \mathbb{R}^q \longrightarrow X$ denotes the parametric jump function for a transition from mode m to mode $\tilde{m}$, where a transition is possible if $x(\hat{t}^-)$ belongs to some set $X^{m \to \tilde{m}}(p) \subseteq X$. The jump function d in (8.2) has to be chosen such that it provides consistent differential states $x(\hat{t}^+)$.

The functions f^m, g^m, s^m, $m \in M$, and $d^{m \to \tilde{m}}$, $m, \tilde{m} \in M$, are supposed to be at least continuously differentiable with respect to all arguments.

Let $z(t; p) = (x(t; p), y(t; p))^\top$ for $t \in [t_0, t_f]$ denote a solution of the switched system for the parameter p with a consistent initial value $z(t_0; p) = z_0(p) = (x_0(p), y_0(p))^\top$ in mode m with $s^m(x_0(p), p) > 0$. In particular, let $\hat{z}(t) := (\hat{x}(t), \hat{y}(t))^\top$ with $\hat{z}(t) = z(t; \hat{p})$ and $\hat{z}_0 = z_0(\hat{p})$ denote the solution for a fixed nominal parameter $\hat{p} \in \mathbb{R}^q$.

We are interested in computing the sensitivities

$$S_x(t) := \frac{\partial x}{\partial p}(t; \hat{p}), \qquad S_y(t) := \frac{\partial y}{\partial p}(t; \hat{p})$$

assuming their existence in $[t_0, t_f]$.

While in mode m with $s^m(x(t; \hat{p}), \hat{p}) > 0$, the sensitivities solve the sensitivity DAE

$$\begin{aligned} S_x'(t) &= A^m(t) S_x(t) + B^m(t) S_y(t) + c^m(t), \\ 0 &= G^m(t) S_x(t) + H^m(t) S_y(t) + k^m(t), \end{aligned}$$

with

$$\begin{aligned} A^m(t) &:= \frac{\partial f^m}{\partial x}(\hat{x}(t), \hat{y}(t), \hat{p}), & G^m(t) &:= \frac{\partial g^m}{\partial x}(\hat{x}(t), \hat{y}(t), \hat{p}), \\ B^m(t) &:= \frac{\partial f^m}{\partial y}(\hat{x}(t), \hat{y}(t), \hat{p}), & H^m(t) &:= \frac{\partial g^m}{\partial y}(\hat{x}(t), \hat{y}(t), \hat{p}), \\ c^m(t) &:= \frac{\partial f^m}{\partial p}(\hat{x}(t), \hat{y}(t), \hat{p}), & k^m(t) &:= \frac{\partial g^m}{\partial p}(\hat{x}(t), \hat{y}(t), \hat{p}), \end{aligned}$$

compare Sect. 7.

We investigate, what happens at a switching point $\hat{t}$ in mode m with parameter $\hat{p}$. Then, we have

$$s^m(x(\hat{t}^-;\hat{p});\hat{p}) = 0. \tag{8.3}$$

In general, this switching point will depend on the parameter. Define

$$r(t,p) := s^m(x(t^-;p),p)$$

for p close to $\hat{p}$. Equation (8.3) implies

$$\hat{r} := r(\hat{t},\hat{p}) = 0.$$

Assumption 8.2 *The switching point $\hat{t}$ in mode m satisfies*

$$0 \neq \frac{\partial r}{\partial t}(\hat{t},\hat{p}) = \frac{d}{dt}s^m(x(\hat{t}^-;\hat{p});\hat{p}) = x'(\hat{t}^-;\hat{p})^\top \nabla_x s^m(x(\hat{t}^-;\hat{p});\hat{p}).$$

If Assumption 8.2 holds, then, by the implicit function theorem, there exist neighborhoods $B_\delta(\hat{p})$, $\delta > 0$, $B_\varepsilon(\hat{t})$, $\varepsilon > 0$, and a continuously differentiable mapping $T : B_\delta(\hat{p}) \longrightarrow B_\varepsilon(\hat{t})$ with

$$\hat{t} = T(\hat{p}) \qquad \text{and} \qquad r(T(p),p) = 0 \quad \forall p \in B_\delta(\hat{p}),$$

and

$$T'(\hat{p}) = -\left(\frac{\partial r}{\partial t}(\hat{t},\hat{p})\right)^{-1} \frac{\partial r}{\partial p}(\hat{t},\hat{p})$$

with

$$\frac{\partial r}{\partial t}(\hat{t},\hat{p}) = x'(\hat{t}^-;\hat{p})^\top \nabla_x s^m(x(\hat{t}^-;\hat{p});\hat{p}),$$
$$\frac{\partial r}{\partial p}(\hat{t},\hat{p}) = \nabla_x s^m(\hat{x}(\hat{t}^-);\hat{p})^\top S_x(\hat{t}^-) + \nabla_p s^m(\hat{x}(\hat{t}^-);\hat{p}).$$

Introducing $T(p)$ into (8.2) yields the relation

$$x(T(p)^+;p) = x(T(p)^-;p) + d^{m\to\tilde{m}}(x(T(p)^-;p),p). \tag{8.4}$$

Herein, we assume that the transition $m \to \tilde{m}$ is stable under small perturbations in p. Differentiation of (8.2) with respect to p and evaluation at $\hat{p}$ yields

$$\begin{aligned} x'(\hat{t}^+;\hat{p})T'(\hat{p}) + S_x(\hat{t}^+) = {} & x'(\hat{t}^-;\hat{p})T'(\hat{p}) + S_x(\hat{t}^-) \\ & + \nabla_x d^{m\to\tilde{m}}(\hat{x}(t^-),\hat{p})^\top \left(x'(\hat{t}^-;\hat{p})T'(\hat{p}) + S_x(\hat{t}^-)\right) \\ & + \nabla_p d^{m\to\tilde{m}}(\hat{x}(t^-),\hat{p})^\top. \end{aligned}$$

Rearranging terms leads to an update rule for the sensitivity S_x at the switching point $\hat{t}$ according to

$$\begin{aligned} S_x(\hat{t}^+) = {} & \left(x'(\hat{t}^-;\hat{p}) - x'(\hat{t}^+;\hat{p}) + \nabla_x d^{m\to\tilde{m}}(\hat{x}(t^-),\hat{p})^\top x'(\hat{t}^-;\hat{p})\right) T'(\hat{p}) \\ & + \left(I + \nabla_x d^{m\to\tilde{m}}(\hat{x}(t^-),\hat{p})^\top\right) S_x(\hat{t}^-) \\ & + \nabla_p d^{m\to\tilde{m}}(\hat{x}(t^-),\hat{p})^\top. \end{aligned} \tag{8.5}$$

If x is continuous at $\hat{t}$, i.e., $d \equiv 0$, then the update rule for S_x reduces to

$$S_x(\hat{t}^+) = S_x(\hat{t}^-) + \left(x'(\hat{t}^-;\hat{p}) - x'(\hat{t}^+;\hat{p})\right) T'(\hat{p}).$$

If, in addition, x' is continuous at $\hat{t}$, then S_x is continuous at $\hat{t}$ as well.

After $x(\hat{t}^+)$ and $S_x(\hat{t}^+)$ have been computed by (8.2) and (8.5), the algebraic component $S_y(\hat{t}^+)$ has to be computed consistently.

Note that this update rule is only valid under Assumption 8.2, which ensures a proper crossing of the switching manifold. If Assumption 8.2 does not hold at $\hat{t}$, it is not clear how to update the sensitivity S_x and in general the state may not depend continuously differentiable on p.

A related parametric sensitivity analysis for mechanical multibody systems using switching functions can be found in [144, Sect. 3.9] and [112]. An adjoint calculus for switched DAEs is derived in [114].

Remark 8.1 Please note that Assumptions 8.1 and 8.2 are crucial in the above analysis. These assumptions are often explicitly or implicitly assumed by standard integrators like DASRT or SCILAB/DASKR. The user needs to be aware of this since codes may fail if the assumptions are not met. As pointed out earlier, it is in general not clear how to continue integration (especially in the context of sensitivity analysis) if the assumptions are not satisfied. In case of the Zeno phenomenon, it is often assumed that the solution stays on the switching manifold. Modifications in the codes are necessary in such cases.

