
S E
M S M
E S

EMS Textbooks in Mathematics

EMS Textbooks in Mathematics is a book series aimed at students or professional mathematicians seeking an introduction into a particular field. The individual volumes are intended to provide not only relevant techniques, results and their applications, but afford insight into the motivations and ideas behind the theory. Suitably designed exercises help to master the subject and prepare the reader for the study of more advanced and specialized literature.

Jørn Justesen and Tom Høholdt, *A Course In Error-Correcting Codes*
Markus Stroppel, *Locally Compact Groups*

Peter Kunkel
Volker Mehrmann

Differential-Algebraic Equations

Analysis and Numerical Solution

European Mathematical Society

Authors:

Peter Kunkel
Fakultät für Mathematik und Informatik
Universität Leipzig
Augustusplatz 10/11
D-04109 Leipzig
Germany

Volker Mehrmann
Institut für Mathematik, MA 4-5
TU Berlin
Strasse des 17. Juni 136
D-10623 Berlin
Germany

2000 Mathematical Subject Classification (primary; secondary):
34A09, 65L80

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie;
detailed bibliographic data are available in the Internet at http://dnb.ddb.de.

ISBN 3-03719-017-5

Contact address:

European Mathematical Society Publishing House
Seminar for Applied Mathematics
ETH-Zentrum FLI C4
CH-8092 Zürich
Switzerland

Phone: +41 (0)1 632 34 36
Email: info@ems-ph.org
Homepage: www.ems-ph.org

Typeset using the authors' TEX files: I. Zimmermann, Freiburg
Printed on acid-free paper produced from chlorine-free pulp. TCF ∞
Printed in Germany

9 8 7 6 5 4 3 2 1

Preface

In the last 30 years, differential-algebraic equations have become a widely accepted tool for the modeling and simulation of constrained dynamical systems in numerous applications, such as mechanical multibody systems, electrical circuit simulation, chemical engineering, control theory, fluid dynamics, and many other areas.

Although there has been a tremendous explosion in the research literature in the area of differential-algebraic equations, there are only very few monographs and essentially no textbooks devoted to this subject. This is mostly due to the fact that the research in this area is still very active and some of the major issues were still under development. This concerns the analysis as well as the numerical solution of such problems and in particular the modeling with differential-algebraic equations in various applications.

At this time, however, we feel that both theory and numerical methods have reached a stage of maturity that should be presented in a regular textbook. In particular, we provide a systematic and detailed analysis of initial and boundary value problems for differential-algebraic equations. We also discuss numerical methods and software for the solution of these problems. This includes linear and nonlinear problems, over- and underdetermined problems as well as control problems, and problems with structure.

We thank R. Janßen and the IBM research center in Heidelberg for giving us the opportunity to start our joint research on differential-algebraic equations in 1988/1989 by placing us for nine months in the same office. We especially appreciate the support by the Deutsche Forschungsgemeinschaft, which we got for our joint research in the years 1993–2001. Major influence for the development of our research came from the biannual workshops on descriptor systems organized by P. C. Müller in Paderborn and from discussions with friends and colleagues at numerous meetings and colloquia in the last fifteen years.

We also thank W. Hackbusch for initiating the idea to write this textbook and our colleagues and students S. Bächle, K. Biermann, B. Benhammouda, F. Ebert, D. Kreßner, J. Liesen, L. Poppe, W. Rath, T. Reis, S. Schlauch, M. Schmidt, C. Schröder, I. Seufer, A. Steinbrecher, R. Stöver, C. Shi, T. Stykel, E. Virnik, J. Weickert, and L. Wunderlich for many discussions and comments and their help in proofreading, programming and keeping software and bibliography files up-to-date.

Contents

Part I

Analysis of differential-algebraic equations

Chapter 1
Introduction

The dynamical behavior of physical processes is usually modeled via differential equations. But if the states of the physical system are in some ways constrained, like for example by conservation laws such as Kirchhoff's laws in electrical networks, or by position constraints such as the movement of mass points on a surface, then the mathematical model also contains algebraic equations to describe these constraints. Such systems, consisting of both differential and algebraic equations are called *differential-algebraic systems, algebro-differential systems*, *implicit differential equations* or *singular systems*.

The most general form of a differential-algebraic equation is

$$F(t, x, \dot{x}) = 0, \tag{1.1}$$

with $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{C}^m$, where $\mathbb{I} \subseteq \mathbb{R}$ is a (compact) interval and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{C}^n$ are open, $m, n \in \mathbb{N}$. The meaning of the quantity $\dot{x}$ is ambiguous as in the case of ordinary differential equations. On one hand, it denotes the derivative of a differentiable function $x : \mathbb{I} \to \mathbb{C}^n$ with respect to its argument $t \in \mathbb{I}$. On the other hand, in the context of (1.1), it is used as an independent variable of F. The reason for this ambiguity is that we want F to determine a differentiable function x that solves (1.1) in the sense that $F(t, x(t), \dot{x}(t)) = 0$ for all $t \in \mathbb{I}$.

In connection with (1.1), we will discuss the question of existence of solutions. Uniqueness of solutions will be considered in the context of initial value problems, when we additionally require a solution to satisfy the condition

$$x(t_0) = x_0 \tag{1.2}$$

with given $t_0 \in \mathbb{I}$ and $x_0 \in \mathbb{C}^n$, and boundary value problems, where the solution is supposed to satisfy

$$b(x(\underline{t}), x(\bar{t})) = 0 \tag{1.3}$$

with $b : \mathbb{D}_x \times \mathbb{D}_x \to \mathbb{C}^d$, $\mathbb{I} = [\underline{t}, \bar{t}]$ and some problem dependent integer d. It will turn out that the properties of differential-algebraic equations reflect the properties of differential equations as well as the properties of algebraic equations, but also that other phenomena occur which result from the mixture of these different types of equations.

Although the basic theory for linear differential-algebraic equations with constant coefficients

$$E\dot{x} = Ax + f(t), \tag{1.4}$$

where $E, A \in \mathbb{C}^{m,n}$ and $f : \mathbb{I} \to \mathbb{C}^m$, has already been established in the nineteenth century by the fundamental work of Weierstraß [223], [224] and Kronecker [121] on matrix pencils, it took until the pioneering work of Gear [90] for the scientific communities in mathematics, computer science, and engineering to realize the large potential of the theory of differential-algebraic equations in modeling dynamical systems. By this work and the subsequent developments in numerical methods for the solution of differential-algebraic equations, it became possible to use differential-algebraic equations in direct numerical simulation. Since then an explosion of the research in this area has taken place and has led to a wide acceptance of differential-algebraic equations in the modeling and simulation of dynamical systems. Despite the wide applicability and the great importance, only very few monographs and essentially no textbooks are so far devoted to this subject, see [29], [100], [105], [182]. Partially, differential-algebraic equations are also discussed in [11], [42], [43], [72], [79], [108], [181].

Until the work of Gear, implicit systems of the form (1.1) were usually transformed into ordinary differential equations

$$\dot{y} = g(t, y) \tag{1.5}$$

via analytical transformations. One way to achieve this is to explicitly solve the constraint equations analytically in order to reduce the given differential-algebraic equation to an ordinary differential equation in fewer variables. But this approach heavily relies on either transformations by hand or symbolic computation software which are both not feasible for medium or large scale systems.

Another possibility is to differentiate the algebraic constraints in order to get an ordinary differential equation in the same number of variables. Due to the necessary use of the implicit function theorem, this approach is often difficult to perform. Moreover, due to possible changes of bases, the resulting variables may have no physical meaning. In the context of numerical solution methods, it was observed in this approach that the numerical solution may drift off from the constraint manifold after a few integration steps. For this reason, in particular in the simulation of mechanical multibody systems, stabilization techniques were developed to address this difficulty. But it is in general preferable to develop methods that operate directly on the given differential-algebraic equation.

In view of the described difficulties, the development of numerical methods that can be directly applied to the differential-algebraic equation has been the subject of a large number of research projects in the last thirty years and many different directions have been taken. In particular, in combination with modern modeling tools (that automatically generate models for substructures and link them together via constraints), it is important to develop generally applicable numerical methods as well as methods that are tailored to a specific physical situation. It would be ideal if such an automatically generated model could be directly transferred to a

numerical simulation package via an appropriate interface so that in practical design problems the engineer can optimize the design via a sequence of modeling and simulation steps. To obtain such a general solution package for differential-algebraic equations is an active area of current research that requires strong interdisciplinary cooperation between researchers working in modeling, the development of numerical methods, and the design of software. A major difficulty in this context is that still not all of the analytical and numerical properties of differential-algebraic systems are completely understood. In particular, the treatment of bifurcations or switches in nonlinear systems and the analysis and numerical solution of heterogeneous (coupled) systems combined of differential-algebraic equations and partial differential equations (sometimes called partial differential-algebraic equations) represent major research tasks.

It is the purpose of this textbook to give a coherent introduction to the theoretical analysis of differential-algebraic equations and to present some appropriate numerical methods for initial and boundary value problems. For the analysis of differential-algebraic equations, there are several paths that can be followed. A very general approach is given by the geometrical analysis initiated by Rheinboldt [190], see also [182], to study differential-algebraic equations as differential equations on manifolds. We will discuss this topic in Section 4.5. Our main approach, however, will be the algebraic path that leads from the theory of matrix pencils by Weierstraß and Kronecker via the fundamental work of Campbell on derivative arrays [44] to canonical forms for linear variable coefficient systems [123], [124] and their extensions to nonlinear systems in the work of the authors ([128], [129], [131], [132]).

This algebraic approach not only gives a systematic approach to the classical analysis of regular differential-algebraic equations, but it also allows the study of generalized solutions and the treatment of over- and underdetermined systems as well as control problems. At the same time, it leads to new discretization methods and new numerical software.

Unfortunately, the simultaneous development of the theory in many different research groups has led to a large number of slightly different existence and uniqueness results, particularly based on different concepts of the so-called index. The general idea of all these index concepts is to measure the degree of smoothness of the problem that is needed to obtain existence and uniqueness results. To set our presentation in perspective, we now briefly discuss the most common approaches.

1.1 Solvability concepts

In order to develop a theoretical analysis for system (1.1), one has to specify the kind of solution that one is interested in, i.e., the function space in which the solution should lie. In this textbook, we will mainly discuss two concepts, namely classical

(continuously differentiable) solutions and weak (distributional) solutions, although other concepts have been studied in the literature, see, e.g., [148], [149].

For the classical case, we use the following solvability definition, the distributional case will be discussed in detail in Sections 2.4 and 3.5.

Definition 1.1. Let $C^k(\mathbb{I}, \mathbb{C}^n)$ denote the vector space of all k-times continuously differentiable functions from the real interval $\mathbb{I}$ into the complex vector space $\mathbb{C}^n$.

1. A function $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ is called a *solution* of (1.1), if it satisfies (1.1) pointwise.
2. The function $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ is called a *solution of the initial value problem* (1.1) with initial condition (1.2), if it furthermore satisfies (1.2).
3. An initial condition (1.2) is called *consistent* with F, if the associated initial value problem has at least one solution.

In the following, a problem is called *solvable* if it has at least one solution. This definition seems natural but it should be noted that in most of the previous literature, the term solvability is used only for systems which have a unique solution when consistent initial conditions are provided. For comparison with Definition 1.1, consider the solvability condition given in [29, Def. 2.2.1].

If the solution of the initial value problem is not unique which is, in particular, the case in the context of control problems, then further conditions have to be specified to single out specific desired solutions. We will discuss such conditions in Section 3.4 and in the context of control problems in Sections 2.5, 3.6, and 4.4.

1.2 Index concepts

In the analysis of linear differential-algebraic equations with constant coefficients (1.4), all properties of the system can be determined by computing the invariants of the associated matrix pair (E, A) under equivalence transformations. In particular, the size of the largest Jordan block to an infinite eigenvalue in the associated Kronecker canonical form [88], called *index*, plays a major role in the analysis and determines (at least in the case of so-called regular pairs) the smoothness that is needed for the inhomogeneity f in (1.4) to guarantee the existence of a classical solution. Motivated by this case, it was first tried to define an analogous index for linear time-varying systems and then for general implicit systems, see [95]. However, it was soon realized that a direct generalization by linearization and consideration of the local linearized constant coefficient system does not lead to a reasonable concept. The reason is that important invariants of constant coefficient systems are not even locally invariant under nonconstant equivalence transformations. This observation led to a multitude of different index concepts even for linear

systems with variable coefficients, see [53]. Among the different approaches, the *differentiation index* and the *perturbation index* are currently the most widely used concepts in the literature. We will give formal definitions in Sections 3.3 and 3.4, respectively.

Loosely speaking, the differentiation index is the minimum number of times that all or part of (1.1) must be differentiated with respect to t in order to determine $\dot{x}$ as a continuous function of t and x. The motivation for this definition is historically based on the procedure to solve the algebraic equations (using their derivatives if necessary) by transforming the implicit system to an ordinary differential equation. Although the concept of the differentiation index is widely used, it has a major drawback, since it is not suited for over- and underdetermined systems. The reason for this is that it is based on a solvability concept that requires unique solvability. In our presentation, we will therefore focus on the concept of the *strangeness index* [123], [128], [129], [132], which generalizes the differentiation index to over- and underdetermined systems. We will not discuss other index concepts such as the *geometric index* [190], the *tractability index* [100], [148], [149] or the *structural index* [161]. A different index concept that is of great importance in the numerical treatment of differential-algebraic equations is the *perturbation index* that was introduced in [105] to measure the sensitivity of solutions with respect to perturbations of the problem. For a detailed analysis and a comparison of various index concepts with the differentiation index, see [53], [92], [147], [150], [156], [189].

At this point, it seems appropriate to introduce some philosophical discussion concerning the counting in the different index definitions. First of all, the motivation to introduce an index is to classify different types of differential-algebraic equations with respect to the difficulty to solve them analytically as well as numerically. In view of this classification aspect, the differentiation index was introduced to determine how far the differential-algebraic equation is away from an ordinary differential equation, for which the analysis and numerical techniques are well established. But purely algebraic equations, which constitute another important special case of (1.1), are equally well analyzed. Furthermore, it would certainly not make sense to turn a uniquely solvable classical linear system $Ax = b$ into a differential equation, since then the solution would not be unique anymore without specifying initial conditions. In view of this discussion, it seems desirable to classify differential-algebraic equations by their distance to a decoupled system of ordinary differential equations and purely algebraic equations. Hence, from our point of view, the index of an ordinary differential equation and that of a system of algebraic equations should be the same.

This differs from the differentiation index, for which an ordinary differential equation has index zero, while an algebraic equation has index one. Although the research community and also people working in applications have widely accepted this way of counting, in the concept of the strangeness index ordinary differential

equations and purely algebraic equations both have index zero. We will present further arguments for this way of counting on several occasions throughout this textbook.

1.3 Applications

We will now discuss some elementary examples of differential-algebraic equations arising in applications such as electrical networks, multibody systems, chemical engineering, semidiscretized Stokes equations and others.

Let us first consider an example arising in electrical circuit simulation. For this topic, there is an extensive literature that includes the classification of properties of the arising differential-algebraic equations depending on the components of the network, see, e.g., [18], [83], [84], [103], [104], [214].

Example 1.2. To obtain a mathematical model for the charging of a capacitor via a resistor, we associate a potential x_i, $i = 1, 2, 3$, with each node of the circuit, see Figure 1.1. The voltage source increases the potential x_3 to x_1 by U, i.e.,

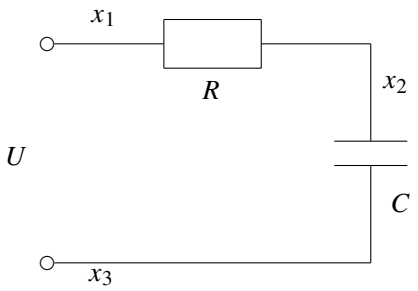


Figure 1.1. A simple electrical network

$x_1 - x_3 - U = 0$. By Kirchhoff's first law, the sum of the currents vanishes in each node. Hence, assuming ideal electronic units, for the second node we obtain that $C(\dot{x}_3 - \dot{x}_2) + (x_1 - x_2)/R = 0$, where R is the size of the resistance of the resistor and C is the capacity of the capacitor. By choosing the zero potential as $x_3 = 0$, we obtain as a mathematical model the differential-algebraic system

$$\begin{aligned} x_1 - x_3 - U &= 0, \\ C(\dot{x}_3 - \dot{x}_2) + (x_1 - x_2)/R &= 0, \\ x_3 &= 0. \end{aligned} \tag{1.6}$$

It is clear that this simple system can be solved for x_3 and x_1 to obtain an ordinary differential equation for x_2 only, combined with algebraic equations for x_1, x_3. This system has differentiation index one.

A second major application area is the simulation of the dynamics of multibody systems, see, e.g., [79], [196], [201], [205].

Example 1.3. A physical pendulum is modeled by the movement of a mass point with mass m in Cartesian coordinates (x, y) under the influence of gravity in a distance l around the origin, see Figure 1.2. With the kinetic energy $T = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2)$ and the potential energy $U = mgy$, where g is the gravity

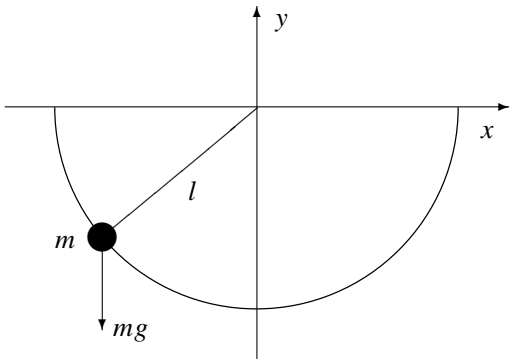


Figure 1.2. A mechanical multibody system

constant, using the constraint equation $x^2 + y^2 - l^2 = 0$, we obtain the Lagrange function

$$L = \frac{1}{2}m(\dot{x}^2 + \dot{y}^2) - mgy - \lambda(x^2 + y^2 - l^2)$$

with Lagrange parameter λ. The equations of motion then have the form

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}}\right) - \frac{\partial L}{\partial q} = 0$$

for the variables $q = x, y, \lambda$, i.e.,

$$\begin{aligned} m\ddot{x} + 2x\lambda &= 0, \\ m\ddot{y} + 2y\lambda + mg &= 0, \\ x^2 + y^2 - l^2 &= 0. \end{aligned} \tag{1.7}$$

It is obvious that this system cannot have differentiation index one, it actually has differentiation index three.

Differential-algebraic equations are also frequently used in the mathematical modeling of chemical reactions, see, e.g., [161].

Example 1.4. Consider the model of a chemical reactor in which a first order isomerization reaction takes place and which is externally cooled.

Denoting by c_0 the given feed reactant concentration, by T_0 the initial temperature, by $c(t)$ and $T(t)$ the concentration and temperature at time t, and by R the reaction rate per unit volume, the model takes the form

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{c} \\ \dot{T} \\ \dot{R} \end{bmatrix} = \begin{bmatrix} k_1(c_0 - c) - R \\ k_1(T_0 - T) + k_2 R - k_3(T - T_C) \\ R - k_3 \exp\left(-\frac{k_4}{T}\right) c \end{bmatrix}, \tag{1.8}$$

where T_C is the cooling temperature (which can be used as control input) and k_1, k_2, k_3, k_4 are constants. If T_C is given, this system has differentiation index one. If T_C is treated as a control variable, the system is underdetermined and the differentiation index is not defined.

Another common source of differential-algebraic equations is the semidiscretization of systems of partial differential equations or coupled systems of partial differential equations and other types of equations, see, e.g., [5], [11], [102], [202].

Example 1.5. The nonstationary Stokes equation is a classical linear model for the laminar flow of a Newtonian fluid [225]. It is described by the partial differential equation

$$u_t = \Delta u + \nabla p, \quad \nabla \cdot u = 0, \tag{1.9}$$

together with initial and boundary conditions. Here u describes the velocity and p the pressure of the fluid. Using the method of lines [209], [211] and discretizing first the space variables with finite element or finite difference methods typically leads to a linear differential-algebraic system of the form

$$\dot{u}_h = Au_h + Bp_h, \quad B^T u_h = 0, \tag{1.10}$$

where u_h and p_h are semi-discrete approximations for u and p. If the nonuniqueness of a free constant in the pressure is fixed by the discretization method, then the differentiation index is well defined for this system. For most discretization methods, it is two, see, e.g., [222].

The study of classical control problems in the modern behavior framework [132], [167] immediately leads to underdetermined DAEs.

Example 1.6. The classical linear control problem to find an input function u that stabilizes the linear control system

$$\dot{x} = Ax + Bu, \quad x(t_0) = x_0 \tag{1.11}$$

can be viewed in the so-called behavior context ([116], [117], [167]) as determining a solution of the underdetermined linear differential-algebraic equation

$$[\, I \;\; 0 \,]\dot{z} = [\, A \;\; B \,]z, \quad [\, I \;\; 0 \,]z(t_0) = x_0 \tag{1.12}$$

such that for $z = \begin{bmatrix} x \\ u \end{bmatrix}$ the part $[\, I \;\; 0 \,]z$ is asymptotically stable.

Differential-algebraic equations also play an important role in the analysis and numerical solution of singular perturbation problems, where they represent the limiting case, see, e.g., [108], [159], [221].

Example 1.7. The van der Pol equation

$$\begin{aligned} \dot{y} &= z, \\ \epsilon \dot{z} &= (1 - y^2)z - y, \end{aligned} \tag{1.13}$$

possesses the differential-algebraic equation

$$\begin{aligned} \dot{y} &= z, \\ 0 &= (1 - y^2)z - y \end{aligned} \tag{1.14}$$

as limiting case for $\epsilon \to 0$. The analysis and understanding of (1.14) is essential in the construction of numerical methods that can solve the equation (1.13) for small parameters ϵ.

Many more application areas could be mentioned here, but these few examples already demonstrate the wide applicability of differential-algebraic equations in the mathematical modeling and the numerical solution of application problems.

1.4 How to use this book in teaching

This book is laid out to be and has been used in teaching graduate courses in several different ways.

Chapters 2, 3, and 4 together form a one semester course (approximately 60 teaching hours) on the analysis of differential-algebraic equations. As a prerequisite for such a course, one would need the level that is reached after a first course on the theory of ordinary differential equations.

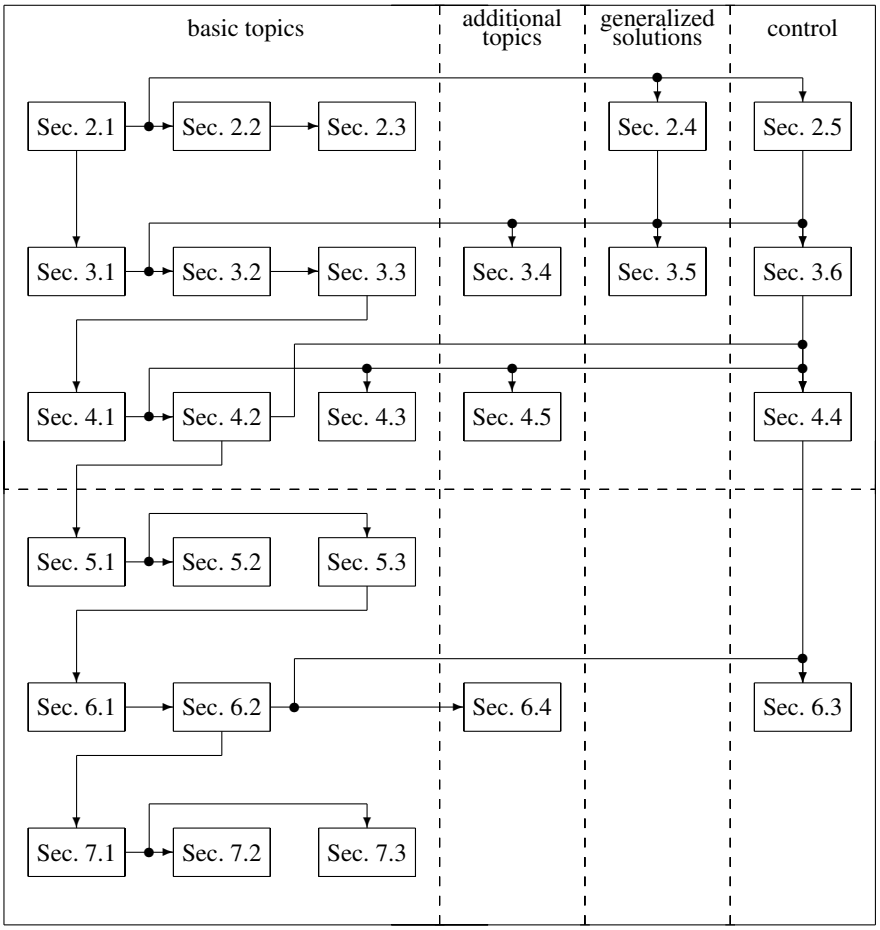
A course with smaller volume is formed by omitting Sections 2.5, 3.6, and 4.4 on control that depend on each other in this order but are not needed for other sections. An even shorter course is obtained by omitting Sections 2.4 and 3.5 on generalized solutions which depend on each other in this order but again are not needed for other sections. Section 3.4 on generalized inverses is useful for the sections on control but not needed for other sections and can therefore also be omitted to shorten the course. Section 4.2 on structured problems and Section 4.5 on differential equations on manifolds again are not needed for other sections and could be omitted.

A combined one semester course (approximately 60 teaching hours) on the analysis and numerical solution of differential-algebraic equations would need as a prerequisite the level that is reached after a first course on the theory of ordinary differential equations as well as a first course on numerical analysis including the

basics of the numerical solution of ordinary differential equations. Such a course would consist of Chapter 2, Section 2.1, Chapter 3, Sections 3.1, 3.2, 3.3, and Chapter 4, Sections 4.1, 4.2, 4.3, concerning the analysis and Chapter 5, Chapter 6, Sections 6.1, 6.2, and Chapter 7 concerning the numerical solution of differential-algebraic equations.

The numerical part of the book, which strongly relies on the analysis part, would represent a separate course (approximately 30 teaching hours) on the numerical solution of initial and boundary value problems for differential-algebraic equations that includes Chapter 5, Chapter 6 Sections 6.1, 6.2 and Chapter 7. A slightly extended course would combine these with Sections 6.3 and 6.4.

The scheme below displays the dependencies between the different sections and may help to organize courses on the basis of this textbook.



Chapter 2
Linear differential-algebraic equations with constant coefficients

In this chapter, we consider linear differential-algebraic equations with constant coefficients of the form

$$E\dot{x} = Ax + f(t), \tag{2.1}$$

with $E, A \in \mathbb{C}^{m,n}$ and $f \in C(\mathbb{I}, \mathbb{C}^m)$, possibly together with an initial condition

$$x(t_0) = x_0. \tag{2.2}$$

Such equations occur for example by linearization of autonomous nonlinear problems with respect to constant (or critical) solutions, where f plays the role of a perturbation.

2.1 Canonical forms

The properties of (2.1) are well understood for more than one century, in particular by the work of Weierstraß [223] and Kronecker [121]. The reason is that (2.1) can be treated by purely algebraic techniques. In the following, we describe the main aspects of this approach.

Scaling (2.1) by a nonsingular matrix $P \in \mathbb{C}^{m,m}$ and the function x according to $x = Q\tilde{x}$ with a nonsingular matrix $Q \in \mathbb{C}^{n,n}$, we obtain

$$\tilde{E}\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{f}(t), \quad \tilde{E} = PEQ, \quad \tilde{A} = PAQ, \quad \tilde{f} = Pf, \tag{2.3}$$

which is again a linear differential-algebraic equation with constant coefficients. Moreover, the relation $x = Q\tilde{x}$ gives a one-to-one correspondence between the corresponding solution sets. This means that we can consider the transformed problem (2.3) instead of (2.1) with respect to solvability and related questions. The following definition of equivalence is now evident.

Definition 2.1. Two pairs of matrices (E_i, A_i), $E_i, A_i \in \mathbb{C}^{m,n}$, $i = 1, 2$, are called (*strongly*) *equivalent* if there exist nonsingular matrices $P \in \mathbb{C}^{m,m}$ and $Q \in \mathbb{C}^{n,n}$, such that

$$E_2 = PE_1Q, \quad A_2 = PA_1Q. \tag{2.4}$$

If this is the case, we write $(E_1, A_1) \sim (E_2, A_2)$.

Similar definitions can be found in the literature in the context of matrix pencils, linear matrix functions, or generalized eigenvalue problems, often with the notation $\lambda E - A$ or $\alpha E - \beta A$ instead of (E, A), see, e.g., [88], [99].

As already suggested by the definition, relation (2.4) fixes an equivalence relation.

Lemma 2.2. *The relation introduced in Definition 2.1 is an equivalence relation.*

Proof. We must show that the relation is reflexive, symmetric, and transitive.

Reflexivity: We have $(E, A) \sim (E, A)$ by $P = I_m$ and $Q = I_n$.

Symmetry: From $(E_1, A_1) \sim (E_2, A_2)$, it follows that $E_2 = PE_1Q$ and $A_2 = PA_1Q$ with nonsingular matrices P and Q. Hence, we have $E_1 = P^{-1}E_2Q^{-1}$, $A_1 = P^{-1}A_2Q^{-1}$ implying that $(E_2, A_2) \sim (E_1, A_1)$.

Transitivity: From $(E_1, A_1) \sim (E_2, A_2)$ and $(E_2, A_2) \sim (E_3, A_3)$ it follows that $E_2 = P_1E_1Q_1$, $A_2 = P_1A_1Q_1$ and $E_3 = P_2E_2Q_2$, $A_3 = P_2A_2Q_2$ with nonsingular matrices P_i and Q_i, $i = 1, 2$. Eliminating E_2, A_2 gives $E_3 = P_2P_1E_1Q_1Q_2$, $A_3 = P_2P_1A_1Q_1Q_2$, such that $(E_1, A_1) \sim (E_3, A_3)$. □

Having defined an equivalence relation, the standard procedure then is to look for a canonical form, i.e., to look for a matrix pair which is equivalent to a given matrix pair and which has a simple form from which we can directly read off the properties and invariants of the corresponding differential-algebraic equation. In the present case, such a canonical form is represented by the so-called *Kronecker canonical form*. It is, however, quite technical to derive this canonical form, see, e.g., [88]. We therefore restrict ourselves here to a special case and only present the general result without proof. In the next chapter, we will derive a canonical form for linear differential-algebraic equations with variable coefficients which will generalize the Kronecker canonical form at least in the sense that existence and uniqueness results can be obtained in the same way as from the Kronecker canonical form for the case of constant coefficients.

Theorem 2.3. *Let $E, A \in \mathbb{C}^{m,n}$. Then there exist nonsingular matrices $P \in \mathbb{C}^{m,m}$ and $Q \in \mathbb{C}^{n,n}$ such that (for all $\lambda \in \mathbb{C}$)*

$$P(\lambda E - A)Q = \operatorname{diag}(\mathcal{L}_{\epsilon_1}, \ldots, \mathcal{L}_{\epsilon_p}, \mathcal{M}_{\eta_1}, \ldots, \mathcal{M}_{\eta_q}, \mathcal{J}_{\rho_1}, \ldots, \mathcal{J}_{\rho_r}, \mathcal{N}_{\sigma_1}, \ldots, \mathcal{N}_{\sigma_s}), \tag{2.5}$$

where the block entries have the following properties:

1. *Every entry $\mathcal{L}_{\epsilon_j}$ is a bidiagonal block of size $\epsilon_j \times (\epsilon_j + 1)$, $\epsilon_j \in \mathbb{N}_0$, of the form*

$$\lambda \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{bmatrix}.$$

2. *Every entry $\mathcal{M}_{\eta_j}$ is a bidiagonal block of size $(\eta_j + 1) \times \eta_j$, $\eta_j \in \mathbb{N}_0$, of the form*

$$\lambda \begin{bmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{bmatrix} - \begin{bmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{bmatrix}.$$

3. *Every entry $\mathcal{J}_{\rho_j}$ is a Jordan block of size $\rho_j \times \rho_j$, $\rho_j \in \mathbb{N}$, $\lambda_j \in \mathbb{C}$, of the form*

$$\lambda \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} - \begin{bmatrix} \lambda_j & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{bmatrix}.$$

4. *Every entry $\mathcal{N}_{\sigma_j}$ is a nilpotent block of size $\sigma_j \times \sigma_j$, $\sigma_j \in \mathbb{N}$, of the form*

$$\lambda \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} - \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix}.$$

The Kronecker canonical form is unique up to permutation of the blocks, i.e., the kind, size and number of the blocks are characteristic for the matrix pair (E, A).

Note that the notation for the blocks in Theorem 2.3 implies that a pair of 1×1-matrices $(0, 0)$ actually consists of two blocks, a block $\mathcal{L}_0$ of size 0×1 and a block $\mathcal{M}_0$ of size 1×0.

Example 2.4. The Kronecker canonical form of the pair

$$(E, A) = \left(\left[\begin{array}{c|cc|ccc} 1 & & & & & \\ \hline & 0 & 1 & & & \\ & 0 & 0 & & & \\ \hline & & & 0 & & \\ \hline & & & & 0 & 1 \\ \hline & & & & & 0 \end{array}\right], \left[\begin{array}{c|cc|ccc} 1 & & & & & \\ \hline & 1 & 0 & & & \\ & 0 & 1 & & & \\ \hline & & & 1 & & \\ \hline & & & & 1 & 0 \\ \hline & & & & & 0 \end{array}\right] \right)$$

consists of one Jordan block $\mathcal{J}_1 = \lambda 1 - 1$, two nilpotent blocks $\mathcal{N}_2$, $\mathcal{N}_1$, and three rectangular blocks $\mathcal{L}_1$, $\mathcal{L}_0$, $\mathcal{M}_0$.

With the help of the Kronecker canonical form, we can now study the behavior of (2.1) by considering single blocks, cp. Exercise 1.

A special case which we want to discuss here in more detail and for which we want to derive the associated part of the Kronecker canonical form is that of so-called *regular* matrix pairs.

Definition 2.5. Let $E, A \in \mathbb{C}^{m,n}$. The matrix pair (E, A) is called *regular* if $m = n$ and the so-called *characteristic polynomial* p defined by

$$p(\lambda) = \det(\lambda E - A) \tag{2.6}$$

is not the zero polynomial. A matrix pair which is not regular is called *singular*.

Lemma 2.6. *Every matrix pair which is strongly equivalent to a regular matrix pair is regular.*

Proof. We only need to discuss square matrices. Let $E_2 = PE_1Q$ and $A_2 = PA_1Q$ with nonsingular P and Q. Then we have

$$\begin{aligned} p_2(\lambda) &= \det(\lambda E_2 - A_2) = \det(\lambda PE_1Q - PA_1Q) \\ &= \det P \det(\lambda E_1 - A_1) \det Q = c\, p_1(\lambda), \end{aligned}$$

with $c \neq 0$. □

Regularity of a matrix pair is closely related to the solution behavior of the corresponding differential-algebraic equation. In particular, regularity is necessary and sufficient for the property that for every sufficiently smooth inhomogeneity f the differential-algebraic equation is solvable and the solution is unique for every consistent initial value. To show sufficiency, we return to the problem of finding an appropriate canonical form, which can be derived on the basis of the Jordan canonical form of a single matrix, see, e.g., [87].

Theorem 2.7. *Let $E, A \in \mathbb{C}^{n,n}$ and (E, A) be regular. Then, we have*

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix} \right), \tag{2.7}$$

where J is a matrix in Jordan canonical form and N is a nilpotent matrix also in Jordan canonical form. Moreover, it is allowed that one or the other block is not present.

Proof. Since (E, A) is regular, there exists a $\lambda_0 \in \mathbb{C}$ with $\det(\lambda_0 E - A) \neq 0$ implying that $\lambda_0 E - A$ is nonsingular. Hence,

$$\begin{aligned} (E, A) &\sim (E, A - \lambda_0 E + \lambda_0 E) \\ &\sim ((A - \lambda_0 E)^{-1} E, I + \lambda_0 (A - \lambda_0 E)^{-1} E). \end{aligned}$$

The next step is to transform $(A - \lambda_0 E)^{-1} E$ to Jordan canonical form. This is given by $\operatorname{diag}(\tilde{J}, \tilde{N})$, where $\tilde{J}$ is nonsingular (i.e., the part belonging to the nonzero eigenvalues) and $\tilde{N}$ is a nilpotent, strictly upper triangular matrix. We obtain

$$(E, A) \sim \left(\begin{bmatrix} \tilde{J} & 0 \\ 0 & \tilde{N} \end{bmatrix}, \begin{bmatrix} I + \lambda_0 \tilde{J} & 0 \\ 0 & I + \lambda_0 \tilde{N} \end{bmatrix} \right).$$

Because of the special form of $\tilde{N}$, the entry $I + \lambda_0 \tilde{N}$ is a nonsingular upper triangular matrix. It follows that

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & (I + \lambda_0 \tilde{N})^{-1} \tilde{N} \end{bmatrix}, \begin{bmatrix} \tilde{J}^{-1} + \lambda_0 I & 0 \\ 0 & I \end{bmatrix} \right),$$

where $(I + \lambda_0 \tilde{N})^{-1} \tilde{N}$ is again a strictly upper triangular matrix and therefore nilpotent. Transformation of the nontrivial entries to Jordan canonical form finally yields (2.7) with the required block structure. □

With the help of (2.7), which is sometimes called *Weierstraß canonical form*, we are now able to write down the solutions of (2.1) in the case of a regular matrix pair explicitly. In particular, we utilize that (2.1) separates into two subproblems when (E, A) is in canonical form. Denoting for both subproblems the unknown function by $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ and the inhomogeneity by $f \in C(\mathbb{I}, \mathbb{C}^n)$, we get for the first subproblem

$$\dot{x} = Jx + f(t), \tag{2.8}$$

which is a linear ordinary differential equation, while for the second subproblem we obtain

$$N\dot{x} = x + f(t). \tag{2.9}$$

Since initial value problems for linear ordinary differential equations are always uniquely solvable for $f \in C(\mathbb{I}, \mathbb{C}^n)$, see, e.g., [65], [106], we only need to consider (2.9) in more detail.

Lemma 2.8. *Consider (2.9) with $f \in C^\nu(\mathbb{I}, \mathbb{C}^n)$, $n \geq 1$. Let ν be the index of nilpotency of N, i.e., let $N^\nu = 0$ and $N^{\nu-1} \neq 0$. Then (2.9) has the unique solution*

$$x = -\sum_{i=0}^{\nu-1} N^i f^{(i)}. \tag{2.10}$$

Proof. One possibility to show that a solution must have the form (2.10) is to separate (2.9) further into the single Jordan blocks and to solve then componentwise. A simpler proof is given by the following formal approach. Let D be the linear operator which maps a (continuously) differentiable function x to its derivative $\dot{x}$.

Then (2.9) becomes $NDx = x + f$ or $(I - ND)x + f = 0$. Because N is nilpotent and N and D commute (N is a constant factor), we obtain

$$x = -(I - ND)^{-1} f = -\sum_{i=0}^{\infty} (ND)^i f = -\sum_{i=0}^{\nu-1} N^i f^{(i)}$$

by using the Neumann series [87]. Inserting this into (2.9) gives

$$N\dot{x} - x - f = -\sum_{i=0}^{\nu-1} N^{i+1} f^{(i+1)} + \sum_{i=0}^{\nu-1} N^i f^{(i)} - f = 0,$$

thus showing that (2.10) is indeed a solution. □

Looking at the result of Lemma 2.8, we can make two important observations. First, the solution is unique without specifying initial values or, in other words, the only possible initial condition at t_0 is given by the value of x from (2.10) at t_0. Second, one must require that f is at least ν times continuously differentiable to guarantee that x is continuously differentiable. In this sense, the quantity ν plays an important role in the theory of regular linear differential-algebraic equations with constant coefficients.

Definition 2.9. Consider a pair (E, A) of square matrices that is regular and has a canonical form as in (2.7). The quantity ν defined by $N^\nu = 0$, $N^{\nu-1} \neq 0$, i.e., by the index of nilpotency of N in (2.7), if the nilpotent block in (2.7) is present and by $\nu = 0$ if it is absent, is called the *index* of the matrix pair (E, A), denoted by $\nu = \operatorname{ind}(E, A)$.

To justify this definition and the notation $\nu = \operatorname{ind}(E, A)$, we must show that ν does not depend on the special transformation to canonical form.

Lemma 2.10. *Suppose that the pair (E, A) of square matrices has the two canonical forms*

$$(E, A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & N_i \end{bmatrix}, \begin{bmatrix} J_i & 0 \\ 0 & I \end{bmatrix} \right), \quad i = 1, 2,$$

where d_i, $i = 1, 2$, is the size of the block J_i. Then $d_1 = d_2$ and, furthermore, $N_1^\nu = 0$ and $N_1^{\nu-1} \neq 0$ if and only if $N_2^\nu = 0$ and $N_2^{\nu-1} \neq 0$.

Proof. For the characteristic polynomials of the two canonical forms, we have

$$p_i(\lambda) = \det \begin{bmatrix} \lambda I - J_i & 0 \\ 0 & \lambda N_i - I \end{bmatrix} = (-1)^{n-d_i} \det(\lambda I - J_i).$$

Hence, p_i is a polynomial of degree d_i. Since the normal forms are strongly equivalent, p_1 and p_2 can only differ by a constant factor according to the proof

of Lemma 2.6. Thus, $d_1 = d_2$ and the block sizes in the canonical forms are the same. Furthermore, from the strong equivalence of the canonical forms it follows that there exist nonsingular matrices

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix},$$

partitioned conformably, such that

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & N_2 \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & N_1 \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$$

and

$$\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} J_2 & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} J_1 & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}.$$

Thus, we obtain the relations

$$P_{11} = Q_{11}, \quad P_{12}N_2 = Q_{12}, \quad P_{21} = N_1 Q_{21}, \quad P_{22}N_2 = N_1 Q_{22},$$

and

$$P_{11}J_2 = J_1 Q_{11}, \quad P_{12} = J_1 Q_{12}, \quad P_{21}J_2 = Q_{21}, \quad P_{22} = Q_{22}.$$

From this, we get $P_{21} = N_1 P_{21} J_2$ and, by successive insertion of P_{21}, finally $P_{21} = 0$ by the nilpotency of N_1. Therefore, $P_{11} = Q_{11}$ and $P_{22} = Q_{22}$ must be nonsingular. In particular, J_1 and J_2 as well as N_1 and N_2 must be similar. The claim now follows, since the Jordan canonical forms of N_1 and N_2 consist of the same nilpotent Jordan blocks. □

This result shows that the block sizes of a canonical form (2.7) and the index, as defined in Definition 2.9, are characteristic values for a pair of square matrices as well as for the associated linear differential-algebraic equation with constant coefficients.

Remark 2.11. It would be much more elegant to use a different counting for the index by taking it to be the smallest $\mu \geq 0$ such that $N^{\mu+1} = 0$. This would allow a rigorous treatment of nilpotent endomorphisms ϕ of a vector space V. With the usual definition, the case that V has zero dimension is not clearly defined. Obviously, if $V = \{0\}$, then every endomorphism maps 0 to 0 and is therefore the identity map, i.e., an isomorphism.

The zero matrix is the matrix that maps every vector to 0 and it is the matrix representation of the endomorphism $\phi\colon V \to V$ with $\phi(v) = 0$ for all $v \in V$. Thus, for $V = \{0\}$ the zero matrix equals the identity matrix.

Using the usual convention that N^0 is the identity matrix, and thus nonzero, immediately leads to difficulties in the classical definition of the index of nilpotency ν, e.g., by

$$\nu = \min\{\ell \in \mathbb{N}_0 \mid \text{kernel } N^{\ell+1} = \text{kernel } N^{\ell}\},$$

since for $V = \{0\}$ this would yield $\nu = 0$ although N is the zero matrix. See also [74] for an extended discussion on properties of matrix representations of linear mappings associated with zero dimensional vector spaces.

We can now summarize the above results.

Theorem 2.12. *Let the pair (E, A) of square matrices be regular and let P and Q be nonsingular matrices which transform (2.1) and (2.2) to Weierstraß canonical form, i.e.,*

$$PEQ = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad PAQ = \begin{bmatrix} J & 0 \\ 0 & I \end{bmatrix}, \quad Pf = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix} \tag{2.11}$$

and set

$$Q^{-1}x = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}, \quad Q^{-1}x_0 = \begin{bmatrix} \tilde{x}_{1,0} \\ \tilde{x}_{2,0} \end{bmatrix}. \tag{2.12}$$

Furthermore, let $\nu = \text{ind}(E, A)$ and $f \in C^{\nu}(\mathbb{I}, \mathbb{C}^n)$. Then we have the following:

1. *The differential-algebraic equation (2.1) is solvable.*

2. *An initial condition (2.2) is consistent if and only if*

$$\tilde{x}_{2,0} = -\sum_{i=0}^{\nu-1} N^i \tilde{f}_2^{(i)}(t_0).$$

   In particular, the set of consistent initial values x_0 is nonempty.

3. *Every initial value problem with consistent initial condition is uniquely solvable.*

Example 2.13. Consider the differential-algebraic equation $E\dot{x} = Ax + f(t)$ with

$$E = \left[\begin{array}{cc|c} 0 & 1 & \\ 0 & 0 & \\ \hline & & 0 \end{array}\right], \quad A = \left[\begin{array}{cc|c} 1 & 0 & \\ 0 & 1 & \\ \hline & & 1 \end{array}\right], \quad f(t) = \left[\begin{array}{c} 0 \\ -t^3 \\ \hline -t \end{array}\right].$$

The solution is unique and given by $x(t) = [\, 3t^2 \;\; t^3 \;\; t \,]^T$ independent of an initial condition.

It remains to consider what happens if a given matrix pair (E, A) is not regular, i.e., if the matrices are not square or the characteristic polynomial (2.6) vanishes identically. In particular, we want to show that in this case the corresponding initial value problem either has more than one solution or there are arbitrarily smooth inhomogeneities for which there is no solution at all. With the well-known principle for linear problems that two solutions of the inhomogeneous problem differ by a solution of the homogeneous problem, this is equivalent to the following statement.

Theorem 2.14. *Let $E, A \in \mathbb{C}^{m,n}$ and suppose that (E, A) is a singular matrix pair.*

1. *If $\operatorname{rank}(\lambda E - A) < n$ for all $\lambda \in \mathbb{C}$, then the homogeneous initial value problem*

$$E\dot{x} = Ax, \quad x(t_0) = 0 \tag{2.13}$$

has a nontrivial solution.