8.3 Contact and Friction in Mechanical Multibody Systems

Mechanical multibody dynamics taking into account contact forces and friction forces are, beyond doubt, the most important examples of switched systems and include particular impact and friction models, compare, e.g., [3, 139]. Suitable discretization schemes for such systems, which typically do not locate impact points exactly but work with a fixed step-size instead, are introduced in [2, 6, 113]. Extensions towards large-scale systems and tailored algorithms for complementarity problems can be found in [5, 137, 138, 140]. Impact models and the interpretation of the mechanical multibody system as a measure differential equation can be found in [65, 88, 102].

The equations of motion of a mechanical multibody system with contact and friction are given by

$$\begin{aligned} q'(t) &= v(t), \\ M(q(t))v'(t) &= f(q(t), v(t)) + F_C(q(t)), \end{aligned}$$

In the above model, $q(t) \in \mathbb{R}^n$ denotes the vector of generalized coordinates, v its velocity, $f(q, v)$ the vector of generalized forces, and $M(q)$ the non-singular mass matrix.

The above multibody system is augmented by an impact model that relates the velocity $v(\hat{t}^-)$ right before an impact to the velocity $v(\hat{t}^+)$ right after the impact. The impact model typically leads to a discontinuity of some components of the velocity vector v at a contact point $\hat{t}$ and hence, the velocity components are only of bounded variation in general. The vector $F_C(q)$ contains the contact and friction forces, which apply only in the case of a contact between the rigid bodies of the multibody system, compare [60, 132]. Whether or not a contact between bodies occurs, is measured by distance functions $s_k : \mathbb{R}^n \longrightarrow \mathbb{R}$ with

$$s_k(q) \geq 0, \qquad k = 1, \ldots, m.$$

Herein, a contact at time $\hat{t}$ occurs, if $s_k(q(\hat{t})) = 0$ for some $k \in \{1, \ldots, m\}$. In case of a contact, the resulting contact and friction force $F_{C,k}(q)$ is an element of the *friction cone*

$$FC_k(q) := \{F^n + F^t \mid F^n = n_k(q)\lambda, F^t = D_k(q)\beta, \lambda \geq 0, \psi(\beta) \leq \mu_k\lambda\}.$$

Herein, $F^n = F^n_k(q)$ denotes the contact force into the normal direction of the contact surface, which can be expressed as $F^n_k(q) = n_k(q)\lambda_k$ with the normal vector $n_k(q) = \nabla s_k(q)$ to the contact manifold $\mathscr{S}_k(q) = \{q \in \mathbb{R}^n \mid s_k(q) = 0\}$. λ_k satisfies the *Signorini contact conditions*

$$0 \leq s_k(q) \qquad \perp \qquad \lambda_k \geq 0,$$

which is a complementarity system for λ_k. The operator $\perp$ in $0 \leq a \perp b \geq 0$ is defined by $a \geq 0$, $b \geq 0$, and $ab = 0$.

The force $F^t = F^t_k(q)$ is the tangential force owing to friction in the contact manifold, which can be expressed as $F^t_k(q) = D_k(q)\beta_k$, where the columns of the matrix $D_k(q)$ span the friction space. For isotropic Coulomb friction, which we assume throughout, the function ψ is given by $\psi(\beta) := \|\beta\|_2$ and $\mu_k \geq 0$ is the friction coefficient at the contact manifold $\mathscr{S}_k(q)$. The norm $\|\cdot\|_2$ causes some difficulties as $\|\beta_k\|_2 \leq \mu_k\lambda_k$ leads to a non-smooth constraint. To overcome this difficulty, the norm $\|\cdot\|_2$ is typically approximated by $\|\cdot\|_1$, which leads to the following polyhedral approximation of the friction cone:

$$\overline{FC}_k(q) := \{F^n + F^t \mid F^n = n_k(q)\lambda, F^t = D_k(q)\beta, \lambda \geq 0, \|\beta\|_1 \leq \mu_k\lambda\},$$

compare [60, 132].

Depending on the choice of the friction cone, the total contact force is then defined by

$$F_C(q) = \sum_{k:s_k(q)=0} F_{C,k}(q)$$

with $F_C(q)$ being an element either of the total friction cone

$$FC(q) = \sum_{k:s_k(q)=0} FC_k(q)$$

or its polyhedral approximation

$$\overline{FC}(q) = \sum_{k:s_k(q)=0} \overline{FC}_k(q).$$

If a contact occurs at time $\hat{t}$, i.e., $s_k(q(\hat{t})) = 0$ for some $k \in \{1, \ldots, m\}$, the *impact model*

$$\nabla s_k(q(\hat{t}))^\top (v(\hat{t}^+) + \varepsilon_k v(\hat{t}^-)) = 0$$

applies. Herein, $\varepsilon_k \in [0, 1]$ denotes the elasticity constant. A fully elastic contact occurs if $\varepsilon_k = 1$. An inelastic contact occurs if $\varepsilon_k = 0$.

An approach to determine the friction force $F_k^t = D_k(q(\hat{t}))\beta_k$ at a contact is based on the *maximum dissipation principle*, compare [132]. Herein, the corresponding friction force F_k^t maximizes the rate of energy dissipation for a given normal force $F_k^n = n_k(q(\hat{t}))\lambda_k$ at a contact. This principle leads to the following convex optimization problem for β:

$$\text{Maximize} \qquad -v(\hat{t}^+)^\top D_k(q(\hat{t}))\beta \qquad \text{s.t.} \qquad \psi(\beta) \le \mu_k\lambda_k. \tag{8.6}$$

Let β_k be a solution of (8.6). If $\lambda_k > 0$, then $\beta = 0$ satisfies the Slater condition for this convex optimization problem, and a necessary and sufficient condition for the solution β_k reads as follows, compare [38, Theorem 6.1.1, Proposition 6.3.1]: There exists a multiplier $\zeta_k \in \mathbb{R}$ such that

$$0 \in D_k(q(\hat{t})))^\top v(\hat{t}^+) + \zeta_k \partial\psi(\beta_k), \tag{8.7}$$

$$0 \le \zeta_k \quad \perp \quad \mu_k\lambda_k - \psi(\beta_k) \ge 0. \tag{8.8}$$

Herein, $\partial\psi = \partial(\|\cdot\|_2)$ denotes the generalized gradient of the locally Lipschitz continuous function ψ, which is given by

$$\partial\psi(\beta) = \begin{cases} \frac{1}{\|\beta\|_2}\beta, & \text{if } \beta \neq 0, \\ \{\alpha \mid \|\alpha\|_2 \le 1\}, & \text{if } \beta = 0. \end{cases}$$

If $\lambda_k = 0$, then $\beta = 0$ is the only feasible point in (8.6) and the conditions (8.7)–(8.8) are satisfied, e.g., by choosing

$$\zeta_k = \|D_k(q(\hat{t})))^\top v(\hat{t}^+)\|_2 \quad \text{and} \quad \alpha = \begin{cases} -\frac{1}{\zeta_k} D_k(q(\hat{t})))^\top v(\hat{t}^+), & \text{if } \zeta_k > 0, \\ 0, & \text{if } \zeta_k = 0. \end{cases}$$

Note that in either case $\alpha \in \partial\psi(0)$. Instead of $\psi(\beta) = \|\beta\|_2$ we may use the approximation $\|\beta\|_1$ in (8.6), which transforms the convex optimization problem in fact into a linear program. To this end, β is replaced by $\beta = \beta^+ - \beta^-$ with $\beta^+, \beta^- \ge 0$:

$$\begin{aligned} &\text{Maximize } -v(\hat{t}^+)^\top D_k(q(\hat{t}))(\beta^+ - \beta^-) \\ &\text{s.t.} \qquad e^\top(\beta^+ + \beta^-) \le \mu_k\lambda_k,\ \beta^+ \ge 0,\ \beta^- \ge 0. \end{aligned} \tag{8.9}$$

Herein, we exploited the relation $\|\beta\|_1 = e^\top(\beta^+ + \beta^-)$ with the vector $e = (1, \ldots, 1)^\top$ of all ones of appropriate dimension. First order necessary and sufficient

conditions for a solution $\beta_k = \beta_k^+ - \beta_k^-$ of (8.9) yield

$$\begin{aligned}
0 &= \left[D_k(q(\hat{t})), -D_k(q(\hat{t}))\right]^\top v(\hat{t}^+) + \zeta_k \begin{pmatrix} e \\ e \end{pmatrix} - \begin{pmatrix} \eta_k^+ \\ \eta_k^- \end{pmatrix}, \\
0 &\le \zeta_k \quad \perp \quad \mu_k \lambda_k - e^\top(\beta_k^+ + \beta_k^-) \ge 0, \\
0 &\le \beta_k^\pm \quad \perp \quad \eta_k^\pm \ge 0.
\end{aligned}$$

Multiplication of the first equation by $(\beta_k^+, \beta_k^-)^\top$ from the left and exploitation of the complementarity conditions yield

$$\begin{aligned}
0 &\le \tilde{\beta}_k \quad \perp \quad \tilde{D}_k(q(\hat{t}))^\top v(\hat{t}^+) + \zeta_k e \ge 0, \\
0 &\le \zeta_k \quad \perp \quad \mu_k \lambda_k - e^\top \tilde{\beta}_k \ge 0
\end{aligned}$$

with

$$\tilde{D}_k(q(\hat{t})) := \left[D_k(q(\hat{t})), -D_k(q(\hat{t}))\right], \quad \tilde{\beta}_k = [\beta_k^+, \beta_k^-]^\top.$$

Note that the matrix $\tilde{D}_k$ is balanced, i.e., if $\tilde{D}_k$ contains a column c, then it contains $-c$ as well. Summarizing, the equations of motion with contact and friction forces satisfy the following complementarity system:

$$\begin{aligned}
q'(t) &= v(t), \\
M(q(t))v'(t) &= f(q(t), v(t)) + \sum_{k=1}^{m} n_k(q(t))\lambda_k(t) + D_k(q(t))\beta_k(t), \\
0 &\le s_k(q(t)) \quad \perp \quad \lambda_k(t) \ge 0, \\
0 &\le \tilde{\beta}_k(t) \quad \perp \quad \tilde{D}_k(q(t))^\top v(t^+) + \zeta_k(t) e \ge 0, \\
0 &\le \zeta_k(t) \quad \perp \quad \mu_k \lambda_k(t) - e^\top \tilde{\beta}_k(t) \ge 0
\end{aligned}$$

and

$$\nabla s_k(q(t))^\top (v(t^+) + \varepsilon_k v(t^-)) = 0 \qquad \text{if } s_k(q(t)) = 0$$

with $k = 1, \ldots, m$. A semi-implicit discretization scheme for the system was suggested in [6, 132]. Let $z^\ell = (q^\ell, v^\ell, \lambda^\ell, \beta^\ell, \zeta^\ell)$ be the state at time t_ℓ and $h > 0$ a step-size. Let the index set of active contacts at t_ℓ be defined by

$$A^\ell := \{k \in \{1, \ldots, m\} \mid s_k(q^\ell + h v^\ell) \le 0\}.$$

Let $\lambda_{A^\ell}^{\ell+1} = (\lambda_k^{\ell+1})_{k \in A^\ell}$ and likewise for $\tilde{\beta}_{A^\ell}^{\ell+1}$ and $\zeta_{A^\ell}^{\ell+1}$.

Then $z^{\ell+1} = (q^{\ell+1}, v^{\ell+1}, \lambda_{A^\ell}^{\ell+1}, \tilde{\beta}_{A^\ell}^{\ell+1}, \zeta_{A^\ell}^{\ell+1})$ solves the following complementarity problem:

$$q^{\ell+1} - q^\ell = hv^{\ell+1}, \tag{8.10}$$

$$M(q^{\ell+1})\left(v^{\ell+1} - v^\ell\right) = h\left(f(q^\ell, v^\ell) + \sum_{k\in A^\ell} n_k(q^\ell)\lambda_k^{\ell+1} + \tilde{D}_k(q^\ell)\tilde{\beta}_k^{\ell+1}\right), \tag{8.11}$$

$$0 \le \lambda_k^{\ell+1} \quad \perp \quad \nabla s_k(q^\ell)^\top (v^{\ell+1} + \varepsilon_k v^\ell) \ge 0, \quad (k \in A^\ell) \tag{8.12}$$

$$0 \le \tilde{\beta}_k^{\ell+1} \quad \perp \quad \tilde{D}_k(q^\ell)^\top v^{\ell+1} + \zeta_k^{\ell+1} e \ge 0, \quad (k \in A^\ell) \tag{8.13}$$

$$0 \le \zeta_k^{\ell+1} \quad \perp \quad \mu_k \lambda_k^{\ell+1} - e^\top \tilde{\beta}_k^{\ell+1} \ge 0. \quad (k \in A^\ell) \tag{8.14}$$

Convergence results and alternative discretizations are discussed in [4, 6, 60, 113].

It remains to discuss, how the nonlinear complementarity problem can be solved. If M is independent of q or if $M(q^{\ell+1})$ was replaced by $M(q^\ell)$, then the problem is actually a linear complementarity problem, compare [3, 46, 138, 139], and it could be solved by Lemke's algorithm [95] or as in [5, 137]. Another approach is to use a semi-smooth Newton method, compare [86, 115, 116]. Herein, the complementarity system (8.10)–(8.14) is rewritten as the nonlinear and non-smooth equation

$$0 = G(z^{\ell+1}) \tag{8.15}$$

$$= \begin{pmatrix} q^{\ell+1} - q^\ell - hv^{\ell+1} \\ M(q^{\ell+1})\left(v^{\ell+1} - v^\ell\right) - h\left(f(q^\ell, v^\ell) + \sum_{k\in A^\ell} n_k(q^\ell)\lambda_k^{\ell+1} + \tilde{D}_k(q^\ell)\tilde{\beta}_k^{\ell+1}\right) \\ \phi_{FB}(\lambda_k^{\ell+1}, \nabla s_k(q^\ell)^\top (v^{\ell+1} + \varepsilon_k v^\ell)) \quad (k \in A^\ell) \\ \phi_{FB}(\tilde{\beta}_k^{\ell+1}, \tilde{D}_k(q^\ell)^\top v^{\ell+1} + \zeta_k^{\ell+1} e) \quad (k \in A^\ell) \\ \phi_{FB}(\zeta_k^{\ell+1}, \mu_k \lambda_k^{\ell+1} - e^\top \tilde{\beta}_k^{\ell+1}) \quad (k \in A^\ell) \end{pmatrix},$$

where $\phi_{FB}(a, b) := \sqrt{a^2 + b^2} - a - b$ denotes the Lipschitz continuous Fischer-Burmeister function, see [55]. Let $\partial G(z)$ denote Clarke's generalized Jacobian of G, compare [38] and [70] for details on how to compute it. Then, a root of G can be obtained by the following basic version of the semi-smooth Newton method.

Algorithm 2 Semi-Smooth Newton Method

(0) Init: Choose tolerance $tol > 0$ and an initial guess for $z^{\ell+1}$, e.g., $z^{(0)} = (q^\ell + hv^\ell, v^\ell, 0, 0, 0)^\top$. Set $j = 0$.
(1) If $\|G(z^{(j)})\| \le tol$, set $z^{\ell+1} = z^{(j)}$ and STOP.

(2) Compute an element $V^{(j)} \in \partial G(z^{(j)})$ and the Newton direction $d^{(j)}$ by solving the linear equation

$$V^{(j)} d = -G(z^{(j)}).$$

(4) Set $z^{(j+1)} = z^{(j)} + d^{(j)}$, $j \leftarrow j+1$, and go to (1).

The following examples summarize results, which have been obtained by applying Algorithm 2 to mechanical multibody systems with contact and friction.