2. *If $\operatorname{rank}(\lambda E - A) = n$ for some $\lambda \in \mathbb{C}$ and hence $m > n$, then there exist arbitrarily smooth inhomogeneities f for which the corresponding differential-algebraic equation is not solvable.*

Proof. For the first claim, suppose that $\operatorname{rank}(\lambda E - A) < n$ for every $\lambda \in \mathbb{C}$. Let now λ_i, $i = 1, \ldots, n+1$, be pairwise different complex numbers. For every λ_i, we then have a $v_i \in \mathbb{C}^n \setminus \{0\}$ with $(\lambda_i E - A)v_i = 0$ and clearly the vectors v_i, $i = 1, \ldots, n+1$, are linearly dependent. Hence, there exist complex numbers α_i, $i = 1, \ldots, n+1$, not all of them being zero, such that

$$\sum_{i=1}^{n+1} \alpha_i v_i = 0.$$

For the function x defined by

$$x(t) = \sum_{i=1}^{n+1} \alpha_i v_i e^{\lambda_i (t-t_0)},$$

we then have $x(t_0) = 0$ as well as

$$E\dot{x}(t) = \sum_{i=1}^{n+1} \alpha_i \lambda_i E v_i e^{\lambda_i (t-t_0)} = \sum_{i=1}^{n+1} \alpha_i A v_i e^{\lambda_i (t-t_0)} = Ax(t).$$

Since x is not the zero function, it is a nontrivial solution of the homogeneous initial value problem (2.13).

For the second claim, suppose that there exists a scalar $\lambda \in \mathbb{C}$ such that $\operatorname{rank}(\lambda E - A) = n$. Since (E, A) is assumed to be singular, we have $m > n$. With $x(t) = e^{\lambda t}\tilde{x}(t)$, we have

$$\dot{x}(t) = e^{\lambda t}\dot{\tilde{x}}(t) + \lambda e^{\lambda t}\tilde{x}(t)$$

such that (2.1) is transformed to

$$E\dot{\tilde{x}} = (A - \lambda E)\tilde{x} + e^{-\lambda t} f(t).$$

Since $A - \lambda E$ has full column rank, there exists a nonsingular matrix $P \in \mathbb{C}^{m,m}$ such that this equation, multiplied from the left by P, gives

$$\begin{bmatrix} E_1 \\ E_2 \end{bmatrix} \dot{\tilde{x}} = \begin{bmatrix} I \\ 0 \end{bmatrix} \tilde{x} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}.$$

Obviously the pair (E_1, I) is regular, implying that

$$E_1 \dot{\tilde{x}} = \tilde{x} + f_1(t), \quad \tilde{x}(t_0) = \tilde{x}_0$$

has a unique solution for every sufficiently smooth inhomogeneity f_1 and for every consistent initial value. But then

$$f_2(t) = E_2 \dot{\tilde{x}}(t)$$

is a consistency condition for the inhomogeneity f_2 that must hold for a solution to exist. This immediately shows that there exist arbitrarily smooth functions f for which this consistency condition is not satisfied. □

Example 2.15. Consider the differential-algebraic equation $E\dot{x} = Ax + f(t)$ with

$$E = \left[\begin{array}{cc|c} 0 & 1 & \\ \hline & & 1 \\ & & 0 \end{array}\right], \quad A = \left[\begin{array}{cc|c} 1 & 0 & \\ \hline & & 0 \\ & & 1 \end{array}\right], \quad f = \left[\begin{array}{c} f_1 \\ \hline f_2 \\ f_3 \end{array}\right].$$

We obtain $\dot{x}_2 = x_1 + f_1$, $\dot{x}_3 = f_2$, and $x_3 = -f_3$, independent of initial values. For a solution to exist, we need $f_2 = -\dot{f}_3$. The solution is not unique, since any continuously differentiable function x_1 can be chosen.

2.2 The Drazin inverse

In this and the following section, we discuss the question whether it is possible to write down an explicit representation of the solutions of (2.1) in terms of the original data E, A, and f. For convenience, we restrict ourselves to the case $m = n$ of square matrices E and A. The presentation follows the work of [42], [43].

Considering linear ordinary differential equations

$$\dot{x} = Ax + f(t), \quad x(t_0) = x_0,$$

we have the well-known formula

$$x(t) = e^{A(t-t_0)}x_0 + \int_{t_0}^{t} e^{A(t-s)} f(s)\, ds \tag{2.14}$$

obtained by variation of constants. Let us first consider the special differential-algebraic equation

$$E\dot{x} = x, \quad x(t_0) = x_0. \tag{2.15}$$

If E is nonsingular, then (2.14) yields

$$x(t) = e^{E^{-1}(t-t_0)}x_0.$$

For general E, we apply the results of the previous section to obtain a solution. Since (E, I) is a regular matrix pair, we can transform it to Weierstraß canonical form (2.7). The easiest way to get this is to transform E to Jordan canonical form. Let therefore

$$E = T^{-1}JT, \quad J = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix},$$

where C is nonsingular and N is nilpotent. The differential-algebraic equation (2.15) then becomes

$$\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} C^{-1} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix},$$

with

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = Tx, \quad \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} = Tx_0.$$

Applying Theorem 2.12 yields that $x_{2,0} = 0$ must hold for the initial condition to be consistent and that the solution is then given by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} e^{C^{-1}(t-t_0)} & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}.$$

Defining

$$J^D = \begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix},$$

we can write this as

$$Tx(t) = e^{J^D(t-t_0)}Tx_0.$$

Transforming back, we obtain

$$x(t) = e^{E^D(t-t_0)}x_0$$

with

$$E^D = T^{-1} J^D T. \tag{2.16}$$

Comparing this formula with that for nonsingular E, we see that E^{-1} is just replaced by the matrix E^D. Although E may in general be singular, there is a matrix that plays the role of an inverse when looking for a solution formula of the above kind of differential-algebraic equations. In such situations, one speaks of so-called *generalized inverses*, see, e.g., [56]. We will therefore first have a closer look at such inverses before we return to solution formulas for differential-algebraic equations. Although we have already defined the desired object, we shall start here with a different, more algebraic definition of the same object.

Definition 2.16. Let $E \in \mathbb{C}^{n,n}$. The quantity $\nu = \operatorname{ind}(E, I)$ as in Definition 2.9 is called *index* (*of nilpotency*) of E and is denoted by $\nu = \operatorname{ind} E$.

Definition 2.17. Let $E \in \mathbb{C}^{n,n}$ have the index ν. A matrix $X \in \mathbb{C}^{n,n}$ satisfying

$$EX = XE, \tag{2.17a}$$

$$XEX = X, \tag{2.17b}$$

$$XE^{\nu+1} = E^{\nu} \tag{2.17c}$$

is called a *Drazin inverse* of E.

Remark 2.18. As we have already discussed in Remark 2.11, the usual index of nilpotency is not rigorously defined in the case that E is a matrix representation of the isomorphism $\phi\colon \{0\} \to \{0\}$. Analogously, we have difficulties to define the Drazin inverse of this matrix E. As an isomorphism between vector spaces, E is invertible, actually the identity matrix, so the Drazin inverse of E should be E itself, and clearly the three conditions (2.17a), (2.17b) and (2.17c) hold. But the definition based on (2.16) is ambiguous, since we must distinguish between zero and identity blocks. Observe that in (2.17c) we again must use the convention that E^0 is the identity matrix and that the identity matrix is different from the zero matrix.

Theorem 2.19. *Every $E \in \mathbb{C}^{n,n}$ has one and only one Drazin inverse E^D.*

Proof. The matrix E^D given in (2.16) satisfies the axioms (2.17), since

$$\nu = \operatorname{ind}(E, I) = \operatorname{ind}(J, I) = \operatorname{ind}\left(\begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} C^{-1} & 0 \\ 0 & I \end{bmatrix}\right) = \operatorname{ind}(N, I),$$

implying that $N^{\nu} = 0$ and $N^{\nu-1} \neq 0$ according to Definition 2.9 and Definition 2.17. To show uniqueness, we suppose that X_1 and X_2 are two Drazin inverses of E. Applying the axioms (2.17), we get

$$\begin{aligned} X_1 &= X_1 E X_1 = \cdots = X_1^{\nu+1} E^{\nu} = X_1^{\nu+1} X_2 E^{\nu+1} = X_1^{\nu+1} E^{\nu+1} X_2 = X_1 E X_2 \\ &= X_1 E^{\nu+1} X_2^{\nu+1} = X_1 X_2^{\nu+1} E^{\nu+1} = X_2^{\nu+1} E^{\nu} = \cdots = X_2 E X_2 = X_2. \end{aligned}$$

□

Lemma 2.20. *For nonsingular $E \in \mathbb{C}^{n,n}$, we have*

$$E^D = E^{-1}, \tag{2.18}$$

and for arbitrary $E \in \mathbb{C}^{n,n}$ and nonsingular $T \in \mathbb{C}^{n,n}$, we have

$$(T^{-1}ET)^D = T^{-1}E^D T. \tag{2.19}$$

Proof. Both assertions follow by direct verification of the axioms (2.17). □

Lemma 2.21. *Consider matrices $E, A \in \mathbb{C}^{n,n}$ with $EA = AE$. Then we have*

$$EA^D = A^D E, \tag{2.20a}$$

$$E^D A = AE^D, \tag{2.20b}$$

$$E^D A^D = A^D E^D. \tag{2.20c}$$

Proof. Since $EA = AE$, for arbitrary $\lambda \in \mathbb{C}$ we have that

$$(\lambda I - E)A = \lambda A - EA = \lambda A - AE = A(\lambda I - E).$$

If λ is not an eigenvalue of E, then the matrix $\lambda I - E$ is nonsingular and we get

$$A = (\lambda I - E)^{-1} A (\lambda I - E).$$

With (2.19), it follows that

$$A^D = (\lambda I - E)^{-1} A^D (\lambda I - E)$$

or

$$(\lambda I - E)A^D = A^D(\lambda I - E),$$

hence (2.20a). By interchanging the roles of E and A, (2.20b) follows. Finally, (2.20c) follows by replacing E by E^D in the proof of (2.20a). □

Theorem 2.22. *Let $E \in \mathbb{C}^{n,n}$ with $\nu = \operatorname{ind} E$. There is one and only one decomposition*

$$E = \tilde{C} + \tilde{N} \tag{2.21}$$

with the properties

$$\tilde{C}\tilde{N} = \tilde{N}\tilde{C} = 0, \quad \tilde{N}^{\nu} = 0, \quad \tilde{N}^{\nu-1} \neq 0, \quad \operatorname{ind} \tilde{C} \leq 1. \tag{2.22}$$

In particular, the following statements hold:

$$\tilde{C}^D \tilde{N} = 0, \qquad \tilde{N}\tilde{C}^D = 0, \tag{2.23a}$$

$$E^D = \tilde{C}^D, \tag{2.23b}$$

$$\tilde{C}\tilde{C}^D\tilde{C} = \tilde{C}, \tag{2.23c}$$

$$\tilde{C}^D\tilde{C} = E^D E, \tag{2.23d}$$

$$\tilde{C} = EE^D E, \qquad \tilde{N} = E(I - E^D E). \tag{2.23e}$$

Proof. We first show that the desired properties of the decomposition imply (2.23a)–(2.23e) in the given order which finally gives an explicit representation of the desired decomposition. Claim (2.23a) follows from

$$\tilde{C}^D = \tilde{C}^D\tilde{C}\tilde{C}^D = \tilde{C}^D\tilde{C}^D\tilde{C} = \tilde{C}\tilde{C}^D\tilde{C}^D$$

and $\tilde{C}\tilde{N} = \tilde{N}\tilde{C} = 0$ by application of Lemma 2.21. For (2.23b), we verify the axioms (2.17) by

$$\begin{aligned}
E\tilde{C}^D &= \tilde{C}\tilde{C}^D + \tilde{N}\tilde{C}^D = \tilde{C}\tilde{C}^D = \tilde{C}^D\tilde{C} = \tilde{C}^D\tilde{C} + \tilde{C}^D\tilde{N} = \tilde{C}^D E,\\
\tilde{C}^D E\tilde{C}^D &= \tilde{C}^D\tilde{C}\tilde{C}^D + \tilde{C}^D\tilde{N}\tilde{C}^D = \tilde{C}^D,\\
\tilde{C}^D E^{\nu+1} &= \tilde{C}^D(\tilde{C} + \tilde{N})^{\nu+1} = \tilde{C}^D\tilde{C}^{\nu+1} = \tilde{C}^{\nu} = (\tilde{C} + \tilde{N})^{\nu} = E^{\nu}.
\end{aligned}$$

Since $\operatorname{ind}\tilde{C} \le 1$, we have $\tilde{C}\tilde{C}^D\tilde{C} = \tilde{C}$ and therefore (2.23c). With

$$E^D E = \tilde{C}^D(\tilde{C} + \tilde{N}) = \tilde{C}^D\tilde{C},$$

we get (2.23d). Finally, we obtain (2.23e) by

$$\begin{aligned}
\tilde{C} &= \tilde{C}\tilde{C}^D\tilde{C} = (E - \tilde{N})\tilde{C}^D\tilde{C} = E\tilde{C}^D\tilde{C} = EE^D E,\\
\tilde{N} &= E - \tilde{C} = E - EE^D E = E(I - E^D E),
\end{aligned}$$

which also shows the uniqueness of the desired decomposition of E.

To show the existence of such a decomposition, it suffices to verify that the matrices $\tilde{C}$ and $\tilde{N}$ from (2.23e) satisfy the properties (2.21), which is trivial, and (2.22). For (2.22), we transform E again to Jordan canonical form. From

$$E = T^{-1}JT, \quad J = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix},$$

with C nonsingular and $\operatorname{ind} N = \nu$, we obtain

$$\begin{aligned}
\tilde{C} &= EE^D E = T^{-1}\begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}\begin{bmatrix} C^{-1} & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}T = T^{-1}\begin{bmatrix} C & 0 \\ 0 & 0 \end{bmatrix}T,\\
\tilde{N} &= E - \tilde{C} = T^{-1}\begin{bmatrix} 0 & 0 \\ 0 & N \end{bmatrix}T,
\end{aligned}$$

and the required properties are obvious. □

Remark 2.23. One can show that a matrix $E \in \mathbb{C}^{n,n}$ is an element of a group $\mathbb{G} \subseteq \mathbb{C}^{n,n}$ with matrix multiplication as the inner operation if and only if $\operatorname{ind} E \le 1$. The inverse of E in such a group $\mathbb{G}$ is exactly the Drazin inverse E^D. In this case, one also speaks of the *group inverse* of E and denotes it by $E^{\#}$, cp. Exercise 12.

2.3 Explicit representation of solutions

In order to develop explicit representations for the solutions of (2.1) in terms of the coefficient matrices E and A and the inhomogeneity f, we first treat the special case that E and A commute, i.e., that

$$EA = AE. \tag{2.24}$$

According to Theorem 2.22, we take $E = \tilde{C} + \tilde{N}$ with the properties of $\tilde{C}$ and $\tilde{N}$ as given there. Note that due to Lemma 2.21 we always have that $E^D E = E E^D$.

Lemma 2.24. *System (2.1) with the property (2.24) is equivalent (in the sense that the solutions are in one-to-one correspondence) to the system*

$$\tilde{C}\dot{x}_1 = Ax_1 + f_1(t), \tag{2.25a}$$

$$\tilde{N}\dot{x}_2 = Ax_2 + f_2(t), \tag{2.25b}$$

where

$$x_1 = E^D E x, \quad x_2 = (I - E^D E)x \tag{2.26}$$

and

$$f_1 = E^D E f, \quad f_2 = (I - E^D E) f. \tag{2.27}$$

Equation (2.25a) together with (2.26) is equivalent to the ordinary differential equation

$$\dot{x}_1 = E^D A x_1 + E^D f_1(t). \tag{2.28}$$

Proof. Multiplying (2.1) in the form

$$(\tilde{C} + \tilde{N})(\dot{x}_1 + \dot{x}_2) = A(x_1 + x_2) + f(t),$$

with $\tilde{C}^D \tilde{C}$ and utilizing (2.20) and (2.23), we obtain (2.25a) since

$$\tilde{C}^D \tilde{C} A = E^D E A = A E^D E = A \tilde{C}^D \tilde{C}. \tag{2.29}$$

Subtracting (2.25a) from (2.1) immediately gives (2.25b). Conversely, adding both parts of (2.25) gives (2.1), since

$$\tilde{C} E^D E \dot{x} + \tilde{N}(I - E^D E)\dot{x} = E E^D E \dot{x} + E(I - E^D E)\dot{x} = E\dot{x}.$$

Multiplying (2.25a) by $\tilde{C}^D = E^D$ and adding $(I - \tilde{C}^D \tilde{C})\dot{x}_1 = 0$ yields (2.28), while multiplying (2.28) by $\tilde{C}$ and using (2.29) yields (2.25a). □

Note that the ordinary differential equation (2.28) has solutions x_1 with values in range$(E^D E)$ as soon as the initial value $x_1(t_0)$ lies in this space.

Since two solutions of a linear (inhomogeneous) problem differ by a solution of the corresponding homogeneous problem, we start with analyzing the solution set of homogeneous differential-algebraic equations.

Lemma 2.25. *Consider the differential-algebraic equation (2.1) with $m = n$ and suppose that (2.24) holds. For every $v \in \mathbb{C}^n$, the function $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ defined by*

$$x(t) = e^{E^D A t} E^D E v \tag{2.30}$$

is a solution of

$$E\dot{x} = Ax. \tag{2.31}$$

Proof. Direct computation yields

$$\begin{aligned} E\dot{x}(t) - Ax(t) &= EE^D A e^{E^D A t} E^D E v - A e^{E^D A t} E^D E v \\ &= A e^{E^D A t} E^D E E^D E v - A e^{E^D A t} E^D E v = 0. \end{aligned}$$ □

The preceding lemma implies that there is a linear space of solutions of (2.31) with dimension $\operatorname{rank}(E^D E)$. In view of Theorem 2.14, we must require regularity of the matrix pair (E, A) in order to show that there are no solutions different from those of the form (2.30). In the case of commuting E and A, we do this as follows.

Lemma 2.26. *Let $E, A \in \mathbb{C}^{n,n}$ satisfy (2.24) and*

$$\operatorname{kernel} E \cap \operatorname{kernel} A = \{0\}. \tag{2.32}$$

Then, we have

$$(I - E^D E) A^D A = (I - E^D E). \tag{2.33}$$

Proof. We again begin with a transformation of E to Jordan canonical form according to

$$T^{-1} E T = \begin{bmatrix} C & 0 \\ 0 & N \end{bmatrix}, \quad T^{-1} A T = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

with C nonsingular and $N^\nu = 0$, $N^{\nu-1} \neq 0$, where $\nu = \operatorname{ind} E$. Since $EA = AE$, we have

$$\begin{bmatrix} CA_{11} & CA_{12} \\ NA_{21} & NA_{22} \end{bmatrix} = \begin{bmatrix} A_{11}C & A_{12}N \\ A_{21}C & A_{22}N \end{bmatrix}.$$

In particular, we see that $A_{21} = NA_{21}C^{-1}$ and $A_{12} = C^{-1}A_{12}N$. Successive insertion until the nilpotency of N can be utilized then shows that $A_{21} = 0$ and $A_{12} = 0$. With this, we obtain

$$\operatorname{kernel} E = \left\{ T \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \mid v_1 = 0,\ Nv_2 = 0 \right\},$$

$$\operatorname{kernel} A = \left\{ T \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \mid A_{11}v_1 = 0,\ A_{22}v_2 = 0 \right\},$$

or

$$\operatorname{kernel} E \cap \operatorname{kernel} A = \left\{ T \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \mid v_1 = 0,\ Nv_2 = 0,\ A_{22}v_2 = 0 \right\}.$$

The aim now is to show that (2.32) implies that A_{22} is nonsingular. Suppose that A_{22} is singular, i.e., suppose that there is a vector $v \neq 0$ with $A_{22}v = 0$. Since $NA_{22} = A_{22}N$, we have

$$A_{22}N^\ell v = N^\ell A_{22}v = 0$$

and therefore

$$N^\ell v \in \text{kernel } A_{22}$$

for all $\ell \in \mathbb{N}_0$. Since N is nilpotent, there exists an $\ell \in \mathbb{N}_0$ such that

$$N^\ell v \neq 0, \quad N^{\ell+1}v = 0$$

and therefore

$$0 \neq N^\ell v \in \text{kernel } N \cap \text{kernel } A_{22}$$

in contradiction to (2.32). Hence, A_{22} is nonsingular and we have

$$T^{-1}(I - E^D E)T = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad T^{-1}A^D AT = \begin{bmatrix} A_{11}^D A_{11} & 0 \\ 0 & I \end{bmatrix},$$

which implies (2.33). □

Note that for commuting matrices E and A, condition (2.32) is equivalent to the regularity of (E, A) in the sense of Definition 2.5, see Exercise 8.

Theorem 2.27. *Let $E, A \in \mathbb{C}^{n,n}$ satisfy (2.24) and (2.32). Then every solution $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ of (2.31) has the form (2.30) for some $v \in \mathbb{C}^n$.*

Proof. We write E as $E = \tilde{C} + \tilde{N}$ according to Theorem 2.22. Using Lemma 2.21, we obtain

$$A\tilde{N} = AE(I - E^D E) = E(I - E^D E)A = \tilde{N}A.$$

Because of Lemma 2.26, we then get the implications

$$\begin{aligned} A\tilde{N}x = 0 &\Longrightarrow A^D A\tilde{N}x = 0 \\ &\Longrightarrow (I - E^D E)A^D A\tilde{N}x = 0 \\ &\Longrightarrow (I - E^D E)\tilde{N}x = 0 \\ &\Longrightarrow \tilde{N}x = 0. \end{aligned}$$

Let now $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ be a solution of (2.31) with the splitting $x = x_1 + x_2$ according to Lemma 2.24. With $\nu = \text{ind } E$ and $f_2 = 0$, we obtain from (2.25b) that

$$0 = \tilde{N}^\nu \dot{x}_2 = \tilde{N}^{\nu-1} A x_2 = A\tilde{N}^{\nu-1} x_2$$

and therefore

$$\tilde{N}^{\nu-1}x_2 = 0.$$

Differentiation yields

$$0 = \tilde{N}^{\nu-1}\dot{x}_2 = \tilde{N}^{\nu-2}Ax_2 = A\tilde{N}^{\nu-2}x_2,$$

hence $\tilde{N}^{\nu-2}x_2 = 0$. Successively applying this procedure finally yields $\tilde{N}x_2 = 0$. A further differentiation then gives $Ax_2 = 0$. Since $x_2 = (I - E^D E)x$, it follows that

$$x_2 = (I - E^D E)x_2 = (I - E^D E)A^D Ax_2 = 0$$

and therefore $x = x_1$. But since x_1 solves the ordinary differential equation (2.28), there exists a $v \in \mathbb{C}^n$ such that

$$x_1(t) = e^{E^D A t}v,$$

and hence

$$x(t) = x_1(t) = E^D E x_1(t) = e^{E^D A t}E^D E v. \qquad \square$$

Lemma 2.25 and Theorem 2.27 describe the solutions of a homogeneous differential-algebraic equation for the case that the coefficient matrices commute and satisfy the regularity condition (2.32). As in the case of linear algebraic equations or ordinary differential equations, we only need a single (so-called particular) solution of the corresponding inhomogeneous problem to be able to describe all solutions.

Theorem 2.28. *Let $E, A \in \mathbb{C}^{n,n}$ satisfy (2.24) and (2.32). Furthermore, let $f \in C^\nu(\mathbb{I}, \mathbb{C}^n)$ with $\nu = \operatorname{ind} E$ and let $t_0 \in \mathbb{I}$. Then, $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ defined by*

$$x(t) = \int_{t_0}^{t} e^{E^D A(t-s)}E^D f(s)\,ds - (I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i)}(t) \tag{2.34}$$

is a particular solution of (2.1).

Proof. The representation (2.34) obviously corresponds to the splitting (2.25) with

$$x_1(t) = \int_{t_0}^{t} e^{E^D A(t-s)}E^D f(s)\,ds,$$

$$x_2(t) = -(I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i)}(t).$$

Thus, we get

$$E\dot{x}_1(t) = EE^D Ax_1(t) + EE^D f(t) = Ax_1(t) + EE^D f(t)$$

and

$$\begin{aligned} E\dot{x}_2(t) &= -E(I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i+1)}(t) \\ &= -(I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^{i+1} f^{(i+1)}(t) \\ &= -(I - E^D E)\sum_{i=0}^{\nu-2}(EA^D)^{i+1} f^{(i+1)}(t) \\ &= -(I - E^D E)A^D A\sum_{i=1}^{\nu-1}(EA^D)^i f^{(i)}(t) \\ &= -A(I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i)}(t) + (I - E^D E)f(t) \\ &= Ax_2(t) + (I - E^D E)f(t), \end{aligned}$$

where we have used (2.33) and $(I - E^D E)E^\nu = 0$. □

As usual in the theory of linear problems, we can merge the preceding two theorems into the following statement.

Theorem 2.29. *Let $E, A \in \mathbb{C}^{n,n}$ satisfy (2.24) and (2.32). Furthermore, let $f \in C^\nu(\mathbb{I}, \mathbb{C}^n)$ with $\nu = \operatorname{ind} E$ and let $t_0 \in \mathbb{I}$. Then every solution $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ of (2.1) has the form*

$$\begin{aligned} x(t) = e^{E^D A(t-t_0)}E^D E v + \int_{t_0}^{t} e^{E^D A(t-s)}E^D f(s)\,ds \\ - (I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i)}(t) \end{aligned} \tag{2.35}$$

for some $v \in \mathbb{C}^n$.

Corollary 2.30. *Let the assumptions of Theorem 2.29 hold. The initial value problem consisting of (2.1) and (2.2) possesses a solution if and only if there exists a $v \in \mathbb{C}^n$ with*

$$x_0 = E^D E v - (I - E^D E)\sum_{i=0}^{\nu-1}(EA^D)^i A^D f^{(i)}(t_0). \tag{2.36}$$

If this is the case, then the solution is unique.

So far, we have required the coefficient matrices of the differential-algebraic equation to commute. Using a nice trick due to Campbell (see [42]), the general regular case (i.e., without (2.24)) can be easily reduced to the special case.

Lemma 2.31. *Let $E, A \in \mathbb{C}^{n,n}$ with (E, A) regular. Let $\tilde{\lambda} \in \mathbb{C}$ be chosen such that $\tilde{\lambda}E - A$ is nonsingular. Then the matrices*

$$\tilde{E} = (\tilde{\lambda}E - A)^{-1}E, \quad \tilde{A} = (\tilde{\lambda}E - A)^{-1}A \tag{2.37}$$

commute.

Proof. By construction, we have $\tilde{\lambda}\tilde{E} - \tilde{A} = I$ which directly yields that $\tilde{E}$ and $\tilde{A}$ commute. □

Since the factor $(\tilde{\lambda}E - A)^{-1}$ represents a simple scaling of (2.1), results similar to Theorem 2.29 and Corollary 2.30 hold for the general case provided that the coefficient matrices form a regular matrix pair, cp. Exercise 9. We only need to perform the replacements

$$E \leftarrow (\tilde{\lambda}E - A)^{-1}E, \quad A \leftarrow (\tilde{\lambda}E - A)^{-1}A, \quad f \leftarrow (\tilde{\lambda}E - A)^{-1}f. \tag{2.38}$$

in (2.35) and (2.36).

Remark 2.32. In Theorems 2.28 and 2.29, we have required that $f \in C^{\nu}(\mathbb{I}, \mathbb{C}^n)$ in order to guarantee that $x \in C^1(\mathbb{I}, \mathbb{C}^n)$. Looking closely at the proof of Theorem 2.28, one recognizes that the ν-th derivative of f actually does not appear in $\dot{x}$. One may therefore relax the smoothness requirements for f when one in turn relaxes the smoothness requirements for a solution x of a differential-algebraic equation as given in Definition 1.1. We will come back to this problem in the case of generalized solutions and also in the case of linear equations with variable coefficients.

Example 2.33. Consider the differential-algebraic equation $0 = x - |t|$. The function x given by $x(t) = |t|$ satisfies this equation, but it is not a solution according to Definition 1.1, since x is not differentiable at $t = 0$.

2.4 Generalized solutions

In Remark 2.32, we have noted that the smoothness requirements for the forcing function f can be mildly relaxed if the solution is allowed to be less smooth. The consistency conditions for the initial values, however, cannot be relaxed when considering classical solutions (i.e., solutions in the sense of Remark 2.32). Another

route that one can take to remove consistency conditions and to relax smoothness requirements is to allow generalized functions (or distributions), see [199], as solutions of (2.1). For the analysis of differential-algebraic equations, this approach is relatively recent. Several different directions can be followed that allow to include nondifferentiable forcing functions f or non-consistent initial values. A very elegant and completely algebraic approach was introduced in [96] to treat the problem by using a particular class of distributions introduced first in [112] in the study of control problems. We essentially follow this approach.

The physical relevance of treating non-differentiable forcing functions can be seen best in the context of switches in electrical circuits.

Example 2.34. The discharging of a capacitor via a resistor, see Figure 2.1, can be modeled by the system

$$x_1 - x_3 = u(t), \quad C(\dot{x}_3 - \dot{x}_2) + \frac{x_1 - x_2}{R} = 0, \quad x_3 = 0,$$

where x_1, x_2, x_3 denote the potentials in the different parts of the circuit. This system can be reduced to the ordinary differential equation

$$\dot{x}_2 = -\frac{1}{RC}x_2 + \frac{1}{RC}u(t).$$

Let the input voltage u be defined by $u(t) = u_0 > 0$ for $t < 0$ and $u(t) = 0$ for $t \geq 0$. Thus, we want to study the behavior of the circuit when we close a switch

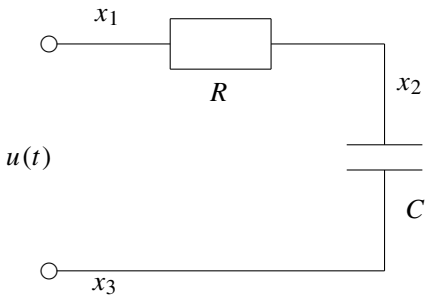


Figure 2.1. Discharging a capacitor

between x_1 and x_3. As initial condition, we take $x_2(0) = u_0$. The differential equation can then be solved separately for $t < 0$ and $t > 0$. Since both parts can be joined together to a continuous function, we may view x_2 defined by

$$x_2(t) = \begin{cases} u_0 & \text{for } t < 0, \\ u_0 e^{-t/RC} & \text{for } t \geq 0. \end{cases}$$

as solution everywhere in $\mathbb{R}$. This procedure can be formalized for linear differential equations working with piecewise continuous inhomogeneities and continuous, piecewise continuously differentiable solutions. Note, however, that such a solution is not differentiable at points where the inhomogeneity jumps.

Example 2.35. A mathematical model for a differentiator, see Figure 2.2, is given by the system

$$x_1 - x_4 = u(t), \quad C(\dot{x}_1 - \dot{x}_2) + \frac{x_3 - x_2}{R} = 0, \quad x_3 = A(x_4 - x_2), \quad x_4 = 0.$$

Typical values for the constants are $R = 10^5$, $C = 10^{-6}$, and $A = 10^5$. Again, we want to study the behavior when the input voltage jumps. Assume that the ideal

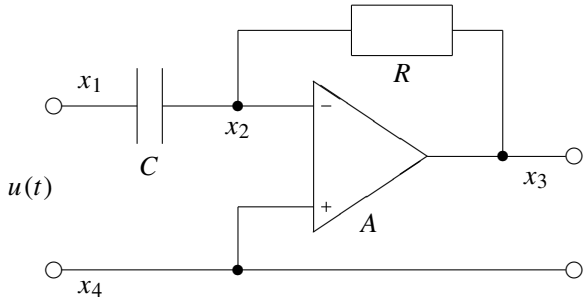


Figure 2.2. An electronic differentiator

input profile is $u(t) = 0$ for $t < 0$ and $u(t) = 1$ for $t \geq 0$. Approximating this u by

$$u(t) = \frac{1}{2}(\tanh \gamma t + 1),$$

with successively larger $\gamma > 0$ and taking consistent initial values at $t = -1$ with $x_2(-1) = x_3(-1) = x_4(-1) = 0$, we get solution profiles as shown in Figures 2.3 and 2.4. In particular, the component x_3 exhibits an impulsive behavior. In the limit $A \to \infty$ (representing an ideal operational amplifier), the third model equation must be replaced by $x_2 = 0$. In this case, the above system can be reduced to

$$x_3 = -RC\dot{u}(t).$$

Of course, for this equation there cannot exist a function as limit for increasing γ. Thus, in order to treat such problems we need solution spaces that are more general than spaces of functions.

Let us recall a few important facts about generalized functions, see, e.g., [199]. Let $\mathcal{D} = C_0^\infty(\mathbb{R}, \mathbb{C})$ be the set of infinitely differentiable functions with values in

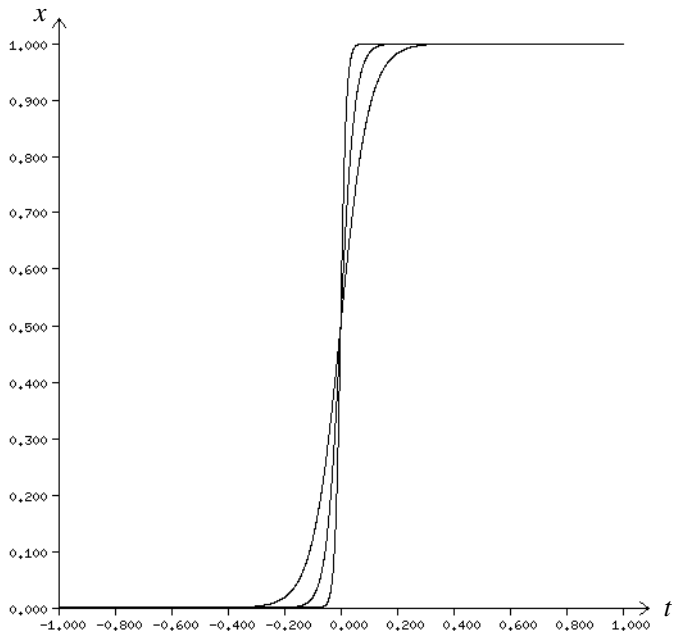


Figure 2.3. Behavior of a non-ideal differentiator — input profiles

$\mathbb{C}$ and compact support in $\mathbb{R}$. We say that a sequence $(\phi_i(t))_{i\in\mathbb{N}}$ converges to zero in $\mathcal{D}$ and write $\phi_i \to 0$, if all functions vanish outside the same bounded interval and the sequences $(\phi_i^{(q)})_{i\in\mathbb{N}}$ of the q-th derivatives converge uniformly to zero for all $q \in \mathbb{N}_0$. The elements of $\mathcal{D}$ are called *test functions*.

Definition 2.36. A linear functional f on $\mathcal{D}$, i.e., a mapping $f : \mathcal{D} \to \mathbb{C}$ with

$$f(\alpha_1\phi_1 + \alpha_2\phi_2) = \alpha_1 f(\phi_1) + \alpha_2 f(\phi_2) \tag{2.39}$$

for all $\alpha_1, \alpha_2 \in \mathbb{C}$ and $\phi_1, \phi_2 \in \mathcal{D}$, is called a *generalized function* or *distribution* if it is continuous in the sense that $f(\phi_i) \to 0$ in $\mathbb{C}$ for all sequences $(\phi_i)_{i\in\mathbb{N}}$ with $\phi_i \to 0$ in $\mathcal{D}$. We denote the space of all distributions acting on $\mathcal{D}$ by $\mathcal{C}$.

For convenience, we also write $\langle f, \phi\rangle$ instead of $f(\phi)$ in order to express the bilinearity of the expression $f(\phi)$ according to (2.39) and

$$(\alpha_1 f_1 + \alpha_2 f_2)(\phi) = \alpha_1 f_1(\phi) + \alpha_2 f_2(\phi) \tag{2.40}$$

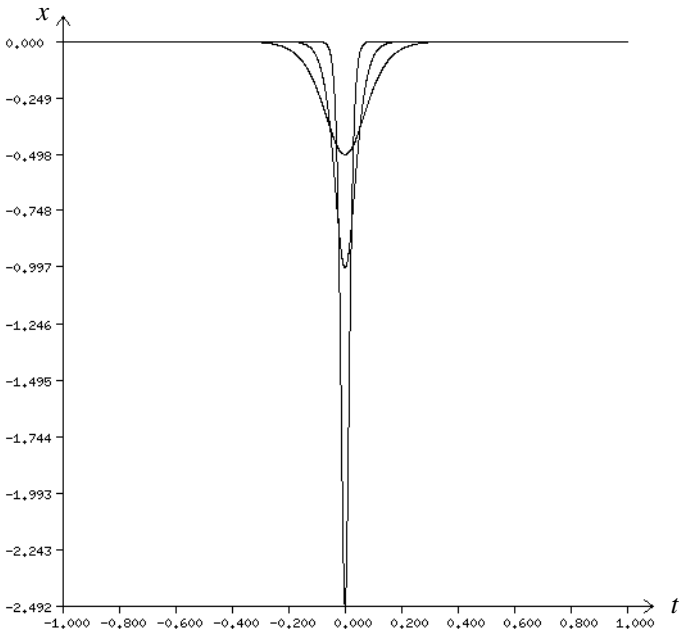


Figure 2.4. Behavior of a non-ideal differentiator — output profiles

for all $\alpha_1, \alpha_2 \in \mathbb{C}$, $f_1, f_2 \in \mathcal{C}$, and $\phi \in \mathcal{D}$.

Every locally Lebesgue integrable function $f : \mathbb{R} \to \mathbb{C}$ defines a distribution by

$$\langle f, \phi \rangle = \int_{\mathbb{R}} f(t)\phi(t)\, dt. \tag{2.41}$$

Distributions that are obtained in this way are called *regular distributions*. As usual, we identify f with the associated distribution. Note that by definition two distributions $f_1, f_2 \in \mathcal{C}$ are equal if $\langle f_1, \phi \rangle = \langle f_2, \phi \rangle$ for all $\phi \in \mathcal{D}$. Even if f_1, f_2 are regular distributions (i.e., if they come from functions $f_1, f_2 : \mathbb{R} \to \mathbb{C}$), this does not imply that $f_1(t) = f_2(t)$ for all $t \in \mathbb{R}$. This is due to the integral involved in (2.41) that allows to alter the values of f_1, f_2 on a set of measure zero. In general, it therefore makes no sense to speak of values of distributions at some point in $\mathbb{R}$.

While (2.40) includes as special case the definition of the sum of two distributions (in the usual way for mappings), there is no meaningful way to define the product of two distributions unless one factor is infinitely often differentiable.

Taking $a \in C^\infty(\mathbb{R}, \mathbb{C})$ and $f \in \mathcal{C}$, the definition

$$\langle af, \phi \rangle = \langle fa, \phi \rangle = \langle f, a\phi \rangle \quad \text{for all } \phi \in \mathcal{D} \tag{2.42}$$

is straightforward.

In order to use distributions in the framework of differential-algebraic equations, we need *derivatives* and *primitives* of distributions. The q-th order derivative $f^{(q)}$, $q \in \mathbb{N}_0$, of a distribution $f \in \mathcal{C}$ is defined by

$$\langle f^{(q)}, \phi \rangle = (-1)^q \langle f, \phi^{(q)} \rangle \quad \text{for all } \phi \in \mathcal{D}. \tag{2.43}$$

The so obtained functional $f^{(q)}$ is obviously linear. It can also be shown, see, e.g. [199], that it is continuous without imposing any restriction on f. Hence, every distribution has derivatives of arbitrary order in $\mathcal{C}$. If $f \in C^q(\mathbb{R}, \mathbb{C})$ with classical derivative $f^{(q)}$, then the associated distributions f and $f^{(q)}$ satisfy (2.43) which is nothing else than partial integration. Note that there are no boundary terms because of the compact support of the test functions.

The *Dirac delta distribution* δ is defined via

$$\langle \delta, \phi \rangle = \phi(0) \quad \text{for all } \phi \in \mathcal{D}. \tag{2.44}$$

Since for given $\phi \in \mathcal{D}$ and sufficiently large $\hat{t} \in \mathbb{R}$

$$\begin{aligned} \phi(0) &= -(\phi(\hat{t}) - \phi(0)) = -\phi(t)\Big|_0^{\hat{t}} = -\int_0^{\hat{t}} \dot{\phi}(t)\, dt \\ &= -\int_0^{\infty} \dot{\phi}(t)\, dt = -\int_{\mathbb{R}} H(t)\dot{\phi}(t)\, dt = -\langle H, \dot{\phi} \rangle = \langle \dot{H}, \phi \rangle, \end{aligned}$$

where

$$H(t) = \begin{cases} 0 & \text{for } t < 0, \\ 1 & \text{for } t \geq 0, \end{cases} \tag{2.45}$$

is the *Heaviside function*, we find that $\delta = \dot{H}$. We will also use the shifted versions H_{t_0} defined by $H_{t_0}(t) = H(t - t_0)$ and $\delta_{t_0} = \dot{H}_{t_0}$.

In order to define a primitive $x \in \mathcal{C}$ for a distribution $f \in \mathcal{C}$, let us consider the equation

$$\dot{x} = f. \tag{2.46}$$

We can rewrite this equation as

$$\langle \dot{x}, \phi \rangle = \langle x, -\dot{\phi} \rangle = \langle f, \phi \rangle \quad \text{for all } \phi \in \mathcal{D}. \tag{2.47}$$

Thus, x is already defined for every $\varphi \in \mathcal{D}$ which is the derivative of a test function, i.e., for which there is a $\phi \in \mathcal{D}$ with $\varphi = \dot{\phi}$. Note that for such a φ

$$\langle 1, \varphi \rangle = \langle 1, \dot{\phi} \rangle = -\langle 0, \phi \rangle = 0 \tag{2.48}$$

holds. In order to extend x to $\mathcal{D}$, we introduce a test function ϕ_1 for which

$$\int_{\mathbb{R}} \phi_1(t)\, dt = 1. \tag{2.49}$$

Then, any test function ψ can be written in the form

$$\psi = \lambda\phi_1 + \varphi, \tag{2.50}$$

where $\lambda \in \mathbb{C}$ and φ is a test function which is the derivative of another test function. Observing that (2.50) implies

$$\langle 1, \psi\rangle = \lambda\langle 1, \phi_1\rangle + \langle 1, \varphi\rangle,$$

and using (2.48) and (2.49), we see that λ and φ must be given by

$$\lambda = \langle 1, \psi\rangle, \quad \varphi = \psi - \langle 1, \psi\rangle\phi_1. \tag{2.51}$$

Indeed, we have $\langle 1, \varphi\rangle = 0$ and the function ϕ defined by

$$\phi(t) = \int_{-\infty}^{t} \varphi(s)\, ds \tag{2.52}$$

is a test function satisfying $\varphi = \dot{\phi}$. Thus, we define a primitive x of f by

$$\langle x, \psi\rangle = \langle x, \lambda\phi_1 + \varphi\rangle = \langle 1, \psi\rangle\langle x, \phi_1\rangle + \langle x, \varphi\rangle = \langle 1, \psi\rangle\langle x, \phi_1\rangle - \langle f, \phi\rangle. \tag{2.53}$$

Since we have only extended the relation (2.47), it is clear that this x satisfies (2.46) in the distributional sense. Of course, one must show that x is linear and continuous. For details, see, e.g., [199], [219]. Again, if x is a primitive of $f \in C(\mathbb{R}, \mathbb{C})$ in the classical sense, then the same holds for the associated distributions.

If $f = 0$ in (2.46), then a corresponding primitive x must satisfy

$$\langle x, \psi\rangle = \langle 1, \psi\rangle\langle x, \phi_1\rangle = \langle c, \psi\rangle$$

with $c = \langle x, \phi_1\rangle$. Hence, all primitives of $f = 0$ correspond to constant functions. This also shows that (like in the classical case) two primitives of the same distribution differ by a distribution which corresponds to a constant function. For arbitrary f in (2.46), we get a special primitive x by

$$\langle x, \psi\rangle = -\langle f, \phi\rangle \tag{2.54}$$

requiring $\langle x, \phi_1\rangle = 0$ in (2.53) as kind of initial condition.

Note finally that all results carry over to vector valued problems with test spaces of the form $\mathcal{D}^n$ and corresponding distribution spaces $\mathcal{C}^n$, by using

$$\langle f, \phi\rangle = \sum_{i=1}^{n} \langle f_i, \phi_i\rangle \tag{2.55}$$

for $f = [\, f_1 \ \cdots \ f_n \,]^T \in \mathcal{C}^n$ and $\phi = [\, \phi_1 \ \cdots \ \phi_n \,]^T \in \mathcal{D}^n$. In addition, we can use multiplication by matrix functions in the form

$$\langle Ax, \phi \rangle = \langle x, A^T \phi \rangle \text{ for all } \phi \in \mathcal{D}^m, \tag{2.56}$$

where $A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$ and $x \in \mathcal{C}^n$.

Our motivation to use generalized solutions was the desire to treat initial conditions that are inconsistent in the classical framework, but we are still faced with the problem that we cannot assign a value at a point $t_0 \in \mathbb{R}$ to a distribution. In $\mathcal{C}^n$, it makes therefore no sense to require a condition like (2.2). The idea now is to restrict the considerations to a subset of $\mathcal{C}^n$ in such a way that we can speak of values at some point. If we require (2.2), then a possible nonsmooth behavior of the solution should be restricted to the point t_0. Away from t_0, the solution should be as smooth as the solution in the classical sense. Thus, nonsmooth behavior of the solution may occur due to inconsistent initial data or nonsmooth behavior of the inhomogeneity. We will discuss both problems in what follows, assuming that nonsmooth behavior only occurs at a single point. For simplicity, we assume without loss of generality that this point is the origin. In the next chapter, we will discuss extensions to the case when nonsmooth behavior appears at more than one point.

Let us begin with the definition of an appropriate subspace of $\mathcal{C}$. For ease of notation, we treat every function $x : \mathbb{I} \to \mathbb{C}$, $\mathbb{I} \subseteq \mathbb{R}$, as being defined on $\mathbb{R}$ by trivially extending it by zero, i.e., by setting $x(t) = 0$ for $t \notin \mathbb{I}$.

Definition 2.37. A generalized function $x \in \mathcal{C}$ is called *impulsive smooth* if it can be written in the form

$$x = x_- + x_+ + x_{\text{imp}}, \tag{2.57}$$

where $x_- \in C^\infty((-\infty, 0], \mathbb{C})$, $x_+ \in C^\infty([0, \infty), \mathbb{C})$ and the *impulsive part* x_{imp} has the form

$$x_{\text{imp}} = \sum_{i=0}^{q} c_i \delta^{(i)}, \quad c_i \in \mathbb{C},\ i = 0, \ldots, q, \tag{2.58}$$

with some $q \in \mathbb{N}_0$. The set of impulsive smooth distributions is denoted by $\mathcal{C}_{\text{imp}}$.

We state here without proof (for details see [219]) that the distributions of the form (2.58) are exactly those distributions x_{imp} for which

$$\langle x_{\text{imp}}, \phi \rangle = 0 \quad \text{for all } \phi \in \mathcal{D} \text{ with } \operatorname{supp} \phi \subset \mathbb{R} \setminus \{0\}$$

holds.

Lemma 2.38. *Impulsive smooth distributions have the following properties:*

1. *A distribution $x \in \mathcal{C}_{\text{imp}}$ uniquely determines the decomposition (2.57).*

2. *With a distribution $x \in \mathcal{C}_{\text{imp}}$, we can assign a function value $x(t)$ for every $t \neq 0$ by*

$$x(t) = \begin{cases} x_-(t) & \text{for } t < 0, \\ x_+(t) & \text{for } t > 0, \end{cases}$$

and limits

$$x(0^-) = \lim_{t \to 0^-} x(t), \quad x(0^+) = \lim_{t \to 0^+} x(t).$$

3. *All derivatives and primitives of $x \in \mathcal{C}_{\text{imp}}$ are again in $\mathcal{C}_{\text{imp}}$.*

4. *The set $\mathcal{C}_{\text{imp}}$ is a (complex) vector space and closed under multiplication with functions $a \in C^\infty(\mathbb{R}, \mathbb{C})$. In particular, we have*

$$ax = ax_- + ax_+ + \sum_{i=0}^{q} \sum_{j=0}^{q-i} (-1)^j \binom{j+i}{j} a^{(j)}(0) c_{i+j} \delta^{(i)} \tag{2.59}$$

for x as in (2.57) with (2.58).

Proof. The proof is left as an exercise, cp. Exercise 14. □

With the obvious generalization of $\mathcal{C}_{\text{imp}}$ to vector valued problems, we can again define the multiplication of an element $x \in \mathcal{C}_{\text{imp}}^n$ with a matrix function $A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$. Decomposing x according to (2.57) and (2.58), where we replace $\mathbb{C}$ by $\mathbb{C}^n$, the distribution $Ax \in \mathcal{C}_{\text{imp}}^m$ is given by

$$Ax = Ax_- + Ax_+ + \sum_{i=0}^{q} \sum_{j=0}^{q-i} (-1)^j \binom{j+i}{j} A^{(j)}(0) c_{i+j} \delta^{(i)}. \tag{2.60}$$

Finally, we introduce a measure for the smoothness of impulsive smooth distributions. Note that we use the typical rules for calculations that involve $\pm\infty$.

Definition 2.39. Let the impulsive part of $x \in \mathcal{C}_{\text{imp}}^n$ have the form

$$x_{\text{imp}} = \sum_{i=0}^{q} c_i \delta^{(i)}, \quad c_i \in \mathbb{C}^n, \ i = 0, \dots, q. \tag{2.61}$$

The *impulse order* of x is defined as $\operatorname{iord} x = -q - 2$ if x can be associated with a continuous function, where q with $0 \leq q \leq \infty$ is the largest integer such that $x \in C^q(\mathbb{R}, \mathbb{C})$. It is defined as $\operatorname{iord} x = -1$ if x can be associated with a function that is continuous everywhere except at $t = 0$ and it is defined as

$$\operatorname{iord} x = \max\{i \in \mathbb{N}_0 \mid 0 \leq i \leq q, \ c_i \neq 0\}$$

otherwise.

Lemma 2.40. *Let $x \in \mathcal{C}^n_{\mathrm{imp}}$ and $A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$. Then*

$$\operatorname{iord} Ax \leq \operatorname{iord} x$$

with equality for $m = n$ and $A(0)$ invertible.

Proof. The claim is a direct consequence of (2.60). □

With these preliminaries, we have the following characterization of generalized solutions of ordinary differential equations

$$\dot{x} = A(t)x + f, \tag{2.62}$$

with $f \in \mathcal{C}^n_{\mathrm{imp}}$ and $A \in C^\infty(\mathbb{R}, \mathbb{C}^{n,n})$. For details and extended results, see, e.g., [180].

Theorem 2.41. *Let $A \in C^\infty(\mathbb{R}, \mathbb{C}^{n,n})$ and let $f \in \mathcal{C}^n_{\mathrm{imp}}$ have impulse order* $\operatorname{iord} f = q \in \mathbb{Z} \cup \{-\infty\}$. *Furthermore, let $t_0 \in \mathbb{R} \setminus \{0\}$ and $x_0 \in \mathbb{C}^n$. Then, we have the following:*

1. *All generalized solutions of (2.62) are in $\mathcal{C}^n_{\mathrm{imp}}$ and have impulse order $q - 1$.*
2. *There exists a unique solution of (2.62) in $\mathcal{C}^n_{\mathrm{imp}}$ satisfying one of the initial conditions*
$$x(t_0) = x_0, \quad x(0^-) = x_0, \quad x(0^+) = x_0. \tag{2.63}$$

Proof. Let $M \in C^\infty(\mathbb{R}, \mathbb{C}^{n,n})$ be the (unique) solution of the fundamental system

$$\dot{M} = A(t)M, \quad M(t_0) = I.$$

Then $M(t)$ is invertible for all $t \in \mathbb{R}$. With M^{-1} defined pointwise by $M^{-1}(t) = M(t)^{-1}$, it follows that $x \in \mathcal{C}^n$ solves (2.62) if and only if $z = M^{-1}x \in \mathcal{C}^n$ solves $\dot{z} = g = M^{-1}f$, i.e., z is a primitive of g. Since $f \in \mathcal{C}^n_{\mathrm{imp}}$, then also $g \in \mathcal{C}^n_{\mathrm{imp}}$ with the same impulse order according to Lemma 2.40.

Consider the decomposition $g = g_+ + g_- + g_{\mathrm{imp}}$, where $g_{\mathrm{imp}} = \sum_{i=0}^{q} c_i \delta^{(i)}$ with $c_i \in \mathbb{C}^n$, $i = 0, \ldots, q$, and $c_q \neq 0$, using the convention that $g_{\mathrm{imp}} = 0$ for $q < 0$. A primitive of g then has the form

$$z = c + \int_{t_0}^{t} (g_-(s) + g_+(s))\, ds + c_0 H + \sum_{i=0}^{q-1} c_{i+1} \delta^{(i)},$$

with some $c \in \mathbb{C}^n$. Hence, every primitive z of g has impulse order $q - 1$ and so every solution $x = Mz$ of (2.62). This finishes the proof of the first part.

To prove the second part, we must show that the transformed initial conditions $z(t_0) = z_0$, $z(0^-) = z_0$, and $z(0^+) = z_0$ fix the constant in the representation of z. But this is obvious due to $z(t_0) = c + c_0 H(t_0)$ in the first case and (setting formally $t_0 = 0$) due to $z(0^-) = c$ and $z(0^+) = c + c_0$ otherwise. □

This result gives a useful generalization of the classical solution theory for linear ordinary differential equations, since it allows discontinuous or even generalized forcing functions f, while still the uniqueness of solutions for the initial value problems is retained.

Now we return to differential-algebraic equations

$$E\dot{x} = Ax + f, \tag{2.64}$$

with a given $f \in \mathcal{C}^m_{\text{imp}}$ and we assume that an initial condition of one of the forms (2.63) is given. Since we can decompose $f = f_- + f_+ + f_{\text{imp}}$, we also consider the differential-algebraic systems

$$E\dot{x} = Ax + f_-, \quad t \in (-\infty, 0] \tag{2.65}$$

and

$$E\dot{x} = Ax + f_+, \quad t \in [0, \infty). \tag{2.66}$$

Note that $f_- \in C^\infty((-\infty, 0], \mathbb{C}^m)$ and $f_+ \in C^\infty([0, \infty), \mathbb{C}^m)$, hence it makes sense to consider initial values for (2.65) and (2.66) at $t_0 = 0$.

We have the immediate extension of Theorem 2.41.

Theorem 2.42. *Consider system (2.64) and assume that $m = n$ and that the pair (E, A) is regular with index $\nu = \operatorname{ind}(E, A)$. Assume further that $\operatorname{iord} f = q \in \mathbb{Z} \cup \{-\infty\}$. Then, we have the following:*

1. *All generalized solutions of (2.64) are in $\mathcal{C}^n_{\text{imp}}$ and have impulse order at most $q + \nu - 1$.*
2. *If $t_0 \neq 0$ and x_0 is consistent for (2.65) or (2.66) respectively, then the initial value problem (2.64) together with $x(t_0) = x_0$ has a unique solution in $\mathcal{C}^n_{\text{imp}}$.*
3. *If $t_0 = 0$ and x_0 is consistent for (2.65) or (2.66) respectively, then the initial value problem (2.64) together with $x(0^-) = x_0$ or $x(0^+) = x_0$ respectively, has a unique solution in $\mathcal{C}^n_{\text{imp}}$.*

Proof. The proof follows immediately by transforming (E, A) to Weierstraß canonical form (2.7) and then considering the different parts in the distributional sense. We obtain the two distributional systems

$$\dot{x}_1 = Jx_1 + f_1 \tag{2.67}$$

and

$$N\dot{x}_2 = x_2 + f_2. \tag{2.68}$$

The initial condition transforms analogously as

$$x_1(t_0) = x_{1,0}, \quad x_2(t_0) = x_{2,0}.$$

For (2.67), we can apply Theorem 2.41 and obtain directly all assertions.

For the analysis of (2.68), we use the decomposition $f_2 = f_{2,-} + f_{2,+} + f_{2,\mathrm{imp}}$. Applying the construction in Lemma 2.8, we obtain that this part has the unique solution

$$x_2 = -\sum_{i=0}^{\nu-1} N^i (f_{2,-}^{(i)} + f_{2,+}^{(i)} + f_{2,\mathrm{imp}}^{(i)}). \tag{2.69}$$

Hence, x_2 is impulsive smooth and the impulse order is at most $q + \nu - 1$. Thus, the first part of the assertion follows.

To prove the second part, observe that consistency of x_0 implies that $x_{2,0} = x_2(t_0)$, i.e., the initial condition does not contradict the only possible solution x_2.

The third part follows analogously. □

We see from this theorem that for a regular matrix pair and consistent initial values, we have in principle the same existence and uniqueness result for distributional forcing terms as in the classical case. The only difference is that we have no smoothness restriction for the inhomogeneity. But up to now, we have not addressed the problem of inconsistent initial conditions. Suppose that (2.64) is given together with $x_-^0 \in C^\infty((-\infty, 0], \mathbb{C}^n)$ to indicate how the system has behaved until $t = 0$. The initial condition $x(0^-) = x_0$ with $x_0 = x_-^0(0)$, however, may not be consistent for (2.65). Setting

$$f_- = E\dot{x}_-^0 - Ax_-^0 \tag{2.70}$$

forces x_-^0 to be a solution for the part (2.65), thus making the initial condition consistent. The problem under consideration therefore should be

$$E\dot{x} = Ax + f, \quad x_- = x_-^0, \tag{2.71}$$

where $f = f_- + f_+ + f_{\mathrm{imp}}$ and f_- satisfies (2.70).

Theorem 2.43. *Let (E, A) be regular with index $\nu = \mathrm{ind}(E, A)$. Let $x_-^0 \in C^\infty((-\infty, 0], \mathbb{C}^n)$ be given and let $f = f_- + f_+ + f_{\mathrm{imp}} \in \mathcal{C}_{\mathrm{imp}}^n$, where f_- satisfies (2.70). Then the following statements hold:*

1. *The problem (2.71) has a unique solution $x \in \mathcal{C}_{\mathrm{imp}}^n$ with $\mathrm{iord}\, x \le \mathrm{iord}\, f + \nu - 1$.*
2. *Let $x = x_- + x_+ + x_{\mathrm{imp}}$ be the unique solution of (2.71) and $f = f_- + f_+ + f_{\mathrm{imp}}$. Then $\tilde{x} = x_+ + x_{\mathrm{imp}}$ is the unique solution of*

$$E\dot{\tilde{x}} = A\tilde{x} + \tilde{f} + Ex_0\delta, \quad \tilde{x}_- = 0, \tag{2.72}$$

where $x_0 = x_-^0(0)$ and $\tilde{f} = f_+ + f_{\mathrm{imp}}$.

Proof. The first part immediately follows from Theorem 2.42. For the second part, we observe that $x = \tilde{x} + x_-$. Since

$$\dot{x} = \dot{x}_- + \dot{x}_+ + \dot{x}_{\text{imp}} + (x_+(0) - x_-(0))\delta, \quad \dot{\tilde{x}} = \dot{x}_+ + \dot{x}_{\text{imp}} + x_+(0)\delta,$$

(2.71) is equivalent to

$$E(\dot{\tilde{x}} + \dot{x}_- - x_-(0)\delta) = A(\tilde{x} + x_-) + f, \quad \tilde{x}_- = 0,$$

hence to (2.72) due to (2.70). □

Remark 2.44. Within the framework of distributions, inconsistent initial conditions can be treated by changing the inhomogeneity to make the system satisfy a given history. The second claim of Theorem 2.43 shows that the impulsive behavior and the future smooth development of the system do not depend on the whole history but only on the (possibly inconsistent) initial condition. In this sense, problem (2.72) is the adequate form to treat inconsistent initial conditions. Observe that the initial condition does not appear as it is stated in the classical formulation (we cannot specify values of distributions) but as part of the inhomogeneity. The physical relevance of the solution x (or equivalently $\tilde{x}$) follows from the following property. The inhomogeneity $f \in \mathcal{C}^n_{\text{imp}}$ can be represented as

$$f = \lim_{\ell \to \infty} f_\ell,$$

where $f_\ell \in C^\infty(\mathbb{R}, \mathbb{C}^n)$ and $f(t) = f_\ell(t)$ for all $t \in \mathbb{R}$ with $|t| \geq 1/\ell$. Here the limit is meant to be taken in $\mathcal{C}^n$, i.e.,

$$\langle f, \phi \rangle = \lim_{\ell \to \infty} \langle f_\ell, \phi \rangle \quad \text{for all } \phi \in \mathcal{D}^n.$$

Let x_ℓ be the (classical) solution of

$$E\dot{x} = Ax + f_\ell(t), \quad x(-1/\ell) = x^0_-(-1/\ell).$$

Then

$$x = \lim_{\ell \to \infty} x_\ell \in \mathcal{C}^n_{\text{imp}}$$

is the unique solution of (2.71). Hence, the generalized solution can be seen as limit of (classical) solutions by "smearing out" nonsmooth behavior of the inhomogeneity. For more details, see [180].

Example 2.45. If we set $u = u_0(1-H)$ in the mathematical model of Example 2.34, we get the differential-algebraic equation

$$x_1 - x_3 = u_0(1 - H), \quad C(\dot{x}_3 - \dot{x}_2) + \frac{x_1 - x_2}{R} = 0, \quad x_3 = 0,$$

which has index $\nu = 1$. Since iord $f = -1$, Theorem 2.42 yields that all solutions x satisfy iord $x \leq -1$. Hence, they can be associated with functions. The initial condition $x(0^-) = [\, u_0 \; u_0 \; 0 \,]^T$ is consistent such that there exists a unique solution of the corresponding initial value problem. Indeed, this is given by the distributions corresponding to

$$x_1(t) = u_0(1 - H(t)), \quad x_2(t) = \begin{cases} u_0 & \text{for } t < 0, \\ u_0 e^{-t/RC} & \text{for } t \geq 0, \end{cases} \quad x_3(t) = 0.$$

In this way, we have shown that there is a mathematically rigorous argument that x_2 solves the differential equation given in Example 2.34, although it is not differentiable for $t = 0$.

Example 2.46. For $u = H$ and $A \to \infty$, the model equations of Example 2.35 take the form

$$x_1 - x_4 = H, \quad C(\dot{x}_1 - \dot{x}_2) + \frac{x_3 - x_2}{R} = 0, \quad x_2 = 0, \quad x_4 = 0.$$

This is a differential-algebraic equation with index $\nu = 2$ and iord $f = -1$. Since the initial value $x(-1) = 0$ is consistent, Theorem 2.42 guarantees that there exists a unique solution x with iord $x \leq 0$. Indeed, we find

$$x = [\, H \;\; 0 \;\; -RC\delta \;\; 0 \,]^T$$

and Figures 2.3 and 2.4 nicely demonstrate the possible limiting behavior when we use smooth approximations to the inhomogeneity, cp. Remark 2.44.

In the case of singular matrix pairs, the analysis can be carried out as in the classical case. Indeed, we obtain the same results as in Theorem 2.14.

Theorem 2.47. *Let (E, A) with $E, A \in \mathbb{C}^{m,n}$ be a singular pair of matrices.*

1. *If* $\operatorname{rank}(\lambda E - A) < n$ *for all* $\lambda \in \mathbb{C}$, *then the distributional version of the homogeneous initial value problem*

$$E\dot{x} = Ax, \quad x_-^0 = 0 \tag{2.73}$$

has a nontrivial solution in $\mathcal{C}_{\text{imp}}^n$.

2. *If* $\operatorname{rank}(\lambda E - A) = n$ *for some* $\lambda \in \mathbb{C}$ *and hence* $m > n$, *then there exist arbitrarily smooth forcing functions* f *in* $\mathcal{C}_{\text{imp}}^m$ *such that (2.64) does not have a generalized solution in* $\mathcal{C}_{\text{imp}}^n$.

Proof. The proof of Theorem 2.14 works verbatim in the case of distributions. □

The formulation (2.72) of an initial value problem suggests that for sufficiently smooth $\tilde{f}$ the smoothness of $\tilde{x}$ will depend on the initial condition. We therefore assume now that in the (possibly nonsquare) problem

$$E\dot{x} = Ax + f + Ex_0\delta, \quad x_- = 0 \tag{2.74}$$

the distribution $f \in \mathcal{C}_{\text{imp}}^m$ has iord $f \leq -1$ and satisfies $f_- = 0$.

Definition 2.48. Let $f \in \mathcal{C}_{\text{imp}}^m$ be given with $f_- = 0$ and iord $f \leq -1$. We say that $x_0 \in \mathbb{C}^n$ is *weakly consistent* with f if there exists a solution $x \in \mathcal{C}_{\text{imp}}^n$ of (2.74) with iord $x \leq -1$. We say that x_0 is *consistent* with f if x_0 is weakly consistent with f and there exists a solution $x \in \mathcal{C}_{\text{imp}}^n$ of (2.74) satisfying $x(0^+) = x_0$.

We have the following theorem which is a reformulation of a result in [96].

Theorem 2.49. *Let the concatenated matrix* $[\, E \;\; A \,]$ *in* (2.74) *have full row rank and let* $f \in \mathcal{C}_{\text{imp}}^m$ *be given with* $f_- = 0$ *and* iord $f \leq -1$. *Then we have the following characterizations:*

1. *All vectors* $x_0 \in \mathbb{C}^n$ *are consistent with* f *if and only if*

$$\text{range } E = \mathbb{C}^m. \tag{2.75}$$

2. *All vectors* $x_0 \in \mathbb{C}^n$ *are weakly consistent with* f *if and only if*

$$\text{range } E + A \,\text{kernel } E = \mathbb{C}^m. \tag{2.76}$$

Proof. For the first part, suppose that all vectors in $\mathbb{C}^n$ are consistent and that there exists a nonzero $z \in \mathbb{C}^m$ such that $z^H E = 0$. Then (2.74) implies that

$$0 = z^H A x_0 + z^H f(0^+)$$

for a corresponding solution x. Since this must hold for all $x_0 \in \mathbb{C}^n$, we get $z^H A = 0$ and $[\, E \;\; A \,]$ cannot have full row rank. For the reverse direction, assume that range $E = \mathbb{C}^m$. Without loss of generality, we may assume that $E = [\, I_m \;\; 0 \,]$ and that

$$A = \begin{bmatrix} A_1 & A_2 \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x_0 = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}$$

are partitioned accordingly. We then obtain the system

$$\dot{x}_1 = A_1 x_1 + A_2 x_2 + f + x_{1,0}\delta.$$

If we choose $x_2 = Hx_{2,0}$, then the inhomogeneity of this differential equation satisfies

$$\text{iord}(A_2 x_2 + f + x_{1,0}\delta) \leq 0.$$

Theorem 2.43 yields $\operatorname{iord} x_1 \leq -1$ for the corresponding solution x_1. Since $x_1(0^+) = x_{1,0}$ (cp. Exercise 15) and $x_2(0^+) = x_{2,0}$, we have that x_0 is consistent.

For the second part, we may assume without loss of generality that the system has the form

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} \delta.$$

The statement to prove is then that all vectors $x_0 \in \mathbb{C}^n$ are weakly consistent with f if and only if A_{22} has full row rank. Suppose that A_{22} does not have full row rank. Then there exists a nonzero vector z such that $z^H A_{22} = 0$ and thus $0 = z^H A_{21} x_{1,0} + z^H f_2$ for all $x_{1,0}$. But then $z^H A_{21} = 0$ and $[\, E \;\; A \,]$ cannot have full row rank. For the other direction, let $x_{1,0}, x_{2,0}$ be given. Let A_{22}^+ be the Moore–Penrose pseudoinverse of A_{22}, see, e.g., [56] or Section 3.4. Then, the distribution

$$x_2 = -A_{22}^+ (A_{21} x_1 + f_1)$$

solves the second block equation for any given distribution x_1, since $A_{22} A_{22}^+ = I$ due to the full row rank of A_{22}. Moreover, $\operatorname{iord} x_2 \leq \max\{\operatorname{iord} x_1, -1\}$. For x_1, it remains to solve

$$\dot{x}_1 = (A_{11} - A_{12} A_{22}^+ A_{21}) x_1 + (f_1 - A_{12} A_{22}^+ f_2) + x_{1,0} \delta.$$

By Theorem 2.43, it follows that $\operatorname{iord} x_1 \leq -1$. Thus, the corresponding x_0 is weakly consistent. Note here that $x_1(0^+) = x_{1,0}$ but

$$x_2(0^+) = -A_{22}^+ (A_{21} x_{1,0} + f_1(0^+))$$

may be different from $x_{2,0}$. □

Remark 2.50. If (E, A) forms a regular matrix pair, then condition (2.76) in Theorem 2.49 means that the pair has $\nu = \operatorname{ind}(E, A) \leq 1$. For a singular pair under the assumption that $[\, E \;\; A \,]$ has full row rank, this condition means that all the nilpotent blocks in the Kronecker canonical form have dimension less than or equal to one, i.e., the index of the regular part is less than or equal to one.

We see from Theorem 2.49 that the consistency of initial values still represents a restriction if we want a regular distribution as solution, i.e., a solution that can be associated with a function. Thus, as noted in Remark 2.32, we can relax the smoothness requirements for f, but not the consistency requirements for the initial values. On the other hand, the second part of Theorem 2.49 shows that for the case of systems with a regular part of index less than or equal to one, generalized solutions exist that can be associated with functions, regardless of the choice of initial conditions.

Example 2.51. Consider the system in distributional form

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} \delta.$$

with $f_- = 0$. Note that this differential-algebraic equation has index $\nu = 1$. Requiring that $x_- = 0$, we obtain $x_2 = -f_2$ and x_1 solves $\dot{x}_1 = x_1 + f_1 + x_{1,0}\delta$. If iord $f \leq -1$, then both components can be associated with functions, regardless of the choice of the initial condition. Thus, Remark 2.50 says that all vectors x_0 are weakly consistent in this case.

If we take the differential-algebraic equation from Example 2.46, then the corresponding system (2.74) reads

$$x_1 - x_4 = H, \quad C(\dot{x}_1 - \dot{x}_2) + \frac{x_3 - x_2}{R} = C(x_{1,0} - x_{2,0})\delta, \quad x_2 = 0, \quad x_4 = 0.$$

The unique solution satisfying $x_- = 0$ is given by

$$x = [\, H \;\; 0 \;\; RC(x_{1,0} - x_{2,0} - 1)\delta \;\; 0 \,]^T.$$

Thus, we have $x_{\text{imp}} = 0$ if and only if $x_{1,0} - x_{2,0} = 1$. In particular, for this differential-algebraic equation with $\nu = 2$ the set of weakly consistent values x_0 is restricted.

2.5 Control problems

Linear control problems with constant coefficients (sometimes also called descriptor systems) have the form

$$E\dot{x} = Ax + Bu + f(t), \tag{2.77a}$$
$$y = Cx + g(t), \tag{2.77b}$$

with $E, A \in \mathbb{C}^{m,n}$, $B \in \mathbb{C}^{m,l}$, $C \in \mathbb{C}^{p,n}$, $f \in C(\mathbb{I}, \mathbb{C}^m)$ and $g \in C(\mathbb{I}, \mathbb{C}^p)$. Here, x represents the state, u the input or control, and y the output of the system. Typically, one also has an initial condition of the form (2.2). The distinction between x and y naturally occurs in applications, since we cannot expect that all quantities necessary to model a physical system can be measured.

Example 2.52. Consider a simple RLC electrical circuit shown in Figure 2.5, cp. [67]. The voltage source v_S is the control input, R, L, and C are the resistance, inductance and capacitance, respectively. The corresponding voltage drops are denoted by v_R, v_L, and v_C, respectively, and I denotes the current.

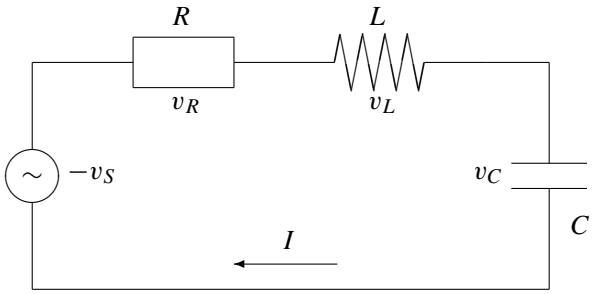


Figure 2.5. A simple RLC circuit

Applying Kirchhoff's laws, we obtain the following circuit equation.

$$\begin{bmatrix} L & 0 & 0 & 0 \\ 0 & 0 & C & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{I} \\ \dot{v}_L \\ \dot{v}_C \\ \dot{v}_R \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -R & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} I \\ v_L \\ v_C \\ v_R \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} v_S.$$

If we measure the voltage at the capacitor as output, we also have the output equation

$$y = [\,0\ 0\ 1\ 0\,] \begin{bmatrix} I \\ v_L \\ v_C \\ v_R \end{bmatrix}.$$

The general theory of control problems for differential-algebraic systems is still a very active research area. For this reason, we discuss only topics that are related directly to the theory of existence, uniqueness and regularization by feedback. For a general behavior approach and its analytical treatment, see [167].

We begin with two major properties of the system, namely consistency and regularity.

Definition 2.53. A control problem of the form (2.77) is called *consistent*, if there exists an input function u, for which the resulting differential-algebraic equation is solvable.

It is called *regular*, if for every sufficiently smooth input function u and inhomogeneity f the corresponding differential-algebraic equation is solvable and the solution is unique for every consistent initial value.

An immediate consequence of Theorem 2.12 is the following sufficient condition for consistency and regularity of control problems.

Corollary 2.54. *If the pair (E, A) of square matrices is regular, then the control problem (2.77) is consistent and regular.*

Proof. Taking the control $u = 0$, we obtain from Theorem 2.12 that the system is solvable and hence we have consistency. The regularity follows trivially by Theorem 2.12. □

Applying Theorem 2.14, we can show that regularity of (2.77) is equivalent to regularity of the matrix pair (E, A) as in the case of differential-algebraic equations.

Corollary 2.55. *If (E, A) with $E, A \in \mathbb{C}^{m,n}$ is a singular matrix pair, then (2.77) is not regular.*

Proof. If $\operatorname{rank}(\lambda E - A) < n$ for all $\lambda \in \mathbb{C}$, then we choose $u = 0$ and $f = 0$. Following the first part of Theorem 2.14, the resulting homogeneous differential-algebraic equation $E\dot{x} = Ax$ together with $x(t_0)$ has more than one solution.

If $\operatorname{rank}(\lambda E - A) = n$ for some $\lambda \in \mathbb{C}$ and hence $m > n$, then we again choose $u = 0$. The second part of Theorem 2.14 then yields that there exist arbitrarily smooth inhomogeneities f for which the resulting differential-algebraic equation is not solvable.

Hence, in both cases, system (2.77) is not regular. □

We have thus shown that the characterization of regularity can be obtained analogously to the analysis of the differential-algebraic equation. In the control context, however, it is possible to modify system properties using feedbacks. Possible feedbacks are *proportional state feedback*

$$u = Fx + w \tag{2.78}$$

and *proportional output feedback*

$$u = Fy + w. \tag{2.79}$$

If we apply these feedbacks, then we obtain the so-called *closed loop systems*

$$E\dot{x} = (A + BF)x + Bw + f(t) \tag{2.80}$$

or

$$E\dot{x} = (A + BFC)x + Bw + f(t) + BFg(t), \tag{2.81}$$

respectively. Thus, these feedbacks can be used to modify the system properties, in particular to make non-regular systems regular or to change the index of the system. This is important in realistic control applications, where the input function u typically is discontinuous like in bang-bang control, see [16]. In such a situation it is essential to choose a feedback so that the closed loop system is regular and of index at most one. There exists a completely algebraic characterization when this is possible.

Theorem 2.56. *Given a matrix quadruple* (E, A, B, C) *as in* (2.77).

1. *There exists an* $F \in \mathbb{C}^{l,n}$ *such that the matrix pair* $(E, A + BF)$ *is regular and of index* $\nu = \operatorname{ind}(E, A + BF)$ *at most one if and only if* E, A *are square and*

$$\operatorname{rank} [\, E \;\; AT \;\; B \,] = n, \tag{2.82}$$

   where T *is a matrix whose columns span* kernel E.

2. *There exists an* $F \in \mathbb{C}^{l,p}$ *such that the matrix pair* $(E, A + BFC)$ *is regular and of index* $\nu = \operatorname{ind}(E, A + BFC)$ *at most one if and only if* E, A *are square and* (2.82) *as well as*

$$\operatorname{rank} \begin{bmatrix} E \\ Z^H A \\ C \end{bmatrix} = n \tag{2.83}$$

   hold, where Z *is a matrix whose columns span* kernel E^H.

Proof. It is clear that the system has to be square, since otherwise the closed loop pair cannot be regular. Let $P, Q \in \mathbb{C}^{n,n}$ be nonsingular matrices such that

$$PEQ = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \quad PAQ = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad PB = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}, \quad CQ = [\, C_1 \;\; C_2 \,].$$

Then condition (2.82) is equivalent to

$$\operatorname{rank} [\, A_{22} \;\; B_2 \,] = n - r \tag{2.84}$$

and condition (2.83) is equivalent to

$$\operatorname{rank} \begin{bmatrix} A_{22} \\ C_2 \end{bmatrix} = n - r. \tag{2.85}$$

It is sufficient to prove only the second part, since the first part follows from the second part for $C = I$. Following Exercise 3, the matrix pair $(E, A + BFC)$ is regular and of index at most one if and only if the matrices are square and $A_{22} + B_2 F C_2$ is either not present or nonsingular. Let U, V be nonsingular with

$$U^{-1} A_{22} V = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Setting

$$A_{22}^{-} = V \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} U^{-1},$$

we have $A_{22} A_{22}^{-} A_{22} = A_{22}$, such that

$$A_{22} + B_2 F C_2 = [\, A_{22} \;\; B_2 \,] \begin{bmatrix} A_{22}^{-} & 0 \\ 0 & F \end{bmatrix} \begin{bmatrix} A_{22} \\ C_2 \end{bmatrix}. \tag{2.86}$$

Suppose that there exists $F \in \mathbb{C}^{l,p}$ such that $A_{22} + B_2FC_2$ is nonsingular. Then the factorization (2.86) of $A_{22} + B_2FC_2$ immediately implies (2.84) and (2.85). Conversely, let (2.84) and (2.85) hold. With the transformation of A_{22} by U, V we obtain

$$\operatorname{rank}(A_{22} + B_2FC_2) = \operatorname{rank}\left(\begin{bmatrix} I & 0 & \tilde{B}_1 \\ 0 & 0 & \tilde{B}_2 \end{bmatrix}\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & F \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & 0 \\ \tilde{C}_1 & \tilde{C}_2 \end{bmatrix}\right).$$

By similar transformations for $\tilde{B}_2$ and $\tilde{C}_2$, which have maximal rank due to (2.84) and (2.85), we have that

$$\operatorname{rank}(A_{22}+B_2FC_2) = \operatorname{rank}\left(\begin{bmatrix} I & 0 & B_{11} & B_{12} \\ 0 & 0 & I & 0 \end{bmatrix}\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & F_{11} & F_{12} \\ 0 & 0 & F_{21} & F_{22} \end{bmatrix}\begin{bmatrix} I & 0 \\ 0 & 0 \\ C_{11} & I \\ C_{21} & 0 \end{bmatrix}\right).$$

The choice $F_{11} = I$, $F_{12} = 0$, $F_{21} = 0$, and $F_{22} = 0$, corresponding to a specific choice of F, finally gives

$$\operatorname{rank}(A_{22} + B_2FC_2) = \operatorname{rank}\begin{bmatrix} I + B_{11}C_{11} & B_{11} \\ C_{11} & I \end{bmatrix} = n - r. \qquad \square$$

For further reading on control problems for differential-algebraic equations with constant coefficients, we refer the reader to [67], [155], [167].

Bibliographical remarks

Linear constant coefficient systems are discussed in the textbooks [29], [42], [43], [100], [105], [108]. Corresponding control problems are studied in [2], [67], [155].

The complete theoretical analysis follows already from the work of Weierstraß [223], [224] and Kronecker [121] on canonical forms of matrix pairs under strong equivalence transformations. The application to differential-algebraic systems can be explicitly found already in the book of Gantmacher [88].

Later work on matrix pairs has taken many different directions. The main interest came from pairs with structure [98], [212] and numerical methods [69], [70], [77], [78], [216], [226]. The explicit solution formulas of Section 2.3 were derived in [57]. Generalizations to the nonsquare case can be found in [41]. See also [42], [43].

Although matrix pairs and also the theory of distributions were well established in the mathematical literature, major developments in the context of distributional

solutions are rather new. They started with the work of Cobb [64] and Verghese, Levy, and Kailath [217] in linear control theory. The analysis was essentially completed in the work of Geerts [96], [97], and Rabier and Rheinboldt [179], [180], [182]. The presentation that we have given here is mostly based on this work.

The regularization via feedback for control problems associated with linear differential-algebraic equations with constant coefficients has been studied in [33], [34], [39], [132].

Exercises

1. Discuss the inhomogeneous differential-algebraic equations belonging to a block $\operatorname{diag}(\mathcal{L}_\varepsilon, \mathcal{M}_\eta)$, $\varepsilon, \eta \in \mathbb{N}_0$, of the Kronecker canonical form with respect to solvability and consistency of initial conditions and inhomogeneities.
2. Check whether the matrix pairs
$$\left(\begin{bmatrix} 1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix}\right), \quad \left(\begin{bmatrix} 2 & -1 & 1 \\ 3 & -2 & 2 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 1 & -1 & 1 \end{bmatrix}\right)$$
are regular or singular and determine their Kronecker canonical forms by elementary row and column transformations.
3. Show that the matrix pair
$$(E, A) = \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right),$$
with $E, A \in \mathbb{C}^{m,n}$ and $r < \min\{m, n\}$, is regular and of index one if and only if A_{22} is square and nonsingular.
4. For $A \in \mathbb{C}^{n,n}$, the matrix exponential is defined by
$$e^A = \sum_{i=0}^{\infty} \frac{1}{i\,!} A^i.$$
   (a) Show that the series is absolutely convergent.
   (b) Show that $T^{-1} e^A T = e^{T^{-1}AT}$ for nonsingular $T \in \mathbb{C}^{n,n}$.
   (c) For $A, B \in \mathbb{C}^{n,n}$ with $AB = BA$, prove that $e^{A+B} = e^A e^B$ and $e^A B = B e^A$.
   (d) Show by counterexamples that the claims from (c) do not hold in general, when the matrices do not commute.
5. Show that the solution of the initial value problem $\dot{x} = Ax$, $x(t_0) = x_0$ is given by $x(t) = e^{A(t-t_0)} x_0$. Use this result for variation of the constant to obtain an explicit representation of the solution of the initial value problem $\dot{x} = Ax + f(t)$, $x(t_0) = x_0$.

6. Determine the Drazin inverse E^D and the corresponding decomposition $E = \tilde{C} + \tilde{N}$ as in (2.21) for
$$E = \begin{bmatrix} 2 & 1 & -1 \\ 0 & 1 & 1 \\ 1 & 2 & 1 \end{bmatrix}$$
and for
$$E = \begin{bmatrix} 1 & 0 & 1 \\ 6 & -6 & 6 \\ 8 & -9 & 8 \end{bmatrix}.$$
7. Prove that $((E^D)^D)^D = E^D$ for all $E \in \mathbb{C}^{n,n}$.
8. Let $E, A \in \mathbb{C}^{n,n}$ satisfy $EA = AE$. Show that kernel $E \cap$ kernel $A = \{0\}$ if and only if (E, A) is a regular matrix pair. Also show that in this case $\operatorname{ind}(E, A) = \operatorname{ind} E$.
9. Let $E, A \in \mathbb{C}^{n,n}$ with (E, A) regular and $\nu = \operatorname{ind}(E, A)$. Furthermore, let $\tilde{\lambda} \in \mathbb{C}$ be chosen such that $\tilde{\lambda}E - A$ is nonsingular. Show that
$$\operatorname{ind}((\tilde{\lambda}E - A)^{-1}E) = \nu.$$
10. Let E, A and $\tilde{\lambda}$ be given as in Exercise 9 and let $EA = AE$. Show that
$$((\tilde{\lambda}E - A)^{-1}E)^D = E^D(\tilde{\lambda}E - A).$$
11. Let E, A and $\tilde{\lambda}$ be given as in Exercise 9 and let $\tilde{E} = (\tilde{\lambda}E - A)^{-1}E$, $\tilde{A} = (\tilde{\lambda}E - A)^{-1}A$. Show that $\tilde{E}^D(\tilde{\lambda}E - A)^{-1}$, $\tilde{A}^D(\tilde{\lambda}E - A)^{-1}$, $\tilde{E}^D\tilde{E}$, $\tilde{E}\tilde{A}^D$, and $\operatorname{ind}\tilde{E}$ are independent of $\tilde{\lambda}$.
12. If $\operatorname{ind} E \leq 1$, then the Drazin inverse E^D of E is also called group inverse of E and denoted by $E^{\#}$. Show that $E \in \mathbb{C}^{n,n}$ is an element of a group $\mathbb{G} \subseteq \mathbb{C}^{n,n}$ with respect to matrix multiplication if and only if $\operatorname{ind} E \leq 1$, and that the inverse of E in such a group is just $E^{\#}$.
13. Let $f \in C(\mathbb{R}, \mathbb{C})$ be given and suppose that there exists a closed interval $\mathbb{I} \subseteq \mathbb{R}$ such that f restricted to $\mathbb{I}$ is continuous and vanishes on the complement. Identify f with its induced distribution. Show that
$$\langle f, \phi \rangle = 0 \quad \text{for all } \phi \in \mathcal{D}$$
implies that $f(t) = 0$ for all $t \in \mathbb{R}$.
14. Prove Lemma 2.38 using Exercise 13.
15. Let $A \in C^\infty(\mathbb{R}, \mathbb{C}^{n,n})$, $f \in \mathcal{C}^n_{\text{imp}}$ with $f_- = 0$ and $\operatorname{iord} f \leq -1$, and let $x_0 \in \mathbb{C}^n$. Show that
$$\dot{x} = A(t)x + f + x_0\delta, \quad x_- = 0$$
has a unique solution x that satisfies $\operatorname{iord} x \leq -1$ and $x(0^+) = x_0$.
16. Analyze the system $E\dot{x} = Ax + f + Ex_0\delta, x_- = 0$ with
$$E = \begin{bmatrix} 2 & 2 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad A = \begin{bmatrix} 0 & -1 & 1 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & -1 & 2 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \quad f = \begin{bmatrix} 3\delta \\ 2\delta \\ \delta + H \\ H \end{bmatrix}.$$

(a) Characterize all the solutions for $x_0 = [\,1\ \ -1\ \ 1\ \ 0\,]^T$.

(b) Does there exist a generalized solution x with $\operatorname{iord} x \leq -1$ for some x_0?

17. Show that condition (2.76) is invariant under (strong) equivalence transformations.

18. Analyze condition (2.76) for the different types of blocks in the Kronecker canonical form.

19. Let a control problem (2.77) be given with a regular pair (E, A) and $f = 0$, $g = 0$. Suppose that we have a periodic input $u(t) = e^{i\omega t}u_0$, where i is the imaginary unit and ω is the frequency. Show that if $i\omega E - A$ is regular, then there exists a periodic output of the system given by $y(t) = W(i\omega)u(t)$. Determine the function W.

20. Consider the control problem

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ 0 \\ -1 \end{bmatrix} u$$

with output

$$y = [\,0\ 0\ 1\ 0\,] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}.$$

Check consistency and regularity for this system.

21. Consider the control problem of Exercise 20. Does there exist a proportional state or output feedback that makes the closed loop system regular and of index at most one?

Chapter 3
Linear differential-algebraic equations with variable coefficients

In this chapter, we discuss general linear differential-algebraic equations with variable coefficients of the form

$$E(t)\dot{x} = A(t)x + f(t), \tag{3.1}$$

where $E, A \in C(\mathbb{I}, \mathbb{C}^{m,n})$ and $f \in C(\mathbb{I}, \mathbb{C}^m)$, again possibly together with an initial condition

$$x(t_0) = x_0. \tag{3.2}$$

3.1 Canonical forms

Comparing with the case of constant coefficients in Chapter 2, in view of Theorem 2.12, an obvious idea in dealing with (3.1) for $m = n$ would be to require regularity of the matrix pair $(E(t), A(t))$ for all $t \in \mathbb{I}$. But, unfortunately, this does not guarantee unique solvability of the initial value problem. Moreover, it turns out that these two properties are completely independent of each other.

Example 3.1. Let E, A and f be given by

$$E(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad f(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbb{I} = \mathbb{R}.$$

Since

$$\det(\lambda E(t) - A(t)) = (1 - \lambda t)(1 + \lambda t) + \lambda^2 t^2 = 1,$$

the matrix pair $(E(t), A(t))$ is regular for all $t \in \mathbb{I}$. A short computation shows that x given by

$$x(t) = c(t) \begin{bmatrix} t \\ 1 \end{bmatrix}$$

is a solution of the corresponding homogeneous initial value problem for every $c \in C^1(\mathbb{I}, \mathbb{C})$ with $c(t_0) = 0$. In particular, there exists more than one solution.

Example 3.2. Let E, A, and f be given by

$$E(t) = \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad \mathbb{I} = \mathbb{R}$$

with $f \in C^2(\mathbb{I}, \mathbb{C}^2)$. Since

$$\det(\lambda E(t) - A(t)) = -\lambda t + \lambda t = 0,$$

the matrix pair $(E(t), A(t))$ is singular for all $t \in \mathbb{I}$. With $x = \left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]$, we can write the corresponding differential-algebraic equation as

$$0 = -x_1(t) + tx_2(t) + f_1(t), \quad \dot{x}_1(t) - t\dot{x}_2(t) = f_2(t).$$

The first equation gives $x_1(t) = tx_2(t) + f_1(t)$. Differentiating this equation and inserting it into the second equation then gives the unique solution

$$x_1(t) = tf_2(t) - t\dot{f}_1(t) + f_1(t), \quad x_2(t) = f_2(t) - \dot{f}_1(t).$$

In this case therefore every initial value problem with consistent initial condition is uniquely solvable.

The reason for this, at first sight, strange behavior is that the equivalence relation (2.4) is not adequate for differential-algebraic equations with variable coefficients. We must admit invertible, time-dependent transformations, which pass a differential-algebraic equation with variable coefficients into an equation of similar form. Given matrix functions P and Q of appropriate size which are pointwise nonsingular, we can scale the equation by multiplying with P from the left and the function x according to $x = Q\tilde{x}$ as in the case of constant coefficients. But to determine the transformed equation, we must now differentiate $x = Q\tilde{x}$ according to $\dot{x} = Q\dot{\tilde{x}} + \dot{Q}\tilde{x}$. Thus, due to the product rule we get an additional term $\dot{Q}\tilde{x}$. In particular, this means that we must consider a different kind of equivalence.

Definition 3.3. Two pairs (E_i, A_i), $E_i, A_i \in C(\mathbb{I}, \mathbb{C}^{m,n})$, $i = 1, 2$, of matrix functions are called (*globally*) *equivalent* if there exist pointwise nonsingular matrix functions $P \in C(\mathbb{I}, \mathbb{C}^{m,m})$ and $Q \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ such that

$$E_2 = PE_1Q, \quad A_2 = PA_1Q - PE_1\dot{Q} \tag{3.3}$$

as equality of functions. We again write $(E_1, A_1) \sim (E_2, A_2)$.

Lemma 3.4. *The relation introduced in Definition 3.3 is an equivalence relation.*

Proof. We show the three required properties.

Reflexivity: We have $(E, A) \sim (E, A)$ by $P = I_m$ and $Q = I_n$.

Symmetry: From $(E_1, A_1) \sim (E_2, A_2)$, it follows that $E_2 = PE_1Q$ and $A_2 = PA_1Q - PE_1\dot{Q}$ with pointwise nonsingular matrix functions P and Q. Hence, we have

$$E_1 = P^{-1}E_2Q^{-1}, \quad A_1 = P^{-1}A_2Q^{-1} + P^{-1}E_2Q^{-1}\dot{Q}Q^{-1}$$

by pointwise definition of the inverse, and it follows that $(E_2, A_2) \sim (E_1, A_1)$, since $\frac{d}{dt}Q^{-1} = -Q^{-1}\dot{Q}Q^{-1}$.

Transitivity: From $(E_1, A_1) \sim (E_2, A_2)$ and $(E_2, A_2) \sim (E_3, A_3)$ it follows that

$$\begin{aligned} E_2 &= P_1 E_1 Q_1, \quad A_2 = P_1 A_1 Q_1 - P_1 E_1 \dot{Q}_1, \\ E_3 &= P_2 E_2 Q_2, \quad A_3 = P_2 A_2 Q_2 - P_2 E_2 \dot{Q}_2, \end{aligned}$$

with pointwise nonsingular matrix functions P_i and Q_i, $i = 1, 2$. Substituting E_2 and A_2 gives

$$E_3 = P_2 P_1 E_1 Q_1 Q_2, \quad A_3 = P_2 P_1 A_1 Q_1 Q_2 - P_2 P_1 E_1 (\dot{Q}_1 Q_2 + Q_1 \dot{Q}_2),$$

such that $(E_1, A_1) \sim (E_3, A_3)$. □

We see from Example 3.1 that regularity of the matrix pair $(E(t), A(t))$ for fixed t is not an invariant under global equivalence, since the pair of matrix functions is equivalent to the singular pair of matrices

$$(E, A) = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right)$$

This raises the question what the relevant invariants under global equivalence are and how a possible canonical form looks like. It turns out that it is very hard to treat this question in full generality, since the matrix functions E and A may depend in a nonlinear way on t. Possible effects, even if the dependence is linear, can be seen from the example $0 = tx + f(t)$ of a scalar differential-algebraic equation. Here, existence of a solution requires the necessary condition $f(0) = 0$. To exclude this and similar effects, we impose additional conditions on the functions E and A. To get a feeling how they must look like, we first consider the action of the equivalence relation (3.3) at a fixed point $t \in \mathbb{I}$. If we take into account that for given matrices $\tilde{P}$, $\tilde{Q}$, and $\tilde{R}$ of appropriate size, using Hermite interpolation, we can always find (even polynomial) matrix functions P and Q, such that at a given value $t = \tilde{t}$ we have $P(\tilde{t}) = \tilde{P}$, $Q(\tilde{t}) = \tilde{Q}$ and $\dot{Q}(\tilde{t}) = \tilde{R}$, we arrive at the following local version of the equivalence relation (3.3).