Example 8.2 Consider a bouncing and rotating ball with radius $r = 1$, mass $m = 1$, and moment of inertia $J = 1$ in the $x-z$-plane with $q = (x, z, \theta)^\top$, $M = \mathrm{diag}(m, m, J)$, $f(q, q') = (0, -mg, 0)^\top$, $g = 9.81$, and $s(q) = z - r$. The friction space is spanned by

$$\tilde{D}(q) = \begin{pmatrix} -1 & 1 \\ 0 & 0 \\ r & -r \end{pmatrix}.$$

Figure 13 shows a simulation of the bouncing ball in the time interval $[0, 10]$ with initial state $q(0) = (0, 10, 0)^\top$, $v(0) = (1, 10, -5)^\top$, friction coefficient $\mu = 0.2$, and elasticity parameter $\varepsilon = 0.675$. The states q, v, λ, β, and ζ as functions of time are depicted in Fig. 14.

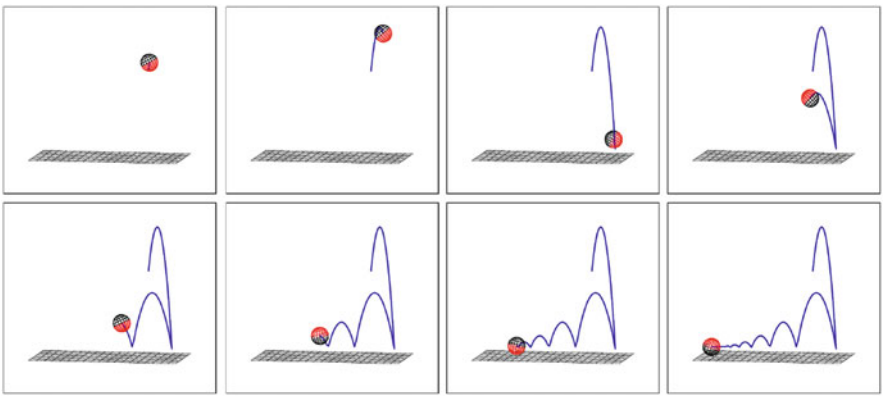


Fig. 13 Snapshot of a bouncing and rotating ball with contact and friction

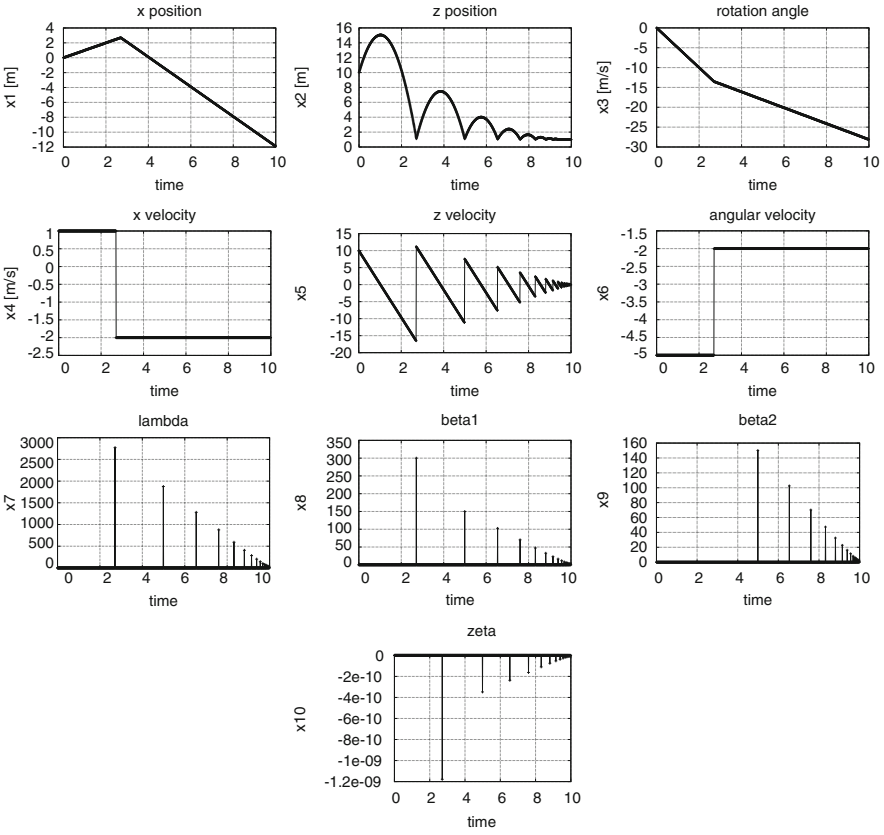


Fig. 14 Snapshot of a bouncing and rotating ball with contact and friction

Example 8.3 Consider a billiard table and the motion of a sphere on the table in the $x - y$-plane. For simplicity, the friction on the table is neglected, but friction forces and contact forces at the boundaries of the table are taken into account with elasticity constant $\varepsilon = 0.9$ and friction coefficient $\mu = 0.5$. The radius of the sphere is $r = 0.04$ [m], its mass is $m = 0.1$ [kg], and its moment of inertia is $J = 1$. The generalized coordinates are $q = (x, y, \theta)^\top$, the mass matrix is $M = \mathrm{diag}(m, m, J)$, the generalized forces are $f(q, q') = (0, 0, 0)^\top$, and the switching function for the

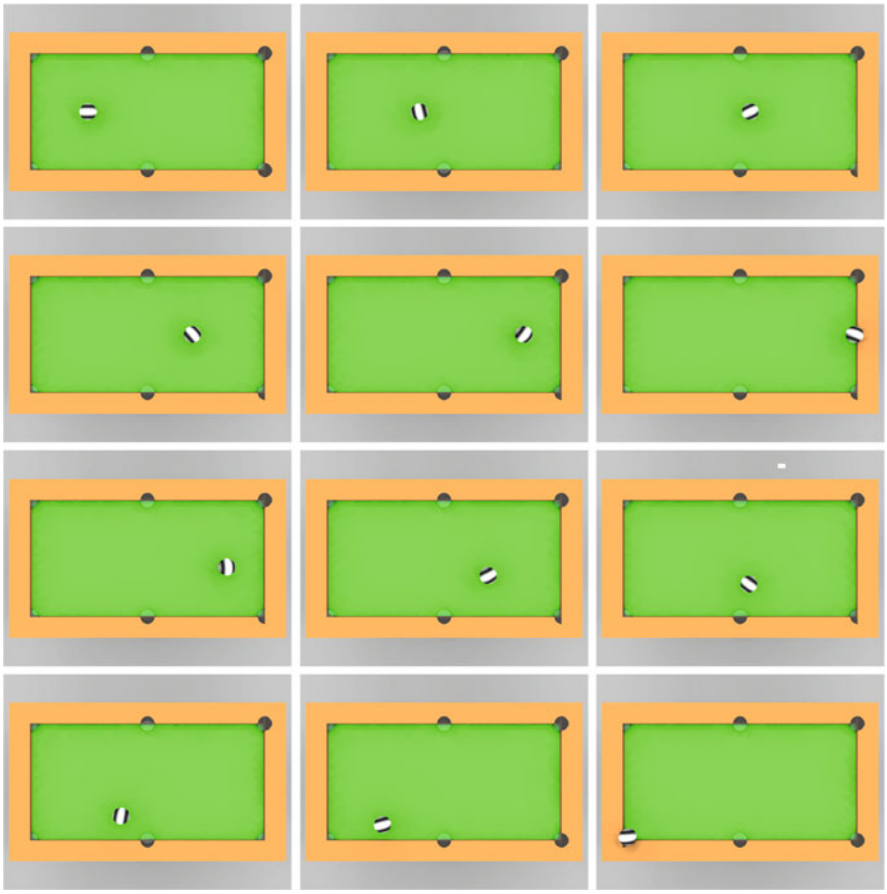


Fig. 15 Snapshot of a billiard table with contact and friction at the borders of the table. For better visibility the sphere was enlarged by a factor of two in the pictures

opposite boundary of the table is $s(q) = y - r$. The friction space is spanned by

$$D(q) = \begin{pmatrix} -1 & 1 \\ 0 & 0 \\ r & -r \end{pmatrix}.$$

Figure 15 shows some snapshots of a simulation of the billiard problem in the time interval $[0, 2.05]$ with initial state $q(0) = (0, 3\ell/4, 0)^\top$, $v(0) = (0, -2, -11)^\top$, where $\ell = 2.24$ denotes the length of the table in [m]. The states q, v, λ, β, and ζ as functions of time are depicted in Fig. 16.