Definition 3.5. Two pairs of matrices (E_i, A_i), $E_i, A_i \in \mathbb{C}^{m,n}$, $i = 1, 2$, are called (*locally*) *equivalent* if there exist matrices $P \in \mathbb{C}^{m,m}$ and $Q, R \in \mathbb{C}^{n,n}$, P, Q nonsingular, such that

$$E_2 = P E_1 Q, \quad A_2 = P A_1 Q - P E_1 R. \tag{3.4}$$

Again, we write $(E_1, A_1) \sim (E_2, A_2)$ and distinguish from the equivalence relation in Definition 3.3 by the type of pairs (matrix or matrix function).

Lemma 3.6. *The relation introduced in Definition (3.5) is an equivalence relation.*

Proof. The proof can be carried out along the lines of the proof of Lemma 3.4. The details are left as an exercise, cp. Exercise 2. □

Note that we obtain the transformation (2.4) belonging to strong equivalence by setting $R = 0$. Hence, there are more transformations available for local equivalence to simplify a given matrix pair. We can therefore expect a simpler canonical form compared with the Kronecker canonical form. For convenience, we say in the following that a matrix is a basis of a vector space if this is valid for its columns. We additionally use the convention that the only basis of the vector space $\{0\} \subseteq \mathbb{C}^n$ is given by the empty matrix $\emptyset_{n,0} \in \mathbb{C}^{n,0}$ with the properties $\operatorname{rank} \emptyset_{n,0} = 0$ and $\det \emptyset_{0,0} = 1$. We usually omit the subscript. For a given matrix T, we use the notation T' to denote a matrix that completes T to a nonsingular matrix, i.e., $[\, T \;\; T' \,]$ constitutes a nonsingular matrix. This also applies to matrix functions. The prime should not be confused with the notation of a derivative. The latter meaning will only be used in Section 4.5 and in Section 5.1, where a special notation adapted to the topic discussed there is introduced.

Theorem 3.7. *Let $E, A \in \mathbb{C}^{m,n}$ and introduce the following spaces and matrices:*

$$T \quad \textit{basis of } \operatorname{kernel} E, \tag{3.5a}$$
$$Z \quad \textit{basis of } \operatorname{corange} E = \operatorname{kernel} E^H, \tag{3.5b}$$
$$T' \quad \textit{basis of } \operatorname{cokernel} E = \operatorname{range} E^H, \tag{3.5c}$$
$$V \quad \textit{basis of } \operatorname{corange}(Z^H A T). \tag{3.5d}$$

Then, the quantities

$$r = \operatorname{rank} E, \qquad (\textit{rank}) \tag{3.6a}$$
$$a = \operatorname{rank}(Z^H A T), \qquad (\textit{algebraic part}) \tag{3.6b}$$
$$s = \operatorname{rank}(V^H Z^H A T'), \qquad (\textit{strangeness}) \tag{3.6c}$$
$$d = r - s, \qquad (\textit{differential part}) \tag{3.6d}$$
$$u = n - r - a, \qquad (\textit{undetermined variables}) \tag{3.6e}$$
$$v = m - r - a - s, \qquad (\textit{vanishing equations}) \tag{3.6f}$$

are invariant under (3.4), and (E, A) is (locally) equivalent to the canonical form

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s \\ d \\ a \\ s \\ v \end{matrix}, \tag{3.7}$$

where all diagonal blocks with the exception of the last one are square and the last block column in both matrices has size u.

Proof. Let (E_i, A_i), $i = 1, 2$, be equivalent. Since

$$\operatorname{rank} E_2 = \operatorname{rank}(PE_1Q) = \operatorname{rank} E_1,$$

it follows that r is invariant. For a and s, we must first show that they do not depend on a particular choice of the bases. Every change of the bases can be represented by

$$\tilde{T} = TM_T, \quad \tilde{Z} = ZM_Z, \quad \tilde{T}' = T'M_{T'}, \quad \tilde{V} = M_Z^{-1}VM_V$$

with nonsingular matrices $M_T, M_Z, M_{T'}, M_V$. From

$$\operatorname{rank}(\tilde{Z}^H A\tilde{T}) = \operatorname{rank}(M_Z^H Z^H ATM_T) = \operatorname{rank}(Z^H AT)$$

and

$$\operatorname{rank}(\tilde{V}^H \tilde{Z}^H A\tilde{T}') = \operatorname{rank}(M_V^H V^H M_Z^{-H} M_Z^H Z^H AT'M_{T'}) = \operatorname{rank}(V^H Z^H AT'),$$

it then follows that a and s are indeed well defined. Let now T_2, Z_2, T_2', V_2 be bases associated with (E_2, A_2), i.e., let

$$\begin{aligned}
&\operatorname{rank}(E_2T_2) = 0, && T_2^H T_2 \text{ nonsingular}, && \operatorname{rank}(T_2^H T_2) = n - r,\\
&\operatorname{rank}(Z_2^H E_2) = 0, && Z_2^H Z_2 \text{ nonsingular}, && \operatorname{rank}(Z_2^H Z_2) = m - r,\\
&\operatorname{rank}(E_2T_2') = r, && T_2'^H T_2' \text{ nonsingular}, && \operatorname{rank}(T_2'^H T_2') = r,\\
&\operatorname{rank}(V_2^H Z_2^H A_2T_2) = 0, && V_2^H V_2 \text{ nonsingular}, && \operatorname{rank}(V_2^H V_2) = k,
\end{aligned}$$

with $k = \dim \operatorname{corange}(Z_2^H A_2T_2)$. Inserting the equivalence relation (3.4) and defining

$$T_1 = QT_2, \quad Z_1^H = Z_2^H P, \quad T_1' = QT_2', \quad V_1^H = V_2^H,$$

we obtain the same relations for (E_1, A_1) with the matrices T_1, Z_1, and T_1'. Hence, T_1, Z_1, T_1' are bases according to (3.5). Because of

$$\begin{aligned}
k &= \dim \operatorname{corange}(Z_2^H A_2T_2)\\
&= \dim \operatorname{corange}(Z_2^H PA_1QT_2 - Z_2^H PE_1RT_2)\\
&= \dim \operatorname{corange}(Z_1^H A_1T_1),
\end{aligned}$$

where we have used $Z_1^H E_1 = 0$, this also applies to V_1. With

$$\operatorname{rank}(Z_2^H A_2T_2) = \operatorname{rank}(Z_2^H PA_1QT_2 - Z_2^H PE_1RT_2) = \operatorname{rank}(Z_1^H A_1T_1)$$

and

$$\begin{aligned}\operatorname{rank}(V_2^H Z_2^H A_2 T_2') &= \operatorname{rank}(V_2^H Z_2^H P A_1 Q T_2' - V_2^H Z_2^H P E_1 R T_2') \\ &= \operatorname{rank}(V_1^H Z_1^H A_1 T_1'),\end{aligned}$$

we finally get the invariance of a and s and therefore also of d, u, and v.

For the construction of the canonical form (3.7), let first Z' be a basis of range E and V' a basis of range$(Z^H A T)$. The matrices $[\ T'\ \ T\ ]$, $[\ Z'\ \ Z\ ]$, $[\ V'\ \ V\ ]$ are then nonsingular. Moreover, ${Z'}^H E T'$ and similarly constructed matrices are also nonsingular. With this, we obtain

$$\begin{aligned}(E, A) &\sim \left(\begin{bmatrix} {Z'}^H E T' & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} {Z'}^H A T' & {Z'}^H A T \\ Z^H A T' & Z^H A T \end{bmatrix}\right) \\ &\sim \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ Z^H A T' & Z^H A T \end{bmatrix}\right) \\ &\sim \left(\left[\begin{array}{c|cc} I_r & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline {V'}^H Z^H A T' & I_a & 0 \\ V^H Z^H A T' & 0 & 0 \end{array}\right]\right) \\ &\sim \left(\left[\begin{array}{c|cc} I_r & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & I_a & 0 \\ V^H Z^H A T' & 0 & 0 \end{array}\right]\right)\end{aligned}$$

and finally (3.7) by a corresponding transformation of $V^H Z^H A T'$. □

Since the quantities defined in (3.6) are invariant under the local equivalence relation (3.4), we call them *local characteristic values* in the following.

Remark 3.8. Theorem 3.7 remains valid if we only require T' to complete T to a nonsingular matrix. This may not be important from the theoretical point of view but may simplify computations done by hand, see, e.g., the proof of Theorem 3.30.

Proof. If $[\ T'\ \ T\ ]$ is nonsingular and $\tilde{T} = T M_T$, then $[\ \tilde{T}'\ \ \tilde{T}\ ]$ is nonsingular if and only if $\tilde{T}'$ has the form

$$\tilde{T}' = T' M_{T'} + T R_T$$

with nonsingular $M_{T'}$ and arbitrary R_T. The critical relation

$$\operatorname{rank}(\tilde{V}^H \tilde{Z}^H A \tilde{T}') = \operatorname{rank}(V^H Z^H A T')$$

still holds, since $V^H Z^H A T = 0$. □

For a pair $(E(t), A(t))$ of matrix functions we can compute the characteristic values (3.6) for every selected $t \in \mathbb{I}$. In this way, we obtain functions $r, a, s : \mathbb{I} \to \mathbb{N}_0$ of characteristic values. A possible simplifying assumption then would be to require these functions to be constant on $\mathbb{I}$, i.e., to require that the sizes of the blocks in the canonical form (3.7) do not depend on $t \in \mathbb{I}$. This restriction then allows for the application of the following property of a matrix function of constant rank, see also [157], [179]. We will discuss at a later stage (see Corollary 3.26) how far this actually is a loss of generality.

Theorem 3.9. *Let $E \in C^\ell(\mathbb{I}, \mathbb{C}^{m,n})$, $\ell \in \mathbb{N}_0 \cup \{\infty\}$, with* rank $E(t) = r$ *for all $t \in \mathbb{I}$. Then there exist pointwise unitary (and therefore nonsingular) functions $U \in C^\ell(\mathbb{I}, \mathbb{C}^{m,m})$ and $V \in C^\ell(\mathbb{I}, \mathbb{C}^{n,n})$, such that*

$$U^H E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \tag{3.8}$$

with pointwise nonsingular $\Sigma \in C^\ell(\mathbb{I}, \mathbb{C}^{r,r})$.

Proof. We begin the proof with the case $\ell = 0$. For this, let $E \in C^0(\mathbb{I}, \mathbb{C}^{m,n})$ with rank $E(t) = r$ for all $t \in \mathbb{I}$.

For $\hat{t} \in \mathbb{I}$, using the singular value decomposition, see, e.g., [99], there exist $\hat{U} \in \mathbb{C}^{m,m}$, $\hat{V} \in \mathbb{C}^{n,n}$ unitary with

$$\hat{U}^H E(\hat{t}) \hat{V} = \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}$$

and $\hat{\Sigma} \in \mathbb{C}^{r,r}$ nonsingular. If we define the matrix functions $E_{11}, E_{12}, E_{21}, E_{22}$ by

$$\hat{U}^H E \hat{V} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix},$$

we have that $E_{11}(\hat{t}) = \hat{\Sigma}$. Hence, $E_{11}(\hat{t})$ is nonsingular and there exists a (relatively) open interval $\hat{\mathbb{I}} \subseteq \mathbb{I}$ with $\hat{t} \in \hat{\mathbb{I}}$ and $E_{11}(t)$ nonsingular for all $t \in \hat{\mathbb{I}}$. On $\hat{\mathbb{I}}$, we therefore get

$$\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix} \to \begin{bmatrix} I_r & E_{11}^{-1} E_{12} \\ E_{21} & E_{22} \end{bmatrix} \to \begin{bmatrix} I_r & E_{11}^{-1} E_{12} \\ 0 & E_{22} - E_{21} E_{11}^{-1} E_{12} \end{bmatrix}$$

by elementary row operations. Since rank $E(t) = r$, we have

$$E_{22} - E_{21} E_{11}^{-1} E_{12} = 0$$

such that

$$\hat{U}^H E \hat{V} = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{21} E_{11}^{-1} E_{12} \end{bmatrix}.$$

We can then define

$$\tilde{T} = \begin{bmatrix} -E_{11}^{-1}E_{12} \\ I_{n-r} \end{bmatrix}, \quad \tilde{T}' = \begin{bmatrix} I_r \\ 0 \end{bmatrix}, \quad \tilde{V} = [\, \tilde{T} \;\; \tilde{T}' \,]$$

and

$$\tilde{Z}^H = [\, -E_{21}E_{11}^{-1} \;\; I_{m-r} \,], \quad \tilde{Z}'^H = [\, I_r \;\; 0 \,], \quad \tilde{U} = [\, \tilde{Z} \;\; \tilde{Z}' \,].$$

Obviously, $\tilde{U}$ and $\tilde{V}$ are continuous on $\hat{\mathbb{I}}$ and

$$\tilde{U}^H \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{21}E_{11}^{-1}E_{12} \end{bmatrix} \tilde{V} = \begin{bmatrix} 0 & 0 \\ 0 & E_{11} \end{bmatrix}.$$

In the following considerations, we concentrate on $\tilde{V}$. The treatment of $\tilde{U}$ will be analogous.

Applying the Gram–Schmidt orthonormalization process to $\tilde{V}$, we get a pointwise upper triangular $R \in C^0(\hat{\mathbb{I}}, \mathbb{C}^{n,n})$ such that

$$\tilde{V}R = [\, \tilde{T} \;\; \tilde{T}' \,] \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

is unitary. Setting

$$[\, T \;\; T' \,] = \hat{V}[\, \tilde{T}' \;\; \tilde{T} \,] \begin{bmatrix} R_{22} & 0 \\ R_{12} & R_{11} \end{bmatrix},$$

the matrix function $[\, T \;\; T' \,]$ is continuous, pointwise unitary, and satisfies

$$\begin{aligned} E[\, T \;\; T' \,] &= E\hat{V}[\, \tilde{T}' \;\; \tilde{T} \,] \begin{bmatrix} R_{22} & 0 \\ R_{12} & R_{11} \end{bmatrix} \\ &= \hat{U} \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{21}E_{11}^{-1}E_{12} \end{bmatrix} \begin{bmatrix} I_r & -E_{11}^{-1}E_{12} \\ 0 & I_{n-r} \end{bmatrix} \begin{bmatrix} R_{22} & 0 \\ R_{12} & R_{11} \end{bmatrix} \\ &= \hat{U} \begin{bmatrix} E_{11} & 0 \\ E_{21} & 0 \end{bmatrix} \begin{bmatrix} R_{22} & 0 \\ R_{12} & R_{11} \end{bmatrix} = \hat{U} \begin{bmatrix} E_{11}R_{22} & 0 \\ E_{21}R_{22} & 0 \end{bmatrix} \end{aligned}$$

on $\hat{\mathbb{I}}$.

Performing this construction for every $\hat{t} \in \mathbb{I}$, we obtain a covering of $\mathbb{I}$ by (relatively) open intervals. This covering contains a finite covering by open intervals according to

$$\mathbb{I} = \bigcup_{j=1}^{N} \hat{\mathbb{I}}_j.$$

Without loss of generality, the covering is minimal, i.e.,

$$\hat{\mathbb{I}}_j \subseteq \hat{\mathbb{I}}_k, \;\; j, k \in \{1, \ldots, N\} \implies j = k.$$

Hence, we may assume that the intervals are ordered according to

$$\hat{\mathbb{I}}_j = (\underline{t}_j, \bar{t}_j),\ j = 1, \ldots, N \implies \underline{t}_j < \underline{t}_k, \bar{t}_j < \bar{t}_k \quad \text{for } j < k.$$

Since we have a covering,

$$\bar{t}_j > \underline{t}_{j+1}, \quad j = 1, \ldots, N-1$$

must hold. Thus, there exist points

$$t_j \in (\underline{t}_{j+1}, \bar{t}_j) = \hat{\mathbb{I}}_j \cap \hat{\mathbb{I}}_{j+1}, \quad j = 1, \ldots, N-1.$$

Together with $\mathbb{I} = [t_0, t_N]$, we then have intervals

$$\mathbb{I}_j = [t_{j-1}, t_j], \quad j = 1, \ldots, N$$

with

$$\mathbb{I}_j \subset \hat{\mathbb{I}}_j, \quad j = 1, \ldots, N.$$

Recall that on every $\hat{\mathbb{I}}_j$, $j = 1, \ldots, N$ the above construction yields a pointwise unitary matrix function $V_j \in C^0(\mathbb{I}_j, \mathbb{C}^{n,n})$ with

$$E V_j = [\ \Sigma_j\ \ 0\ ],$$

where Σ_j has pointwise full column rank r. In a neighborhood of t_1, we therefore have pointwise unitary matrix functions

$$V_1 = [\ T_1'\ \ T_1\ ], \quad V_2 = [\ T_2'\ \ T_2\ ]$$

with

$$E(t_1)T_1(t_1) = 0, \quad E(t_1)T_2(t_1) = 0.$$

Observing that

$$V_1(t_1) = V_2(t_1)V_2(t_1)^H V_1(t_1) = V_2(t_1)Q,$$

where Q is obviously unitary, and using the block structure of V_1 and V_2, we have

$$[\ T_1'(t_1)\ \ T_1(t_1)\ ] = [\ T_2'(t_1)\ \ T_2(t_1)\ ] \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}.$$

Since the columns of $T_1(t_1)$ and $T_2(t_1)$ form orthonormal bases of kernel $E(t_1)$ and the columns of $T_1'(t_1)$ and $T_2'(t_1)$ are orthogonal to kernel $E(t_1)$, it follows that $Q_{12} = 0$ and $Q_{21} = 0$. Defining now a matrix function V on $[t_0, t_2]$ via

$$V(t) = \begin{cases} V_1(t) & \text{for } t \in \mathbb{I}_1, \\ V_2(t)Q & \text{for } t \in \mathbb{I}_2, \end{cases}$$

we have

$$EV = EV_1 = [\ \Sigma_1\ \ 0\ ] \qquad \text{on } \mathbb{I}_1,$$

$$EV = EV_2 Q = [\ \Sigma_2 Q_{11}\ \ 0\ ] \quad \text{on } \mathbb{I}_2.$$

Moreover, the matrix function V is pointwise unitary and continuous on $[t_0, t_2]$. Repeating this construction for $t_2, \ldots, t_{N-1}$ finally gives a pointwise unitary $V \in C^0(\mathbb{I}, \mathbb{C}^{n,n})$ with

$$EV = [\ \Sigma_V\ \ 0\ ]$$

on $\mathbb{I}$ such that Σ_V has pointwise full column rank r.

In the same way, we get a pointwise unitary $U \in C^0(\mathbb{I}, \mathbb{C}^{m,m})$ with

$$U^H E = \begin{bmatrix} \Sigma_U \\ 0 \end{bmatrix}$$

on $\mathbb{I}$ such that Σ_U has pointwise full row rank r. Together this yields

$$\begin{bmatrix} \Sigma_U V \\ 0 \end{bmatrix} = U^H E V = [\ U^H \Sigma_V\ \ 0\ ].$$

Hence, we have that

$$U^H E V = \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma \in C^0(\mathbb{I}, \mathbb{C}^{r,r})$ is pointwise nonsingular.

For $\ell \in \mathbb{N} \cup \{\infty\}$ we proceed as follows. Given $t_0 \in \mathbb{I}$, there exist $U_0 \in \mathbb{C}^{m,m}$ and $V_0 \in \mathbb{C}^{n,n}$ unitary with

$$U_0^H E(t_0) V_0 = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix}$$

such that $\Sigma_0 \in \mathbb{C}^{r,r}$ is nonsingular. Let $U_0 = [\ Z_0'\ \ Z_0\ ]$ and $V_0 = [\ T_0'\ \ T_0\ ]$. Obviously,

$$S(t) = E(t) T_0'$$

has full column rank for all t in a sufficiently small neighborhood of t_0. Hence, if Π denotes the (pointwise defined) orthogonal projection onto range E, we have

$$\Pi = S(S^H S)^{-1} S^H$$

in this neighborhood. This local representation shows that Π is as smooth as S, which in turn is as smooth as E, i.e., $\Pi \in C^\ell(\mathbb{I}, \mathbb{C}^{m,m})$. Recall that Π, as an orthogonal projection, satisfies $\Pi\Pi = \Pi$ and $\Pi^H = \Pi$.

The initial value problem for the ordinary differential equation

$$\dot{W} = (\dot{\Pi}(t)\Pi(t) - \Pi(t)\dot{\Pi}(t))W, \tag{3.9}$$

with initial value $W(t_0) = I_m$, is known to have a unique solution $W \in C^\ell(\mathbb{I}, \mathbb{C}^{m,m})$, see, e.g., [65], [220]. Moreover, since

$$\begin{aligned}\tfrac{d}{dt}(W^H W) &= \dot{W}^H W + W^H \dot{W} \\ &= W^H(\dot{\Pi}\Pi - \Pi\dot{\Pi})^H W + W^H(\dot{\Pi}\Pi - \Pi\dot{\Pi})W \\ &= W^H(\Pi\dot{\Pi} - \dot{\Pi}\Pi)W + W^H(\dot{\Pi}\Pi - \Pi\dot{\Pi})W = 0,\end{aligned}$$

we have $W^H W = I_m$ and W is pointwise unitary. Since $\dot{\Pi} = \dot{\Pi}\Pi + \Pi\dot{\Pi}$ and therefore $\Pi\dot{\Pi}\Pi = 0$, we get

$$\begin{aligned}\tfrac{d}{dt}(\Pi W) - (\dot{\Pi}\Pi - \Pi\dot{\Pi})\Pi W &= \dot{\Pi}W + \Pi\dot{W} - \dot{\Pi}\Pi W + \Pi\dot{\Pi}\Pi W \\ &= \dot{\Pi}W + \Pi(\dot{\Pi}\Pi - \Pi\dot{\Pi})W - \dot{\Pi}\Pi W \\ &= (\dot{\Pi} - \dot{\Pi}\Pi - \Pi\dot{\Pi})W = 0.\end{aligned}$$

Thus, ΠW also solves (3.9) but with respect to the initial value $\Pi(t_0)$. Since W is a fundamental solution, we immediately have

$$\Pi W = W\Pi(t_0).$$

Defining now

$$Z' = WZ_0', \quad Z = WZ_0$$

yields that $U = [\, Z' \;\; Z \,] = W[\, Z_0' \;\; Z_0 \,]$ is pointwise unitary with

$$\begin{aligned}Z' &= WZ_0' = W\Pi(t_0)Z_0' = \Pi WZ_0' = \Pi Z', \\ Z &= WZ_0 = W(I_m - \Pi(t_0))Z_0 = (W - \Pi W)Z_0 = (I_m - \Pi)Z.\end{aligned}$$

In particular, the columns of Z' and Z form orthonormal bases of range E and corange E, respectively, and are as smooth as E. By symmetry, the corresponding claim holds for $V = [\, T' \;\; T \,]$ with similarly constructed T' and T. □

Corollary 3.10. *If $E \in C^1(\mathbb{I}, \mathbb{C}^{m,n})$, then suitable functions $U = [\, Z' \;\; Z \,]$ and $V = [\, T' \;\; T \,]$ in (3.8) can be obtained as solution of the ordinary differential*

equation

$$\begin{aligned}
\begin{bmatrix} Z'(t)^H E(t) \\ T(t)^H \end{bmatrix} \dot{T}(t) &= -\begin{bmatrix} Z'(t)^H \dot{E}(t) T(t) \\ 0 \end{bmatrix}, \\
\begin{bmatrix} T'(t)^H E(t)^H \\ Z(t)^H \end{bmatrix} \dot{Z}(t) &= -\begin{bmatrix} T'(t)^H \dot{E}(t)^H Z(t) \\ 0 \end{bmatrix}, \\
\begin{bmatrix} T(t)^H \\ T'(t)^H \end{bmatrix} \dot{T}'(t) &= -\begin{bmatrix} \dot{T}(t)^H T'(t) \\ 0 \end{bmatrix}, \\
\begin{bmatrix} Z(t)^H \\ Z'(t)^H \end{bmatrix} \dot{Z}'(t) &= -\begin{bmatrix} \dot{Z}(t)^H Z'(t) \\ 0 \end{bmatrix},
\end{aligned} \tag{3.10}$$

with initial values

$$T(t_0) = T_0, \quad T'(t_0) = T_0', \quad Z(t_0) = Z_0, \quad Z'(t_0) = Z_0', \tag{3.11}$$

so that

$$[\, Z_0' \;\; Z_0 \,]^H E(t_0) [\, T_0' \;\; T_0 \,] = \begin{bmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Sigma_0 \textit{ nonsingular.}$$

In particular, suitable functions U and V can be determined numerically, provided that E and $\dot{E}$ are available in form of implementable subroutines.

Proof. Let $U = [\, Z' \; Z \,]$ and $V = [\, T' \; T \,]$ be given according to Theorem 3.9. Then, all matrix functions multiplying the derivatives on the left hand sides of (3.10) are pointwise nonsingular. Moreover, we have

$$E(t)T(t) = 0, \quad E(t)^H Z(t) = 0, \quad T(t)^H T'(t) = 0, \quad Z(t)^H Z'(t) = 0$$

for all $t \in \mathbb{I}$. Hence, differentiation with respect to t gives the first block of each equation in (3.10). The other blocks of (3.10) consist of equations of the form

$$W(t)^H \dot{W}(t) = 0.$$

Note that a given matrix function W with pointwise orthonormal columns does not necessarily satisfy this relation. Defining Q as solution of the (linear) initial value problem

$$\dot{Q} = -W(t)^H \dot{W}(t) Q, \quad Q(t_0) = I,$$

we find that

$$\begin{aligned}
\tfrac{d}{dt}(Q^H Q) &= \dot{Q}^H Q + Q^H \dot{Q} = -Q^H \dot{W}^H W Q - Q^H W^H \dot{W} Q \\
&= -Q^H (\dot{W}^H W + W^H \dot{W}) Q = -Q^H \tfrac{d}{dt}(W^H W) Q = 0
\end{aligned}$$

and Q is pointwise unitary. Hence, the columns of Q and WQ both pointwise form an orthonormal basis of the same space. Instead of $W^H\dot{W} = 0$, the relation

$$(WQ)^H \tfrac{d}{dt}(WQ) = Q^H W^H(\dot{W}Q + W\dot{Q}) = Q^H W^H(\dot{W}Q - WW^H\dot{W}Q) = 0.$$

holds. In particular, the other blocks of (3.10) only select a special renormalization of the given U and V which does not change the property (3.8). □

Using Theorem 3.9, we can construct the following global canonical form for the equivalence relation (3.3). Since we must satisfy $R = \dot{Q}$ in (3.4), we expect the canonical form to be more complicated than the local canonical form (3.7). Note that we use here the term canonical form in a way that differs from the terminology of abstract algebra.

Theorem 3.11. *Let $E, A \in C(\mathbb{I}, \mathbb{C}^{m,n})$ be sufficiently smooth and suppose that*

$$r(t) \equiv r, \quad a(t) \equiv a, \quad s(t) \equiv s \tag{3.12}$$

for the local characteristic values of $(E(t), A(t))$. Then, (E, A) is globally equivalent to the canonical form

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s \\ d \\ a \\ s \\ v \end{matrix}. \tag{3.13}$$

All entries A_{ij} are again matrix functions on $\mathbb{I}$ and the last block column in both matrix functions of (3.13) has size $u = n - s - d - a$.

Proof. In the following, the word "new" on top of the equivalence operator denotes that the indexing is adapted to the new block structure of the matrices such that the same symbol can denote a different entry. Using Theorem 3.9, we obtain the following sequence of equivalent pairs of matrix functions.

$$\begin{aligned}
(E, A) &\sim \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \\
&\overset{\text{new}}{\sim} \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\right) \\
&\sim \left(\begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12}V_1 \\ U_1^H A_{21} & U_1^H A_{22} V_1 \end{bmatrix} - \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 & 0 \\ 0 & \dot{V}_1 \end{bmatrix}\right) \\
&\overset{\text{new}}{\sim} \left(\left[\begin{array}{c|cc} I_r & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} A_{11} & A_{12} & A_{13} \\ \hline A_{21} & I_a & 0 \\ A_{31} & 0 & 0 \end{array}\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&\sim \left(\left[\begin{array}{c|cc} V_2 & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|cc} A_{11}V_2 & A_{12} & A_{13} \\ \hline A_{21}V_2 & I_a & 0 \\ U_2^H A_{31} V_2 & 0 & 0 \end{array}\right] - \left[\begin{array}{c|cc} I_r & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \left[\begin{array}{c|cc} \dot{V}_2 & 0 & 0 \\ \hline 0 & 0 & 0 \\ 0 & 0 & 0 \end{array}\right] \right) \\
&\overset{\text{new}}{\sim} \left(\left[\begin{array}{cc|cc} V_{11} & V_{12} & 0 & 0 \\ V_{21} & V_{22} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ \hline A_{31} & A_{32} & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right) \\
&\overset{\text{new}}{\sim} \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ \hline 0 & A_{32} & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right) \\
&\sim \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ \hline 0 & A_{32} & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & -A_{32} & I_a & 0 \\ 0 & 0 & 0 & I_u \end{array}\right] \right) \\
&\overset{\text{new}}{\sim} \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12} & A_{13} & A_{14} \\ 0 & A_{22} & A_{23} & A_{24} \\ \hline 0 & 0 & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right) \\
&\sim \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12} & 0 & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ \hline 0 & 0 & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right) \\
&\sim \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & Q_2 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12}Q_2 & 0 & A_{14} \\ 0 & A_{22}Q_2 - \dot{Q}_2 & 0 & A_{24} \\ \hline 0 & 0 & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right) \\
&\overset{\text{new}}{\sim} \left(\left[\begin{array}{cc|cc} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ \hline 0 & 0 & I_a & 0 \\ \hline I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right).
\end{aligned}
$$

In the last step, Q_2 is chosen as the solution of the initial value problem

$$\dot{Q}_2 = A_{22}(t)Q_2, \quad Q_2(t_0) = I_d$$

on $\mathbb{I}$. The unique solvability of such problems (see, e.g., [65]) ensures that Q_2 is pointwise nonsingular. □

We now apply this result to the two examples from the beginning of this chapter.

Example 3.12. For Example 3.1, we immediately see that $r = \operatorname{rank} E = 1$ on $\mathbb{I}$. With the choice of bases

$$T = \begin{bmatrix} t \\ 1 \end{bmatrix}, \quad T' = \begin{bmatrix} 1 \\ -t \end{bmatrix}, \quad Z = \begin{bmatrix} 1 \\ -t \end{bmatrix},$$

we find that $a = \operatorname{rank}(Z^H A T) = 0$ holds with $V = [\,1\,]$ and therefore $s = \operatorname{rank}(V^H Z^H A T') = 1$.

Example 3.13. For Example 3.2, we also have $r = \operatorname{rank} E = 1$ on $\mathbb{I}$. With the choice of bases

$$T = \begin{bmatrix} t \\ 1 \end{bmatrix}, \quad T' = \begin{bmatrix} 1 \\ -t \end{bmatrix}, \quad Z = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

we find that $a = \operatorname{rank}(Z^H A T) = 0$ holds with $V = [\,1\,]$ and therefore $s = \operatorname{rank}(V^H Z^H A T') = 1$.

For the pairs of matrix functions of both examples, we obtain the same characteristic values (r, a, s). This shows, that essential information must be hidden in the matrix functions A_{12}, A_{14}, and A_{24} of (3.13). Writing down the differential-algebraic equation associated with (3.13), we obtain

$$\dot{x}_1 = A_{12}(t)x_2 + A_{14}(t)x_4 + f_1(t), \tag{3.14a}$$
$$\dot{x}_2 = A_{24}(t)x_4 + f_2(t), \tag{3.14b}$$
$$0 = x_3 + f_3(t), \tag{3.14c}$$
$$0 = x_1 + f_4(t), \tag{3.14d}$$
$$0 = f_5(t). \tag{3.14e}$$

We recognize an algebraic equation (3.14c) for x_3 (algebraic part) and a consistency condition (3.14e) for the inhomogeneity (vanishing equations). Furthermore, (3.14b) looks like a differential equation (differential part) with a possible free choice in x_4 (undetermined variables). The intrinsic problem, however, is the coupling between the algebraic equation (3.14d) and the differential equation (3.14a) for x_1. The idea now is to differentiate (3.14d) in order to eliminate $\dot{x}_1$ from (3.14a)

which then becomes purely algebraic. This step of differentiating and eliminating corresponds to passing from the pair (3.13) of matrix functions to the pair

$$\left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s \\ d \\ a \\ s \\ v \end{matrix}. \tag{3.15}$$

Since the algebraic equation (3.14d) is preserved, we can reverse this procedure by differentiating (3.14d) and adding it then to the new first equation. This also shows that the solution set of the corresponding differential-algebraic equation is not altered.

For the so obtained new pair (3.15) of matrix functions, we can again determine the corresponding characteristic values (r, a, s) and, if they are constant on the whole interval $\mathbb{I}$, transform it to global canonical form. But before we can proceed in this way, we must show that the new values are characteristic for the original problems. After all, it might happen that two different but equivalent global canonical forms can lead to new pairs with different characteristic values.

Theorem 3.14. *Assume that the pairs (E, A) and $(\tilde{E}, \tilde{A})$ of matrix functions are (globally) equivalent and in global canonical form (3.13). Then the modified pairs $(E_{\mathrm{mod}}, A_{\mathrm{mod}})$ and $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ obtained by passing from (3.13) to (3.15) are also (globally) equivalent.*

Proof. By assumption, there exist sufficiently smooth pointwise nonsingular matrix functions P and Q, such that

$$P\tilde{E} = EQ, \tag{3.16a}$$

$$P\tilde{A} = AQ - E\dot{Q}. \tag{3.16b}$$

From (3.16a), we deduce

$$\begin{bmatrix} P_{11} & P_{12} & 0 & 0 \\ P_{21} & P_{22} & 0 & 0 \\ P_{31} & P_{32} & 0 & 0 \\ P_{41} & P_{42} & 0 & 0 \\ P_{51} & P_{52} & 0 & 0 \end{bmatrix} = \begin{bmatrix} Q_{11} & Q_{12} & Q_{13} & Q_{14} \\ Q_{21} & Q_{22} & Q_{23} & Q_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

if we partition P, Q according to (3.13). With this, we obtain for the last three block rows of (3.16b) that

$$\begin{bmatrix} P_{34} & 0 & P_{33} & 0 \\ P_{44} & 0 & P_{43} & 0 \\ P_{54} & 0 & P_{53} & 0 \end{bmatrix} = \begin{bmatrix} Q_{31} & Q_{32} & Q_{33} & Q_{34} \\ Q_{11} & Q_{12} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

The matrix functions P and Q are therefore of the form

$$P = \begin{bmatrix} Q_{11} & 0 & P_{13} & P_{14} & P_{15} \\ Q_{21} & Q_{22} & P_{23} & P_{24} & P_{25} \\ 0 & 0 & Q_{33} & Q_{31} & P_{35} \\ 0 & 0 & 0 & Q_{11} & P_{45} \\ 0 & 0 & 0 & 0 & P_{55} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & 0 & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 \\ Q_{31} & 0 & Q_{33} & 0 \\ Q_{41} & Q_{42} & Q_{43} & Q_{44} \end{bmatrix}.$$

In particular, the functions Q_{11}, Q_{22}, Q_{33}, Q_{44}, and P_{55} must be pointwise nonsingular. From the first two block rows of (3.16b), we then get

$$\begin{bmatrix} Q_{11} & 0 \\ Q_{21} & Q_{22} \end{bmatrix} \begin{bmatrix} \tilde{A}_{12} & \tilde{A}_{14} \\ 0 & \tilde{A}_{24} \end{bmatrix} = \begin{bmatrix} A_{12} & A_{14} \\ 0 & A_{24} \end{bmatrix} \begin{bmatrix} Q_{22} & 0 \\ Q_{42} & Q_{44} \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ \dot{Q}_{22} & 0 \end{bmatrix}.$$

Thus, it follows that $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ is equivalent to

$$\begin{aligned} &\left(\begin{bmatrix} Q_{11} & 0 & 0 & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 & 0 \\ 0 & 0 & 0 & I_s & 0 \\ 0 & 0 & 0 & 0 & I_v \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \right. \\ &\qquad \left. \begin{bmatrix} Q_{11} & 0 & 0 & 0 & 0 \\ Q_{21} & Q_{22} & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 & 0 \\ 0 & 0 & 0 & I_s & 0 \\ 0 & 0 & 0 & 0 & I_v \end{bmatrix} \begin{bmatrix} 0 & \tilde{A}_{12} & 0 & \tilde{A}_{14} \\ 0 & 0 & 0 & \tilde{A}_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \\ &\sim \left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & Q_{42} & 0 & Q_{44} \end{bmatrix} \right. \\ &\qquad \left. - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \dot{Q}_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \end{aligned}$$

$$
\begin{aligned}
\sim \Bigg(& \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & Q_{22}^{-1} & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & -Q_{44}^{-1} Q_{42} Q_{22}^{-1} & 0 & Q_{44}^{-1} \end{bmatrix}, \\
& \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \dot{Q}_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & Q_{22}^{-1} & 0 & 0 \\ 0 & 0 & I_a & 0 \\ 0 & -Q_{44}^{-1} Q_{42} Q_{22}^{-1} & 0 & Q_{44}^{-1} \end{bmatrix} \\
& \qquad - \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & Q_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \frac{d}{dt} Q_{22}^{-1} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -\frac{d}{dt}(Q_{44}^{-1} Q_{42} Q_{22}^{-1}) & 0 & \frac{d}{dt} Q_{44}^{-1} \end{bmatrix} \Bigg) \\
\sim \Bigg(& \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & X & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \Bigg)
\end{aligned}
$$

with

$$
X = -(\dot{Q}_{22} Q_{22}^{-1} + Q_{22} \tfrac{d}{dt} Q_{22}^{-1}) = -\tfrac{d}{dt}(Q_{22} Q_{22}^{-1}) = -\dot{I}_d = 0. \qquad \square
$$

The statement of Theorem 3.14 allows for the following inductive procedure. Starting from $(E_0, A_0) = (E, A)$, we define a sequence (E_i, A_i), $i \in \mathbb{N}_0$, of pairs of matrix functions by transforming (E_i, A_i) to global canonical form (3.13) and then passing to (3.15) to obtain (E_{i+1}, A_{i+1}). Observe that we must assume a constant rank condition of the form (3.12) in every step of this procedure. Although (E, A) does not uniquely determine the resulting pairs (E_i, A_i), Theorem 3.14 implies that it uniquely determines a sequence of invariants (r_i, a_i, s_i), which are characteristic for (E, A), where (r_i, a_i, s_i) denote the characteristic values of (E_i, A_i). The relation $r_{i+1} = r_i - s_i$, which can directly be deduced by comparing the left matrices in (3.13) and (3.15), guarantees that after a finite number of steps the strangeness s_i must vanish. If this is the case, the sequence (r_i, a_i, s_i) becomes stationary because the pairs (3.13) and (3.15) are then the same. The index when this happens is also characteristic for the given pair (E, A).

Definition 3.15. Let (E, A) be a pair of sufficiently smooth matrix functions. Let the sequence (r_i, a_i, s_i), $i \in \mathbb{N}_0$, be well defined. In particular, let (3.12) hold for every entry (E_i, A_i) of the above sequence. Then, we call

$$
\mu = \min\{i \in \mathbb{N}_0 \mid s_i = 0\} \tag{3.17}
$$

the *strangeness index* of (E, A) and of (3.1). In the case that $\mu = 0$ we call (E, A) and (3.1) *strangeness-free*.

Remark 3.16. In Section 3.2, we will show that the requirement that the sequence (r_i, a_i, s_i), $i \in \mathbb{N}_0$, is well defined in the whole interval $\mathbb{I}$ can be significantly relaxed. This then immediately yields a new definition of the strangeness index under some weaker assumptions.

We summarize the above discussion in the following theorem.

Theorem 3.17. *Let the strangeness index μ of (E, A) as in (3.17) be well defined (i.e., let (3.12) hold for every entry (E_i, A_i) of the above sequence) and let $f \in C^\mu(\mathbb{I}, \mathbb{C}^m)$. Then the differential-algebraic equation (3.1) is equivalent (in the sense that there is a one-to-one correspondence between the solution spaces via a pointwise nonsingular matrix function) to a differential-algebraic equation of the form*

$$\dot{x}_1 = A_{13}(t)x_3 + f_1(t), \quad d_\mu \tag{3.18a}$$
$$0 = x_2 + f_2(t), \quad a_\mu \tag{3.18b}$$
$$0 = f_3(t), \quad v_\mu \tag{3.18c}$$

where $A_{13} \in C(\mathbb{I}, \mathbb{C}^{d_\mu, u_\mu})$ and the inhomogeneities f_1, f_2, f_3 are determined by $f^{(0)}, \ldots, f^{(\mu)}$.

Proof. By the above discussion, it follows that the pair (E_μ, A_μ) is strangeness-free. Thus, (3.13) reduces to three block rows and columns leading directly to (3.18). Additionally, all transformations are reversible and do not change the structure of the solution set in the described sense. □

Theorem 3.17 allows to read off the existence and uniqueness of solutions of (3.1).

Corollary 3.18. *Let the strangeness index μ of (E, A) as in (3.17) be well defined and let $f \in C^{\mu+1}(\mathbb{I}, \mathbb{C}^m)$. Then we have:*

1. *The problem (3.1) is solvable if and only if the v_μ functional consistency conditions*

$$f_3 = 0$$

are fulfilled.

2. *An initial condition (3.2) is consistent if and only if in addition the a_μ conditions*

$$x_2(t_0) = -f_2(t_0)$$

are implied by (3.2).

3. *The corresponding initial value problem is uniquely solvable if and only if in addition*

$$u_\mu = 0$$

holds.

Observe that the stronger assumption on the smoothness of the inhomogeneity, i.e., that $f \in C^{\mu+1}(\mathbb{I}, \mathbb{C}^m)$ rather than $f \in C^\mu(\mathbb{I}, \mathbb{C}^m)$, is only used to guarantee that x_2 is continuously differentiable. The structure of (3.18), however, suggests that it is sufficient to require only continuity for the parts x_2 and x_3 of the solution. We will examine this problem in more detail in the context of an operator formulation of differential-algebraic equations, see Section 3.4.

Remark 3.19. Comparing Theorem 3.17 with Theorem 2.12 (the strangeness index is trivially well defined for pairs of constant matrix functions), we can obviously replace the condition of regularity for pair (E, A) by the condition

$$u_\mu = 0, \quad v_\mu = 0, \tag{3.19}$$

which automatically implies that $m = n$. Furthermore, μ plays the role of $\nu - 1$. Again, we must treat the case $\nu = 0$ separately.

Instead of actually performing the transition from (3.13) to (3.15), we can apply all equivalence transformations to (E, A). The aim then would be to construct a pair of matrix functions which is (globally) equivalent to (E, A) and from which we can read off the whole sequence of characteristic values (r_i, a_i, s_i) of (E, A).

Lemma 3.20. *Let the strangeness index μ of (E, A) as in (3.17) be well defined. Let the process leading to Theorem 3.17 yield a sequence (E_i, A_i), $i \in \mathbb{N}_0$, with characteristic values $(r_i, a_i, s_i, d_i, u_i, v_i)$ according to (3.6) and*

$$(E_i, A_i) \sim \left(\begin{bmatrix} I_{s_i} & 0 & 0 & 0 \\ 0 & I_{d_i} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12}^{(i)} & 0 & A_{14}^{(i)} \\ 0 & 0 & 0 & A_{24}^{(i)} \\ 0 & 0 & I_{a_i} & 0 \\ I_{s_i} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s_i \\ d_i \\ a_i \\ s_i \\ v_i \end{matrix}, \tag{3.20}$$

where the last block column in both matrix functions has size u_i. Defining

$$b_0 = a_0, \qquad b_{i+1} = \operatorname{rank} A_{14}^{(i)}, \tag{3.21a}$$

$$c_0 = a_0 + s_0, \quad c_{i+1} = \operatorname{rank}[\, A_{12}^{(i)} \;\; A_{14}^{(i)} \,], \tag{3.21b}$$

$$w_0 = v_0, \qquad w_{i+1} = v_{i+1} - v_i, \tag{3.21c}$$

we have

$$r_{i+1} = r_i - s_i, \tag{3.22a}$$
$$a_{i+1} = a_i + s_i + b_{i+1} = c_0 + \cdots + c_{i+1} - s_{i+1}, \tag{3.22b}$$
$$s_{i+1} = c_{i+1} - b_{i+1}, \tag{3.22c}$$
$$d_{i+1} = r_{i+1} - s_{i+1} = d_i - s_{i+1}, \tag{3.22d}$$
$$w_{i+1} = s_i - c_{i+1}, \tag{3.22e}$$
$$u_{i+1} = u_0 - b_1 - \cdots - b_{i+1}, \tag{3.22f}$$
$$v_{i+1} = v_0 + w_1 + \cdots + w_{i+1}. \tag{3.22g}$$

Proof. The proof is left as an exercise, cp. Exercise 12. □

In the following, for convenience, we denote unspecified blocks in a matrix by $*$.

Theorem 3.21. *Let the strangeness index μ of (E, A) as in (3.17) be well defined. Then, (E, A) is (globally) equivalent to a pair of the form*

$$\left(\begin{bmatrix} I_{d_\mu} & 0 & W \\ 0 & 0 & F \\ 0 & 0 & G \end{bmatrix}, \begin{bmatrix} 0 & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{a_\mu} \end{bmatrix} \right), \tag{3.23}$$

with

$$F = \begin{bmatrix} 0 & F_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & F_1 \\ & & & 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 & G_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & G_1 \\ & & & 0 \end{bmatrix}, \tag{3.24}$$

where F_i and G_i have sizes $w_i \times c_{i-1}$ and $c_i \times c_{i-1}$, respectively, with w_i, c_i as in (3.21), and $W = [\,0 \;\; * \;\; \cdots \;\; *\,]$ is partitioned accordingly. In particular, F_i and G_i together have full row rank, i.e.,

$$\operatorname{rank} \begin{bmatrix} F_i \\ G_i \end{bmatrix} = c_i + w_i = s_{i-1} \leq c_{i-1}. \tag{3.25}$$

Proof. To start an induction argument, we first permute (3.13), such that the identities of the right matrix come into the bottom right corner. Additionally changing the order of blocks and adapting the notation, we get

$$(E, A) \sim \left(\begin{bmatrix} I_{d_0} & 0 & 0 \\ 0 & 0 & \tilde{U}_0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{c_0} \end{bmatrix} \right)$$

with $\tilde{U}_0 = [\, I_{s_0} \;\; 0 \,]$ and $c_0 = a_0 + s_0$. We therefore take as inductive assumption that

$$(E, A) \sim (\tilde{E}_i, \tilde{A}_i) = \left(\begin{bmatrix} I_{d_i} & 0 & * \\ 0 & 0 & \tilde{U}_i \\ 0 & 0 & \tilde{F}_i \\ 0 & 0 & \tilde{G}_i \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{a_i+s_i} \end{bmatrix} \right)$$

with

$$\tilde{F}_i = \begin{bmatrix} 0 & 0 & F_i & & * \\ 0 & 0 & 0 & \ddots & \\ \vdots & \vdots & & \ddots & F_1 \\ 0 & 0 & & & 0 \end{bmatrix}, \quad \tilde{G}_i = \begin{bmatrix} 0 & 0 & G_i & & * \\ 0 & 0 & 0 & \ddots & \\ \vdots & \vdots & & \ddots & G_1 \\ 0 & 0 & & & 0 \end{bmatrix}$$

and $\tilde{U}_i = [\, I_{s_i} \;\; 0 \;\; * \;\; \cdots \;\; * \,]$, partitioned accordingly.

In addition, the inductive assumption includes that the last block column of $\tilde{E}_i$ would become zero if we performed the step from (3.13) to (3.15). Comparing with (3.21), this gives $b_{i+1} = \operatorname{rank} A_{22}$. We must then carry out the equivalence transformation that separates the corresponding null-spaces. It is sufficient to consider the first two block rows only, because there will be no transformations acting on the other block rows. Thus, we obtain

$$\left(\left[\begin{array}{c|c|c} I_{d_i} & 0 & * \\ \hline 0 & 0 & \tilde{U}_i \end{array}\right], \left[\begin{array}{c|c|c} 0 & A_{12} & 0 \\ \hline A_{21} & A_{22} & 0 \end{array}\right] \right)$$

$$\overset{\text{new}}{\sim} \left(\left[\begin{array}{c|cc|c} I_{d_i} & 0 & 0 & * \\ \hline 0 & 0 & 0 & U_1 \\ 0 & 0 & 0 & U_2 \end{array}\right], \left[\begin{array}{c|cc|c} 0 & A_{12} & A_{13} & 0 \\ \hline A_{21} & 0 & 0 & 0 \\ A_{31} & 0 & I_{b_{i+1}} & 0 \end{array}\right] \right).$$

Comparing again with (3.21) shows that $s_{i+1} = \operatorname{rank} A_{21}$. This yields

$$\left(\left[\begin{array}{c|c|c} I_{d_i} & 0 & * \\ \hline 0 & 0 & \tilde{U}_i \end{array}\right], \left[\begin{array}{c|c|c} 0 & A_{12} & 0 \\ \hline A_{21} & A_{22} & 0 \end{array}\right] \right)$$

$$\overset{\text{new}}{\sim} \left(\left[\begin{array}{cc|cc|c} I_{d_{i+1}} & 0 & 0 & 0 & * \\ 0 & I_{s_{i+1}} & 0 & 0 & * \\ \hline 0 & 0 & 0 & 0 & U_1 \\ 0 & 0 & 0 & 0 & U_2 \\ \hline 0 & 0 & 0 & 0 & U_3 \end{array}\right], \left[\begin{array}{cc|cc|c} 0 & 0 & A_{13} & A_{14} & 0 \\ 0 & 0 & A_{23} & A_{24} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & I_{s_{i+1}} & 0 & 0 & 0 \\ \hline A_{51} & A_{52} & 0 & I_{b_{i+1}} & 0 \end{array}\right] \right).$$

The entries A_{14}, A_{24}, A_{51}, and A_{52} can be eliminated by appropriate row and column operations. Adding the unchanged third and fourth block rows of the initial pair and performing appropriate row and column permutations, we get

$$(E, A) \overset{\text{new}}{\sim} (\tilde{E}_{i+1}, \tilde{A}_{i+1}) = \left(\begin{bmatrix} I_{d_{i+1}} & 0 & * \\ 0 & 0 & \tilde{U}_{i+1} \\ 0 & 0 & \tilde{F}_{i+1} \\ 0 & 0 & \tilde{G}_{i+1} \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 \\ A_{21} & A_{22} & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{a_{i+1}+s_{i+1}} \end{bmatrix} \right)$$

with

$$\tilde{F}_{i+1} = \begin{bmatrix} 0 & 0 & U_1 \\ 0 & 0 & \tilde{F}_i \end{bmatrix}, \quad \tilde{G}_{i+1} = \begin{bmatrix} 0 & 0 & U_2 \\ 0 & 0 & U_3 \\ 0 & 0 & \tilde{G}_i \end{bmatrix},$$

and $\tilde{U}_{i+1} = [\, I_{s_{i+1}} \; 0 \; * \,]$, partitioned accordingly. Since U_1, U_2, and U_3 are obtained from $\tilde{U}_i$ by row transformations only, their leading s_i columns have full rank thus proving (3.25). If we had performed the step from (3.13) to (3.15), then the entry $I_{s_{i+1}}$ in $\tilde{U}_{i+1}$ would be replaced by zero. Together with the vanishing third block column of $\tilde{E}_i$ this yields that the third block column of $\tilde{E}_{i+1}$ would vanish as well. This completes the induction argument and we obtain (3.23) for $i = \mu$ in $(\tilde{E}_i, \tilde{A}_i)$, since $s_\mu = 0$. □

The canonical form (3.23) can be seen as a generalization of the Kronecker canonical form to the case of matrix functions, at least if we are interested in properties of the corresponding differential-algebraic equation. One can therefore expect that it plays a central role in the analysis of pairs of matrix functions and linear differential-algebraic equations with variable coefficients.

Remark 3.22. In the construction of the canonical form (3.23), we make use of two types of transformations. These are the global equivalence transformations of the form (3.3), which generate a one-to-one correspondence between the two solution sets, and the process of adding derivatives of some block rows to other block rows, which is the step from (3.13) to (3.15), that does not alter the solution set.

We now illustrate the results on the Examples 3.1 and 3.2. We use here the abbreviation "dif" on top of the equivalence operator to mark the step of passing from (3.13) to (3.15).

Example 3.23. For the problem of Example 3.1 we have

$$
\begin{aligned}
(E, A) &= \left(\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 1 & -t \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ -1 & t \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \right) \\
&\overset{\text{dif}}{\sim} \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \right).
\end{aligned}
$$

Hence, the characteristic values are given by

$$
\begin{aligned}
&r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 1, \quad v_0 = 0, \\
&r_1 = 0, \quad a_1 = 1, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 1, \quad v_1 = 1,
\end{aligned}
$$

together with $\mu = 1$. The corresponding differential-algebraic equation therefore consists of one algebraic equation in combination with one consistency condition for the right hand side and one undetermined solution component. In particular, the solution of the homogeneous initial value problem is not unique, in agreement with the results of Example 3.1.

Example 3.24. For the problem of Example 3.2 we have

$$
\begin{aligned}
(E, A) &= \left(\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix}, \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & t \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \right) \sim \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right) \\
&\overset{\text{dif}}{\sim} \left(\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).
\end{aligned}
$$

Hence, the characteristic values are given by

$$
\begin{aligned}
&r_0 = 1, \quad a_0 = 0, \quad s_0 = 1, \quad d_0 = 0, \quad u_0 = 1, \quad v_0 = 0, \\
&r_1 = 0, \quad a_1 = 2, \quad s_1 = 0, \quad d_1 = 0, \quad u_1 = 0, \quad v_1 = 0,
\end{aligned}
$$

together with $\mu = 1$. The corresponding differential-algebraic equation therefore consists of two algebraic equations. In particular, the solution is unique without supplying an initial condition, in agreement with the results of Example 3.2.

It remains to discuss the question how restrictive the constant rank assumptions (3.12) are that we had to apply in each step of the described inductive transformation procedure. The answer lies in the following observation on the rank of continuous matrix functions.

Theorem 3.25. *Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $M \in C(\mathbb{I}, \mathbb{C}^{m,n})$. Then there exist open intervals $\mathbb{I}_j \subseteq \mathbb{I}$, $j \in \mathbb{N}$, with*

$$\overline{\bigcup_{j\in\mathbb{N}} \mathbb{I}_j} = \mathbb{I}, \qquad \mathbb{I}_i \cap \mathbb{I}_j = \emptyset \quad \text{for } i \neq j, \tag{3.26}$$

and integers $r_j \in \mathbb{N}_0$, $j \in \mathbb{N}$, such that

$$\operatorname{rank} M(t) = r_j \quad \text{for all } t \in \mathbb{I}_j. \tag{3.27}$$

Proof. See, e.g., [56, Ch. 10]. □

Applying this property of a continuous matrix function to the construction leading to Theorem 3.17, one immediately obtains the following result.

Corollary 3.26. *Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $E, A \in C(\mathbb{I}, \mathbb{C}^{m,n})$ be sufficiently smooth. Then there exist open intervals $\mathbb{I}_j$, $j \in \mathbb{N}$, as in Theorem 3.25, such that the strangeness index of (E, A) restricted to $\mathbb{I}_j$ is well defined for every $j \in \mathbb{N}$.*

As a consequence, the strangeness index is defined on a dense subset of the given closed interval, and we can transform to the global canonical form (3.23) on each component $\mathbb{I}_j$ separately.

3.2 Local and global invariants

In Section 3.1, we have introduced characteristic values (invariants) of matrix pairs together with a canonical form which we called local, because we obtained the matrix pairs as evaluation of a pair of matrix functions at a fixed point. We then required these (local) invariants to be global, cp. (3.12), and developed a whole sequence of invariants and a corresponding global canonical form. We call these *global invariants*, since they are only defined by a process involving transformations by matrix functions. We have shown that the involved constant rank assumptions are satisfied on a dense subset of the given interval. At an *exceptional point*, where any of the ranks changes, the only available information is given by the local invariants of the pair evaluated there. In particular, we cannot associate a strangeness index with these points. In this section, we will develop purely local invariants which will

allow to determine the global invariants including the strangeness index, wherever they are defined.

Since the process in Section 3.1 included differentiation, the idea (which is due to Campbell, see [47]) is to differentiate the original differential-algebraic equation (3.1). In this way, we get so-called *derivative arrays* or *inflated differential-algebraic equations*

$$M_\ell(t)\dot{z}_\ell = N_\ell(t)z_\ell + g_\ell(t), \tag{3.28}$$

where

$$\begin{aligned}
(M_\ell)_{i,j} &= \tbinom{i}{j}E^{(i-j)} - \tbinom{i}{j+1}A^{(i-j-1)}, \quad i,j = 0,\dots,\ell,\\
(N_\ell)_{i,j} &= \begin{cases} A^{(i)} & \text{for } i = 0,\dots,\ell,\ j = 0,\\ 0 & \text{otherwise,}\end{cases}\\
(z_\ell)_j &= x^{(j)},\ j = 0,\dots,\ell,\\
(g_\ell)_i &= f^{(i)},\ i = 0,\dots,\ell,
\end{aligned} \tag{3.29}$$

using the convention that $\binom{i}{j} = 0$ for $i < 0$, $j < 0$ or $j > i$. In more detail, we have

$$\begin{aligned}
M_\ell &= \begin{bmatrix} E & & & & \\ \dot{E}-A & E & & & \\ \ddot{E}-2\dot{A} & 2\dot{E}-A & E & & \\ \vdots & & \ddots & \ddots & \\ E^{(\ell)}-\ell A^{(\ell-1)} & \cdots & \cdots & \ell\dot{E}-A & E \end{bmatrix},\\
N_\ell &= \begin{bmatrix} A & 0 & \cdots & 0 \\ \dot{A} & 0 & \cdots & 0 \\ \ddot{A} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ A^{(\ell)} & 0 & \cdots & 0 \end{bmatrix}.
\end{aligned} \tag{3.30}$$

For every $\ell \in \mathbb{N}_0$ and every $t \in \mathbb{I}$, we can determine the local characteristic values of the pair $(M_\ell(t), N_\ell(t))$ if it is defined. To do this, we first show that these local quantities are invariant under global equivalence transformations of the pair (E, A) of matrix functions. For this, we need some elementary lemmas.

Lemma 3.27. *For all integers i, j, k, l with $i \geq 0$, $i \geq j \geq 0$, $i - j \geq k \geq 0$, we have the identities*

$$\tbinom{i}{j}\tbinom{i-j}{k}\tbinom{i-j-k}{l} + \tbinom{i}{j+1}\tbinom{i-j-1}{k}\tbinom{i-j-k-1}{l} = \tbinom{i}{k}\tbinom{i-k}{l}\tbinom{i-k-l+1}{j+1}, \tag{3.31a}$$

$$\binom{i}{j+1}\binom{i-j-1}{k}\binom{i-j-k-1}{l} = \binom{i}{k}\binom{i-k}{l}\binom{i-k-l}{j+1}, \tag{3.31b}$$

$$\binom{i}{k-1}\binom{i-k+1}{l} + \binom{i}{k}\binom{i-k}{l-1} + \binom{i}{k}\binom{i-k}{l} = \binom{i+1}{k}\binom{i+1-k}{l}. \tag{3.31c}$$

Proof. The proof follows by straightforward calculation. □

Lemma 3.28. *Let $D = ABC$ be the product of three sufficiently smooth matrix valued functions of appropriate dimensions. Then*

$$D^{(i)} = \sum_{j=0}^{i} \sum_{k=0}^{i-j} \binom{i}{j}\binom{i-j}{k} A^{(j)} B^{(k)} C^{(i-j-k)}. \tag{3.32}$$

Proof. The proof follows immediately by induction using (3.31c). □

Theorem 3.29. *Consider two pairs (E, A) and $(\tilde{E}, \tilde{A})$ of sufficiently smooth matrix functions that are equivalent via the transformation*

$$\tilde{E} = PEQ, \quad \tilde{A} = PAQ - PE\dot{Q} \tag{3.33}$$

according to Definition 3.3, with sufficiently smooth P and Q. Let (M_ℓ, N_ℓ) and $(\tilde{M}_\ell, \tilde{N}_\ell)$, $\ell \in \mathbb{N}_0$, be the corresponding inflated pairs constructed as in (3.29) and introduce the block matrix functions

$$\begin{aligned} (\Pi_\ell)_{i,j} &= \binom{i}{j} P^{(i-j)}, \quad (\Theta_\ell)_{i,j} = \binom{i+1}{j+1} Q^{(i-j)}, \\ (\Psi_\ell)_{i,j} &= \begin{cases} Q^{(i+1)} & \text{for } i = 0, \ldots, \ell, \ j = 0, \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{3.34}$$

Then

$$[\ \tilde{M}_\ell(t)\ \ \tilde{N}_\ell(t)\ ] = \Pi_\ell(t)[\ M_\ell(t)\ \ N_\ell(t)\ ] \begin{bmatrix} \Theta_\ell(t) & -\Psi_\ell(t) \\ 0 & \Theta_\ell(t) \end{bmatrix} \tag{3.35}$$

for every $t \in \mathbb{I}$. In particular, the corresponding matrix pairs are locally equivalent.

Proof. All matrix functions M_ℓ, N_ℓ, $\tilde{M}_\ell$, $\tilde{N}_\ell$, Π_ℓ, Θ_ℓ, and Ψ_ℓ are block lower triangular with the same block structure. Observe, furthermore, that N_ℓ, $\tilde{N}_\ell$, and Ψ_ℓ have nonzero blocks only in the first block column. From Lemma 3.28 we obtain (leaving out the argument t)

$$\begin{aligned} \tilde{E}^{(i)} &= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \binom{i}{k_1}\binom{i-k_1}{k_2} P^{(k_1)} E^{(k_2)} Q^{(i-k_1-k_2)} \\ \tilde{A}^{(i)} &= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \Big[\binom{i}{k_1}\binom{i-k_1}{k_2} P^{(k_1)} A^{(k_2)} Q^{(i-k_1-k_2)} \\ &\qquad - \binom{i}{k_1}\binom{i-k_1}{k_2} P^{(k_1)} E^{(k_2)} Q^{(i+1-k_1-k_2)}\Big], \end{aligned}$$

and

$$(\Pi_\ell M_\ell \Theta_\ell)_{i,j} = \sum_{l_1=j}^{i} \sum_{l_2=j}^{l_1} (\Pi_\ell)_{i,l_1} (M_\ell)_{l_1,l_2} (\Theta_\ell)_{l_2,j}$$
$$= \sum_{l_1=j}^{i} \sum_{l_2=j}^{l_1} \tbinom{i}{l_1} P^{(i-l_1)} \big[\tbinom{l_1}{l_2} E^{(l_1-l_2)} - \tbinom{l_1}{l_2+1} A^{(l_1-l_2-1)} \big] \tbinom{l_2+1}{j+1} Q^{(l_2-j)}$$

by inserting the definitions. Shifting and inverting the summations and applying (3.31a) and (3.31b) then leads to

$$\begin{aligned}
&(\Pi_\ell M_\ell \Theta_\ell)_{i,j} \\
&\quad = \sum_{k_1=0}^{i-j} \sum_{l_2=j}^{k_1+j} \tbinom{i}{k_1+j} P^{(i-j-k_1)} \big[\tbinom{k_1+j}{l_2} E^{(k_1+j-l_2)} \\
&\qquad - \tbinom{k_1+j}{l_2+1} A^{(k_1+j-l_2-1)} \big] \tbinom{l_2+1}{j+1} Q^{(l_2-j)} \\
&\quad = \tbinom{i}{j} \sum_{k_1=0}^{i-j} \sum_{k_2=0}^{i-j-k_1} \tbinom{i-j}{k_1} P^{(k_1)} \tbinom{i-j-k_1}{k_2} E^{(k_2)} Q^{(i-j-k_1-k_2)} \\
&\qquad - \tbinom{i}{j+1} \sum_{k_1=0}^{i-j-1} \sum_{k_2=0}^{i-j-1-k_1} \big[\tbinom{i-j-1}{k_1} P^{(k_1)} \tbinom{i-j-k_1-1}{k_2} A^{(k_2)} Q^{(i-j-1-k_1-k_2)} \\
&\qquad - \tbinom{i-j-1}{k_1} P^{(k_1)} \tbinom{i-j-1-k_1}{k_2} E^{(k_2)} Q^{(i-j-k_1-k_2)} \big] \\
&\quad = \tbinom{i}{j} \tilde{E}^{(i-j)} - \tbinom{i}{j+1} \tilde{A}^{(i-j-1)} = (\tilde{M}_\ell)_{i,j}.
\end{aligned}$$

In a similar way it follows that

$$\begin{aligned}
&(\Pi_\ell N_\ell \Theta_\ell)_{i,0} - (\Pi_\ell M_\ell \Psi_\ell)_{i,0} \\
&\quad = \sum_{l_1=0}^{i} (\Pi_\ell)_{i,l_1} (N_\ell)_{l_1,0} (\Theta_\ell)_{0,0} - \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} (\Pi_\ell)_{i,l_1} (M_\ell)_{l_1,l_2} (\Psi_\ell)_{l_2,0} \\
&\quad = \sum_{k_1=0}^{i} \tbinom{i}{k_1} P^{(k_1)} A^{(i-k_1)} Q^{(0)} + \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-1-k_1} \tbinom{i}{k_1} P^{(k_1)} \tbinom{i-k_1}{k_2} A^{(k_2)} Q^{(i-k_1-k_2)} \\
&\qquad - \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} P^{(k_1)} \tbinom{i-k_1}{k_2} E^{(k_2)} Q^{(i-k_1-k_2+1)} \\
&\quad = \tilde{A}^{(i)} = (\tilde{N}_\ell)_{i,0}.
\end{aligned}$$

□

Thus, we have shown that the local characteristic values of the inflated pair $(M_\ell(t), N_\ell(t))$, in the following denoted by $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell, \tilde{d}_\ell, \tilde{u}_\ell, \tilde{v}_\ell)$, are well defined for equivalent pairs of matrix functions. This observation immediately raises the question, whether these quantities are related to the global characteristic values $(r_i, a_i, s_i, d_i, u_i, v_i)$ derived in the Section 3.1.

Theorem 3.30. *Let the strangeness index μ of (E, A) as in (3.17) be well defined with (global) characteristic values (r_i, a_i, s_i), $i \in \mathbb{N}_0$. Moreover, let $(M_\ell(t), N_\ell(t))$, $\ell \in \mathbb{N}_0$, be the corresponding inflated matrix pair at a fixed $t \in \mathbb{I}$ with (local) characteristic values $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$. Then,*

$$\tilde{r}_\ell = (\ell+1)m - \sum_{i=0}^{\ell} c_i - \sum_{i=0}^{\ell} v_i, \tag{3.36a}$$

$$\tilde{a}_\ell = c_\ell - s_\ell, \tag{3.36b}$$

$$\tilde{s}_\ell = s_\ell + \sum_{i=0}^{\ell-1} c_i, \tag{3.36c}$$

with c_i defined in (3.21b).

Proof. By Theorem 3.29, we may assume without loss of generality that the pair (E, A) is already in the global canonical form (3.23). For fixed $t \in \mathbb{I}$, we must determine the local characteristic quantities of $(M_\ell(t), N_\ell(t))$. For convenience, we omit the argument t. We first observe that, with the help of the entries I_{d_μ} coming from (3.23), we can eliminate all other entries in the corresponding block rows and columns of (M_ℓ, N_ℓ). Recall that this is an allowed equivalence transformation of the form (3.4). It is then clear that the entries I_{d_μ} always contribute to the rank of M_ℓ and the (transformed) kernel and corange vectors must have zero entries at the corresponding places. In addition, we can omit the zero columns of the normal form (3.23), since these do not affect the characteristic values $\tilde{a}_\ell$ and $\tilde{s}_\ell$. It is therefore sufficient to consider

$$(\tilde{E}, \tilde{A}) = \left(\begin{bmatrix} F \\ G \end{bmatrix}, \begin{bmatrix} 0 \\ I_{a_\mu} \end{bmatrix} \right).$$

Denoting the corresponding inflated pairs by $(\tilde{M}_\ell, \tilde{N}_\ell)$, we perform a suitable block row and column permutation. In particular, we write the block rows and columns in opposite order and then exchange block rows and columns in such a way that the entries coming from F and those from G are separated. Thus, with respect to local equivalence, we have

$$(\tilde{M}_\ell, \tilde{N}_\ell) \sim \left(\begin{bmatrix} \mathfrak{f}_\ell & Y_\ell \\ \mathfrak{g}_\ell & X_\ell - I \\ 0 & \mathfrak{h}_\ell \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \mathfrak{e}_\ell \end{bmatrix} \right),$$

where

$$\mathfrak{f}_\ell = \begin{bmatrix} F \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathfrak{g}_\ell = \begin{bmatrix} G \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathfrak{h}_\ell = \begin{bmatrix} 0 & \cdots & 0 & F \\ 0 & \cdots & 0 & G \end{bmatrix}, \quad \mathfrak{e}_\ell = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ 0 & \cdots & 0 & I_{a_\mu} \end{bmatrix},$$

and

$$X_\ell = \begin{bmatrix} \ell\dot{G} & \cdots & \cdots & G^{(\ell)} \\ G & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & G & \dot{G} \end{bmatrix}, \quad Y_\ell = \begin{bmatrix} \ell\dot{F} & \cdots & \cdots & F^{(\ell)} \\ F & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & F & \dot{F} \end{bmatrix}.$$

Since G and all its derivatives are nilpotent due to their structure, X_ℓ is nilpotent. Hence, $X_\ell - I$ is invertible. Since, in addition, $X_\ell - I$ has upper block Hessenberg form, we can decompose it according to

$$X_\ell - I = U_\ell(L_\ell - I)$$

with

$$U_\ell = \begin{bmatrix} I - H_\ell & * & \cdots & * \\ & \ddots & \ddots & \vdots \\ & & \ddots & * \\ & & & I - H_1 \end{bmatrix},$$

$$L_\ell = \begin{bmatrix} 0 & & & & \\ (I - H_{\ell-1})^{-1}G & 0 & & & \\ & \ddots & \ddots & & \\ & & & (I - H_1)^{-1}G & 0 \end{bmatrix}.$$

Typically, the arising entries here and in the following are sums of products, where the first factor is F or G or a derivative of them and the other factors are G or derivatives of it. For convenience, we call a term which is a sum of products with at least k factors a *k-term*. Note that k-terms with $k \geq 1$ are automatically nilpotent due to the structure of F and G and their derivatives. In this sense, it follows by induction that H_i, $i = 1, \ldots, \ell$, and the entries $*$ of U_ℓ are 1-terms.

Let $\tilde{Y}_\ell$ denote the upper block triangular part of the upper block Hessenberg matrix Y_ℓ. The relation

$$R_\ell(L_\ell - I) = \tilde{Y}_\ell + \tilde{D}_\ell$$

uniquely determines an upper block triangular matrix R_ℓ and a matrix $\tilde{D}_\ell$ which only has nontrivial entries in the lower block diagonal (similar to L_ℓ). Induction shows that the nontrivial entries of R_ℓ are 1-terms and those of $\tilde{D}_\ell$ are 2-terms.

With these preparations, we get

$$\begin{aligned}(\tilde{M}_\ell, \tilde{N}_\ell) &\sim \left(\begin{bmatrix} \mathfrak{f}_\ell & Y_\ell \\ \mathfrak{g}_\ell & U_\ell(L_\ell - I) \\ 0 & \mathfrak{h}_\ell \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \mathfrak{e}_\ell \end{bmatrix}\right) \\ &\sim \left(\begin{bmatrix} \mathfrak{f}_\ell - R_\ell U_\ell^{-1}\mathfrak{g}_\ell & Y_\ell - \tilde{Y}_\ell - \tilde{D}_\ell \\ U_\ell^{-1}\mathfrak{g}_\ell & L_\ell - I \\ 0 & \mathfrak{h}_\ell \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \mathfrak{e}_\ell \end{bmatrix}\right),\end{aligned}$$

where all nontrivial entries besides the identities have the form F or G plus 2-terms. Hence, these blocks differ from F or G only in the entries $*$ in (3.24).

Therefore, we have (by reordering the block rows)

$$(\tilde{M}_\ell, \tilde{N}_\ell) \sim \left(\left[\begin{array}{c|ccc|c} \tilde{F}_0 & & & & \\ \tilde{G}_0 & -I_{a_\mu} & & & \\ \hline & \tilde{F}_1 & & & \\ & \tilde{G}_1 & \ddots & & \\ & & \ddots & \tilde{F}_{\ell-1} & \\ & & & \tilde{G}_{\ell-1} & -I_{a_\mu} \\ \hline & & & & \tilde{F}_\ell \\ & & & & \tilde{G}_\ell \end{array}\right], \left[\begin{array}{c|ccc|c} 0 & & & & \\ 0 & 0 & & & \\ \hline & 0 & & & \\ & 0 & \ddots & & \\ & & \ddots & 0 & \\ & & & 0 & 0 \\ \hline & & & & 0 \\ & & & & I_{a_\mu} \end{array}\right]\right),$$

with

$$\tilde{F}_i = \begin{bmatrix} 0 & F_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & F_1 \\ & & & 0 \end{bmatrix}, \quad \tilde{G}_i = \begin{bmatrix} 0 & G_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & G_1 \\ & & & 0 \end{bmatrix}, \quad i = 0, \ldots, \ell.$$

For a further refinement of the structure of the matrix pair, let

$$W_j = \begin{bmatrix} & & \\ & & \\ I_{c_{j-1}} & & \\ & \ddots & \\ & & I_{c_0} \end{bmatrix}, \quad W_j' = \begin{bmatrix} I_{c_\mu} & & \\ & \ddots & \\ & & I_{c_j} \\ & & \\ & & \end{bmatrix}, \quad j = 0, \ldots, \mu + 1.$$

Then, the columns of W_j form an orthonormal basis of corange G_j and the relations $[\ W_j'\ \ W_j\ ] = I_{a_\mu}$ and range $\tilde{G}_j W_j' =$ range W_{j+1}' hold. By induction, it follows that $(\tilde{M}_\ell, \tilde{N}_\ell) \sim (\hat{M}_\ell, \hat{N}_\ell)$ with

$$(\hat{M}_\ell, \hat{N}_\ell) = \left(\begin{bmatrix} \tilde{F}_0 W_0' & & & & & & \\ \tilde{G}_0 W_0' & W_1 & W_1' & & & & \\ \hline & & \tilde{F}_1 W_1' & & & & \\ & & \tilde{G}_1 W_1' & W_2 & W_2' & & \\ \hline & & & & \ddots & \ddots & \\ & & & & & \ddots & \\ \hline & & & & & \tilde{F}_{\ell-1} W_{\ell-1}' & \\ & & & & & \tilde{G}_{\ell-1} W_{\ell-1}' & W_\ell \quad W_\ell' \\ \hline & & & & & & \tilde{F}_\ell W_\ell' \\ & & & & & & \tilde{G}_\ell W_\ell' \end{bmatrix}, \right.$$

$$\left. \begin{bmatrix} 0 & & & & & & \\ 0 & 0 & 0 & & & & \\ \hline & & 0 & & & & \\ & & 0 & 0 & 0 & & \\ \hline & & & \ddots & \ddots & & \\ & & & & \ddots & \ddots & \\ \hline & & & & 0 & & \\ & & & & 0 & 0 & 0 \\ \hline & & & & & 0 & 0 \\ & & & & & W_\ell & W_\ell' \end{bmatrix} \right).$$

We see that the block rows of $\hat{M}_\ell$ are decoupled such that

$$\operatorname{rank} \hat{M}_\ell = \operatorname{rank} \begin{bmatrix} \tilde{F}_0 W_0' & 0 \\ \tilde{G}_0 W_0' & W_1 \end{bmatrix} + \cdots + \operatorname{rank} \begin{bmatrix} \tilde{F}_{\ell-1} W_{\ell-1}' & 0 \\ \tilde{G}_{\ell-1} W_{\ell-1}' & W_\ell \end{bmatrix} + \operatorname{rank} \begin{bmatrix} \tilde{F}_\ell W_\ell' \\ \tilde{G}_\ell W_\ell' \end{bmatrix}.$$

Recognizing that

$$\operatorname{rank} \begin{bmatrix} \tilde{F}_i W_i' & 0 \\ \tilde{G}_i W_i' & W_{i+1} \end{bmatrix} = a_\mu + (w_{i+1} + \cdots + w_\mu), \quad i = 0, \ldots, \ell - 1,$$

and

$$\operatorname{rank} \begin{bmatrix} \tilde{F}_\ell W_\ell' \\ \tilde{G}_\ell W_\ell' \end{bmatrix} = (c_{\ell+1} + \cdots + c_\mu) + (w_{\ell+1} + \cdots + w_\mu),$$

we obtain

$$\begin{aligned}
\tilde{r}_\ell &= \operatorname{rank} M_\ell = (\ell+1)d_\mu + \operatorname{rank} \hat{M}_\ell \\
&= (\ell+1)d_\mu + \ell a_\mu + (c_{\ell+1} + \cdots + c_\mu) \\
&\qquad + (w_1 + \cdots + w_\mu) + (w_2 + \cdots + w_\mu) + \cdots + (w_{\ell+1} + \cdots + w_\mu) \\
&= \tilde{r}_{\ell-1} + d_\mu + a_\mu - c_\ell - (c_1 + \cdots + c_\mu) \\
&\qquad - (w_1 + \cdots + w_\ell) + (s_0 + \cdots + s_{\mu-1}) \\
&= \tilde{r}_{\ell-1} + m - c_0 - v_0 + c_0 - c_\ell - (w_1 + \cdots + w_\ell) \\
&= \tilde{r}_{\ell-1} + m - c_\ell - v_\ell
\end{aligned}$$

as asserted.

To compute $\tilde{a}_\ell$ and $\tilde{s}_\ell$, we first must determine the corange and the kernel of $\hat{M}_\ell$. Let

$$U_j = \begin{bmatrix} & & \\ I_{w_{j-1}} & & \\ & \ddots & \\ & & I_{w_0} \end{bmatrix}, \quad j = 0, \ldots, \mu+1.$$

Then, the corange of $\hat{M}_\ell$ is given by

$$\hat{Z}_\ell = \left[\begin{array}{c|c|c} U_1 & & \\ 0 & & \\ \hline & U_2 & \\ & 0 \quad \ddots & \\ & \ddots \quad U_\ell & \\ & \quad 0 & \\ \hline & & U_{\ell+1} \\ & & \quad W_{\ell+1} \end{array}\right].$$

The computation of the kernel is more complicated, since $\hat{M}_\ell$ does not decouple with respect to the block columns. Let the columns of $\tilde{K}_i$, $i = 0, \ldots, \ell$, span

$$\operatorname{kernel} \begin{bmatrix} \tilde{F}_i W_i' \\ \tilde{G}_i W_i' \end{bmatrix}$$

and let $\tilde{K}_i'$ be given, such that $[\ \tilde{K}_i' \ \ \tilde{K}_i\ ]$ is invertible. Then the kernel of $\hat{M}_\ell$ and a

possible complement (cp. Remark 3.8) are given by

$$\hat{T}_\ell = \left[\begin{array}{c|ccc|c} \tilde{K}_0 & * & \cdots & * & * \\ 0 & 0 & \cdots & 0 & 0 \\ \hline & \tilde{K}_1 & & & \vdots \\ & 0 & \ddots & & \vdots \\ & & \ddots & \tilde{K}_{\ell-1} & * \\ & & & 0 & 0 \\ \hline & & & & \tilde{K}_\ell \end{array}\right],$$

$$\hat{T}'_\ell = \left[\begin{array}{cc|cccccc|c} \tilde{K}'_0 & 0 & & & & & & & \\ 0 & I & & & & & & & \\ \hline & & \tilde{K}'_1 & 0 & & & & & \\ & & 0 & I & \ddots & & & & \\ & & & \ddots & \ddots & \ddots & & & \\ & & & & \ddots & \tilde{K}'_{\ell-1} & 0 & & \\ & & & & & 0 & I & & \\ \hline & & & & & & & & \tilde{K}'_\ell \end{array}\right].$$

Hence, we have that

$$\hat{Z}_\ell^H \hat{N}_\ell \hat{T}_\ell = \begin{bmatrix} 0 & 0 & 0 \\ 0 & W_{\ell+1}^H W_\ell & W_{\ell+1}^H W'_\ell \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & \tilde{K}_\ell \end{bmatrix}.$$

We can choose $\tilde{K}_\ell$ together with a possible complement $\tilde{K}'_\ell$ as

$$\tilde{K}_\ell = \left[\begin{array}{c|ccc} I_{c_\mu} & 0 & \cdots & 0 \\ \hline & K_\mu & & * \\ & & \ddots & \\ & & & K_{\ell+1} \end{array}\right], \quad \tilde{K}'_\ell = \left[\begin{array}{ccc} 0 & \cdots & 0 \\ \hline K'_\mu & & \\ & \ddots & \\ & & K'_{\ell+1} \end{array}\right],$$

where the columns of K_i, $i = 1, \ldots, \mu$, span the common kernel of F_i and G_i and $[\ K'_i \ \ K_i\ ]$ is nonsingular. From (3.25), it follows that

$$\operatorname{rank} K_i = \dim \operatorname{kernel} \begin{bmatrix} F_i \\ G_i \end{bmatrix} = c_{i-1} - c_i - w_i = c_{i-1} - s_{i-1}.$$

Since

$$W_{\ell+1}^H W_\ell' \tilde{K}_\ell = \begin{bmatrix} & I_{c_\ell} \\ & \end{bmatrix} \begin{bmatrix} I_{c_\mu} & 0 & \cdots & 0 \\ & K_\mu & & * \\ & & \ddots & \\ & & & K_{\ell+1} \end{bmatrix} = \begin{bmatrix} & K_{\ell+1} \\ & \end{bmatrix},$$

we obtain that

$$\tilde{a}_\ell = \operatorname{rank}(\hat{Z}_\ell^H \hat{N}_\ell \hat{T}_\ell) = \operatorname{rank}(W_{\ell+1}^H W_\ell' \tilde{K}_\ell) = \operatorname{rank} K_{\ell+1} = c_\ell - s_\ell.$$

The columns of

$$\hat{V}_\ell = \begin{bmatrix} I & \\ & V_\ell \end{bmatrix}, \quad V_\ell = \begin{bmatrix} K_{\ell+1}' & & & \\ & I_{c_{\ell-1}} & & \\ & & \ddots & \\ & & & I_{c_0} \end{bmatrix},$$

span corange$(\hat{Z}_\ell^H \hat{N}_\ell \hat{T}_\ell)$. Thus, we have that

$$\begin{aligned} \hat{V}_\ell^H \hat{Z}_\ell^H \hat{N}_\ell \hat{T}_\ell' &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & V_\ell^H W_{\ell+1}^H W_\ell & V_\ell^H W_{\ell+1}^H W_\ell' \end{bmatrix} \begin{bmatrix} * & & \\ & I & \\ & & \tilde{K}_\ell' \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & V_\ell^H W_{\ell+1}^H W_\ell & V_\ell^H W_{\ell+1}^H W_\ell' \tilde{K}_\ell' \end{bmatrix}, \end{aligned}$$

with

$$V_\ell^H W_{\ell+1}^H W_\ell = \begin{bmatrix} I_{c_{\ell-1}} & & \\ & \ddots & \\ & & I_{c_0} \end{bmatrix}, \quad V_\ell^H W_{\ell+1}^H W_\ell' \tilde{K}_\ell' = \begin{bmatrix} & K_{\ell+1}'^H K_{\ell+1}' \\ & \end{bmatrix}.$$

It finally follows that

$$\begin{aligned} \tilde{s}_\ell &= \operatorname{rank}(\hat{V}_\ell^H \hat{Z}_\ell^H \hat{N}_\ell \hat{T}_\ell') = \operatorname{rank}[\, V_\ell^H W_{\ell+1}^H W_\ell \quad V_\ell^H W_{\ell+1}^H W_\ell' \tilde{K}_{\ell+1}' \,] r \\ &= \operatorname{rank}(V_\ell^H W_{\ell+1}^H W_\ell) + \operatorname{rank} K_{\ell+1}' = (c_0 + \cdots + c_{\ell-1}) + s_\ell. \end{aligned}$$

□

We also have the converse of this result, i.e., the knowledge of the sequence $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$ allows for the determination of the sequence (r_i, a_i, s_i) of the (global) characteristic values of (E, A).

Corollary 3.31. *Let the strangeness index μ of (E, A) as in (3.17) be well defined and let $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$, $\ell = 0, \ldots, \mu$, be the sequence of the (local) characteristic values of $(M_\ell(t), N_\ell(t))$ for some $t \in \mathbb{I}$. Then the sequence (r_i, a_i, s_i) of the (global) characteristic values of (E, A) can be obtained from*

$$c_0 = \tilde{a}_0 + \tilde{s}_0, \qquad c_{i+1} = (\tilde{a}_{i+1} - \tilde{a}_i) + (\tilde{s}_{i+1} - \tilde{s}_i), \tag{3.37a}$$

$$v_0 = m - c_0 - \tilde{r}_0, \qquad v_{i+1} = m - c_{i+1} - (\tilde{r}_{i+1} - \tilde{r}_i), \tag{3.37b}$$

$$s_i = c_i - \tilde{a}_i, \tag{3.37c}$$

$$a_i = c_0 + \cdots + c_i - s_i, \tag{3.37d}$$

$$r_i = m - a_i - s_i - v_i. \tag{3.37e}$$

Proof. The claim follows by trivial rearrangements of (3.36) together with the relations of Lemma 3.20. □

With these results, we have shown that for well-defined strangeness index μ the complete structural information on the global characteristic values of (E, A) can be obtained from the local information of the inflated pairs $(M_\ell(t), N_\ell(t))$, $\ell = 0, \ldots, \mu$. In view of Theorem 3.17, an immediate question is whether it is possible to derive a system of the form (3.18) by using only local information from $(M_\mu(t), N_\mu(t))$.

Let $(\tilde{E}, \tilde{A})$ be a normal form of (E, A) according to (3.23) with inflated pairs $(\tilde{M}_\mu, \tilde{N}_\mu)$. For convenience, in the following we omit the subscript μ.

Note first that, if the columns of Z span the corange of M, multiplication of (3.28) for $\ell = \mu$, now reading $M(t)\dot{z} = N(t)z + g(t)$, by $Z(t)^H$ gives $0 = Z(t)^H N(t)z + Z(t)^H g(t)$. But recall that the only nontrivial entries in N are in the first block column belonging to the original unknown x. Hence, we get purely algebraic equations for x. Comparing with (3.18b), we are looking for a_μ such equations. Indeed, (3.36c) gives $\tilde{a}_\mu + \tilde{s}_\mu = a_\mu$. Since the rank of $\tilde{M}$ is constant, Theorem 3.9 yields the existence of a continuous matrix function $\tilde{Z}$, whose columns form a basis of corange $\tilde{M}$. By definition, $\tilde{Z}^H \tilde{N}$ has rank $\tilde{a}_\mu + \tilde{s}_\mu$. Applying again Theorem 3.9, it follows that there exists a continuous matrix function $\tilde{Z}_2$ of size $(\mu + 1)m \times a_\mu$ with

$$\tilde{Z}_2^H = [\tilde{Z}_{2,0}^H \tilde{Z}_{2,1}^H \cdots \tilde{Z}_{2,\mu}^H], \quad \tilde{Z}_{2,0}^H = [0\,0\,I_{a_\mu}], \quad \tilde{Z}_{2,\ell}^H = [0\,0\,*], \quad \ell = 1, \ldots, \mu,$$

such that

$$\operatorname{rank}(\tilde{Z}_2^H \tilde{N}[\, I_n \;\; 0 \;\; \cdots \;\; 0\,]^H) = \operatorname{rank}\left(\begin{bmatrix} 0 \\ 0 \\ I_{a_\mu} \end{bmatrix}^H \begin{bmatrix} * & * & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I_{a_\mu} \end{bmatrix} \right) = a_\mu.$$

As already mentioned, with $\tilde{Z}_2$ we obtain the complete set of algebraic equations. Next, we must get d_μ differential equations which complete these algebraic equations to a strangeness-free differential-algebraic equation. In each step of the iterative procedure of the previous section, the number of equations with derivatives of the unknown function is reduced. Therefore, the differential equations we look for must be already present in the original system. If we set

$$\tilde{T}_2 = \begin{bmatrix} I_{d_\mu} & 0 \\ 0 & I_{u_\mu} \\ 0 & 0 \end{bmatrix},$$

then we get

$$\tilde{Z}_2^H \tilde{N} [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H \tilde{T}_2 = 0$$

and

$$\operatorname{rank}(\tilde{E}\tilde{T}_2) = \operatorname{rank}\left(\begin{bmatrix} I_{d_\mu} & 0 & * \\ 0 & 0 & F \\ 0 & 0 & G \end{bmatrix} \begin{bmatrix} I_{d_\mu} & 0 \\ 0 & I_{u_\mu} \\ 0 & 0 \end{bmatrix} \right) = d_\mu.$$

Obviously, the desired differential equations correspond to the first block of $\tilde{E}\tilde{T}_2$. The construction obtained in this way must now be carried over to the matrix functions of the original problem. From

$$\tilde{M} = \Pi M \Theta, \quad \tilde{N} = \Pi N \Theta - \Pi M \Psi$$

according to (3.35), it follows that

$$\tilde{Z}_2^H \tilde{M} = \tilde{Z}_2^H \Pi M \Theta = 0$$

and, due to the special structure of N,

$$\begin{aligned} a_\mu &= \operatorname{rank}(\tilde{Z}_2^H \tilde{N} [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H) \\ &= \operatorname{rank}(\tilde{Z}_2^H \Pi N \Theta [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H) \\ &= \operatorname{rank}(\tilde{Z}_2^H \Pi N [\, Q^H \;\; * \;\; \cdots \;\; * \,]^H) \\ &= \operatorname{rank}(\tilde{Z}_2^H \Pi N [\, Q^H \;\; 0 \;\; \cdots \;\; 0 \,]^H) \\ &= \operatorname{rank}(\tilde{Z}_2^H \Pi N [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H Q). \end{aligned}$$

So there exists a smooth matrix valued function $Z_2 = \Pi^H \tilde{Z}_2$ of size $(\mu+1)m \times a_\mu$ such that

$$Z_2^H M = 0, \quad \operatorname{rank}(Z_2^H N [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H) = a_\mu. \tag{3.38}$$

From

$$\begin{aligned} 0 &= \tilde{Z}_2^H \tilde{N} [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H \tilde{T}_2 \\ &= \tilde{Z}_2^H \Pi N \Theta [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H \tilde{T}_2 \\ &= \tilde{Z}_2^H \Pi N [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^H Q \tilde{T}_2 \end{aligned}$$

and

$$d_\mu = \operatorname{rank}(\tilde{E}\tilde{T}_2) = \operatorname{rank}(PEQ\tilde{T}_2) = \operatorname{rank}(EQ\tilde{T}_2),$$

it follows that there exists a smooth matrix function $T_2 = Q\tilde{T}_2$ of size $n \times (d_\mu + u_\mu)$ such that

$$Z_2^H N[\, I_n \;\; 0 \;\; \cdots \;\; 0\,]^H T_2 = 0, \quad \operatorname{rank}(ET_2) = d_\mu. \tag{3.39}$$

Thus, there exists a smooth matrix function Z_1 of size $m \times d_\mu$ such that

$$\operatorname{rank}(Z_1^H E T_2) = d_\mu. \tag{3.40}$$

In summary, we have constructed a pair of matrix functions

$$(\hat{E}, \hat{A}) = \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right), \tag{3.41}$$

with entries

$$\hat{E}_1 = Z_1^H E, \quad \hat{A}_1 = Z_1^H A, \quad \hat{A}_2 = Z_2^H N[\, I_n \;\; 0 \;\; \cdots \;\; 0\,]^H, \tag{3.42}$$

which has the same size as the original pair (E, A). Moreover, we can show that $(\hat{E}, \hat{A})$ is indeed strangeness-free with the same characteristic values as (3.18). This will turn out to be important in the context of the index reduction methods discussed in Section 6.

Theorem 3.32. *Let the strangeness index μ of (E, A) as in (3.17) be well defined with global characteristic values (r_i, a_i, s_i), $i = 1, \ldots, \mu$. Then, every pair $(\hat{E}, \hat{A})$, constructed as in (3.41), has a well-defined strangeness index $\hat{\mu} = 0$. The global characteristic values $(\hat{r}, \hat{a}, \hat{s})$ of $(\hat{E}(t), \hat{A}(t))$ are given by*

$$(\hat{r}, \hat{a}, \hat{s}) = (d_\mu, a_\mu, 0) \tag{3.43}$$

uniformly in $t \in \mathbb{I}$.