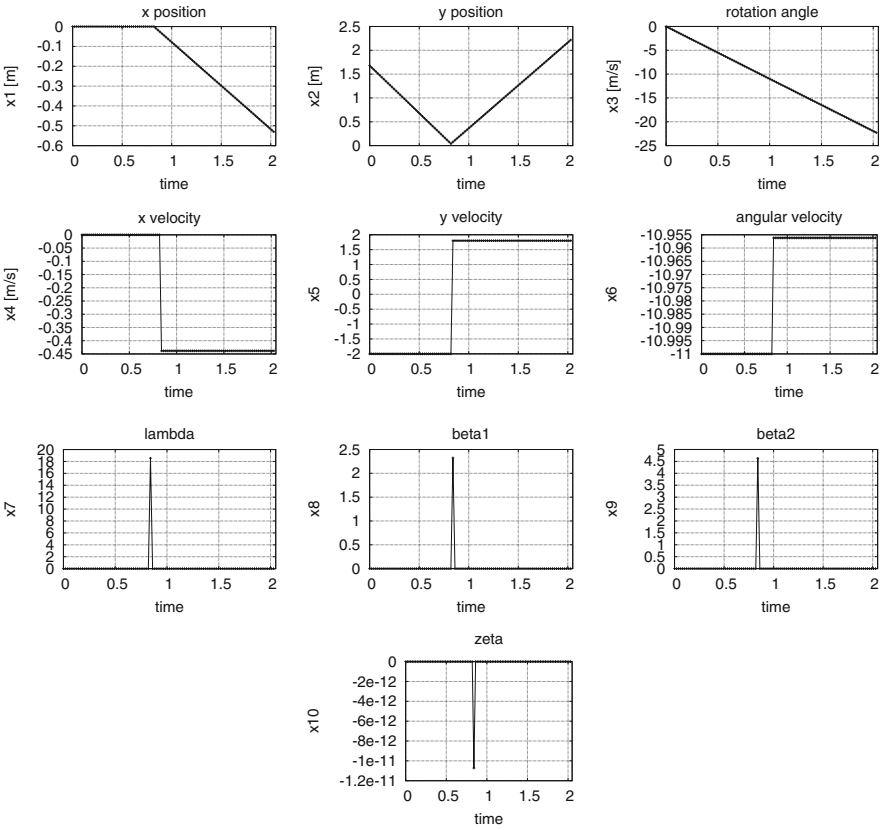


Fig. 16 Snapshot of a billiard table with contact and friction at the borders of the table

9 Conclusions

Simulation is a well-established and indispensable tool in industrial design procedures. Moreover, efficient simulation techniques are required in other disciplines such as controller design, parameter identification, or optimal control. This paper aims to provide an overview on different aspects in the simulation of DAE initial value problems. The focus was set on a choice of methods and concepts that are relevant in industrial simulation environments for coupled systems of potentially large size. These concepts build upon basic integration schemes and add features like sensitivity analysis (needed, e.g., in optimization procedures), contact dynamics, real-time schemes, or co-simulation techniques. Each of these topics is a field of research in its own right with many contributions. Only some of the many contributions could be mentioned in this overview paper and we refer the interested reader to the specialized literature in the bibliography.

References

1. 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)
2. Anitescu, M.: Optimization-based simulation of nonsmooth rigid multibody dynamics. Math. Program. **105**(1(A)), 113–143 (2006)
3. Anitescu, M., Potra, F.A.: Formulating dynamic multi-rigid-body contact problems with friction as solvable linear complementarity problems. Nonlinear Dyn. **14**(3), 231–247 (1997)
4. Anitescu, M., Potra, F.A.: A time-stepping method for stiff multibody dynamics with contact and friction. Int. J. Numer. Methods Eng. **55**(7), 753–784 (2002)
5. Anitescu, M., Tasora, A.: An iterative approach for cone complementarity problems for nonsmooth dynamics. Comput. Optim. Appl. **47**(2), 207–235 (2010)
6. Anitescu, M., Potra, F.A., Stewart, D.E.: Time-stepping for three-dimensional rigid body dynamics. Comput. Methods Appl. Mech. Eng. **177**(3–4), 183–197 (1999)
7. Arnold, M.: Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2. BIT **38**(3), 415–438 (1998)
8. Arnold, M.: Multi-rate time integration for large scale multibody system models. In: IUTAM Symposium on Multiscale Problems in Multibody System Contacts: Proceedings of the IUTAM Symposium held in Stuttgart, Germany, February 20–23, 2006, pp. 1–10. Springer, Dordrecht (2007)
9. Arnold, M.: Stability of sequential modular time integration methods for coupled multibody system models. J. Comput. Nonlinear Dyn. **5**, 031003 (2010)
10. Arnold, M.: Modular time integration of block-structured coupled systems without algebraic loops. In: Schöps, S., Bartel, A., Günther, M., ter Maten, E.J.W., Müller, P.C. (eds.) Progress in Differential-Algebraic Equations. Differential-Algebraic Equations Forum, pp. 97–106. Springer, Berlin/Heidelberg (2014)
11. Arnold, M., Günther, M.: Preconditioned dynamic iteration for coupled differential-algebraic systems. BIT Numer. Math. **41**(1), 001–025 (2001)
12. Arnold, M., Murua, A.: Non-stiff integrators for differential-algebraic systems of index 2. Numer. Algorithm. **19**(1–4), 25–41 (1998)
13. Arnold, M., Strehmel, K., Weiner, R.: Half-explicit Runge–Kutta methods for semi-explicit differential-algebraic equations of index 1. Numer. Math. **64**(1), 409–431 (1993)
14. Arnold, M., Burgermeister, B., Eichberger, A.: Linearly implicit time integration methods in real-time applications: DAEs and stiff ODEs. Multibody Syst. Dyn. **17**(2–3), 99–117 (2007)
15. Arnold, M., Burgermeister, B., Führer, C., Hippmann, G., Rill, G.: Numerical methods in vehicle system dynamics: state of the art and current developments. Veh. Syst. Dyn. **49**(7), 1159–1207 (2011)
16. Arnold, M., Clauß, C., Schierz, T.: Error analysis and error estimates for co-simulation in FMI for model exchange and co-simulation v2.0. Arch. Mech. Eng. **LX**, 75–94 (2013)
17. Arnold, M., Hante, S., Köbis, M.A.: Error analysis for co-simulation with force-displacement coupling. Proc. Appl. Math. Mech. **14**(1), 43–44 (2014)
18. Ascher, U.M., Petzold, L.R.: Projected implicit Runge-Kutta methods for differential-algebraic equations. SIAM J. Numer. Anal. **28**(4), 1097–1120 (1991)
19. Balzer, M., Burger, M., Däuwel, T., Ekevid, T., Steidel, S., Weber, D.: Coupling DEM particles to MBS wheel loader via co-simulation. In: Proceedings of the 4th Commercial Vehicle Technology Symposium (CVT 2016), pp. 479–488 (2016)
20. Bartel, A., Brunk, M., Günther, M., Schöps, S.: Dynamic iteration for coupled problems of electric circuits and distributed devices. SIAM J. Sci. Comput. **35**(2), B315–B335 (2013)
21. Bartel, A., Brunk, M., Schöps, S.: On the convergence rate of dynamic iteration for coupled problems with multiple subsystems. J. Comput. Appl. Math. **262**, 14–24 (2014). Selected Papers from NUMDIFF-13
22. Baumgarte, J.: Stabilization of constraints and integrals of motion in dynamical systems. Comput. Methods Appl. Mech. Eng. **1**, 1–16 (1972)