Proof. For convenience, in the following we again omit the argument t. By construction, the columns of T_2 form a basis of kernel $\hat{A}_2$. Because $\hat{E}_1$ has full row rank, we can split T_2 without loss of generality into $T_2 = [\, T_1' \; T_3\,]$ in such a way that $\hat{E}_1 T_1'$ is nonsingular. Choosing T_2' such that $\hat{A}_2 T_2'$ is also nonsingular, we obtain a nonsingular matrix $[\, T_1' \; T_2' \; T_3\,]$. To determine the characteristic quantities of $(\hat{E}, \hat{A})$, we multiply with this matrix from the right. In particular, we get the following local

equivalences:

$$
\begin{aligned}
(\hat{E}, \hat{A}) &= \left(\begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} \hat{E}_1 T_1' & \hat{E}_1 T_2' & \hat{E}_1 T_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 T_1' & \hat{A}_1 T_2' & \hat{A}_1 T_3 \\ 0 & \hat{A}_2 T_2' & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} \hat{E}_1 T_1' & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} * & * & * \\ 0 & \hat{A}_2 T_2' & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_{d_\mu} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} * & * & * \\ 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 \end{bmatrix} \right) \\
&\sim \left(\begin{bmatrix} I_{d_\mu} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 \\ 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 \end{bmatrix} \right).
\end{aligned}
$$

From the last pair, we obtain $\hat{r} = d_\mu$, $\hat{a} = a_\mu$, and $\hat{s} = 0$. □

Setting $\hat{f}_1 = Z_1^H f$ and $\hat{f}_2 = Z_2^H g$, we can deduce from the inflated differential-algebraic equation $M(t)\dot{z} = N(t)z + g(t)$ the equations $\hat{E}_1(t)\dot{x} = \hat{A}_1(t)x + \hat{f}_1(t)$ and $0 = \hat{A}_2(t)x + \hat{f}_2(t)$. In contrast to the construction that led to (3.18) the unknown function x is not transformed. Compared with the canonical from (3.18), we are missing an equation $0 = \hat{f}_3(t)$ associated with the consistency of the inhomogeneity. Of course, it would suffice to select some Z_3 with v_μ linearly independent columns satisfying $Z_3^H M = 0$ and $Z_3^H N = 0$ and set $\hat{f}_3 = Z_3^H g$. Because of $\dim \operatorname{corange}(M, N) = v_0 + \cdots + v_\mu$, such a Z_3 certainly exists. But, in general, it is not clear whether there is an invariant way how one can select Z_3. One possibility would be to enlarge the system by letting Z_3 have $v_0 + \cdots + v_\mu$ linearly independent columns (which then span $\operatorname{corange}(M, N)$). In order to check solvability, this would be an appropriate way, but it would be an artificial extension of the system and we would not have a result analogous to Theorem 3.18. If the system is solvable, however, then every choice of Z_3 will give $\hat{f}_3 = 0$. We may therefore simply set $\hat{f}_3 = 0$ in this case and obtain a system

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t) \tag{3.44}$$

similar to that of Theorem 3.17. To set $\hat{f}_3 = 0$ can be seen as a *regularization*, since we replace an unsolvable problem by a solvable one. It should be emphasized that due to the construction of (3.44), every solution of the original differential-algebraic

equation (3.1) is also a solution of (3.44). We will come back to this discussion in Section 4.3 for the case of nonlinear differential-algebraic equations.

Let us study again our two examples from the beginning of the section.

Example 3.33. For the problem of Example 3.1 with $\mu = 1$, we get

$$M(t) = \left[\begin{array}{cc|cc} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ \hline 0 & 2t & -t & t^2 \\ 0 & 2 & -1 & t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

We have rank $M(t) = 2$ independent of $t \in \mathbb{I}$ and we can choose

$$Z_2^H(t) = [\,1 \;\; -t \;\mid\; 0 \;\; 0\,], \quad Z_3^H(t) = [\,0 \;\; 0 \;\mid\; 1 \;\; -t\,].$$

With this choice, we get

$$\hat{E}(t) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{A}(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad \hat{f}(t) = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and we can directly read off the solution as given in Example 3.1.

Example 3.34. For the problem of Example 3.2 with $\mu = 1$, we get

$$M(t) = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 1 & -t & 0 & 0 \\ \hline 1 & -t & 0 & 0 \\ 0 & -1 & 1 & -t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right].$$

We have rank $M(t) = 2$ independent of $t \in \mathbb{I}$ and we can choose

$$Z_2^H(t) = \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ \hline 0 & 1 & -1 & 0 \end{array}\right].$$

With this choice, we get

$$\hat{E}(t) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \hat{A}(t) = \begin{bmatrix} -1 & t \\ 0 & -1 \end{bmatrix}, \quad \hat{f}(t) = \begin{bmatrix} f_1(t) \\ f_2(t) - \dot{f}_1(t) \end{bmatrix}.$$

Again, we can directly read off the solution as given in Example 3.2.

3.3 The differentiation index

Another idea for the construction of a global characteristic quantity for systems of differential-algebraic equations is based on the construction of an ordinary differential equation for the unknown function x from the inflated system (3.28) for some

sufficiently large $\ell \in \mathbb{N}_0$. This approach, which is most common in the literature, has the disadvantage that it is only feasible for square systems with unique solutions, since a differential-algebraic equation with free solution components cannot lead to an ordinary differential equation because of the infinite dimension of the solution space of the homogeneous problem.

In order to compare the analysis that we have presented so far with these other concepts, we must therefore restrict ourselves to the case $m = n$, i.e., throughout this section we assume that $E, A \in C(\mathbb{I}, \mathbb{C}^{n,n})$. Recall that we still assume that the functions that we consider are sufficiently smooth, but that we will not specify the degree of smoothness. Most of the following results are due to Campbell, see, e.g., [48].

The first basic notion that we need is that of 1-fullness of a block matrix.

Definition 3.35. A block matrix $M \in \mathbb{C}^{kn,ln}$ is called 1-*full* (with respect to the block structure built from $n \times n$-matrices) if and only if there exists a nonsingular matrix $R \in \mathbb{C}^{kn,kn}$ such that

$$RM = \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}. \tag{3.45}$$

A corresponding matrix function $M \in C(\mathbb{I}, \mathbb{C}^{kn,ln})$ is called *smoothly* 1-*full* if and only if there is a pointwise nonsingular matrix function $R \in C(\mathbb{I}, \mathbb{C}^{kn,kn})$ such that (3.45) holds as equality of functions.

Lemma 3.36. *Consider a block matrix function $M \in C(\mathbb{I}, \mathbb{C}^{kn,ln})$ as in Definition 3.35 and suppose it has constant rank. Then M is smoothly* 1-*full if and only if it is pointwise* 1-*full.*

Proof. One direction of the claim is trivial. For the other direction, let $M \in C(\mathbb{I}, \mathbb{C}^{kn,ln})$ be pointwise 1-full with constant rank. Writing

$$M = \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix}$$

with $M_{11} \in C(\mathbb{I}, \mathbb{C}^{n,n})$, the first block column must have pointwise full column rank. Therefore, by Theorem 3.9 there exists a smooth, pointwise nonsingular matrix function R_1 such that

$$R_1 M = \begin{bmatrix} I_n & \tilde{M}_{12} \\ 0 & \tilde{M}_{22} \end{bmatrix}.$$

By assumption, $\tilde{M}_{22}$ has constant rank. Hence, again by Theorem 3.9, there exist smooth, pointwise nonsingular matrix functions P and Q with

$$P \tilde{M}_{22} Q = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}.$$

Setting

$$R_2 = \begin{bmatrix} I_n & 0 \\ 0 & P \end{bmatrix},$$

we obtain

$$R_2 R_1 M \begin{bmatrix} I_n & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} I_n & \hat{M}_{12} & \hat{M}_{13} \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Since M is pointwise 1-full, the matrix on the right hand side must also be pointwise 1-full, hence we must have $\hat{M}_{13} = 0$. Finally, setting

$$R_3 = \begin{bmatrix} I_n & -\hat{M}_{12} & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

we end up with

$$R_3 R_2 R_1 M \begin{bmatrix} I_n & 0 \\ 0 & Q \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & \tilde{H} \end{bmatrix}, \quad \tilde{H} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix},$$

or (3.45) with $R = R_3 R_2 R_1$ and $H = \tilde{H} Q^{-1}$. □

Note that the constant rank assumption in Lemma 3.36 cannot be removed, cp. Exercise 20.

Definition 3.37. Let a pair (E, A) be given with inflated pairs (M_ℓ, N_ℓ) constructed as in (3.28). The smallest number $\nu \in \mathbb{N}_0$ (if it exists), for which M_ν is pointwise 1-full and has constant rank, is called the *differentiation index* of (E, A) or (3.1), respectively.

If the differentiation index ν is well defined for (E, A), then there exists a smooth, pointwise nonsingular matrix function $R \in C(\mathbb{I}, \mathbb{C}^{(\nu+1)n,(\nu+1)n})$ with

$$R M_\nu = \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}. \tag{3.46}$$

From

$$M_\nu(t)\dot{z}_\nu = N_\nu(t) z_\nu + g_\nu(t)$$

according to (3.28), we then obtain with

$$\dot{x} = [\, I_n \;\; 0 \,] R(t) M_\nu(t) \dot{z}_\nu = [\, I_n \;\; 0 \,] R(t) N_\nu(t) [\, I_n \;\; 0 \,]^H x + [\, I_n \;\; 0 \,] R(t) g_\nu(t)$$

indeed an ordinary differential equation which is often called *underlying ordinary differential equation*. Observe that it is clear from the construction that solutions of

the original problem (3.1) are also solutions of this underlying ordinary differential equation.

In order to compare the concepts of strangeness and differentiation index, we must first show that the differentiation index is characteristic for a pair of matrix functions.

Theorem 3.38. *The differentiation index is invariant under (global) equivalence transformations.*

Proof. We use the notation of Theorem 3.29. We must show that $\tilde{M}_\nu$ is pointwise 1-full and has constant rank if the same is valid for M_ν. Since

$$\tilde{M}_\nu = \Pi_\nu M_\nu \Theta_\nu$$

with pointwise nonsingular Π_ν and Θ_ν, $\tilde{M}_\nu$ has constant rank if and only if M_ν has constant rank. Let now M_ν be pointwise 1-full according to (3.46). Then we have

$$R\Pi_\nu^{-1}\tilde{M}_\nu\Theta_\nu^{-1} = \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}$$

or

$$R\Pi_\nu^{-1}\tilde{M}_\nu = \begin{bmatrix} I_n & 0 \\ 0 & H \end{bmatrix}\begin{bmatrix} Q & 0 \\ \Theta_{21} & \Theta_{22} \end{bmatrix} = \begin{bmatrix} Q & 0 \\ H\Theta_{21} & H\Theta_{22} \end{bmatrix}$$

with pointwise nonsingular Q. Block row elimination yields

$$\begin{bmatrix} Q^{-1} & 0 \\ -H\Theta_{21}Q^{-1} & I \end{bmatrix} R\Pi_\nu^{-1}\tilde{M}_\nu = \begin{bmatrix} I_n & 0 \\ 0 & H\Theta_{22} \end{bmatrix}$$

and thus $\tilde{M}_\nu$ is pointwise 1-full. □

Up to now, we have shown that the differentiation index is characteristic for a given pair of matrix functions and the related differential-algebraic equation and that, if the differentiation index is well defined, then we can derive an ordinary differential equation from the corresponding inflated differential-algebraic equation in such a way that all solutions of the differential-algebraic equation are also solutions of this ordinary differential equation. We will now show that this statement can be reversed in the sense that for every differential-algebraic equation that has a solution behavior similar to that of an ordinary differential equation the differentiation index is well defined. We begin with the construction of a canonical form similar to (3.23).

Theorem 3.39. *Let (E, A) be a pair of sufficiently smooth matrix functions and suppose that the interval $\mathbb{I}$ is compact. Suppose that (3.1) is solvable for every sufficiently smooth f and that the solution is unique for every $t_0 \in \mathbb{I}$ and every*

consistent initial condition $x_0 \in \mathbb{C}^n$ given at t_0. Suppose furthermore that the solution depends smoothly on f and the initial condition. Then we have

$$(E, A) \sim \left(\begin{bmatrix} I_{\hat{d}} & W \\ 0 & G \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & I_{\hat{a}} \end{bmatrix} \right), \tag{3.47}$$

where $\hat{d}$ is the dimension of the solution space of the corresponding homogeneous differential-algebraic equation, $\hat{a} = n - \hat{d}$, and

$$G(t)\dot{x}_2 = x_2 + f_2(t) \tag{3.48}$$

is uniquely solvable for every sufficiently smooth f_2 without specifying initial conditions.

Proof. We first show that

$$(E, A) \sim \left(\begin{bmatrix} I_{\hat{d}} & E_{12} \\ 0 & E_{22} \end{bmatrix}, \begin{bmatrix} 0 & A_{12} \\ 0 & A_{22} \end{bmatrix} \right), \tag{3.49}$$

where

$$E_{22}(t)\dot{x}_2 = A_{22}(t)x_2 + f_2(t)$$

is uniquely solvable for every sufficiently smooth f_2.

If the homogeneous equation

$$E(t)\dot{x} = A(t)x$$

has only the trivial solution, then the first block is missing (i.e., $\hat{d} = 0$) and the claim holds trivially by assumption. In any case, the solution space is finite dimensional, since otherwise we could not select a unique solution by prescribing initial conditions. Let $\{\phi_1, \ldots, \phi_{\hat{d}}\}$ be a basis of the solution space and $\Phi = [\phi_1 \ \cdots \ \phi_{\hat{d}}]$. Then we have

$$\operatorname{rank} \Phi(t) = \hat{d} \quad \text{for all } t \in \mathbb{I},$$

since, if we had $\operatorname{rank} \Phi(t) < \hat{d}$ for some $t_0 \in \mathbb{I}$, then there would exist coefficients $\alpha_1, \ldots, \alpha_{\hat{d}} \in \mathbb{C}$, not all being zero, with

$$\alpha_1\phi_1(t_0) + \cdots + \alpha_{\hat{d}}\phi_{\hat{d}}(t_0) = 0$$

and $\alpha_1\phi_1 + \cdots + \alpha_{\hat{d}}\phi_{\hat{d}}$ would be a nontrivial solution of the homogeneous initial value problem.

Hence, by Theorem 3.9 there exists a smooth, pointwise nonsingular matrix function U with

$$U^H \Phi = \begin{bmatrix} I_{\hat{d}} \\ 0 \end{bmatrix}.$$

Defining

$$\Phi' = U \begin{bmatrix} 0 \\ I_{\hat{a}} \end{bmatrix}$$

yields a pointwise nonsingular matrix function $Q = [\ \Phi\ \ \Phi'\ ]$. Since $E\dot{\Phi} = A\Phi$, we obtain

$$(E, A) \sim ([\ E\Phi\ \ E\Phi'\ ], [\ A\Phi\ \ A\Phi'\ ] - [\ E\dot{\Phi}\ \ E\dot{\Phi}'\ ]) = ([\ E_1\ \ E_2\ ], [\ 0\ \ A_2\ ]).$$

In this relation, E_1 has full column rank $\hat{d}$. To see this, suppose that rank $E_1(\hat{t}) < \hat{d}$ for some $\hat{t} \in \mathbb{I}$. Then there would exist a vector $w \neq 0$ with

$$E_1(\hat{t})w = 0.$$

Defining in this situation

$$f(t) = \begin{cases} \frac{1}{t-\hat{t}} E_1(t)w & \text{for } t \neq \hat{t}, \\ \frac{d}{dt}(E_1(t)w) & \text{for } t = \hat{t}, \end{cases}$$

we would obtain a smooth inhomogeneity f. The function x given by

$$x(t) = \begin{bmatrix} \log(|t - \hat{t}|)w \\ 0 \end{bmatrix}$$

would then solve

$$[\ E_1(t)\ \ E_2(t)\ ]\dot{x} = [\ 0\ \ A_2(t)\ ]x + f(t)$$

on $\mathbb{I} \setminus \{\hat{t}\}$ in contradiction to the assumption of unique solvability, which includes by definition that solutions are defined on the entire interval $\mathbb{I}$.

Hence, since E_1 has full column rank, there exists a smooth, pointwise nonsingular matrix function P, with

$$PE_1 = \begin{bmatrix} I_{\hat{d}} \\ 0 \end{bmatrix},$$

and thus

$$(E, A) \sim \left(\begin{bmatrix} I_{\hat{d}} & E_{12} \\ 0 & E_{22} \end{bmatrix}, \begin{bmatrix} 0 & A_{12} \\ 0 & A_{22} \end{bmatrix} \right).$$

The equation

$$E_{22}(t)\dot{x}_2 = A_{22}(t)x_2$$

only admits the trivial solution. To see this, suppose that $x_2 \neq 0$ is a nontrivial solution and x_1 a solution of the ordinary differential equation

$$\dot{x}_1 + E_{12}(t)\dot{x}_2(t) = A_{22}(t)x_2(t).$$

Then we obtain

$$[\, E_1(t) \;\; E_2(t) \,] \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = [\, 0 \;\; A_2(t) \,] \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}.$$

Transforming back gives

$$E(t)Q(t) \begin{bmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \end{bmatrix} = (A(t)Q(t) - E(t)\dot{Q}(t)) \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$$

or $E(t)\dot{x}(t) = A(t)x(t)$ with

$$x = Q \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \neq 0, \quad x \notin \operatorname{span}\{\phi_1, \ldots, \phi_{\hat{d}}\},$$

since $x_2 \neq 0$. But this contradicts the construction of $\phi_1, \ldots, \phi_{\hat{d}}$. Thus, we have shown (3.49).

For smooth, pointwise nonsingular S, we can then further transform (3.49) according to

$$\begin{aligned} (E, A) &\sim \left(\begin{bmatrix} I_{\hat{d}} & E_{12} \\ 0 & E_{22} \end{bmatrix}, \begin{bmatrix} 0 & A_{12} \\ 0 & A_{22} \end{bmatrix} \right) \\ &\sim \left(\begin{bmatrix} I_{\hat{d}} & E_{12}S \\ 0 & E_{22}S \end{bmatrix}, \begin{bmatrix} 0 & A_{12}S \\ 0 & A_{22}S \end{bmatrix} - \begin{bmatrix} 0 & E_{12}\dot{S} \\ 0 & E_{22}\dot{S} \end{bmatrix} \right), \end{aligned}$$

and, if also $A_{22}S - E_{22}\dot{S}$ were pointwise nonsingular, then

$$(E, A) \overset{\text{new}}{\sim} \left(\begin{bmatrix} I_{\hat{d}} & E_{12} \\ 0 & E_{22} \end{bmatrix}, \begin{bmatrix} 0 & A_{12} \\ 0 & I_{\hat{a}} \end{bmatrix} \right) \sim \left(\begin{bmatrix} I_{\hat{d}} & W \\ 0 & G \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & I_{\hat{a}} \end{bmatrix} \right).$$

It therefore remains to show that there exists a smooth, pointwise nonsingular S, for which $A_{22}S - E_{22}\dot{S}$ is pointwise nonsingular as well.

Let $\tilde{S}$ be the unique (smooth) solution of

$$A_{22}(t)\tilde{S} - E_{22}(t)\dot{\tilde{S}} = I_{\hat{a}}.$$

By the Weierstraß approximation theorem (see, e.g., [138]), there exists an elementwise polynomial matrix function $\hat{S}$, such that

$$\Delta = \tilde{S} - \hat{S}$$

and its derivative are elementwise arbitrary small in the L_∞-norm and

$$A_{22}\hat{S} - E_{22}\dot{\hat{S}} = A_{22}\tilde{S} - A_{22}\Delta - E_{22}\dot{\tilde{S}} + E_{22}\dot{\Delta} = I_{\hat{a}} - A_{22}\Delta + E_{22}\dot{\Delta} = K$$

remains pointwise nonsingular (recall that $\mathbb{I}$ is assumed to be compact). Let $\lambda_1(t), \ldots, \lambda_{\hat{a}}(t) \in \mathbb{C}$ be the eigenvalues of $\hat{S}(t)$. Since $\hat{S}$ is (real) analytic, the set

$$\Lambda = \{\lambda_i(t) \mid t \in \mathbb{I},\ i = 1, \ldots, \hat{a}\} \subseteq \mathbb{C}$$

has no interior as a subset of $\mathbb{C}$ (see, e.g., [119, Ch. II]). Hence, there exists a $c \in \mathbb{C}$, arbitrarily small in modulus, such that $S = \hat{S} - cI_{\hat{a}}$ as well as

$$A_{22}S - E_{22}\dot{S} = A_{22}\hat{S} - cA_{22} - E_{22}\dot{\hat{S}} = K - cA_{22}$$

are pointwise nonsingular. □

Remark 3.40. Note that the argument in the final step of the proof of Theorem 3.39 relies on the fact that we are working in the complex field. If the pair of matrix functions (E, A) is real, then we can nevertheless treat it as a complex problem and obtain an equivalent complex pair.

Following Corollary 3.26, we can decompose $\mathbb{I}$ according to (3.26) in such a way that we can transform (E, A) on every interval $\mathbb{I}_j$ to the (global) canonical form (3.23). Comparing (3.23) with (3.47), for the characteristic values on $\mathbb{I}_j$ we obviously have the relations $d_\mu = \hat{d}$, $u_\mu = 0$, $v_\mu = 0$, and $a_\mu = \hat{a}$.

Lemma 3.41. *Under the assumptions of Theorem 3.39, we have*

$$\operatorname{rank}[\, M_\ell \ \ N_\ell \,] = (\ell + 1)n \tag{3.50}$$

for arbitrary $\ell \in \mathbb{N}_0$, i.e., M_ℓ and N_ℓ together have pointwise full row rank.

Proof. Suppose that there exists $\hat{t} \in \mathbb{I}$ with

$$\operatorname{rank}[\, M_\ell(\hat{t}) \ \ N_\ell(\hat{t}) \,] < (\ell + 1)n.$$

Then there exists a vector $w \neq 0$ satisfying

$$w^H M_\ell(\hat{t}) = 0, \quad w^H N_\ell(\hat{t}) = 0.$$

In view of (3.28), for a solution of (3.1) to exist, we then have the consistency condition

$$w^H g_\ell(\hat{t}) = 0,$$

which is a contradiction to the assumptions. □

Lemma 3.42. *Consider a pair of matrix functions (E, A) that is sufficiently smooth and has a well-defined strangeness index μ as in (3.17). Furthermore, let*

$$(E, A) \sim \left(\begin{bmatrix} I_{d_\mu} & W \\ 0 & G \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & I_{a_\mu} \end{bmatrix} \right) \tag{3.51}$$

according to (3.23)*, i.e., let* $u_\mu = v_\mu = 0$*. Then, the associated inflated pair* (M_ℓ, N_ℓ) *satisfies*

$$\operatorname{corank} M_\ell = a_\mu \text{ for } \ell \geq \mu. \tag{3.52}$$

Proof. We may assume without loss of generality that (E, A) is in the canonical form (3.51). Because of the special structure of (E, A) with the identity I_{d_μ} as sole entry in the first block column of E, the rank defect in M_ℓ can only be caused by the second block row. Therefore, we may even assume that $(E, A) = (G, I)$, with G as in (3.24).

We then consider the infinite matrix function

$$M = \begin{bmatrix} G & & & \\ \dot{G} - I & G & & \\ \ddot{G} & 2\dot{G} - I & G & \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix},$$

built according to (3.29). To determine its corange, we look for a matrix function Z of maximal rank with $Z^H M = 0$, i. e.,

$$[\, Z_0^H \;\; Z_1^H \;\; Z_2^H \;\; \cdots \,] \left\{ \begin{bmatrix} G & & & \\ \dot{G} & G & & \\ \ddot{G} & 2\dot{G} & G & \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} - \begin{bmatrix} 0 & & & \\ I & 0 & & \\ & I & 0 & \\ & & \ddots & \ddots \end{bmatrix} \right\} = 0,$$

where $Z^H = [\, Z_0^H \;\; Z_1^H \;\; Z_2^H \;\; \cdots \,]$. This is equivalent to

$$[Z_1^H Z_2^H \cdots] = Z_0^H [G\, 0 \cdots] \left\{ \begin{bmatrix} I & & & \\ & I & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} - \begin{bmatrix} \dot{G} & G & & \\ \ddot{G} & 2\dot{G} & G & \\ \vdots & \ddots & \ddots & \ddots \\ \vdots & & \ddots & \ddots \end{bmatrix} \right\}^{-1}. \tag{3.53}$$

Since G and all its derivatives are strictly upper triangular, all $(\mu+1)$-fold products of G and its derivatives vanish. Hence, all terms in the above relations are actually finite and we in fact work only formally with a matrix function of infinite size. This also shows that the inverse in (3.53) exists, since it is the inverse of the sum of the identity and a nilpotent matrix function. Via the same argument, it follows by induction that Z_j is a sum of at least j-fold products, i.e., that

$$Z_j = 0 \text{ for } j \geq \mu + 1.$$

Since Z is parameterized with respect to Z_0, the choice $Z_0 = I$ yields maximal rank for Z showing that the corank of M equals the size of its blocks. But this is just (3.52) for M_ℓ. □

Lemma 3.43. *Under the assumptions of Lemma 3.42, we have that*

$$M_\ell \text{ is (pointwise) 1-full for } \ell \geq \mu + 1. \tag{3.54}$$

Proof. We must show the existence of a nonsingular matrix R satisfying (3.45) at a given fixed point $t \in \mathbb{I}$ (for convenience again we omit the argument t). The essential property is to fulfill the relation belonging to the first block row of the right hand side in (3.45). The remaining part then follows by completion of the first block row of R to a nonsingular matrix followed by a block row elimination to obtain the zeros in the first block column.

Again we may assume that (E, A) is in the canonical form (3.51). Because of the entry I_{d_μ}, this part is already in the required form. Thus, we may even assume that $(E, A) = (G, I)$ with G as in (3.24). As in the proof of Lemma 3.42, we use the infinite matrix function M. The first block row of (3.45) then reads

$$[\, R_{0,0}\, R_{0,1}\, R_{0,2}\, \cdots] \left\{ \begin{bmatrix} G & & & \\ \dot{G} & G & & \\ \ddot{G} & 2\dot{G} & G & \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix} - \begin{bmatrix} 0 & & & \\ I & 0 & & \\ & I & 0 & \\ & & \ddots & \ddots \end{bmatrix} \right\} = [\, I\, 0\, 0\, \cdots].$$

This is equivalent to

$$[\, R_{0,1}\ R_{0,2}\ \cdots] = [\, R_{0,0}G - I\ \ 0\ \cdots] \left\{ \begin{bmatrix} I & & & \\ & I & & \\ & & \ddots & \\ & & & \ddots \end{bmatrix} - \begin{bmatrix} \dot{G} & G & & \\ \ddot{G} & 2\dot{G} & G & \\ \vdots & \ddots & \ddots & \ddots \\ \vdots & & \ddots & \ddots \end{bmatrix} \right\}^{-1}.$$

Setting for example $R_{0,0} = I$, we obtain a possible first block row for the desired R. The claim follows, since $R_{0,j} = 0$ for $j \geq \mu + 2$. □

Lemma 3.44. *Consider a matrix function $M \in C(\mathbb{I}, \mathbb{C}^{kn,ln})$ with constant rank r and smooth kernel $T \in C(\mathbb{I}, \mathbb{C}^{ln,ln-r})$ according to Theorem 3.9. Then, M is (pointwise) 1-full if and only if*

$$[\, I_n\ \ 0\,]T = 0. \tag{3.55}$$

Proof. Let M be pointwise 1-full. Then, by Lemma 3.36, M is smoothly 1-full with (3.45) and T must have the form

$$T = \begin{bmatrix} 0 \\ \tilde{T} \end{bmatrix}.$$

Hence, (3.55) holds.

For the converse, let (3.55) hold. Writing $M = [\, M_1 \;\; M_2 \,]$ according to the block structure in (3.55), it follows that M_1 has full column rank, since otherwise there exists a $w \neq 0$ with $M_1 w = 0$ or

$$M \begin{bmatrix} w \\ 0 \end{bmatrix} = 0.$$

Hence, there exists a smooth, pointwise nonsingular matrix function R_1 with

$$R_1 M = \begin{bmatrix} I_n & M_{12} \\ 0 & M_{22} \end{bmatrix}.$$

For every $w \neq 0$ with $M_{22} w = 0$, we must have $M_{12} w = 0$, since

$$M \begin{bmatrix} -M_{12} w \\ w \end{bmatrix} = 0.$$

Thus,

$$\text{kernel } M_{22} \subseteq \text{kernel } M_{12},$$

implying that

$$M_{12}(I - M_{22}^{+} M_{22}) = 0,$$

where the superscript $+$ denotes the (pointwise) Moore–Penrose pseudoinverse (see also Section 3.4), such that $I - M_{22}^{+} M_{22}$ is the smooth orthogonal projection onto the kernel of M_{22}. It follows that $M_{12} = \tilde{R} M_{22}$ with $\tilde{R} = M_{12} M_{22}^{+}$. Forming the smooth, pointwise nonsingular matrix function

$$R_2 = \begin{bmatrix} I_n & -\tilde{R} \\ 0 & I \end{bmatrix},$$

we finally obtain

$$R_2 R_1 M = \begin{bmatrix} I_n & -\tilde{R} \\ 0 & I \end{bmatrix} \begin{bmatrix} I_n & M_{12} \\ 0 & M_{22} \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ 0 & M_{22} \end{bmatrix}.$$

□

Theorem 3.45. *Let the assumptions of Theorem 3.39 hold. Then the differentiation index ν of (E, A) is well defined.*

Proof. We prove that there exists $\ell \in \mathbb{N}_0$ for which M_ℓ is (pointwise) 1-full and has constant rank. Let $\mathbb{I}$ be decomposed according to Corollary 3.26 and let μ_j be the strangeness index on $\mathbb{I}_j$. By Theorem 3.39, the characteristic values on $\mathbb{I}_j$ must satisfy

$$d_{\mu_j} = \hat{d}, \quad a_{\mu_j} = \hat{a}.$$

Lemma 3.42 then yields

$$\operatorname{corank} M_{\mu_j} = \hat{a} \quad \text{on } \mathbb{I}_j.$$

Using (3.52) on every interval $\mathbb{I}_j$, we obtain

$$\operatorname{corank} M_{\hat{\mu}+1} = \hat{a} \quad \text{on } \bigcup_{j\in\mathbb{N}} \mathbb{I}_j$$

for

$$\hat{\mu} = \max_{j\in\mathbb{N}} \mu_j \le n - 1.$$

Since the rank function is lower semi-continuous (for a given matrix there is a neighborhood such that every matrix in this neighborhood has the same or a higher rank), it follows that

$$\operatorname{corank} M_{\hat{\mu}+1} \ge \hat{a} \quad \text{on } \mathbb{I}.$$

Assuming without loss of generality that (E, A) is in the canonical form (3.47), we obviously have

$$\operatorname{rank} N_{\hat{\mu}+1} = \hat{a} \quad \text{on } \mathbb{I}.$$

Lemma 3.41 then implies that

$$\operatorname{corank} M_{\hat{\mu}+1} \le \hat{a} \quad \text{on } \mathbb{I}.$$

Thus, we have that

$$\operatorname{corank} M_{\hat{\mu}+1} = \hat{a} \quad \text{on } \mathbb{I}.$$

In particular, $M_{\hat{\mu}+1}$ has constant rank on $\mathbb{I}$. Therefore, there exists a continuous matrix function T whose columns pointwise span kernel $M_{\hat{\mu}+1}$. Since $M_{\hat{\mu}+1}$ is (pointwise) 1-full, we get

$$[\, I_n \;\; 0 \,]T = 0 \quad \text{on } \bigcup_{j\in\mathbb{N}} \mathbb{I}_j$$

by applying Lemma 3.44. Since T is continuous and $[\, I_n \;\; 0 \,]T$ vanishes on a dense subset of $\mathbb{I}$, it follows that

$$[\, I_n \;\; 0 \,]T = 0 \quad \text{on } \mathbb{I},$$

and hence

$$M_{\hat{\mu}+1} \text{ is (pointwise) 1-full on } \mathbb{I}.$$

It follows that the differentiation index ν is well defined with $\nu \le \hat{\mu} + 1$. □

Corollary 3.46. *Under the assumptions of Theorem 3.39, the relation*

$$\nu = \begin{cases} 0 & \text{for } \hat{a} = 0, \\ \hat{\mu} + 1 & \text{for } \hat{a} \neq 0 \end{cases} \tag{3.56}$$

holds, where

$$\hat{\mu} = \max_{j \in \mathbb{N}} \mu_j \leq n - 1 \tag{3.57}$$

and μ_j is the strangeness index on $\mathbb{I}_j$ as defined in Corollary 3.26.

Proof. If $\hat{a} = 0$, then the corresponding differential-algebraic equation is equivalent to an ordinary differential equation. In this case, $\nu = 0$ and $\hat{\mu} = 0$ holds. If $\hat{a} \neq 0$, then by the definition of $\hat{\mu}$, there exists a $j \in \mathbb{N}$ such that $\hat{\mu}$ is the strangeness index of (E, A) restricted to $\mathbb{I}_j$. For $R_{0,\hat{\mu}+1}$ as introduced in the proof of Lemma 3.43, induction then shows that

$$R_{0,\hat{\mu}+1} = -G^{\hat{\mu}}.$$

But by (3.24) and (3.25), we have that $R_{0,\hat{\mu}+1} \neq 0$ pointwise, see also Exercise 17. Hence, $M_{\hat{\mu}}$ cannot be 1-full on $\mathbb{I}_j$. □

Corollary 3.47. *Let (E, A) be sufficiently smooth with well-defined strangeness index μ as in (3.17) and suppose that $u_\mu = v_\mu = 0$. Then the differentiation index of (E, A) is well defined with*

$$\nu = \begin{cases} 0 & \text{for } a_\mu = 0, \\ \mu + 1 & \text{for } a_\mu \neq 0. \end{cases} \tag{3.58}$$

Proof. Theorem 3.17 shows that (E, A) satisfies the assumptions of Theorem 3.39 and therefore those of Theorem 3.45. Hence, the differentiation index is well defined. The assertion then follows from (3.56), since $\hat{\mu} = \mu$ and $\hat{a} = a_\mu$. □

As in the proof of Theorem 3.45, one can show that already $M_{\hat{\mu}}$ has constant corank $\hat{a}$. Looking into the reduction procedure of Section 3.2 leading to (3.44), we have used the constant corank of M_μ to obtain the correct algebraic equations. In the same way, we can determine here $\hat{a}$ algebraic equations. Comparing with (3.13) they are related to the block given by (G, I) with G as in (3.24). The other block then represents the correct differential equation which, together with the algebraic equations, then constitute a differential-algebraic equation that should be strangeness-free. To investigate this connection in more detail, we first formulate a hypothesis that simply guarantees that the reduction procedure of Section 3.2 can be performed and no consistency condition for the inhomogeneity or free solution components are present.

Hypothesis 3.48. *There exist integers $\hat{\mu}$, $\hat{a}$, and $\hat{d}$ such that the inflated pair $(M_{\hat{\mu}}, N_{\hat{\mu}})$ associated with the given pair of matrix functions (E, A) has the following properties:*

1. *For all $t \in \mathbb{I}$ we have* rank $M_{\hat{\mu}}(t) = (\hat{\mu}+1)n - \hat{a}$ *such that there exists a smooth matrix function Z_2 of size $(\hat{\mu}+1)n \times \hat{a}$ and pointwise maximal rank satisfying $Z_2^H M_{\hat{\mu}} = 0$.*
2. *For all $t \in \mathbb{I}$ we have* rank $\hat{A}_2(t) = \hat{a}$, *where $\hat{A}_2 = Z_2^H N_{\hat{\mu}}[I_n\ 0\ \cdots\ 0]^H$ such that there exists a smooth matrix function T_2 of size $n \times \hat{d}$, $\hat{d} = n - \hat{a}$, and pointwise maximal rank satisfying $\hat{A}_2 T_2 = 0$.*
3. *For all $t \in \mathbb{I}$ we have* rank $E(t)T_2(t) = \hat{d}$ *such that there exists a smooth matrix function Z_1 of size $n \times \hat{d}$ and pointwise maximal rank satisfying* rank $\hat{E}_1 T_2 = \hat{d}$ *with $\hat{E}_1 = Z_1^H E$.*

Remark 3.49. Since Gram–Schmidt orthonormalization is a continuous process, we may assume without loss of generality that the columns of the matrix functions Z_1, Z_2, and T_2 are pointwise orthonormal. This will turn out to be important for the numerical methods discussed in Section 6.

Recall that we have already shown by the discussion in Section 3.2 that Hypothesis 3.48 is invariant under global equivalence transformations and that it holds for systems with well-defined strangeness index μ and $u_\mu = 0$, $v_\mu = 0$ by setting $\hat{\mu} = \mu$, $\hat{a} = a_\mu$, and $\hat{d} = d_\mu$.

Our next aim is to prove that (up to some technical assumptions) Hypothesis 3.48 is equivalent to the requirement that the differentiation index is well defined. As (3.56) and (3.58) suggest, the main difference between both concepts will be that, in general, we will need one differentiation less when dealing with Hypothesis 3.48. We start with the simpler direction of the claimed equivalence.

Theorem 3.50. *Consider a pair (E, A) of sufficiently smooth matrix functions with a well-defined differentiation index ν. Then, (E, A) satisfies Hypothesis 3.48 with*

$$\hat{\mu} = \max\{0, \nu - 1\}, \quad \hat{d} = n - \hat{a}, \quad \hat{a} = \begin{cases} 0 & \text{for } \nu = 0, \\ \operatorname{corank} M_{\nu-1}(t) & \text{otherwise.} \end{cases} \tag{3.59}$$

Proof. The claim is trivial for $\nu = 0$. We therefore assume $\nu \geq 1$. By definition, $\hat{a} = \operatorname{corank} M_\nu$ is constant on $\mathbb{I}$. Lemma 3.42 then yields

$$\operatorname{corank} M_{\nu-1} = \hat{a} \quad \text{on } \bigcup\nolimits_{j\in\mathbb{N}} \mathbb{I}_j.$$

Since $\operatorname{corank} M_{\nu-1} \leq \operatorname{corank} M_\nu$ by construction, we get $\operatorname{corank} M_{\nu-1} \leq \hat{a}$ and equality holds on a dense subset of $\mathbb{I}$. Because the rank function is lower semi-continuous, equality then holds on the whole interval $\mathbb{I}$. Together with (3.50), this

implies that we can choose Z_2 and T_2 according to the requirements of Hypothesis 3.48. If (E, A) is in the normal form (3.51), then we obtain $T_2 = [\ I \quad 0\ ]^H$ and rank $ET_2 = n - \hat{a}$ on a dense subset of $\mathbb{I}$ and therefore on the whole interval $\mathbb{I}$. The claim follows, since all relevant quantities are invariant under global equivalence. □

It is more complicated to show that Hypothesis 3.48 implies that the differentiation index is well defined. In principle, we must prove that the assumptions of Theorem 3.39 are satisfied. For this, we need an existence and uniqueness theorem on the basis of Hypothesis 3.48. In contrast to the previous sections, where we knew the structure of the solution space by assuming that the strangeness index is well defined, we here only know that the reduction to

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t), \tag{3.60}$$

where

$$(\hat{E}, \hat{A}) = \left(\begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} \right), \tag{3.61}$$

with entries

$$\hat{E}_1 = Z_1^H E, \quad \hat{A}_1 = Z_1^H A, \quad \hat{A}_2 = Z_2^H N_{\hat{\mu}}[\ I_n\ 0\ \cdots\ 0\ ]^H, \tag{3.62}$$

preserves solutions. We do not yet know whether all solutions of (3.60) also solve the original problem (3.1).

Theorem 3.51. *Let (E, A) be a pair of sufficiently smooth matrix functions that satisfies Hypothesis 3.48. Then x solves (3.1) if and only if it solves (3.60).*

Proof. As already mentioned, if x solves (3.1) it is immediately clear by construction that it also solves (3.60). For the other direction, let x be a solution of (3.60). According to Corollary 3.26, we restrict the problem to an interval $\mathbb{I}_j$ and transform (E, A) on this interval to the canonical form $(\tilde{E}, \tilde{A})$ given in (3.23). Due to Hypothesis 3.48 and Theorem 3.32, the quantities $\hat{d}$ and $\hat{a}$ must coincide with the corresponding blocksizes of (3.23). In particular, the second block row and block column in (3.23) is missing. Let the derivative arrays belonging to (E, A) and $(\tilde{E}, \tilde{A})$ be denoted by (M_ℓ, N_ℓ) and $(\tilde{M}_\ell, \tilde{N}_\ell)$, respectively. In the notation of Hypothesis 3.48 and Theorem 3.29, we have

$$\begin{aligned}
\tilde{Z}_1^H \tilde{E} &= Z_1^H P^{-1} P E Q = Z_1 E Q, \\
\tilde{Z}_1^H \tilde{A} &= Z_1^H P^{-1}(P A Q - P E \dot{Q}) = Z_1 A Q - Z_1 E \dot{Q}, \\
\tilde{Z}_2^H \tilde{N}_\mu[\ I_n\ 0\ \cdots\ 0\ ]^H &= Z_2^H \Pi_\mu^{-1}(\Pi_\mu N_\mu \Theta_\mu - \Pi_\mu M_\mu \Psi_\mu)[\ I_n\ 0\ \cdots\ 0\ ]^H \\
&= Z_2^H N_\mu \Theta_\mu [\ I_n\ 0\ \cdots\ 0\ ]^H = Z_2^H N_\mu [\ Q\ *\ \cdots\ *\ ]^H \\
&= Z_2^H N_\mu [\ I_n\ 0\ \cdots\ 0\ ]^H Q.
\end{aligned}$$

This shows that the reduced problem transforms covariantly with Q. Thus, it is sufficient to consider the problem in the canonical form (3.23). Hence, we may assume that

$$E = \begin{bmatrix} I_{\hat{d}} & W \\ 0 & G \end{bmatrix}, \quad A = \begin{bmatrix} 0 & 0 \\ 0 & I_{\hat{a}} \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix},$$

where G has the nilpotent structure of (3.24). Using again formally infinite matrix functions, by (3.53) we may choose

$$Z_2^H = [\, 0 \;\; I_{\hat{a}} \,|\, 0 \;\; Z_{2,1}^H \,|\, 0 \;\; Z_{2,2}^H \,|\, \cdots \,]$$

with

$$[\, Z_{2,1}^H \;\; Z_{2,2}^H \;\; \cdots \,] = [\, G \;\; 0 \;\; \cdots \,](I - X)^{-1},$$

where

$$X = \begin{bmatrix} \dot{G} & G & & \\ \ddot{G} & 2\dot{G} & G & \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}.$$

Accordingly, we may choose

$$Z_1^H = [\, I_d \;\; 0 \;\; 0 \,].$$

The corresponding reduced problem thus reads

$$\begin{bmatrix} I_{\hat{d}} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & I_{\hat{a}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix},$$

with

$$\begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 + GV^H(I - X)^{-1}g \end{bmatrix}, \quad V = \begin{bmatrix} I_{\hat{a}} \\ 0 \\ \vdots \end{bmatrix}, \quad g = \begin{bmatrix} \dot{f}_2 \\ \ddot{f}_2 \\ \vdots \end{bmatrix}.$$

The reduced problem immediately yields $x_2 = -\hat{f}_2$. With the block up-shift matrix

$$S = \begin{bmatrix} 0 & & & \\ I & 0 & & \\ & I & 0 & \\ & & \ddots & \ddots \end{bmatrix},$$

we find that $\dot{g} = S^H g$ and $S^H X = \dot{X} + X S^H$. Subtracting S^H on both sides of the latter relation yields

$$(I - X)^{-1} S^H = S^H (I - X)^{-1} - (I - X)^{-1} \dot{X} (I - X)^{-1}.$$

We then get that

$$\begin{aligned} V^H(I-X)^{-1} &= V^H \sum_{i\geq 0} X^i = V^H + V^H X \sum_{i\geq 0} X^i \\ &= V^H + (\dot{G}V^H + GV^H S^H)(I-X)^{-1}, \end{aligned}$$

and thus

$$\begin{aligned} \hat{f}_2 - G\dot{\hat{f}}_2 &= f_2 + GV^H(I-X)^{-1}g - G\dot{f}_2 - G\dot{G}V^H(I-X)^{-1}g \\ &\quad - G^2V^H(I-X)^{-1}\dot{X}(I-X)^{-1}g - G^2V^H(I-X)^{-1}\dot{g} \\ &= f_2 + GV^H g + G\dot{G}V^H(I-X)^{-1}g + G^2V^H S^H(I-X)^{-1}g \\ &\quad - G\dot{f}_2 - G\dot{G}V^H(I-X)^{-1}g - G^2V^H S^H(I-X)^{-1}g \\ &\quad + G^2V^H(I-X)^{-1}S^H g - G^2V^H(I-X)^{-1}S^H g \\ &= f_2 + GV^H g - G\dot{f}_2 = f_2. \end{aligned}$$

Hence, we have that

$$G\dot{x}_2 = -G\dot{\hat{f}}_2 = f_2 - \hat{f}_2 = x_2 + f_2.$$

This shows that the transformed x solves the transformed differential-algebraic equation (3.1) on $\mathbb{I}_j$. Thus, x solves (3.1) on $\mathbb{I}_j$ for every $j \in \mathbb{N}$ and therefore on a dense subset of $\mathbb{I}$. Since all functions are continuous, the given x solves (3.1) on the entire interval $\mathbb{I}$. □

As a consequence of these results, we can characterize consistency of initial values and existence and uniqueness of solutions.

Theorem 3.52. *Let (E, A) satisfy Hypothesis 3.48 with values $\hat{\mu}$, $\hat{d}$ and $\hat{a}$. In particular, suppose that $E, A \in C^{\hat{\mu}+1}(\mathbb{I}, \mathbb{C}^{n,n})$ and $f \in C^{\hat{\mu}+1}(\mathbb{I}, \mathbb{C}^n)$.*

1. *An initial condition (3.2) is consistent if and only if (3.2) implies the $\hat{a}$ conditions*

$$\hat{A}_2(t_0)x_0 + \hat{f}_2(t_0) = 0. \tag{3.63}$$

2. *Every initial value problem with consistent initial condition has a unique solution.*

Proof. The claims are an immediate consequence of Theorem 3.52 and Theorem 3.17 applied to the reduced problem (3.60). □

In particular, Theorem 3.52 implies that (3.1) and (3.60) have the same solutions under Hypothesis 3.48. But (3.60) has strangeness index $\hat{\mu} = 0$ and characteristic values

$$\hat{u} = n - \hat{r} - \hat{a} = 0, \quad \hat{v} = n - \hat{r} - \hat{a} - \hat{s} = 0, \tag{3.64}$$

cp. Theorem 3.32 and Theorem 3.7. Corollary 3.47 then yields that the differentiation index is well defined for (3.60). Thus, (3.60) and therefore also (3.1) satisfy the assumptions of Theorem 3.39. This argument proves the following corollary.

Corollary 3.53. *Let (E, A) be a pair of sufficiently smooth matrix functions that satisfies Hypothesis 3.48 with characteristic quantities $\hat{\mu}$, $\hat{d}$, and $\hat{a}$. Then the differentiation index ν is well defined for (E, A). If $\hat{\mu}$ is chosen minimally, then (3.56) holds.*

Together with Theorem 3.50, we have shown that the requirements of Hypothesis 3.48 and that of a well-defined differentiation index are equivalent up to some (technical) smoothness requirements.