23. Becker, U.: Efficient time integration and nonlinear model reduction for incompressible hyperelastic materials. Ph.D. thesis, TU Kaiserslautern (2012)
24. Becker, U., Simeon, B., Burger, M.: On rosenbrock methods for the time integration of nearly incompressible materials and their usage for nonlinear model reduction. J. Comput. Appl. Math. **262**, 333–345 (2014). Selected Papers from NUMDIFF-13
25. Bock, H.G.: Randwertproblemmethoden zur Parameteridentifizierung in Systemen nichtlinearer Differentialgleichungen, vol. 183. Bonner Mathematische Schriften, Bonn (1987)
26. Brasey, V., Hairer, E.: Half-explicit RungeKutta methods for differential-algebraic systems of index 2. SIAM J. Numer. Anal. **30**(2), 538–552 (1993)
27. Brenan, K.E., Engquist, B.E.: Backward differentiation approximations of nonlinear differential/algebraic systems. Math. Comput. **51**(184), 659–676 (1988)
28. Brenan, K.E., Petzold, L.R.: The numerical solution of higher index differential/algebraic equations by implicit methods. SIAM J. Numer. Anal. **26**(4), 976–996 (1989)
29. Brenan, K.E., Campbell, S.L., Petzold, L.R.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. Classics in Applied Mathematics, vol. 14. SIAM, Philadelphia (1996)
30. Brown, P.N., Hindmarsh, A.C., Petzold, L.R.: Consistent initial condition calculation for differential-algebraic systems. SIAM J. Sci. Comput. **19**(5), 1495–1512 (1998)
31. Burgermeister, B., Arnold, M., Esterl, B.: DAE time integration for real-time applications in multi-body dynamics. Z. Angew. Math. Mech. **86**(10), 759–771 (2006)
32. Burgermeister, B., Arnold, M., Eichberger, A.: Smooth velocity approximation for constrained systems in real-time simulation. Multibody Syst. Dyn. **26**(1), 1–14 (2011)
33. Büskens, C., Gerdts, M.: Differentiability of consistency functions for DAE systems. J. Optim. Theory Appl. **125**(1), 37–61 (2005)
34. Campbell, S.L., Gear, C.W.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)
35. Campbell, S.L., Kelley, C.T., Yeomans, K.D.: Consistent initial conditions for unstructured higher index DAEs: a computational study. In: Computational Engineering in Systems Applications, France, pp. 416–421 (1996)
36. Cao, Y., Li, S., Petzold, L.R., Serban, R.: Adjoint sensitivity analysis for differential-algebraic equations: the adjoint DAE system and its numerical solution. SIAM J. Sci. Comput. **24**(3), 1076–1089 (2003)
37. Caracotsios, M., Stewart, W.E.: Sensitivity analysis of initial-boundary-value problems with mixed PDEs and algebraic equations. Comput. Chem. Eng. **19**(9), 1019–1030 (1985)
38. Clarke, F.H.: Optimization and Nonsmooth Analysis. Wiley, New York (1983)
39. Cuadrado, J., Cardenal, J., Bayo, E.: Modeling and solution methods for efficient real-time simulation of multibody dynamics. Multibody Syst. Dyn. **1**(3), 259–280 (1997)
40. Curtiss, C.F., Hirschfelder, J.O.: Integration of stiff equations. Proc. Nat. Acad. Sci. U.S.A. **38**, 235–243 (1952)
41. Deuflhard, P., Hairer, E., Zugck, J.: One-step and extrapolation methods for differential-algebraic systems. Numer. Math. **51**(5), 501–516 (1987)
42. Diehl, M., Bock, H.G., Schlöder, J.P., Findeisen, R., Nagy, Z., Allgöwer, F.: Real-time optimization and nonlinear model predictive control of processes governed by differential-algebraic equations. J. Process Control **12**(4), 577–585 (2002)
43. Diehl, M., Bock, H.G., Schlöder, J.P.: A real-time iteration scheme for nonlinear optimization in optimal feedback control. SIAM J. Control Optim. **43**(5), 1714–1736 (2005)
44. Dopico, D., Lugris, U., Gonzalez, M., Cuadrado, J.: Two implementations of IRK integrators for real-time multibody dynamics. Int. J. Numer. Methods Eng. **65**(12), 2091–2111 (2006)
45. Duff, I.S., Gear, C.W.: Computing the structural index. SIAM J. Algebr. Discrete Methods **7**(4), 594–603 (1986)
46. Ebrahimi, S., Eberhard, P.: A linear complementarity formulation on position level for frictionless impact of planar deformable bodies. Z. Angew. Math. Mech. **86**(10), 807–817 (2006)

47. Eich, E.: Convergence results for a coordinate projection method applied to mechanical systems with algebraic constraints. SIAM J. Numer. Anal. **30**(5), 1467–1482 (1993)
48. Eichberger, A., Rulka, W.: Process save reduction by macro joint approach: the key to real time and efficient vehicle simulation. Veh. Syst. Dyn. **41**(5), 401–413 (2004)
49. Engelhardt, L., Burger, M., Bitsch, G.: Real-time simulation of multibody systems for on-board applications. In: Proceedings of the First Joint International Conference on Multibody System Dynamics (IMSD2010) (2010)
50. Esterl, B., Butz, T., Simeon, B., Burgermeister, B.: Real-time capable vehicletrailer coupling by algorithms for differential-algebraic equations. Veh. Syst. Dyn. **45**(9), 819–834 (2007)
51. Estévez Schwarz, D.: Consistent initialization for index-2 differential-algebraic equations and its application to circuit simulation. Ph.D. thesis, Mathematisch-Naturwissenschaftlichen Fakultät II, Humboldt-Universität Berlin (2000)
52. Feehery, W.F., Tolsma, J.E., Barton, P.I.: Efficient sensitivity analysis of large-scale differential-algebraic systems. Appl. Numer. Math. **25**, 41–54 (1997)
53. Feng, A., Holland, C.D., Gallun, S.E.: Development and comparison of a generalized semi-implicit Runge–Kutta method with Gear's method for systems of coupled differential and algebraic equations. Comput. Chem. Eng. **8**(1), 51–59 (1984)
54. Fiacco, A.V.: Introduction to Sensitivity and Stability Analysis in Nonlinear Programming. Mathematics in Science and Engineering, vol. 165. Academic Press, New York (1983)
55. Fischer, A.: A special Newton-type optimization method. Optimization **24**, 269–284 (1992)
56. Führer, C.: Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen: Theorie, numerische Ansätze und Anwendungen. Ph.D. thesis, Fakultät für Mathematik und Informatik, Technische Universität München (1988)
57. Führer, C., Leimkuhler, B.J.: Numerical solution of differential-algebraic equations for constraint mechanical motion. Numer. Math. **59**, 55–69 (1991)
58. Gallrein, A., Baecker, M., Burger, M., Gizatullin, A.: An advanced flexible realtime tire model and its integration into Fraunhofer's driving simulator. SAE Technical Paper 2014-01-0861 (2014)
59. Garavello, M., Piccoli, B.: Hybrid necessary principle. SIAM J. Control Optim. **43**(5), 1867–1887 (2005)
60. Gavrea, B.I., Anitescu, M., Potra, F.A.: Convergence of a class of semi-implicit time-stepping schemes for nonsmooth rigid multibody dynamics. SIAM J. Optim. **19**(2), 969–1001 (2008)
61. Gear, C.W.: Simultaneous numerical solution of differential-algebraic equations. IEEE Trans. Circuit Theory **18**(1), 89–95 (1971)
62. Gear, C.W.: Differential-algebraic equation index transformations. SIAM J. Sci. Stat. Comput. **9**, 39–47 (1988)
63. Gear, C.W., Petzold, L.R.: ODE methods for the solution of differential/algebraic systems. SIAM J. Numer. Anal. **21**(4), 716–728 (1984)
64. Gear, C.W., Leimkuhler, B., Gupta, G.K.: Automatic integration of Euler-Lagrange equations with constraints. J. Comput. Appl. Math. **12**(13), 77–90 (1985)
65. Geier, T., Foerg, M., Zander, R., Ulbrich, H., Pfeiffer, F., Brandsma, A., van der Velde, A.: Simulation of a push belt CVT considering uni- and bilateral constraints. Z. Angew. Math. Mech. **86**(10), 795–806 (2006)
66. Gerdts, M.: Optimal control and real-time optimization of mechanical multi-body systems. Z. Angew. Math. Mech. **83**(10), 705–719 (2003)
67. Gerdts, M.: Parameter optimization in mechanical multibody systems and linearized runge-kutta methods. In: Buikis, A., Ciegis, R., Flitt, A.D. (eds.) Progress in Industrial Mathematics at ECMI 2002. Mathematics in Industry, vol. 5, pp. 121–126. Springer, Heidelberg (2004)
68. Gerdts, M.: Optimal Control of ODEs and DAEs. Walter de Gruyter, Berlin/Boston (2012)
69. Gerdts, M., Büskens, C.: Consistent initialization of sensitivity matrices for a class of parametric DAE systems. BIT Numer. Math. **42**(4), 796–813 (2002)
70. Gerdts, M., Kunkel, M.: A nonsmooth Newton's method for discretized optimal control problems with state and control constraints. J. Ind. Manag. Optim. **4**(2), 247–270 (2008)

71. Gopal, V., Biegler, L.T.: A successive linear programming approach for initialization and reinitialization after discontinuities of differential-algebraic equations. SIAM J. Sci. Comput. **20**(2), 447–467 (1998)
72. Griewank, A., Walther, A.: Evaluating Derivatives. Principles and Techniques of Algorithmic Differentiation, 2nd edn. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA (2008)
73. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems. Springer Series in Computational Mathematics, vol. 14, 2nd edn. Springer, Berlin/Heidelberg/New York (1996)
74. Hairer, E., Lubich, C., Roche, M.: Error of Rosenbrock methods for stiff problems studied via differential algebraic equations. BIT **29**(1), 77–90 (1989)
75. Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods. Lecture Notes in Mathematics, vol. 1409. Springer, Berlin/Heidelberg/New York (1989)
76. Hairer, E., Norsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer Series in Computational Mathematics, vol. 8, 2nd edn. Springer, Berlin/Heidelberg/New York (1993)
77. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations. Reprint of the Second 2006 edition. Springer, Berlin (2010)
78. Hansen, B.: Computing consistent initial values for nonlinear index-2 differential-algebraic equations. Seminarberichte Humboldt-Universität Berlin, 92-1, 142–157 (1992)
79. Heim, A.: Parameteridentifizierung in differential-algebraischen Gleichungssystemen. Master's thesis, Mathematisches Institut, Technische Universität München (1992)
80. 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**(3), 363–396 (2005)
81. INTEC GmbH. SIMPACK – Analysis and Design of General Mechanical Systems. Weßling
82. Jackiewicz, Z., Kwapisz, M.L Convergence of waveform relaxation methods for differential-algebraic systems. SIAM J. Numer. Anal. **33**(6), 2303–2317 (1996)
83. Jackson, K.R.: A survey of parallel numerical methods for initial value problems for ordinary differential equations. IEEE Trans. Magn. **27**(5), 3792–3797 (1991)
84. Jay, L.: Collocation methods for differential-algebraic equations of index 3. Numer. Math. **65**, 407–421 (1993)
85. Jay, L.: Convergence of Runge-Kutta methods for differential-algebraic systems of index 3. Appl. Numer. Math. **17**, 97–118 (1995)
86. Jiang, H.: Global convergence analysis of the generalized Newton and Gauss-Newton methods of the Fischer-Burmeister equation for the complementarity problem. Math. Oper. Res. **24**(3), 529–543 (1999)
87. Kiehl, M.: Sensitivity analysis of ODEs and DAEs - theory and implementation guide. Optim. Methods Softw. **10**(6), 803–821 (1999)
88. Kleinert, J., Simeon, B., Dreßler, K.: Nonsmooth contact dynamics for the large-scale simulation of granular material. Technical report, Fraunhofer ITWM, Kaiserslautern, Germany. J. Comput. Appl. Math. (2015, in press). http://dx.doi.org/10.1016/j.cam.2016.09.037
89. Kübler, R., Schiehlen, W.: Two methods of simulator coupling. Math. Comput. Model. Dyn. Syst. **6**(2), 93–113 (2000)
90. Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations. Analysis and Numerical Solution, vol. viii, 377 p. European Mathematical Society Publishing House, Zürich (2006)
91. Küsters, F., Ruppert, M.G.-M., Trenn, S.: Controllability of switched differential-algebraic equations. Syst. Control Lett. **78**, 32–39 (2015)
92. Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Differential-Algebraic Equations Forum. Springer, Berlin (2013)

93. Leimkuhler, B., Petzold, L.R., Gear, C.W.: Approximation methods for the consistent initialization of differential-algebraic equations. SIAM J. Numer. Anal. **28**(1), 205–226 (1991)
94. Lelarasmee, E., Ruehli, A.E., Sangiovanni-Vincentelli, A.L.: The waveform relaxation method for time-domain analysis of large scale integrated circuits. IEEE Trans. Comput. Aided Des. Integr. Circuits Syst. **1**(3), 131–145 (1982)
95. Lemke, C.E.: The dual method of solving the linear programming problem. Naval Res. Log. Q. **1**, 36–47 (1954)
96. Leyendecker, S., Ober-Blöbaum, S.: A variational approach to multirate integration for constrained systems. In: Multibody Dynamics. Computational Methods and Applications. Selected Papers Based on the Presentations at the ECCOMAS Thematic Conference, Brussels, Belgium, 4–7 July 2011, pp. 97–121. Springer, Dordrecht (2013)
97. Liberzon, D., Trenn, S.: Switched nonlinear differential algebraic equations: solution theory, Lyapunov functions, and stability. Automatica **48**(5), 954–963 (2012)
98. Linn, J., Stephan, T., Carlson, J.S., Bohlin, R.: Fast simulation of quasistatic rod deformations for VR applications. In: Bonilla, L.L., Moscoso, M., Platero, G., Vega, J.M. (eds.) Progress in Industrial Mathematics at ECMI 2006. Springer, New York (2007)
99. Lötstedt, P., Petzold, L.R.: Numerical solution of nonlinear differential equations with algebraic constraints I: convergence results for backward differentiation formulas. Math. Comput. **46**, 491–516 (1986)
100. Lubich, C., Engstler, C., Nowak, U., Pöhle, U.: Numerical integration of constrained mechanical systems using MEXX*. Mech. Struct. Mach. **23**(4), 473–495 (1995)
101. Maly, T., Petzold, L.R.: Numerical methods and software for sensitivity analysis of differential-algebraic systems. Appl. Numer. Math. **20**(1), 57–79 (1996)
102. Michael, J., Gerdts, M.: A method to model impulsive multi-body-dynamics using Riemann-Stieltjes- Integrals. In: 8th Vienna International Conference on Mathematical Modelling, International Federation of Automatic Control, pp. 629–634 (2015)
103. Michael, J., Chudej, K., Gerdts, M., Pannek, J.: Optimal rendezvous path planning to an uncontrolled tumbling target. In: IFAC Proceedings Volumes (IFAC-PapersOnline), 19th IFAC Symposium on Automatic Control in Aerospace, ACA 2013, Wurzburg, Germany, 2–6 September 2013, vol. 19, pp. 347–352 (2013)
104. Miekkala, U., Nevanlinna, O.: Convergence of dynamic iteration methods for initial value problems. SIAM J. Sci. Stat. Comput. **8**(4), 459–482 (1987)
105. Murua, A.: Partitioned half-explicit Runge–Kutta methods for differential-algebraic systems of index 2. Computing **59**(1), 43–61 (1997)
106. Negrut, D., Sandu, A., Haug, E.J., Potra, F.A., Sandu, C.: A Rosenbrock-Nystrom state space implicit approach for the dynamic analysis of mechanical systems: II –the method and numerical examples. J. Multi-body Dyn. **217**(4), 273–281 (2003)
107. Ostermann, A.: A class of half-explicit Runge–Kutta methods for differential-algebraic systems of index 3. Appl. Numer. Math. **13**(1), 165–179 (1993)
108. Pantelides, C.C.: The consistent initialization of differential-algebraic systems. SIAM J. Sci. Stat. Comput. **9**(2), 213–231 (1988)
109. Petzold, L.R.: A description of DASSL: a differential/algebraic system solver. Rep. Sand 82-8637, Sandia National Laboratory, Livermore (1982)
110. Petzold, L.R.: Differential/algebraic equations are not ODE's. SIAM J. Sci. Stat. Comput. **3**(3), 367–384 (1982)
111. Petzold, L.R.: Recent developments in the numerical solution of differential/algebraic systems. Comput. Methods Appl. Mech. Eng. **75**, 77–89 (1989)
112. Pfeiffer, A.: Numerische Sensitivitätsanalyse unstetiger multidisziplinärer Modelle mit Anwendungen in der gradientenbasierten Optimierung. Fortschritt-Berichte VDI Reihe 20, Nr. 417. VDI–Verlag, Düsseldorf (2008)
113. Potra, F.A., Anitescu, M., Gavrea, B., Trinkle, J.: A linearly implicit trapezoidal method for integrating stiff multibody dynamics with contact, joints, and friction. Int. J. Numer. Methods Eng. **66**(7), 1079–1124 (2006)