Example 3.54. Consider the differential-algebraic equation

$$\begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad \mathbb{I} = [-1, 1],$$

see, e.g., [179]. Obviously, E has a rank drop at $t = 0$ such that the strangeness index μ as in (3.17) is not well defined. Nevertheless, the system has the unique solution

$$x_1(t) = -(f_1(t) + t\dot{f}_2(t)), \quad x_2(t) = -f_2(t)$$

in the entire interval $\mathbb{I} = [-1, 1]$ and $t = 0$ seems to be not an exceptional point. Rewriting the system by means of the product rule as

$$\tfrac{d}{dt}(tx_2) - x_2 = x_1 + f_1(t), \quad 0 = x_2 + f_2(t)$$

yields the (unique) solution

$$x_1(t) = f_2(t) - (f_1(t) + \tfrac{d}{dt}(tf_2(t))), \quad x_2(t) = -f_2(t),$$

which makes sense if tf_2 is continuously differentiable, e.g., for $f_2(t) = |t|$.

This shows that $t = 0$ still is an exceptional point which is reflected by changes in the characteristic values according to

$$\begin{aligned} r_0 &= 1, \quad a_0 = 0, \quad s_0 = 1, \\ r_1 &= 0, \quad a_1 = 2, \quad s_1 = 0 \end{aligned}$$

for $t \neq 0$ and

$$r_0 = 0, \quad a_0 = 2, \quad s_0 = 0$$

for $t = 0$. This behavior corresponds to the splitting of the interval $\mathbb{I}$ into

$$[-1, 1] = \{-1\} \cup (-1, 0) \cup \{0\} \cup (0, 1) \cup \{1\}$$

as in Corollary 3.26. Examining the derivative array

$$(M_1(t), N_1(t)) = \left(\left[\begin{array}{cc|cc} 0 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline -1 & 1 & 0 & t \\ 0 & -1 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] \right),$$

we first find that M_1 has a constant corank $\hat{a} = 2$. Choosing

$$Z_2^H(t) = \left[\begin{array}{cc|cc} 1 & 0 & 0 & t \\ 0 & 1 & 0 & 0 \end{array}\right],$$

we then get

$$\hat{A}_2(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \hat{f}_2(t) = \begin{bmatrix} f_1(t) + t\dot{f}_2(t) \\ f_2(t) \end{bmatrix}.$$

Thus, (E, A) satisfies Hypothesis 3.48 with $\hat{\mu} = 1$, $\hat{a} = 2$, and $\hat{d} = 0$. This also shows that the differentiation index is well defined with $\nu = 2$. Observe also that the last block column of Z_2^H vanishes for $t = 0$.

Remark 3.55. In this section, we have shown that both the concept of the differentiation index and the concept of Hypothesis 3.48 are equivalent approaches for the investigation of regular linear differential-algebraic equations, at least up to some technical differences in smoothness assumptions. The principle difference is that the differentiation index aims at a reformulation of the given problem as an ordinary differential equation, whereas Hypothesis 3.48 aims at a reformulation as a differential-algebraic equation with the same solutions but better analytical properties. These properties essentially are that we have a separated part which states all constraints of the problem and a complement part which would yield an ordinary differential equation for a part of the unknown if we locally solved the constraints for the other part of the unknown. We refer to these parts as *differential part* and *algebraic part*, respectively. In this respect, Hypothesis 3.48 only contains the requirement that such a reformulation is possible. In particular, Part 1 of Hypothesis 3.48 says that we expect $\hat{a}$ constraints and that Z_2 points to these constraints in the derivative array. Part 2 then requires that these constraints are linearly independent. This also excludes otherwise possible consistency conditions on the inhomogeneity. The quantity T_2 then points to the differential part of the unknown. Finally, Part 3 requires that there is a sufficiently large part in the original problem that yields the ordinary differential equation for the differential part of the unknown if we would eliminate the algebraic part of the unknown. These equations are selected by Z_1.

3.4 Differential-algebraic operators and generalized inverses

If for a pair (E, A) of a given differential-algebraic equation, the characteristic values u and v do not both vanish, then we are not only faced with the problem of inconsistent initial values but also with inconsistent inhomogeneities and nonunique solvability. In the case of linear equations

$$Ax = b \tag{3.65}$$

with $A \in \mathbb{C}^{m,n}$ and $b \in \mathbb{C}^m$, which can be seen as a special case of (3.1), this problem is overcome by embedding (3.65) into the minimization problem

$$\frac{1}{2}\|x\|_2^2 = \min! \quad \text{s. t.} \quad \frac{1}{2}\|Ax - b\|_2^2 = \min! \tag{3.66}$$

which has a unique solution in any case. This unique solution, also called *least squares solution*, can be written in the form

$$x = A^+ b \tag{3.67}$$

with the help of the *Moore–Penrose pseudoinverse* A^+ of A.

A more abstract interpretation of (3.67) is that the matrix A induces a homomorphism $A\colon \mathbb{C}^n \to \mathbb{C}^m$ by $x \mapsto Ax$. For fixed A, the mapping which maps b to the unique solution x of (3.66) is found to be linear. A matrix representation of this homomorphism with respect to canonical bases is then given by A^+. It is well-known that A^+ satisfies the four Penrose axioms

$$AA^+A = A, \tag{3.68a}$$
$$A^+AA^+ = A^+, \tag{3.68b}$$
$$(AA^+)^H = AA^+, \tag{3.68c}$$
$$(A^+A)^H = A^+A, \tag{3.68d}$$

see, e.g., [25], [56], [101]. On the other hand, for given $A \in \mathbb{C}^{m,n}$ the four axioms fix a unique matrix $A^+ \in \mathbb{C}^{n,m}$, whose existence follows for example by the unique solvability of (3.66).

The aim in this section is to introduce a least squares solution for linear differential-algebraic equations with well-defined strangeness index. As for linear equations, this least squares solution should be unique, even if there is no unique solution of the original problem, and even if the initial condition or the inhomogeneity are inconsistent.

Following the lines of the construction of the Moore–Penrose pseudoinverse for matrices as sketched above, we must deal with homomorphisms between function spaces, preferably some linear spaces of continuous functions or appropriate

subspaces. In view of (3.66), the norm of choice would be given by

$$\|x\| = \sqrt{(x, x)}, \quad (x, y) = \int_{\underline{t}}^{\bar{t}} x(t)^H y(t)\, dt, \tag{3.69}$$

where $\mathbb{I} = [\underline{t}, \bar{t}\,]$ is a compact interval. Since spaces of continuous functions cannot be closed with respect to this norm, we are neither in the pure setting of Banach spaces nor of Hilbert spaces. See [25, Ch. 8] for details on generalized inverses of operators on Hilbert spaces. In this section we therefore build up a scenario for defining a Moore–Penrose pseudoinverse which is general enough to be applicable in the setting of linear spaces of continuous functions.

Looking at (3.68), we find two essential ingredients in imposing the four Penrose axioms. These are the binary operation of matrix multiplication and the conjugate transposition of square matrices. In the language of mappings, they must be interpreted as composition of homomorphisms (we shall still call it multiplication) and the adjoint of endomorphisms. While the first item is trivial in any setting, the notion of the adjoint is restricted to the presence of a Hilbert space structure. The most general substitute we can find here is the concept of conjugates.

Definition 3.56. Let $\mathbb{X}$ be a (complex) vector space equipped with an inner product $(\cdot, \cdot)\colon \mathbb{X} \times \mathbb{X} \to \mathbb{C}$ and let $A\colon \mathbb{X} \to \mathbb{X}$ be an endomorphism. An endomorphism $A^*\colon \mathbb{X} \to \mathbb{X}$ is called a *conjugate* of A if and only if

$$(Ax, y) = (x, A^* y) \tag{3.70}$$

holds for all $x, y \in \mathbb{X}$.

For a unique definition of a Moore–Penrose pseudoinverse we of course need at least uniqueness of a conjugate. In addition we also need the inversion rule for the conjugate of a product.

Lemma 3.57. *Let $\mathbb{X}$ be a (complex) vector space equipped with an inner product $(\cdot, \cdot)\colon \mathbb{X} \times \mathbb{X} \to \mathbb{C}$.*

1. *For every endomorphism $A\colon \mathbb{X} \to \mathbb{X}$, there is at most one endomorphism $A^*\colon \mathbb{X} \to \mathbb{X}$ being conjugate to A.*

2. *Let the endomorphisms $A^*, B^*\colon \mathbb{X} \to \mathbb{X}$ be conjugate to the endomorphisms $A, B\colon \mathbb{X} \to \mathbb{X}$, respectively. Then AB has a conjugate $(AB)^*$ which is given by*

$$(AB)^* = B^* A^*. \tag{3.71}$$

Proof. Let A^* and $\tilde{A}^*$ be two conjugates of A. For all $x, y \in \mathbb{X}$, we then have

$$(x, A^* y) = (Ax, y) = (x, \tilde{A}^* y)$$

or $(x, (\tilde{A}^* - A^*)y) = 0$. Choosing $x = (\tilde{A}^* - A^*)y$ and using the definiteness of the inner product, we get $(\tilde{A}^* - A^*)y = 0$ for all $y \in \mathbb{X}$ or $\tilde{A}^* = A^*$. The second assertion follows from

$$(x, B^*A^*y) = (Bx, A^*y) = (ABx, y). \qquad \square$$

With these preparations, we can define the Moore–Penrose pseudoinverse of a homomorphism acting between two inner product spaces.

Definition 3.58. Let $\mathbb{X}$ and $\mathbb{Y}$ be (complex) inner product spaces and $D\colon \mathbb{X} \to \mathbb{Y}$ be a homomorphism. A homomorphism $D^+\colon \mathbb{Y} \to \mathbb{X}$ is called *Moore–Penrose pseudoinverse* of D if and only if DD^+ and D^+D possess conjugates $(DD^+)^*$ and $(D^+D)^*$, respectively, and the relations

$$DD^+D = D, \tag{3.72a}$$
$$D^+DD^+ = D^+, \tag{3.72b}$$
$$(DD^+)^* = DD^+, \tag{3.72c}$$
$$(D^+D)^* = D^+D \tag{3.72d}$$

hold.

As for matrices, the four axioms (3.72) guarantee uniqueness of the Moore–Penrose pseudoinverse, whereas existence in general cannot be shown.

Lemma 3.59. *Let $\mathbb{X}$ and $\mathbb{Y}$ be (complex) inner product spaces and $D\colon \mathbb{X} \to \mathbb{Y}$ be a homomorphism. Then D has at most one Moore–Penrose pseudoinverse $D^+\colon \mathbb{Y} \to \mathbb{X}$.*

Proof. Let $D^+, D^-\colon \mathbb{Y} \to \mathbb{X}$ be two Moore–Penrose pseudoinverses of D. Then we have

$$\begin{aligned} D^+ &= D^+DD^+ = D^+DD^-DD^+ \\ &= (D^+D)^*(D^-D)^*D^+ = (D^-DD^+D)^*D^+ \\ &= (D^-D)^*D^+ = D^-DD^+ = D^-(DD^+)^* \\ &= D^-(DD^-DD^+)^* = D^-(DD^+)^*(DD^-)^* \\ &= D^-DD^+DD^- = D^-DD^- = D^-. \end{aligned}$$ $\square$

According to (3.66) and (3.69), we consider the minimization problem

$$\frac{1}{2}\|x\|^2 = \min! \quad \text{s. t.} \quad \frac{1}{2}\|Dx - f\|^2 = \min!, \tag{3.73}$$

with D defined by

$$Dx(t) = E(t)\dot{x}(t) - A(t)x(t) \tag{3.74}$$

according to (3.1) or more explicitly

$$\frac{1}{2}\int_{\underline{t}}^{\bar{t}} \|x(t)\|_2^2\, dt = \min! \quad \text{s. t.} \quad \frac{1}{2}\int_{\underline{t}}^{\bar{t}} \|E(t)\dot{x}(t) - A(t)x(t) - f(t)\|_2^2\, dt = \min!. \tag{3.75}$$

In order to assure that D represents a reasonable linear operator related to (3.1) together with (3.2) in the form $x(\underline{t}) = x_0$, we make two assumptions. First, we assume that $x_0 = 0$ which can easily be achieved by shifting $x(t)$ to $x(t) - x_0$ and simultaneously changing the inhomogeneity from $f(t)$ to $f(t) + A(t)x_0$. Second, we restrict ourselves to strangeness-free pairs (E, A). This is also easily achieved for the class of pairs (E, A) with well-defined strangeness index, since we can perform the construction of Section 3.2 which led to a strangeness-free pair $(\hat{E}, \hat{A})$ with the extra requirement that the functions Z_1 and Z_2 have pointwise orthonormal columns. This is possible due to Theorem 3.9, since the functions Z_1 and Z_2 are unique up to pointwise unitary transformations on the columns and thus, $(\hat{E}, \hat{A})$ is unique up to a pointwise unitary transformation from the left. Since unitary transformations leave (3.73) invariant, the solution of (3.73) will not depend on the specific choice of $(\hat{E}, \hat{A})$. Recall that we had the problem to define the part $\hat{f}_3$ of the inhomogeneity in a reasonable way. Again this will not present a difficulty, since the results that we derive below will not depend on $\hat{f}_3$.

In contrast to the original problem, where we could use arbitrary pointwise nonsingular transformations P and Q, the minimization problem (3.73) requires these to be pointwise unitary. This means that we cannot make use of the canonical form (3.13). We replace this by a corresponding canonical form under unitary transformations.

Theorem 3.60. *Consider a strangeness-free pair of matrix functions (E, A). Then there exist pointwise unitary matrix functions $P \in C(\mathbb{I}, \mathbb{C}^{m,m})$ and $Q \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ such that (3.1) transforms to*

$$\tilde{E}(t)\dot{\tilde{x}} = \tilde{A}(t)\tilde{x} + \tilde{f}(t), \tag{3.76}$$

via $\tilde{E} = PEQ$, $\tilde{A} = PAQ - PE\dot{Q}$, $\tilde{x} = Q^H x$, $\tilde{f} = Pf$, with

$$\tilde{E} = \begin{bmatrix} \Sigma_E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & \Sigma_A & 0 \\ 0 & 0 & 0 \end{bmatrix}, \tag{3.77}$$

where Σ_E and Σ_A are pointwise nonsingular and all blocksizes are allowed to be zero.

Proof. We proceed as in the proof of Theorem 3.11 by application of Theorem 3.9, with the only difference that we are now not allowed to transform invertible sub-

matrices to identity matrices. We therefore obtain

$$(E, A) \sim \left(\begin{bmatrix} \Sigma_E & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & \Sigma_A & 0 \\ A_{31} & 0 & 0 \end{bmatrix} \right)$$

and A_{31} must vanish because (E, A) is assumed to be strangeness-free. □

Having fixed the class of differential-algebraic equations that we want to consider, we now return to the minimization problem (3.73). To describe the problem completely, we need to specify the spaces $\mathbb{X}$ and $\mathbb{Y}$. Requiring x to be continuously differentiable, in general, yields a continuous $f = Dx$. But even in the uniquely solvable case, f being continuous cannot guarantee the solution x to be continuously differentiable as the case $E = 0$ shows. We circumvent this problem by setting

$$\begin{aligned} \mathbb{X} &= \{x \in C(\mathbb{I}, \mathbb{C}^n) \mid E^+ E x \in C^1(\mathbb{I}, \mathbb{C}^n),\ E^+ E x(\underline{t}) = 0\}, \\ \mathbb{Y} &= C(\mathbb{I}, \mathbb{C}^m) \end{aligned} \tag{3.78}$$

and defining $D\colon \mathbb{X} \to \mathbb{Y}$ indirectly via the canonical form (3.77) by

$$D = P^H \tilde{D} Q^H, \tag{3.79}$$

where $\tilde{D}\colon \tilde{\mathbb{X}} \to \tilde{\mathbb{Y}}$ with

$$\tilde{D}\tilde{x}(t) = \tilde{E}(t)\dot{\tilde{x}}(t) - \tilde{A}(t)\tilde{x}(t) \tag{3.80}$$

and

$$\begin{aligned} \tilde{\mathbb{X}} &= \{\tilde{x} \in C(\mathbb{I}, \mathbb{C}^n) \mid \tilde{E}^+ \tilde{E} \tilde{x} \in C^1(\mathbb{I}, \mathbb{C}^n),\ \tilde{E}^+ \tilde{E} \tilde{x}(\underline{t}) = 0\}, \\ \tilde{\mathbb{Y}} &= C(\mathbb{I}, \mathbb{C}^m). \end{aligned} \tag{3.81}$$

As usual, all actions for functions are to be understood pointwise. In this way, the matrix functions P and Q can be seen as operators $P\colon \mathbb{Y} \to \tilde{\mathbb{Y}}$ and $Q\colon \tilde{\mathbb{X}} \to \mathbb{X}$. The latter property holds, because for $\tilde{x} \in \tilde{\mathbb{X}}$ and $x = Q\tilde{x}$ we get

$$\begin{aligned} E^+ E x &= (P^H \tilde{E} Q^H)^+ (P^H \tilde{E} Q^H) x \\ &= Q \tilde{E}^+ P P^H \tilde{E} \tilde{x} = Q \tilde{E}^+ \tilde{E} \tilde{x} \in C^1(\mathbb{I}, \mathbb{C}^n), \end{aligned}$$

since $Q \in C^1(\mathbb{I}, \mathbb{C}^{n,n})$ and P, Q are pointwise unitary, and hence $x \in \mathbb{X}$.

Partitioning

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix}, \quad \tilde{f} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{f}_3 \end{bmatrix} \tag{3.82}$$

according to the block structure of (3.77) and observing the special form of $\tilde{E}$, the condition $\tilde{x} \in \tilde{\mathbb{X}}$ implies $\tilde{x}_1$ to be continuously differentiable. But only this part of $\dot{\tilde{x}}$ actually appears on the right hand side of (3.80). Thus, (3.79) indeed defines an operator $D\colon \mathbb{X} \to \mathbb{Y}$ allowing the use of less smooth functions x compared with Definition 1.1. In addition, it is obvious that D is a homomorphism. In accordance with the theory of differential equations we call D a *differential-algebraic operator*. Compare this construction with that of a so-called *modified pair* of matrix functions which can be found in [100]. Since

$$\|x\| = \|Q^H x\| = \|\tilde{x}\|, \quad \|Dx - f\| = \|P(P^H \tilde{D} Q^H x - f)\| = \|\tilde{D}\tilde{x} - \tilde{f}\|, \tag{3.83}$$

the minimization problem (3.73) transforms covariantly with the application of the operators P and Q. Hence, we can first solve the minimization problem for differential-algebraic equations in the form (3.77) and then transform the solution back to get a solution of the original problem. Moreover, having found the Moore–Penrose pseudoinverse $\tilde{D}^+$ of $\tilde{D}$ the relation

$$D^+ = Q\tilde{D}^+ P \tag{3.84}$$

immediately gives the Moore–Penrose pseudoinverse of D.

Inserting the explicit form of $\tilde{E}$ and $\tilde{A}$ into (3.75) for the transformed problem yields the least squares problem

$$\begin{aligned} &\frac{1}{2}\int_{\underline{t}}^{\bar{t}} (\tilde{x}_1(t)^H \tilde{x}_1(t) + \tilde{x}_2(t)^H \tilde{x}_2(t) + \tilde{x}_3(t)^H \tilde{x}_3(t))\, dt = \min! \quad \text{s. t.} \\ &\frac{1}{2}\int_{\underline{t}}^{\bar{t}} (\tilde{w}_1(t)^H \tilde{w}_1(t) + \tilde{w}_2(t)^H \tilde{w}_2(t) + \tilde{w}_3(t)^H \tilde{w}_3(t))\, dt = \min! \end{aligned} \tag{3.85}$$

with

$$\begin{aligned} \tilde{w}_1(t) &= \Sigma_E(t)\dot{\tilde{x}}_1(t) - A_{11}(t)\tilde{x}_1(t) - A_{12}(t)\tilde{x}_2(t) - A_{13}(t)\tilde{x}_3(t) - \tilde{f}_1(t), \\ \tilde{w}_2(t) &= -A_{21}(t)\tilde{x}_1(t) - \Sigma_A(t)\tilde{x}_2(t) - \tilde{f}_2(t), \\ \tilde{w}_3(t) &= -\tilde{f}_3(t). \end{aligned}$$

For given $\tilde{f} \in \tilde{\mathbb{Y}}$, the minimization is to be taken over the whole of $\tilde{\mathbb{X}}$ as in (3.81), which can be written as

$$\tilde{\mathbb{X}} = \{\tilde{x} \in C(\mathbb{I}, \mathbb{C}^n) \mid \tilde{x}_1 \in C^1(\mathbb{I}, \mathbb{C}^d),\ \tilde{x}_1(\underline{t}) = 0\}, \tag{3.86}$$

where d denotes the size of $\tilde{x}_1$. The constraint equation in (3.85) is easily satisfied by choosing an arbitrary continuous function $\tilde{x}_3$, taking $\tilde{x}_1$ to be the solution of the

linear initial value problem

$$\begin{aligned}\dot{\tilde{x}}_1(t) &= \Sigma_E(t)^{-1}(A_{11}(t) - A_{12}(t)\Sigma_A(t)^{-1}A_{21}(t))\tilde{x}_1(t)\\ &\quad + \Sigma_E(t)^{-1}(A_{13}(t)\tilde{x}_3(t) + \tilde{f}_1(t)\\ &\quad - A_{12}(t)\Sigma_A(t)^{-1}\tilde{f}_2(t)), \quad \tilde{x}_1(\underline{t}) = 0,\end{aligned} \tag{3.87}$$

and finally setting

$$\tilde{x}_2(t) = -\Sigma_A(t)^{-1}(A_{21}(t)\tilde{x}_1(t) + \tilde{f}_2(t)). \tag{3.88}$$

Thus, we remain with the problem of minimizing $\frac{1}{2}\|x\|^2$ under the constraints (3.87) and (3.88). This can be interpreted as a linear quadratic optimal control problem, where $\tilde{x}_3$ takes the role of the control. But, compared with the standard linear quadratic optimal control problem treated in the literature (see, e.g., [16], [113], [168]), the constraints are more general due to the occurrence of inhomogeneities. The following theorem gives the essential properties of this problem in a new simplified and adapted notation which should not be mixed up with the one used so far.

Theorem 3.61. *Let*

$$\begin{gathered}A \in C(\mathbb{I}, \mathbb{C}^{d,d}), \quad B \in C(\mathbb{I}, \mathbb{C}^{d,l}), \quad C \in C(\mathbb{I}, \mathbb{C}^{p,d}),\\ f \in C(\mathbb{I}, \mathbb{C}^d), \quad g \in C(\mathbb{I}, \mathbb{C}^p).\end{gathered} \tag{3.89}$$

Then the linear quadratic control problem

$$\begin{aligned}&\frac{1}{2}\int_{\underline{t}}^{\bar{t}} (x(t)^H x(t) + y(t)^H y(t) + u(t)^H u(t))\, dt = \min! \quad s.\,t.\\ &\qquad \dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t),\ x(\underline{t}) = 0\\ &\qquad y(t) = C(t)x(t) + g(t)\end{aligned} \tag{3.90}$$

has a unique solution

$$x \in C^1(\mathbb{I}, \mathbb{C}^d), \quad y \in C(\mathbb{I}, \mathbb{C}^p), \quad u \in C(\mathbb{I}, \mathbb{C}^l).$$

This solution coincides with the corresponding part of the unique solution of the boundary value problem

$$\begin{aligned}&\dot{\lambda}(t) = (I + C(t)^H C(t))x(t) - A(t)^H \lambda(t) + C(t)^H g(t),\ \lambda(\bar{t}) = 0,\\ &\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t),\ x(\underline{t}) = 0,\\ &y(t) = C(t)x(t) + g(t),\\ &u(t) = B(t)^H \lambda(t)\end{aligned} \tag{3.91}$$

which can be obtained by the successive solution of the initial value problems

$$\begin{aligned}
\dot P(t) &= I + C(t)^H C(t) - P(t)A(t) \\
&\quad - A(t)^H P(t) - P(t)B(t)B(t)^H P(t), \quad P(\bar t) = 0, \\
\dot v(t) &= C(t)^H g(t) - P(t) f(t) \\
&\quad - A(t)^H v(t) - P(t)B(t)B(t)^H v(t), \quad v(\bar t) = 0, \\
\dot x(t) &= A(t)x(t) + B(t)B(t)^H (P(t)x(t) + v(t)) + f(t), \quad x(\underline t) = 0, \\
\lambda(t) &= P(t)x(t) + v(t), \\
y(t) &= C(t)x(t) + g(t), \\
u(t) &= B(t)^H \lambda(t).
\end{aligned} \tag{3.92}$$

Proof. Eliminating y with the help of the algebraic constraint and using a Lagrange multiplier λ, see, e.g., [113], cp. also Exercise 21, problem (3.92) is equivalent to

$$J[x, \dot x, u, \lambda] = \min!,$$

where (omitting arguments)

$$\begin{aligned}
J[x, \dot x, u, \lambda] = \int_{\underline t}^{\bar t} \Big[\frac{1}{2} (x^H x + (Cx + g)^H (Cx + g) + u^H u) \\
+ \mathfrak{Re}(\lambda^H (\dot x - Ax - Bu - f)) \Big] \, dt
\end{aligned}$$

and $x, \lambda \in C^1(\mathbb{I}, \mathbb{C}^d)$, $u \in C(\mathbb{I}, \mathbb{C}^l)$. Variational calculus then yields

$$\begin{aligned}
&J[x + \varepsilon\delta x, \dot x + \varepsilon\delta\dot x, u + \varepsilon\delta u, \lambda + \varepsilon\delta\lambda] \\
&= \int_{\underline t}^{\bar t} \Big[\frac{1}{2} \big((x + \varepsilon\delta x)^H (x + \varepsilon\delta x) + (u + \varepsilon\delta u)^H (u + \varepsilon\delta u) \big) \\
&\qquad + (C(x + \varepsilon\delta x) + g)^H (C(x + \varepsilon\delta x) + g) \\
&\qquad + \mathfrak{Re}\big((\lambda + \varepsilon\delta\lambda)^H \big((\dot x + \varepsilon\delta\dot x) - A(x + \varepsilon\delta x) - B(u + \varepsilon\delta u) - f \big) \big) \Big] \, dt \\
&= J[x, \dot x, u, \lambda] \\
&\qquad + \varepsilon \, \mathfrak{Re} \Big[\lambda^H \delta x \Big|_{\underline t}^{\bar t} + \int_{\underline t}^{\bar t} (x^H + (Cx + g)^H C - \lambda^H A - \dot\lambda^H) \delta x \, dt \\
&\qquad + \int_{\underline t}^{\bar t} (u^H - \lambda^H B) \delta u \, dt + \int_{\underline t}^{\bar t} \delta\lambda^H (\dot x - Ax - Bu - f) \, dt \Big]
\end{aligned}$$

$$+\varepsilon^2\Big[\frac{1}{2}\int_{\underline{t}}^{\bar{t}}(\delta x^H(I+C^HC)\delta x+\delta u^H\delta u)\,dt$$
$$+\mathfrak{Re}\int_{\underline{t}}^{\bar{t}}\delta\lambda^H(\delta\dot{x}-A\delta x-B\delta u)\,dt\Big]$$

after sorting and integration by parts. A necessary condition for a minimum at (x,u,λ) is that for all variations the coefficient of $\varepsilon\in\mathbb{R}$ vanishes. This at once yields (3.91).

Suppose that $(x+\varepsilon\delta x, u+\varepsilon\delta u, \lambda+\varepsilon\delta\lambda)$ is a second minimum. Without loss of generality, we may take $\varepsilon>0$. It follows that $(\delta x,\delta u,\delta\lambda)$ must solve the corresponding homogeneous problem. In particular, we must have

$$\delta\dot{x}=A\delta x+B\delta u,\quad \delta x(\underline{t})=0$$

and

$$\delta\dot{\lambda}=(I+C^HC)\delta x-A^H\delta\lambda,\quad \delta\lambda(\bar{t})=0.$$

This yields

$$J[x+\varepsilon\delta x,\dot{x}+\varepsilon\delta\dot{x},\lambda+\varepsilon\delta\lambda,u+\varepsilon\delta u]$$
$$=J[x,\dot{x},\lambda,u]+\varepsilon^2\int_{\underline{t}}^{\bar{t}}\frac{1}{2}(\delta x^H(I+C^HC)\delta x+\delta u^H\delta u)\,dt,$$

and it follows that $\delta x=0$, $\delta u=0$ and therefore $\delta\lambda=0$. Hence, there exists at most one solution of the linear quadratic control problem (3.90) and thus also of the boundary value problem (3.91). It remains to show that (3.91) indeed has a solution. To do this, we set

$$\lambda=Px+v,\quad \dot{\lambda}=P\dot{x}+\dot{P}x+\dot{v}$$

with some $P\in C^1(\mathbb{I},\mathbb{C}^{d,d})$, $v\in C^1(\mathbb{I},\mathbb{C}^d)$. Inserting into (3.91), we obtain

$$P\dot{x}+\dot{P}x+\dot{v}=(I+C^HC)x-A^H(Px+v)+C^Hg$$

and

$$P\dot{x}=PAx+PBB^H(Px+v)+Pf.$$

Combining these equations, we find

$$(PA+A^HP+PBB^HP-(I+C^HC)+\dot{P})x$$
$$+(PBB^Hv+Pf+A^Hv-C^Hg+\dot{v})=0.$$

Now we choose P and v to be the solutions of the initial value problems

$$\dot{P}=I+C^HC-PA-A^HP-PBB^HP,\quad P(\bar{t})=0,$$
$$\dot{v}=C^Hg-Pf-A^Hv-PBB^Hv,\qquad v(\bar{t})=0.$$

This choice is possible, since the second equation is linear (and hence always has a unique solution) and the first equation is a Riccati differential equation of a kind for which one can show that a Hermitian solution exists for any interval of the form $\mathbb{I}$, see, e.g., [120, Ch. 10].

It remains to show that (3.92) indeed solves (3.91). This is trivial for the third and fourth equation of (3.91). For the second equation, we of course have $x(\underline{t}) = 0$ but also

$$\begin{aligned}&\dot{x} - Ax - Bu - f\\ &\quad = Ax + BB^H Px + BB^H v + f - Ax - BB^H Px - BB^H v - f = 0.\end{aligned}$$

For the first equation, we have $\lambda(\overline{t}) = P(\overline{t})x(\overline{t}) + v(\overline{t}) = 0$ and also

$$\begin{aligned}&\dot{\lambda} - (I + C^H C)x + A^H \lambda - C^H g\\ &\quad = P\dot{x} + \dot{P}x + \dot{v} - (I + C^H C)x + A^H Px + A^H v - C^H g\\ &\quad = PAx + PBB^H Px + PBB^H v + Pf\\ &\qquad + (I + C^H C)x - PAx - A^H Px - PBB^H Px\\ &\qquad + C^H g - Pf - A^H v - PBB^H v\\ &\qquad - (I + C^H C)x + A^H Px + A^H v - C^H g = 0.\end{aligned}$$

□

We can immediately apply this result to the constrained minimization problem (3.85).

Corollary 3.62. *Problem* (3.85) *with constraints* (3.87) *and* (3.88) *has a unique solution* $\tilde{x} \in \tilde{\mathbb{X}}$.

Proof. The assertion follows from Theorem 3.61 by the following substitutions (again without arguments)

$$\begin{gathered}A = \Sigma_E^{-1}(A_{11} - A_{12}\Sigma_A^{-1}A_{21}), \quad B = \Sigma_E^{-1}A_{13}, \quad C = -\Sigma_A^{-1}A_{21},\\ f = \Sigma_E^{-1}(\tilde{f}_1 - A_{12}\Sigma_A^{-1}\tilde{f}_2), \quad g = -\Sigma_A^{-1}\tilde{f}_2.\end{gathered} \tag{3.93}$$

The unique solution has the form (3.82), where $\tilde{x}_1$ as part x of (3.92) is continuously differentiable with $\tilde{x}_1(\underline{t}) = 0$. □

With these preparations, we can define an appropriate operator

$$\tilde{D}^+ \colon \tilde{\mathbb{Y}} \to \tilde{\mathbb{X}} \tag{3.94}$$

as follows. For

$$\tilde{f} = \begin{bmatrix}\tilde{f}_1\\ \tilde{f}_2\\ \tilde{f}_3\end{bmatrix} \in \tilde{\mathbb{Y}}$$

the image

$$\tilde{x} = \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \tilde{D}^+ \tilde{f}$$

is given by the unique solution of (3.85) with (3.87) and (3.88). Note that $\tilde{D}^+ \tilde{f} \in \tilde{\mathbb{X}}$, because $\tilde{x}_1$ as part x of (3.92) is continuously differentiable and $\tilde{x}_1(\underline{t}) = 0$. Moreover, because the Riccati differential equation in (3.92) does not depend on the inhomogeneities, the operator $\tilde{D}^+$ is linear, hence a homomorphism.

Theorem 3.63. *The operator $\tilde{D}^+$, defined in (3.94), is the Moore–Penrose pseudoinverse of $\tilde{D}$, i.e., the endomorphisms $\tilde{D}\tilde{D}^+$ and $\tilde{D}^+\tilde{D}$ have conjugates such that (3.72) holds for $\tilde{D}$ and $\tilde{D}^+$.*

Proof. Let

$$\tilde{f} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{f}_3 \end{bmatrix} \in \tilde{\mathbb{Y}}, \quad \tilde{D}\tilde{D}^+\tilde{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix}.$$

With (3.80), using for simplicity the notation of Theorem 3.61 and Corollary 3.62, we get

$$\begin{aligned}
\hat{f}_1 &= \Sigma_E \dot{\tilde{x}}_1 - A_{11}\tilde{x}_1 - A_{12}\tilde{x}_2 - A_{13}\tilde{x}_3 \\
&= \Sigma_E \dot{x} - A_{11}x - A_{12}y - A_{13}u \\
&= \Sigma_E(Ax + BB^H(Px + v) + f) - A_{11}x \\
&\qquad - A_{12}(Cx + g) - A_{13}B^H(Px + v) \\
&= A_{11}x - A_{12}\Sigma_A^{-1}A_{21}x + \Sigma_E BB^H(Px + v) + \tilde{f}_1 - A_{12}\Sigma_A^{-1}\tilde{f}_2 \\
&\qquad - A_{11}x + A_{12}\Sigma_A^{-1}A_{21}x + A_{12}\Sigma_A^{-1}\tilde{f}_2 - \Sigma_E BB^H(Px + v) = \tilde{f}_1, \\
\hat{f}_2 &= -A_{21}\tilde{x}_1 - \Sigma_A\tilde{x}_2 = -A_{21}x - \Sigma_A y \\
&= \Sigma_A(Cx - y) = \Sigma_A(Cx - Cx - g) = \tilde{f}_2, \\
\hat{f}_3 &= 0,
\end{aligned}$$

and $\tilde{D}\tilde{D}^+$ is obviously conjugate to itself. Since $\tilde{D}\tilde{D}^+$ projects onto the first two components and since $\tilde{f}_3$ has no influence on the solution of (3.85), we also have $\tilde{D}^+\tilde{D}\tilde{D}^+ = \tilde{D}^+$. Since $\tilde{D}\tilde{x}$ has a vanishing third component for all $\tilde{x} \in \tilde{\mathbb{X}}$, the projector $\tilde{D}\tilde{D}^+$ acts as identity on $\tilde{D}$, i.e., $\tilde{D}\tilde{D}^+\tilde{D} = \tilde{D}$. The remainder of the proof deals with the fourth Penrose axiom.

Let

$$\tilde{x} = \begin{bmatrix} x \\ y \\ u \end{bmatrix} \in \tilde{\mathbb{X}}, \quad \tilde{D}^+\tilde{D}\tilde{x} = \begin{bmatrix} \hat{x} \\ \hat{y} \\ \hat{u} \end{bmatrix}.$$

We must now apply $\tilde{D}^+$ to the inhomogeneity

$$\tilde{D}\tilde{x} = \begin{bmatrix} \Sigma_E \dot{x} - A_{11}x - A_{12}y - A_{13}u \\ -A_{21}x - \Sigma_A y \\ 0 \end{bmatrix}.$$

Therefore, we must set

$$\begin{aligned} f &= \Sigma_E^{-1}(\Sigma_E \dot{x} - A_{11}x - A_{12}y - A_{13}u + A_{12}\Sigma_A^{-1}(A_{21}x + \Sigma_A y)) \\ &= \dot{x} - Ax - Bu, \\ g &= \Sigma_A^{-1}(A_{21}x + \Sigma_A y) \\ &= -Cx + y. \end{aligned}$$

Recalling that the solution P of the Riccati differential equation in (3.92) does not depend on the inhomogeneity, we must solve

$$\begin{aligned} \dot{v} &= C^H(-Cx + y) - A^H v - PBB^H v - P(\dot{x} - Ax - Bu), \quad v(\bar{t}) = 0, \\ \dot{\hat{x}} &= A\hat{x} + BB^H(P\hat{x} + v) + (\dot{x} - Ax - Bu), \quad \hat{x}(\underline{t}) = 0, \\ \hat{y} &= C\hat{x} - Cx + y, \\ \hat{u} &= B^H(P\hat{x} + v), \end{aligned}$$

where A, B, C, D are as in (3.93). Setting $v = w - Px$, $\dot{v} = \dot{w} - P\dot{x} - \dot{P}x$, we obtain

$$\begin{aligned} \dot{w} &= P\dot{x} + (I + C^H C)x - PAx - A^H Px - PBB^H Px \\ &\quad - C^H Cx + C^H y - A^H w + A^H Px - PBB^H w + PBB^H Px \\ &\quad - P\dot{x} + PAx + PBu \\ &= -(A^H + PBB^H)w + (x + C^H y + PBu), \quad w(\bar{t}) = 0. \end{aligned} \tag{3.95}$$

Let $W(t, s)$ be the *Wronskian matrix* belonging to $A + BB^H P$ in the sense that

$$\dot{W}(t, s) = (A + BB^H P)W(t, s), \quad W(s, s) = I.$$

Then $W(t, s)^{-H}$ is the Wronskian matrix belonging to $-(A^H + PBB^H)$. With the help of $W(t, s)$ we can represent the solution of the initial value problem (3.95) in the form

$$w = \int_{\bar{t}}^{t} W(t, s)^{-H}(x + C^H y + PBu)\, ds$$

or

$$v = -Px + \int_{\bar{t}}^{t} W(t, s)^{-H}(x + C^H y + PBu)\, ds.$$

Here and in the following, the arguments which must be inserted start with t and a Wronskian matrix changes it from the first to the second argument. Setting $\hat{x} = x + z$, we obtain

$$\begin{aligned}\dot{z} &= -\dot{x} + Ax + Az + BB^H Px + BB^H Pz \\ &\quad + BB^H w - BB^H Px + \dot{x} - Ax - Bu \\ &= (A + BB^H P)z + (BB^H w - Bu),\ z(\underline{t}) = 0\end{aligned}$$

or

$$z = \int_{\underline{t}}^{t} W(t,s)(BB^H w - Bu)\, ds.$$

Thus, we get $\hat{x}$, $\hat{y}$, $\hat{u}$ according to

$$\hat{x} = x + z,\ \hat{y} = y + Cz,\ \hat{u} = B^H(Pz + w).$$

Let now, in addition,

$$\begin{bmatrix}\tilde{x}\\ \tilde{y}\\ \tilde{u}\end{bmatrix} \in \tilde{\mathbb{X}}$$

be given and let

$$\tilde{D}^+\tilde{D}\begin{bmatrix}\tilde{x}\\ \tilde{y}\\ \tilde{u}\end{bmatrix} = \begin{bmatrix}\hat{\tilde{x}}\\ \hat{\tilde{y}}\\ \hat{\tilde{u}}\end{bmatrix}.$$

Then we have

$$\begin{aligned}&\int_{\underline{t}}^{\bar{t}} (\tilde{x}^H\hat{x} + \tilde{y}^H\hat{y} + \tilde{u}^H\hat{u})\, dt \\ &\quad = \int_{\underline{t}}^{\bar{t}} \Big[\tilde{x}^H x + \tilde{x}^H \int_{\underline{t}}^{t} W(t,s)(BB^H \int_{\bar{t}}^{s} W(s,r)^{-H}(x + C^H y + PBu)\, dr - Bu)\, ds \\ &\qquad + \tilde{y}^H y + \tilde{y}^H C \int_{\underline{t}}^{t} W(t,s)(BB^H \int_{\bar{t}}^{s} W(s,r)^{-H}(x + C^H y + PBu)\, dr - Bu)\, ds \\ &\qquad + \tilde{u}^H B^H P \int_{\underline{t}}^{t} W(t,s)(BB^H \int_{\bar{t}}^{s} W(s,r)^{-H}(x + C^H y + PBu)\, dr - Bu)\, ds \\ &\qquad + \tilde{u}^H B^H \int_{\bar{t}}^{t} W(t,s)^{-H}(x + C^H y + PBu)\, ds\Big]\, dt\end{aligned}$$

$$
\begin{aligned}
&= \int_{\underline{t}}^{\bar{t}} (\tilde{x}^H x + \tilde{y}^H y)\, dt \\
&\quad - \int_{\underline{t}}^{\bar{t}} \int_{\underline{t}}^{t} (\tilde{x}^H + \tilde{y}^H C + \tilde{u}^H B^H P) W(t,s) B u \, ds \, dt \\
&\quad + \int_{\underline{t}}^{\bar{t}} \int_{\bar{t}}^{t} \tilde{u}^H B^H W(t,s)^{-H} (x + C^H y + PBu) \, ds \, dt \\
&\quad + \int_{\underline{t}}^{\bar{t}} \int_{\bar{t}}^{t} \int_{\bar{t}}^{s} (\tilde{x}^H + \tilde{y}^H C + \tilde{u}^H B^H P) W(t,s) B \\
&\qquad\qquad \cdot B^H W(s,r)^{-H} (x + C^H y + PBu) \, dr \, ds \, dt.
\end{aligned}
$$

By transposition and changing the order of the integrations, we finally have

$$
\int_{\underline{t}}^{\bar{t}} (\tilde{x}^H \hat{x} + \tilde{y}^H \hat{y} + \tilde{u}^H \hat{u})\, dt = \int_{\underline{t}}^{\bar{t}} (x^H \hat{\tilde{x}} + y^H \hat{\tilde{y}} + u^H \hat{\tilde{u}})\, dt,
$$

which is nothing else than that $\tilde{D}^+ \tilde{D}$ is conjugate to itself. □

It then follows from Theorem 3.63 that (3.79) yields the Moore–Penrose pseudoinverse of D, i.e., we have shown the existence and uniqueness of an operator D^+ satisfying (3.72) and thus we have fixed a unique (classical) least squares solution for a large class of differential-algebraic equations (including higher index problems) with possibly inconsistent initial data or inhomogeneities or free solution components.

Using D^+ for solving differential-algebraic equations with undetermined solutions components, however, bears at least two disadvantages. First, the undetermined component $\tilde{x}_3$ need not satisfy the given initial condition and, second, instead of an initial value problem we must solve a boundary value problem, which means that values of the coefficients in future times influence the solution at the present time. A simple way out of this problem is to choose the undetermined part to be zero. In the following, we investigate this approach in the context of generalized inverses. To do this, we consider the matrix functions given by

$$
\Pi = E^+ E + F^+ F, \quad F = (I - EE^+) A (I - E^+ E). \tag{3.96}
$$

Transforming to the form (3.77), we then have

$$
\begin{aligned}
\tilde{F} &= (I - \tilde{E}\tilde{E}^+)\tilde{A}(I - \tilde{E}^+\tilde{E}) \\
&= (I - PEQQ^H E^+ P^H)(PAQ - PE\dot{Q})(I - Q^H E^+ P^H PEQ) \\
&= P(I - EE^+)(A - E\dot{Q}Q^H)(I - E^+E)Q \\
&= P(I - EE^+)A(I - E^+E)Q = PFQ.
\end{aligned}
$$

Thus, F transforms like E and therefore

$$\begin{aligned}\tilde{\Pi} &= \tilde{E}^+\tilde{E} + \tilde{F}^+\tilde{F}\\ &= Q^H E^+ P^H P E Q + Q^H F^+ P^H P F Q\\ &= Q^H (E^+E + F^+F) Q = Q^H \Pi Q.\end{aligned}$$

A simple calculation then yields

$$\tilde{\Pi} = \begin{bmatrix} I_d & 0 & 0\\ 0 & I_a & 0\\ 0 & 0 & 0\end{bmatrix}.$$

This, in particular, shows that Π is pointwise an orthogonal projector. Note that $I - \tilde{\Pi}$ indeed projects onto the undetermined component $\tilde{x}_3$. Hence we are led to the problem

$$\begin{aligned}&\frac{1}{2}\int_{\underline{t}}^{\bar{t}} \|(I - \Pi(t))x(t)\|_2^2\, dt = \min! \quad \text{s. t.}\\ &\frac{1}{2}\int_{\underline{t}}^{\bar{t}} \|E(t)\dot{x}(t) - A(t)x(t) - f(t)\|_2^2\, dt = \min!\end{aligned} \tag{3.97}$$

replacing (3.75). The preceding results state that again the problem transforms covariantly with the application of P and Q so that we only need to solve (3.97) for differential-algebraic equations in the form (3.77). Since (3.97) here implies $\tilde{x}_3 = 0$ by construction, we remain with a reduced differential-algebraic equation (3.87) and (3.88) that is uniquely solvable. We can therefore carry over all results obtained so far as long as they do not depend on the specific choice of $\tilde{x}_3$. Recognizing that only for the fourth Penrose axiom this choice was utilized, we find that (3.97) fixes a so-called (1,2,3)-pseudoinverse $\tilde{D}^-$ of $\tilde{D}$ satisfying the axioms (3.72a), (3.72b) and (3.72c). Keeping the spaces as before, we arrive at the following result.

Theorem 3.64. *The operator $\tilde{D}^-$ defined via (3.97) is a $(1,2,3)$-pseudoinverse of $\tilde{D}$, i.e., the endomorphism $\tilde{D}\tilde{D}^-$ has a conjugate such that (3.72a)–(3.72c) hold for $\tilde{D}$ and $\tilde{D}^-$.*

Again defining the operator D^- by $D^- = Q\tilde{D}^- P$ then gives a (1,2,3)-pseudoinverse of the operator D.

Remark 3.65. Recall that in Section 3.2, following Theorem 3.32, we have discussed how to fix the inhomogeneity belonging to the third block row in (3.77) when coming from a higher index differential-algebraic equation. Fortunately, both generalized inverses D^+, D^- yield solutions of the corresponding minimization problems that do not depend on this part of the inhomogeneity. However, we

cannot determine the residual associated with the third block row in (3.77). But this is not really a problem, since we can specify a residual with respect to the original differential-algebraic equation.

Remark 3.66. In the case that $A_{13} = 0$ (including A_{13} empty, i.e., no corresponding block in the canonical form (3.77) exists), we immediately have $D^- = D^+$. Observing that for $E = 0$ the existence of (3.77) requires rank A to be constant on $\mathbb{I}$, we find $D^+ = D^- = -A^+$. In particular, this shows that both D^+ and D^- are indeed generalizations of the Moore–Penrose pseudoinverse of matrices.