114. Pytlak, R., Suski, D.: On solving hybrid optimal control problems with higher index DAEs. Institute of Automatic Control and Robotics, Warsaw University of Technology, Warsaw, Poland (2015, Preprint)
115. Qi, L.: Convergence analysis of some algorithms for solving nonsmooth equations. Math. Oper. Res. **18**(1), 227–244 (1993)
116. Qi, L., Sun, J.: A nonsmooth version of Newton's method. Math. Program. **58**(3), 353–367 (1993)
117. Rentrop, P., Roche, M., Steinebach, G.: The application of Rosenbrock-Wanner type methods with stepsize control in differential-algebraic equations. Numer. Math. **55**(5), 545–563 (1989)
118. Rill, G.: A modified implicit Euler algorithm for solving vehicle dynamic equations. Multibody Syst. Dyn. **15**(1), 1–24 (2006)
119. Rill, G., Chucholowski, C.: Real time simulation of large vehicle systems. In: Proceedings of Multibody Dynamics 2007 (ECCOMAS Thematic Conference) (2007)
120. Roche, M.: Rosenbrock methods for differential algebraic equations. Numer. Math. **52**(1), 45–63 (1988)
121. Rosenbrock, H.H.: Some general implicit processes for the numerical solution of differential equations. Comput. J. **5**(4), 329–330 (1963)
122. Rulka, W., Pankiewicz, E.: MBS approach to generate equations of motions for hil-simulations in vehicle dynamics. Multibody Syst. Dyn. **14**(3), 367–386 (2005)
123. Sandu, A., Negrut, D., Haug, E.J., Potra, F.A., Sandu, C.: A Rosenbrock-Nystrom state space implicit approach for the dynamic analysis of mechanical systems: I—theoretical formulation. J. Multi-body Dyn. **217**(4), 263–271 (2003)
124. Schaub, M., Simeon, B.: Blended Lobatto methods in multibody dynamics. Z. Angew. Math. Mech. **83**(10), 720–728 (2003)
125. Schierz, T., Arnold, M.: Stabilized overlapping modular time integration of coupled differential-algebraic equations. Appl. Numer. Math. **62**(10), 1491–1502 (2012). Selected Papers from NUMDIFF-12
126. Schneider, F., Burger, M., Arnold, M., Simeon, B.: A new approach for force-displacement co-simulation using kinematic coupling constraints. Submitted to Z. Angew. Math. Mech. (2016)
127. 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**(2), 440–467 (1998)
128. Schwartz, W., Frik, S., Leister, G.: Simulation of the IAVSD Road Vehicle Benchmark Bombardier Iltis with FASIM, MEDYNA, NEWEUL and SIMPACK. Technical Report IB 515/92-20, Robotik und Systemdynamik, Deutsche Forschungsanstalt für Luft- und Raumfahrt (1992)
129. Schweizer, B., Lu, D.: Stabilized index-2 co-simulation approach for solver coupling with algebraic constraints. Multibody Syst. Dyn. **34**(2), 129–161 (2014)
130. Schweizer, B., Li, P., Lu, D.: Implicit co-simulation methods: stability and convergence analysis for solver coupling approaches with algebraic constraints. Z. Angew. Math. Mech. **96**(8), 986–1012 (2016)
131. Stetter, H.J.: Analysis of Discretization Methods for Ordinary Differential Equations. Springer Tracts in Natural Philosophy, vol. 23. Springer, Berlin/Heidelberg/New York (1973)
132. Stewart, D.E.: Rigid-body dynamics with friction and impact. SIAM Rev. **42**(1), 3–39 (2000)
133. Stewart, D.E., Anitescu, M.: Optimal control of systems with discontinuous differential equations. Numer. Math. **114**(4), 653–695 (2010)
134. Strehmel, K., Weiner, R.: Numerik gewöhnlicher Differentialgleichungen. Teubner, Stuttgart (1995)
135. Strehmel, K., Weiner, R., Dannehl, I.: On error behaviour of partitioned linearly implicit Runge–Kutta methods for stiff and differential algebraic systems. BIT **30**(2), 358–375 (1990)
136. Sussmann, H.J.: A nonsmooth hybrid maximum principle. In: Stability and Stabilization of Nonlinear Systems. Proceedings of the 1st Workshop on Nonlinear Control Network, Held in Gent, Belgium, 15–16 March 1999, pp. 325–354. Springer, London (1999)

137. Tasora, A., Anitescu, M.: A fast NCP solver for large rigid-body problems with contacts, friction, and joints. In: Multibody Dynamics. Computational Methods and Applications. Revised, extended and selected papers of the ECCOMAS Thematic Conference on Multibody Dynamics 2007, Milano, Italy, 25–28 June 2007, pp. 45–55. Springer, Dordrecht (2009)
138. Tasora, A., Anitescu, M.: A matrix-free cone complementarity approach for solving large-scale, nonsmooth, rigid body dynamics. Comput. Methods Appl. Mech. Eng. **200**(5–8), 439–453 (2011)
139. Tasora, A., Anitescu, M.: A complementarity-based rolling friction model for rigid contacts. Meccanica **48**(7), 1643–1659 (2013)
140. Tasora, A., Negrut, D., Anitescu, M.: GPU-based parallel computing for the simulation of complex multibody systems with unilateral and bilateral constraints: an overview. In: Multibody Dynamics. Computational Methods and Applications. Selected papers based on the presentations at the ECCOMAS Conference on Multibody Dynamics, Warsaw, Poland, June 29–July 2, 2009, pp. 283–307. Springer, New York, NY (2011)
141. Trenn, S.: Solution concepts for linear DAEs: a survey. In: Ilchmann, A., Reis, T. (eds.) Surveys in Differential-Algebraic Equations I. Differential-Algebraic Equations Forum, pp. 137–172. Springer, Berlin (2013)
142. van der Schaft, A., Schumacher, H.: An Introduction to Hybrid Dynamical Systems. Springer, London (1989)
143. Veitl, A., Gordon, T., van de Sand, A., Howell, M., Valasek, M., Vaculin, O., Steinbauer, P.: Methodologies for coupling simulation models and codes in mechatronic system analysis and design. In: Proceedings of the 16th IAVSD Symposium on Dynamics of Vehicles on Roads and Tracks. Pretoria. Supplement to Vehicle System Dynamics, vol. 33, pp. 231–243. Swets & Zeitlinger (1999)
144. von Schwerin, R.: Multibody System Simulation: Numerical Methods, Algorithms, and Software. Lecture Notes in Computational Science and Engineering, vol. 7. Springer, Berlin/Heidelberg/New York (1999)
145. Wensch, J.: An eight stage fourth order partitioned Rosenbrock method for multibody systems in index-3 formulation. Appl. Numer. Math. **27**(2), 171–183 (1998)
146. Wensch, J., Strehmel, K., Weiner, R.: A class of linearly-implicit Runge–Kutta methods for multibody systems. Appl. Numer. Math. **22**(13), 381–398 (1996). Special Issue Celebrating the Centenary of Runge–Kutta Methods
147. Wolfbrandt, A., Steihaug, T.: An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations. Math. Comput. **33**(146), 521–534 (1979)

Index

A. Ilchmann, T. Reis (eds.), *Surveys in Differential-Algebraic Equations IV*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-319-46618-7