Remark 3.67. The boundedness of the linear operators $D\colon \mathbb{X} \to \mathbb{Y}$ and $D^+, D^-\colon \mathbb{Y} \to \mathbb{X}$, where $\mathbb{X}$ and $\mathbb{Y}$ are seen as the given vector spaces equipped with the norms $\|x\|_{\mathbb{X}} = \|x\|_{L_2} + \|\frac{d}{dt}(E^+Ex)\|_{L_2}$ and $\|y\|_{\mathbb{Y}} = \|y\|_{L_2}$, allows for their extension to the closure of $\mathbb{X}$ and $\mathbb{Y}$ with respect to these norms, see, e.g., [80, Lemma 4.3.16]. In particular, $\mathbb{Y}$ becomes the Hilbert space $L_2(\mathbb{I}, \mathbb{C}^m)$. Other choices of norms are possible as well.

Example 3.68. Consider the initial value problem

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}, \quad \begin{bmatrix} x_1(t_0) \\ x_2(t_0) \end{bmatrix} = \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}.$$

It has strangeness index $\mu = 1$, whereas the differentiation index is not defined. To obtain a strangeness-free differential-algebraic equation with the same solution space according to Section 3.2, we compute

$$M = \begin{bmatrix} E & 0 \\ \dot{E} - A & E \end{bmatrix}, \quad N = \begin{bmatrix} A & 0 \\ \dot{A} & 0 \end{bmatrix}, \quad g = \begin{bmatrix} f \\ \dot{f} \end{bmatrix}$$

and obtain (with shifted initial values)

$$M(t) = \left[\begin{array}{cc|cc} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ \hline 0 & 2t & -t & t^2 \\ 0 & 2 & -1 & t \end{array}\right], \quad N(t) = \left[\begin{array}{cc|cc} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right],$$

$$g(t) = \left[\begin{array}{c} f_1(t) - x_{1,0} \\ f_2(t) - x_{2,0} \\ \hline \dot{f}_1(t) \\ \dot{f}_2(t) \end{array}\right].$$

Since rank $M(t) = 2$ for all $t \in \mathbb{R}$, the procedure of Section 3.2 reduces to the computation of an orthogonal projection onto the corange of $M(t)$ given, e.g., by

$$Z(t)^H = \frac{1}{\sqrt{1+t^2}} \left[\begin{array}{cc|cc} 1 & -t & 0 & 0 \\ 0 & 0 & 1 & -t \end{array}\right].$$

Replacing E, A, and f by $Z^H M$, $Z^H N$, and $Z^H g$ yields the strangeness-free differential-algebraic equation

$$\begin{aligned} 0 &= \frac{1}{\sqrt{1+t^2}}(-x_1 + tx_2 + f_1(t) - x_{1,0} - tf_2(t) + tx_{2,0}),\\ 0 &= \frac{1}{\sqrt{1+t^2}}(\dot{f}_1(t) - t\dot{f}_2(t)), \end{aligned}$$

together with homogeneous initial conditions. Denoting the coefficient functions again by E, A and f, we have $E(t) = 0$ and

$$A(t) = \frac{1}{\sqrt{1+t^2}}\begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \frac{1}{\sqrt{1+t^2}}\begin{bmatrix} f_1(t) - x_{1,0} - tf_2(t) + tx_{2,0} \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix}.$$

According to Remark 3.66, the least squares solution of this purely algebraic equation is given by $x = -A^+ f$. Transforming back, we obtain the least squares solution of the given original problem

$$x(t) = -\frac{1}{\sqrt{1+t^2}}\begin{bmatrix} -1 & 0 \\ t & 0 \end{bmatrix}\frac{1}{\sqrt{1+t^2}}\begin{bmatrix} f_1(t) - x_{1,0} - tf_2(t) + tx_{2,0} \\ \dot{f}_1(t) - t\dot{f}_2(t) \end{bmatrix} + \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}$$

or

$$x(t) = \frac{1}{1+t^2}\begin{bmatrix} f_1(t) - x_{1,0} - tf_2(t) + tx_{2,0} \\ -t(f_1(t) - x_{1,0} - tf_2(t) + tx_{2,0}) \end{bmatrix} + \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix}.$$

Returning to the notation of Section 3.2, the least squares solution of (3.1) with initial condition $x(t_0) = 0$ has been defined by

$$x = D^+\hat{f}, \quad D\colon \mathbb{X} \to \mathbb{Y}, \quad Dx(t) = \hat{E}(t)\dot{x}(t) - \hat{A}(t)x(t). \tag{3.98}$$

The vector spaces $\mathbb{X}$ and $\mathbb{Y}$, given now by

$$\begin{aligned} \mathbb{X} &= \{x \in C(\mathbb{I}, \mathbb{C}^n) \mid \hat{E}^+\hat{E}x \in C^1(\mathbb{I}, \mathbb{C}^n),\ \hat{E}^+\hat{E}x(t_0) = 0\},\\ \mathbb{Y} &= C(\mathbb{I}, \mathbb{C}^m), \end{aligned} \tag{3.99}$$

become Banach spaces by introducing the norms

$$\|x\|_{\mathbb{X}} = \|x\|_{\mathbb{Y}} + \|\tfrac{d}{dt}(\hat{E}^+\hat{E}x)\|_{\mathbb{Y}}, \quad \|f\|_{\mathbb{Y}} = \max_{t\in\mathbb{I}} \|f(t)\|_\infty. \tag{3.100}$$

The operator D^+ (and also D^-) is continuous with respect to these norms, i.e.,

$$\|D^+\|_{\mathbb{X}\leftarrow\mathbb{Y}} = \sup_{\hat{f}\in\mathbb{Y}\setminus\{0\}} \frac{\|D^+\hat{f}\|_{\mathbb{X}}}{\|\hat{f}\|_{\mathbb{Y}}} < \infty. \tag{3.101}$$

In this sense, the discussed minimization problems are well-posed. The original differential-algebraic equation is well-posed in the same sense, when the minimization

problems reduce to the original differential-algebraic equation, i.e., when D is invertible. This is the case if and only if the differential-algebraic equation is strangeness-free with $m = n$ and has no undetermined solution components.

Let x be the least squares solution of (3.1), i.e., let $x = D^+(\hat{f} + \hat{A}x_0) + x_0$, and let $\hat{x}$ be the least squares solution of the perturbed problem

$$E(t)\dot{\hat{x}} = A(t)\hat{x} + f(t) + \eta(t), \quad \hat{x}(t_0) = \hat{x}_0 \tag{3.102}$$

with $\eta \in C^\mu(\mathbb{I}, \mathbb{C}^m)$, i.e., let $x = D^+(\hat{f} + \hat{\eta} + \hat{A}x_0) + x_0$ with

$$\hat{\eta} = \begin{bmatrix} \hat{\eta}_1 \\ \hat{\eta}_2 \\ 0 \end{bmatrix}, \quad \hat{\eta}_1 = Z_1^H \eta, \quad \hat{\eta}_2 = Z_2^H \begin{bmatrix} \eta \\ \vdots \\ \eta^{(\mu)} \end{bmatrix} \in \mathbb{Y}.$$

Then the estimate

$$\begin{aligned} \|\hat{x} - x\|_{\mathbb{X}} = \|D^+\hat{\eta} + (\hat{x}_0 - x_0)\|_{\mathbb{X}} &\le \|D^+\|_{\mathbb{X}\leftarrow\mathbb{Y}} \|\hat{\eta}\|_{\mathbb{Y}} + \|\hat{x}_0 - x_0\|_\infty \\ &\le C(\|\hat{x}_0 - x_0\|_\infty + \|\eta\|_{\mathbb{Y}} + \|\dot{\eta}\|_{\mathbb{Y}} + \cdots + \|\eta^{(\mu)}\|_{\mathbb{Y}}) \end{aligned} \tag{3.103}$$

holds. By the definition of the strangeness index, it is clear that (3.103) cannot hold for a smaller value of μ.

The estimate (3.103) shows that the strangeness index is directly related to the perturbation index that was introduced in [108].

Definition 3.69. Given a solution $x \in C^1(\mathbb{I}, \mathbb{C}^n)$ of (3.1). The differential-algebraic equation (3.1) is said to have *perturbation index* $\kappa \in \mathbb{N}$ along x, if κ is the smallest number such that for all sufficiently smooth $\hat{x} \in C^1(\mathbb{I}, \mathbb{C}^n)$ the defect $\eta \in C(\mathbb{I}, \mathbb{C}^m)$ defined by (3.102) satisfies the estimate

$$\|\hat{x} - x\|_{\mathbb{X}} \le C(\|\hat{x}_0 - x_0\|_\infty + \|\eta\|_{\mathbb{Y}} + \|\dot{\eta}\|_{\mathbb{Y}} + \cdots + \|\eta^{(\kappa-1)}\|_{\mathbb{Y}}), \tag{3.104}$$

with a constant C independent of $\hat{x}$. It is said to have *perturbation index* $\kappa = 0$ if the estimate

$$\|\hat{x} - x\|_{\mathbb{X}} \le C(\|\hat{x}_0 - x_0\|_\infty + \max_{t\in\mathbb{I}} \| \textstyle\int_{t_0}^t \eta(s)\, ds\|_\infty) \tag{3.105}$$

holds.

The above discussion then has shown the following relation between the strangeness and the perturbation index.

Theorem 3.70. *Let the strangeness index μ of (3.1) be well defined as in (3.17) and let x be a solution of (3.1). Then the perturbation index κ of (3.1) along x is well defined with*

$$\kappa = \begin{cases} 0 & \text{for } \mu = 0,\ a_\mu = 0, \\ \mu + 1 & \text{otherwise.} \end{cases} \tag{3.106}$$

Remark 3.71. The reason for the two cases in the definition of the perturbation index and hence in the relation (3.106) to the strangeness index is that in this way the perturbation index equals the differentiation index if it is defined. Counting in the way of the strangeness index according to the estimate (3.103), there would be no need in the extension (3.105).

3.5 Generalized solutions

As in the case of constant coefficients, we may again weaken the differentiability constraints and also the consistency constraints. The first complete analysis of the distributional version for the linear variable coefficient case was given in [179], [180] on the basis of a reformulation and extension of the construction given in [125], that we have discussed in detail in Sections 3.1 and 3.2. Here we use the theory described in these sections directly to derive the distributional theory.

We consider the distributional version of the linear time varying system

$$E(t)\dot{x} = A(t)x + f, \tag{3.107}$$

with impulsive smooth distribution $f \in \mathcal{C}_{\mathrm{imp}}^m$. Looking for solutions $x \in \mathcal{C}_{\mathrm{imp}}^n$, we must require that $E, A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$ in order to have well-defined products $E\dot{x}$ and Ax. To apply the techniques of the previous sections, we observe that Theorem 3.9 guarantees that all constructions based on the decomposition (3.8) stay in spaces of infinitely often differentiable functions. In particular, systems like (3.18), (3.44), or (3.60) can be obtained in the same way with infinitely often differentiable matrix functions. The only difference is that we do not need smoothness requirements for the inhomogeneity. For convenience, we restrict ourselves here to the following modification of Theorem 3.17.

Theorem 3.72. *Let $E, A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$ and let the strangeness index μ of (E, A) be well defined. Furthermore, let $f \in \mathcal{C}_{\mathrm{imp}}^m$ with impulse order* $\operatorname{iord} f = q \in \mathbb{Z} \cup \{-\infty\}$. *Then the differential-algebraic equation (3.107) is equivalent (in the sense that there is a one-to-one correspondence between the solution spaces via a pointwise nonsingular infinitely often differentiable matrix function) to a differential-algebraic equation of the form*

$$\dot{x}_1 = A_{13}(t)x_3 + f_1, \quad d_\mu \tag{3.108a}$$
$$0 = x_2 + f_2, \quad a_\mu \tag{3.108b}$$
$$0 = f_3, \quad v_\mu \tag{3.108c}$$

where $A_{13} \in C^\infty(\mathbb{I}, \mathbb{C}^{d_\mu, u_\mu})$ *and*

$$\operatorname{iord} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} \leq q + \mu. \tag{3.109}$$

Proof. All the constructions that lead to the canonical form (3.13) can be executed using infinitely often differentiable matrix functions due to Theorem 3.9. Then the form (3.108) follows directly from Theorem 3.17, where the inhomogeneities f_1, f_2, f_3 are determined from $f^{(0)}, \ldots, f^{(\mu)}$ via transformations with infinitely often differentiable matrix functions. □

Corollary 3.73. *Let* $E, A \in C^\infty(\mathbb{R}, \mathbb{C}^{m,n})$ *satisfy the assumptions of Theorem 3.72. Then, we have:*

1. *Problem (3.107) has a solution in* $\mathcal{C}^n_{\text{imp}}$ *if and only if the* v_μ *distributional conditions*

$$f_3 = 0 \tag{3.110}$$

are fulfilled.

2. *Let* $t_0 \neq 0$ *and* $x_0 \in \mathbb{C}^n$. *There exists a solution* $x \in \mathcal{C}^n_{\text{imp}}$ *satisfying one of the initial conditions*

$$x(t_0) = x_0, \quad x(0^-) = x_0, \quad x(0^+) = x_0 \tag{3.111}$$

if and only if in addition to (3.110) the corresponding condition out of

$$x_2(t_0) = -f_2(t_0), \quad x_2(0^-) = -f_2(0^-), \quad x_2(0^+) = -f_2(0^+) \tag{3.112}$$

is implied by the initial condition.

3. *The corresponding initial value problem has a unique solution in* $\mathcal{C}^n_{\text{imp}}$ *if and only if in addition*

$$u_\mu = 0 \tag{3.113}$$

holds.

Moreover, all solutions x *satisfy* $\operatorname{iord} x \leq \max\{q + \mu, \operatorname{iord} x_3\}$.

Proof. The proof follows directly as in Section 2.4. □

To compare this result with Theorem 2.42, we recall that in the variable coefficient case regularity corresponds to the condition (3.19). In this case, (3.110) is an empty condition and (3.113) holds by assumption. Thus, Corollary 3.73 states that a regular problem is always uniquely solvable for consistent initial conditions

with $\operatorname{iord} x \leq q + \mu$. Moreover, the assertions of Theorem 2.43 hold in the variable coefficient case if the system is regular and if we replace ν by $\mu + 1$. This immediately implies that inconsistent initial conditions should again be treated as impulses in the inhomogeneity according to

$$E(t)\dot{x} = A(t)x + f + E(t)x_0\delta, \quad x_- = 0 \tag{3.114}$$

with $f_- = 0$.

Theorem 3.74. *Consider a differential-algebraic equation in the form (3.114) and suppose that the strangeness index μ of (E, A) as in (3.17) is well defined with $v_\mu = 0$. Let $f \in \mathcal{C}^m_{\text{imp}}$ be given with $f_- = 0$ and $\operatorname{iord} f \leq -1$.*

1. *All vectors $x_0 \in \mathbb{C}^n$ are consistent with f if and only if $\mu = 0$ and $a_\mu = 0$.*

2. *All vectors $x_0 \in \mathbb{C}^n$ are weakly consistent with f if and only if $\mu = 0$.*

Proof. For the first part, let $a_\mu \neq 0$. Then (3.108b) is present and the condition $x_2(0^+) = -f_2(0^+)$ restricts the set of consistent initial values. Hence, for all initial values to be consistent, $a_\mu = 0$ must hold, implying that $\mu = 0$.

For the converse, let $\mu = 0$ and $a_\mu = 0$. Equation (3.114) transforms covariantly with infinitely differentiable global equivalence transformations, since setting $x = Q\tilde{x}$, $x_0 = Q(0)\tilde{x}_0$ and multiplying (3.114) by P with pointwise nonsingular $P \in C^\infty(\mathbb{R}, \mathbb{C}^{m,m})$ and $Q \in C^\infty(\mathbb{R}, \mathbb{C}^{n,n})$ yields

$$\tilde{E}(t)\dot{\tilde{x}} = \tilde{A}(t)\tilde{x} + \tilde{f} + \tilde{E}(t)\tilde{x}_0\delta,$$

with

$$\tilde{E} = PEQ, \quad \tilde{A} = PAQ - PE\dot{Q}, \quad \tilde{f} = Pf.$$

Hence, we may assume that (E, A) is in global canonical form (3.23). For $\mu = 0$ and $a_\mu = 0$, the corresponding differential-algebraic equation reads (omitting the argument t)

$$\begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} 0 & A_{12} \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} + \tilde{f} + \tilde{x}_{1,0}\delta,$$

with $\operatorname{iord} \tilde{f} \leq -1$. Choosing $\tilde{x}_2 = \tilde{x}_{2,0}$, the part $\tilde{x}_1$ solves

$$\dot{\tilde{x}}_1 = A_{12}\tilde{x}_{2,0} + \tilde{f} + \tilde{x}_{1,0}\delta,$$

with $\operatorname{iord}(A_{12}\tilde{x}_{2,0} + \tilde{f}) \leq -1$. According to Exercise 15 of Chapter 2, the solution $\tilde{x}_1$ satisfies $\tilde{x}_1(0^+) = \tilde{x}_{1,0}$ and $\operatorname{iord} \tilde{x}_1 \leq -1$. Hence, there exists a solution $\tilde{x}$ with $\operatorname{iord} \tilde{x} \leq -1$. Transforming back, we get a solution $x = Q\tilde{x}$ of (3.114) with $\operatorname{iord} x \leq -1$.

For the second part, let $\mu > 0$. Then the transformation of (3.114) to the global canonical form (3.23) yields

$$\begin{bmatrix} 0 & G_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & G_1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} (\dot{\tilde{x}}_2)_\mu \\ \vdots \\ (\dot{\tilde{x}}_2)_0 \end{bmatrix} = \begin{bmatrix} (\tilde{x}_2)_\mu \\ \vdots \\ (\tilde{x}_2)_0 \end{bmatrix} + \begin{bmatrix} \tilde{f}_\mu \\ \vdots \\ \tilde{f}_0 \end{bmatrix} + \begin{bmatrix} 0 & G_\mu & & * \\ & \ddots & \ddots & \\ & & \ddots & G_1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} (\tilde{x}_{2,0})_\mu \\ \vdots \\ (\tilde{x}_{2,0})_0 \end{bmatrix} \delta$$

as part of the transformed problem (3.114). We first obtain

$$(\tilde{x}_2)_0 = -\tilde{f}_0$$

with $\operatorname{iord}(\tilde{x}_2)_0 \leq -1$. Differentiating and inserting into the second but last block equation gives

$$(\tilde{x}_2)_1 = -\tilde{f}_1 - G_1 \dot{\tilde{f}}_0 - G_1(\tilde{x}_{2,0})_0 \delta.$$

Since $G_1(0)$ has full row rank, there exists an initial value $(\tilde{x}_{2,0})_0$ such that $\operatorname{iord}(\tilde{x}_2)_1 = 0$. Hence, there exists an $x_0 \in \mathbb{C}^n$ which is not weakly consistent.

Conversely, let $\mu = 0$. The problem in global canonical form then reads

$$\begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} 0 & A_{12} & 0 \\ 0 & 0 & I \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} + \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix} + \begin{bmatrix} \tilde{x}_{1,0} \\ 0 \end{bmatrix} \delta,$$

with $\operatorname{iord}(\tilde{f}_1, \tilde{f}_2) \leq -1$. Choosing $\tilde{x}_2 = 0$, we obtain $\tilde{x}_3 = -\tilde{f}_2$ with $\operatorname{iord} \tilde{x}_3 \leq -1$ and $\tilde{x}_1$ solves

$$\dot{\tilde{x}}_1 = \tilde{f}_1 + \tilde{x}_{1,0}\delta.$$

Here, $\operatorname{iord} \tilde{x}_1 \leq -1$ follows by Theorem 2.41. □

A similar result under the weaker assumption that (E, A) satisfies Hypothesis 3.48 does not hold. Of course, due to Corollary 3.26, we can reduce this case to the case of Theorem 3.74 if

$$0 \in \bigcup_{j \in \mathbb{N}} \mathbb{I}_j$$

because then there exists a neighborhood of $0 \in \mathbb{R}$, where the strangeness index of (E, A) is well defined. If, however,

$$0 \in \mathbb{R} \setminus \bigcup_{j \in \mathbb{N}} \mathbb{I}_j,$$

then new effects may occur. Examples for this are given in Exercises 22 and 23.

As already announced in Chapter 2, we want to discuss the case that nonsmooth behavior of the solution due to nonsmooth behavior of the inhomogeneity (or inconsistent initial conditions) occurs at (possibly) more than one point. Suppose that the set

$$\mathbb{T} = \{t_j \in \mathbb{R} \mid t_j < t_{j+1},\ j \in \mathbb{Z}\}$$

has no accumulation point. Then there exists an immediate extension of impulsive smooth distributions. Consider the formal composition

$$x = \hat{x} + x_{\mathrm{imp}}, \quad \hat{x} = \sum_{j\in\mathbb{Z}} x_j \tag{3.115}$$

of distributions with

$$x_j \in C^\infty([t_j, t_{j+1}], \mathbb{C}) \quad \text{for all } j \in \mathbb{Z} \tag{3.116}$$

and impulsive part

$$x_{\mathrm{imp}} = \sum_{j\in\mathbb{Z}} x_{\mathrm{imp},j}, \qquad x_{\mathrm{imp},j} = \sum_{i=0}^{q_j} c_{ij}\delta_{t_j}^{(i)}, \quad c_{ij} \in \mathbb{C},\ q_j \in \mathbb{N}_0. \tag{3.117}$$

The first question is whether every x of this form is actually a distribution, since formally there are two possibilities to be faced with infinite sums of complex numbers (such that there would be a need for restrictions to guarantee convergence). The first one is that j runs over all integers and the second one is that q_j need not be bounded as a sequence in $\mathbb{Z}$. Testing such an x with a $\phi \in \mathcal{D}$ gives

$$\langle x, \phi\rangle = \sum_{j\in\mathbb{Z}} \langle x_j, \phi\rangle + \sum_{j\in\mathbb{Z}} \sum_{i=0}^{q_j} c_{ij}\langle \delta_{t_j}^{(i)}, \phi\rangle.$$

Since ϕ has bounded support and $\mathbb{T}$ has no accumulation point, all but finitely many terms $\langle x_j, \phi\rangle$ and $\langle \delta_{t_j}^{(i)}, \phi\rangle$ vanish. Thus, only a finite sum remains. Since the continuity of a distribution is defined with respect to sequences $\phi_k \to 0$, where all $\phi_k \in \mathcal{D}$ vanish outside the same bounded interval, it is clear that x is a continuous linear form. This shows that all compositions of the form (3.115) are distributions with impulsive behavior restricted to the set $\mathbb{T}$. We denote the set of such distributions by $\mathcal{C}_{\mathrm{imp}}(\mathbb{T})$.

Lemma 3.75. *Impulsive smooth distributions in $\mathcal{C}_{\mathrm{imp}}(\mathbb{T})$, where $\mathbb{T} \subseteq \mathbb{R}$ has no accumulation point, have the following properties:*

1. *A distribution $x \in \mathcal{C}_{\mathrm{imp}}(\mathbb{T})$ uniquely determines the decomposition (3.115).*

2. *With a distribution $x \in \mathcal{C}_{\mathrm{imp}}(\mathbb{T})$, we can assign a function value $x(t)$ for every $t \in \mathbb{R} \setminus \mathbb{T}$ by*
$$x(t) = x_j(t) \quad \text{for } t \in (t_j, t_{j+1})$$
and limits
$$x(t_j^-) = \lim_{t \to t_j^-} x_{j-1}(t), \quad x(t_j^+) = \lim_{t \to t_j^+} x_j(t)$$
for every $t_j \in \mathbb{T}$.

3. *All derivatives and primitives of $x \in \mathcal{C}_{\mathrm{imp}}(\mathbb{T})$ are again in $\mathcal{C}_{\mathrm{imp}}(\mathbb{T})$.*

4. *The set $\mathcal{C}_{\mathrm{imp}}(\mathbb{T})$ is a (complex) vector space and closed under multiplication with functions $A \in C^\infty(\mathbb{R}, \mathbb{C})$.*

Proof. The proof follows the lines of the proof of Lemma 2.38. □

Theorem 3.72 and, with slight modifications concerning the initial conditions, Corollary 3.73 hold for $f \in \mathcal{C}^m_{\mathrm{imp}}(\mathbb{T})$ with solutions $x \in \mathcal{C}^n_{\mathrm{imp}}(\mathbb{T})$. The assertions for the impulse order are valid for every $t_j \in \mathbb{T}$ separately. In view of (3.114), we can now allow for initial conditions $x(t_j^+) = x_{j,0}$ at every $t_j \in \mathbb{T}$ by replacing (3.107) with

$$E(t)\dot{x} = A(t)x + f_j + f_{\mathrm{imp},j} + E(t)x_{j,0}\delta, \quad x_{j-1} = 0, \tag{3.118}$$

where f_j, $f_{\mathrm{imp},j}$, and x_{j-1} are the corresponding parts of the compositions (3.115) of f and x. Equation (3.118) then determines the impulsive part $x_{\mathrm{imp},j}$ and the smooth part x_j of the solution provided that (E, A) satisfies one of the properties of Theorem 3.72.

In view of Corollary 3.26, a question would be whether we can treat jumps in the index and in the characteristic values between the intervals $\mathbb{I}_j$ of (3.26) within the framework of (impulsive smooth) distributions. The first difficulty here is that the set
$$\mathbb{T} = \mathbb{I} \setminus \bigcup_{j \in \mathbb{N}} \mathbb{I}_j$$
with closed interval $\mathbb{I}$ does not need to be countable. This means that it is not straightforward to define impulsive smooth distributions with impulses allowed at every point of $\mathbb{T}$. The second problem is that jumps in characteristic values may affect the solvability within the set of impulsive smooth distributions in an inconvenient way. See Exercises 22 and 23 and the following example for possible effects.

Example 3.76. Consider the initial value problem
$$tx = 0, \quad x(0^-) = 0$$

with solution space $\mathcal{C}_{\text{imp}}$. For this differential-algebraic equation neither the strangeness index is well defined nor Hypothesis 3.48 holds. A possible decomposition according to Corollary 3.26 is given by

$$\mathbb{R} = \overline{(-\infty, 0) \cup (0, \infty)}.$$

Thus, jumps of characteristic values may occur exactly at the same point where we allow impulsive behavior. Obviously, all distributions of the form $x = c\delta$ with $c \in \mathbb{C}$ solve the initial value problem. Moreover, there is no initial condition of the form (3.111) that fixes a unique solution. On the other hand, there is a unique solution of the initial value problem in $C^1(\mathbb{R}, \mathbb{C})$, namely $x = 0$. Hence, we may loose unique solvability when we turn to distributional solutions.

We conclude this section with the remark that up to now there do not exist any detailed investigations of phenomena caused by such jumps of the characteristic values, neither within the framework of classical solutions (where the above example seems to be trivial as long as we do not include inhomogeneities) nor within the framework of (impulsive smooth) distributions.

3.6 Control problems

As in the case of constant coefficients, we can also study linear control problems with variable coefficients via the analysis of differential-algebraic equations. Consider the system

$$E(t)\dot{x} = A(t)x + B(t)u + f(t), \tag{3.119a}$$
$$y = C(t)x + g(t), \tag{3.119b}$$

where $E, A \in C(\mathbb{I}, \mathbb{C}^{m,n})$, $B \in C(\mathbb{I}, \mathbb{C}^{m,l})$, $C \in C(\mathbb{I}, \mathbb{C}^{p,n})$, $f \in C(\mathbb{I}, \mathbb{C}^m)$ and $g \in C(\mathbb{I}, \mathbb{C}^p)$. Again, x represents the state, u the input, and y the output of the system. Typically one also has an initial condition of the form (3.2). Note that we use here the symbol u for the input (as is common in the control literature), since this will not lead to any confusion with the characteristic values $u_i, \hat{u}$ as in the canonical form (3.13) or in (3.64).

In order to analyze the properties of the system, we proceed in a different way compared with the constant coefficient case. We perform a behavior approach as it was suggested by Willems, see, e.g., [167].

Introducing a new vector

$$z = \begin{bmatrix} x \\ u \end{bmatrix},$$

we can rewrite the state equation (3.119a) as

$$\mathcal{E}(t)\dot{z} = \mathcal{A}(t)z + f(t), \tag{3.120}$$

with

$$\mathcal{E} = [\, E \;\; 0 \,], \quad \mathcal{A} = [\, A \;\; B \,]. \tag{3.121}$$

Note that the derivative of the original input u occurs only formally. In principle, we could also include the output equation in this system by introducing a behavior vector

$$z = \begin{bmatrix} x \\ u \\ y \end{bmatrix}.$$

Since the output equation (3.119b) together with its derivatives explicitly determines y and its derivatives, it is obvious that the output equation will not contribute to the analysis. So we keep it unchanged and only analyze the state equation.

System (3.120) is a general nonsquare linear differential-algebraic equation with variable coefficients for which we can apply the theory based on the strangeness index that we have developed in the previous sections.

If we carry out the analysis for this system and ignore the fact that z is composed of parts that may have quite different orders of differentiability (note that in practice the input u may not even be continuous), then we extract strangeness-free systems like (3.18) or (3.41). Observe that we cannot apply the theory based on the differentiation index, since in general it will not be defined for (3.120).

The associated inflated system has the form

$$M_\ell(t)\dot{z}_\ell = N_\ell(t)z_\ell + h_\ell(t), \tag{3.122}$$

where

$$\begin{aligned}
(M_\ell)_{i,j} &= \tbinom{i}{j}\mathcal{E}^{(i-j)} - \tbinom{i}{j+1}\mathcal{A}^{(i-j-1)}, \quad i, j = 0, \ldots, \ell, \\
(N_\ell)_{i,j} &= \begin{cases} \mathcal{A}^{(i)} & \text{for } i = 0, \ldots, \ell, \;\; j = 0, \\ 0 & \text{otherwise,} \end{cases} \\
(z_\ell)_j &= z^{(j)}, \quad j = 0, \ldots, \ell, \\
(h_\ell)_i &= f^{(i)}, \quad i = 0, \ldots, \ell,
\end{aligned} \tag{3.123}$$

and the analysis of Section 3.2 yields the following canonical form for (3.120).

Theorem 3.77. *Let the strangeness index μ as in (3.17) be well defined for the system given by $(\mathcal{E}, \mathcal{A})$ in (3.120). Setting*

$$\hat{a} = a_\mu, \quad \hat{d} = d_\mu, \quad \hat{v} = v_0 + \cdots + v_\mu, \tag{3.124}$$

the inflated pair (M_μ, N_μ) *as in* (3.122)*, associated with* $(\mathcal{E}, \mathcal{A})$*, has the following properties:*

1. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} M_\mu(t) = (\mu+1)m - \hat{a} - \hat{v}$*. This implies the existence of a smooth matrix function* Z *of size* $(\hat{\mu}+1)n \times (\hat{a}+\hat{v})$ *and pointwise maximal rank satisfying* $Z^H M_{\hat{\mu}} = 0$*.*

2. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} Z^H N_\mu [I_{n+l}\ 0\ \cdots\ 0]^H = \hat{a}$*. This implies that without loss of generality* Z *can be partitioned as* $Z = [\, Z_2\ Z_3\,]$*, with* Z_2 *of size* $(\mu+1)m \times \hat{a}$ *and* Z_3 *of size* $(\mu+1)m \times \hat{v}$*, such that* $\hat{A}_2 = Z_2^H N_\mu [I_{n+l} 0 \cdots 0]^H$ *has full row rank* $\hat{a}$ *and* $Z_3^H N_\mu [I_{n+l}\ 0\ \cdots\ 0]^H = 0$*. Furthermore, there exists a smooth matrix function* T_2 *of size* $(n+l) \times \hat{d}$*,* $\hat{d} = m - \hat{a}$*, and pointwise maximal rank satisfying* $\hat{A}_2 T_2 = 0$*.*

3. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} \mathcal{E}(t) T_2(t) = \hat{d}$*. This implies the existence of a smooth matrix function* Z_1 *of size* $m \times \hat{d}$ *and pointwise maximal rank satisfying* $\operatorname{rank} \hat{E}_1 = \hat{d}$ *with* $\hat{E}_1 = Z_1^H E$*.*

Furthermore, system (3.120) *has the same solution set as the strangeness-free system*

$$\begin{bmatrix} \hat{E}_1(t) \\ 0 \\ 0 \end{bmatrix} \dot{z} = \begin{bmatrix} \hat{A}_1(t) \\ \hat{A}_2(t) \\ 0 \end{bmatrix} z + \begin{bmatrix} \hat{f}_1(t) \\ \hat{f}_2(t) \\ \hat{f}_3(t) \end{bmatrix}, \tag{3.125}$$

where $\hat{A}_1 = Z_1^H \mathcal{A}$*,* $\hat{f}_1 = Z_1^H f$*,* $\hat{f}_i = Z_i^H h_\mu$ *for* $i = 2, 3$*.*

Note that the third block row in (3.125) has $\hat{v}$ equations, which in general is larger than v_μ as occurring in (3.41).

Remark 3.78. We have formulated Theorem 3.77 along the lines of Hypothesis 3.48 with quantities $\hat{a}$, $\hat{d}$, and $\hat{v}$. Instead of requiring a well-defined strangeness index for $(\mathcal{E}, \mathcal{A})$, we can obviously weaken the assumptions by requiring that the claims of Theorem 3.77 hold. This would correspond to an extension of Hypothesis 3.48 to general over- and underdetermined systems. We will discuss this extension in Section 4.3 for nonlinear differential-algebraic equations.

Considering the construction in Section 3.2 which led to the canonical form (3.125), we observe that the constructed submatrices $\hat{A}_1$ and $\hat{A}_2$ have been obtained from the block matrix

$$\begin{bmatrix} A & B \\ \dot{A} & \dot{B} \\ \vdots & \vdots \\ A^{(\mu)} & B^{(\mu)} \end{bmatrix}$$

by transformations from the left. This has two immediate consequences.

First of all this means that derivatives of the input function u are nowhere needed, just derivatives of the coefficient matrices, i.e., although formally the derivatives of u occur in the inflated pair, they are not used for the form (3.125), and hence, we do not need any additional smoothness requirements for the input function u.

Second, it follows from the construction of $\hat{A}_1$ and $\hat{A}_2$ that the partitioning into the part stemming from the original states x and the original controls u is not mixed up. Including again the output equation, from the inflated pair we have extracted the system

$$E_1(t)\dot{x} = A_1(t)x + B_1(t)u + \hat{f}_1(t), \quad \hat{d} \tag{3.126a}$$

$$0 = A_2(t)x + B_2(t)u + \hat{f}_2(t), \quad \hat{a} \tag{3.126b}$$

$$0 = \hat{f}_3(t), \quad \hat{v} \tag{3.126c}$$

$$y = C(t)x + g(t), \quad p \tag{3.126d}$$

where

$$E_1 = \hat{E}_1 \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad A_i = \hat{A}_i \begin{bmatrix} I_n \\ 0 \end{bmatrix}, \quad B_i = \hat{A}_i \begin{bmatrix} 0 \\ I_l \end{bmatrix}, \quad i = 1, 2.$$

Note also that the initial condition is not changed.

It turns out, however, that the characteristic quantities obtained in the canonical form (3.125) are not sufficient to analyze consistency and regularity of the control system (as defined in Definition 2.53) and to determine whether the system can be regularized by feedback, see [40]. To determine the structural information, we need to perform additional (global) equivalence transformations. In principle, we can obtain an equivalent differential-algebraic equation of the form (3.18), since (3.126) is strangeness-free as differential-algebraic equation for the unknown z. But this would mean that we must use transformations that mix x and u. To avoid that, we are restricted in the choice of possible equivalence transformations.

Setting

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) = \left(\begin{bmatrix} E_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_1 & B_1 & 0 \\ A_2 & B_2 & 0 \\ 0 & 0 & 0 \\ C & 0 & -I_p \end{bmatrix} \right) \tag{3.127}$$

to include the output equation, we first observe that E_1 has pointwise full row rank $\hat{d}$. A smooth transformation according to Theorem 3.9 gives

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) \overset{\text{new}}{\sim} \left(\left[\begin{array}{c|c|c|c} I_{\hat{d}} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|c|c|c} A_{11} & A_{12} & B_1 & 0 \\ \hline A_{21} & A_{22} & B_2 & 0 \\ 0 & 0 & 0 & 0 \\ C_1 & C_2 & 0 & -I_p \end{array}\right] \right).$$

Assuming that A_{22}, which is of size $\hat{a} \times (n - \hat{d})$, has constant rank $\hat{a} - \phi$, we can transform further to obtain

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) \overset{\text{new}}{\sim} \left(\left[\begin{array}{c|c|c|c|c} I_{\hat{d}} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|c|c|c|c} A_{11} & A_{12} & A_{13} & B_1 & 0 \\ \hline A_{21} & I_{\hat{a}-\phi} & 0 & B_2 & 0 \\ A_{31} & 0 & 0 & B_3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 \\ \hline C_1 & C_2 & C_3 & 0 & -I_p \end{array}\right] \right).$$

Since $(\hat{\mathcal{E}}, \hat{\mathcal{A}})$ is strangeness-free by construction and $[\, A_2 \;\; B_2 \,]$ has full row rank, B_3 of size $\phi \times l$ has full row rank. It follows that

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) \overset{\text{new}}{\sim} \left(\left[\begin{array}{c|cc|cc|c} I_{\hat{d}} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|c|c|cc|c} A_{11} & A_{12} & A_{13} & B_{11} & B_{12} & 0 \\ \hline A_{21} & I_{\hat{a}-\phi} & 0 & B_{21} & B_{22} & 0 \\ A_{31} & 0 & 0 & I_\phi & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline C_1 C_2 & C_3 & 0 & 0 & -I_p & \end{array}\right] \right).$$

For considerations concerning feedbacks, we perform an additional transformation in the last block row which corresponds to the output equation. Assuming C_3, which is of size $p \times (n - \hat{d} - \hat{a} + \phi)$, to have constant rank ω, we obtain

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) \overset{\text{new}}{\sim} \left(\left[\begin{array}{c|c|c|c|c|c|c} I_{\hat{d}} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{c|c|c|c|c|c|c|c} A_{11} & A_{12} & A_{13} & A_{14} & B_{11} & B_{12} & 0 & 0 \\ \hline A_{21} & I_{\hat{a}-\phi} & 0 & 0 & B_{21} & B_{22} & 0 & 0 \\ A_{31} & 0 & 0 & 0 & I_\phi & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline C_{11} & C_{12} & I_\omega & 0 & 0 & 0 & -I_\omega & 0 \\ C_{21} & C_{22} & 0 & 0 & 0 & 0 & 0 & -I_{p-\omega} \end{array}\right] \right).$$

Finally, we apply some block row and column eliminations, where we restrict the latter ones to those acting only on columns that belong to the same variable x, u, or y, and use a scaling as in the last step of the proof of Theorem 3.11. In this way,

we arrive at

$$(\hat{\mathcal{E}}, \hat{\mathcal{A}}) \overset{\text{new}}{\sim} \left(\left[\begin{array}{c|cc|cc|c|cc} I_{\hat{d}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \right.$$

$$\left. \left[\begin{array}{c|c|cc|c|c|cc} 0 & 0 & A_{13} & A_{14} & 0 & B_{12} & 0 & 0 \\ \hline 0 & I_{\hat{a}-\phi} & 0 & 0 & 0 & B_{22} & 0 & 0 \\ A_{31} & 0 & 0 & 0 & I_{\phi} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{\omega} & 0 & 0 & 0 & -I_{\omega} & 0 \\ C_{21} & C_{22} & 0 & 0 & 0 & 0 & 0 & -I_{p-\omega} \end{array}\right] \right),$$

where the fourth and sixth block column have widths $n - \hat{d} - \hat{a} + \phi - \omega$ and $l - \phi$, respectively. The corresponding control system now reads

$$\begin{aligned} \dot{x}_1 &= A_{13}(t)x_3 + A_{14}(t)x_4 + B_{12}(t)u_2 + f_1(t), && \hat{d} && \text{(3.128a)} \\ 0 &= x_2 + B_{22}(t)u_2 + f_2(t), && \hat{a} - \phi && \text{(3.128b)} \\ 0 &= A_{31}(t)x_1 + u_1 + f_3(t), && \phi && \text{(3.128c)} \\ 0 &= f_4(t), && \hat{v} && \text{(3.128d)} \\ y_1 &= x_3 + g_1(t), && \omega && \text{(3.128e)} \\ y_2 &= C_{21}(t)x_1 + C_{22}(t)x_2 + g_2(t), && p - \omega. && \text{(3.128f)} \end{aligned}$$

With this transformed system, we are now able to characterize consistency, regularity and regularizability via feedback, at least in the case where the solution x does not depend on derivatives of the input u. Extending Definition 2.53 in the obvious way to linear systems with variable coefficients, we obtain the following corollary.

Corollary 3.79. *Let the strangeness index μ be well defined for the system given by $(\mathcal{E}, \mathcal{A})$ in (3.120). Furthermore, let the quantities ϕ and ω defined by the above procedure be constant on $\mathbb{I}$. Then we have the following:*

1. *The system (3.119) is consistent if and only if $f_4 = 0$. The corresponding equations in (3.128c) describe redundancies in the system that can be omitted.*

2. *If the system is consistent and if $\phi = 0$, then for a given input function u an initial condition is consistent if and only if it implies (3.128b). Solutions of the corresponding initial value problem will in general not be unique.*

3. *The system is regular and strangeness-free (as a free system, i.e., $u = 0$) if and only if $\hat{v} = \phi = 0$ and $\hat{d} + \hat{a} = n$.*

Proof. Observe first that the global equivalence transformations that lead to (3.128) do not alter the solution behavior of the control system, since we do not mix the different variables x, u, and y. In particular, the solution sets of (3.119) and (3.128) are in one-to-one correspondence via pointwise nonsingular matrix functions applied to x, u, and y separately. Thus, it is sufficient to study (3.128).

If $\hat{v} \neq 0$ and $f_4 \neq 0$, then clearly the system has no solution, regardless how we choose the input function. Conversely if either $\hat{v} = 0$ or $f_4 = 0$, then we can determine an input u for which the system is solvable as follows. Setting $u_2 = 0$, $x_3 = 0$, and $x_4 = 0$, (3.128a) is an ordinary differential equation for x_1. Having thus fixed x_1, we obtain x_2 from (3.128b) and u_1 from (3.128c).

A consistent system with $\phi = 0$ reduces to (3.128a) and (3.128b), which is a strangeness-free differential-algebraic equation for every input function and (3.128b) represents the algebraic part. Thus, the second part follows by Corollary 3.18. Note that the solution will in general not be unique.

For the third part, we first assume that $\hat{v} = \phi = 0$ and $\hat{d} + \hat{a} = n$ in (3.128). In this case, (3.128) reduces to

$$\dot{x}_1 = B_{12}(t)u_2 + f_1(t), \quad 0 = x_2 + B_{22}(t)u_2 + f_2(t),$$

which is uniquely solvable for every input function u_2 and every inhomogeneity. Moreover, it is strangeness-free for $u_2 = 0$.

Conversely, let the system be regular and strangeness-free for $u = 0$. Since (3.128d) restricts the possible inhomogeneities, we must have $\hat{v} = 0$. If $\phi \neq 0$, then (3.128c) gives either consistency conditions for the inhomogeneity or a nonvanishing strangeness in contradiction to the assumptions. Hence, we must have $\phi = 0$. Finally, if $\hat{d} + \hat{a} \neq n$, then fixing u gives a nonsquare differential-algebraic equation for x resulting in either consistency conditions for the inhomogeneity or free solution components. Thus, we must also have $\hat{d} + \hat{a} = n$. □

As we have already seen in Section 2.5, we can also modify some of the system properties by feedback. The characteristic values in the canonical form (3.125), however, are invariant under proportional feedback.

Theorem 3.80. *Consider a linear variable coefficient control system of the form (3.119) and suppose that the strangeness index μ for (3.120) is well defined. Then, the characteristic values $\hat{d}$, $\hat{a}$, and $\hat{v}$ are invariant under proportional state feedback $u = F(t)x + w$ and proportional output feedback $u = F(t)y + w$.*

Proof. Proportional state feedback is just a change of basis in the behavior approach, i.e., in (3.120) we set

$$\begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} I_n & 0 \\ F(t) & I_l \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{u} \end{bmatrix}.$$

Proportional output feedback is an equivalence transformation in the more general behavior approach that includes also the output variables into the vector z, i.e., we set

$$\begin{bmatrix} x \\ u \\ y \end{bmatrix} = \begin{bmatrix} I_n & 0 & 0 \\ 0 & I_l & F(t) \\ 0 & 0 & I_p \end{bmatrix} \begin{bmatrix} \tilde{x} \\ \tilde{u} \\ \tilde{y} \end{bmatrix}$$

and premultiply by the nonsingular matrix

$$\begin{bmatrix} I_m & B(t)F(t) \\ 0 & I_p \end{bmatrix}.$$

It follows that the characteristic quantities μ, $\hat{d}$, $\hat{a}$, and $\hat{v}$ are invariant under both types of feedbacks. □

Corollary 3.81. *Let the assumptions of Corollary 3.79 hold. There exists a state feedback $u = F(t)x + w$ such that the closed loop system*

$$E(t)\dot{x} = (A(t) + B(t)F(t))x + B(t)w + f(t) \tag{3.129}$$

is regular (as a free system, i.e., $w = 0$) if and only if $\hat{v} = 0$ and $\hat{d} + \hat{a} = n$.

Proof. Observe that applying the feedback to the original system and then computing the corresponding reduced system gives the same as first reducing the original system and then applying the feedback. Thus, (3.129) is regular as a free system if and only if (3.126) with inserted feedback is regular and strangeness-free as a free system. Hence, it is sufficient to study the system in the form (3.128).

Since the output equation is not involved here, we can formally set $\omega = 0$ and x_3 does not appear in (3.128). Assuming $\hat{v} = 0$ and $\hat{d} + \hat{a} = n$ (implying that x_4 and u_1 have the same size), the feedback

$$u_1 = x_4 - A_{31}(t)x_1 + w_1, \quad u_2 = w_2$$

gives the closed loop system

$$\begin{aligned} \dot{x}_1 &= A_{14}(t)x_4 + B_{12}(t)w_2 + f_1(t), \\ 0 &= x_2 + B_{22}(t)w_2 + f_2(t), \\ 0 &= x_4 + w_1 + f_3(t), \end{aligned}$$

which is obviously regular and strangeness-free for $w = 0$. For the converse, observe that the condition $\hat{v} = 0$ is necessary, since otherwise the possible inhomogeneities are restricted. The condition $\hat{d} + \hat{a} = n$ is necessary, since otherwise for any given feedback F the closed loop system is nonsquare and either restricts the possible inhomogeneities or is not uniquely solvable. □

We also have the characterization when the system can be regularized by output feedback.

Corollary 3.82. *Let the assumptions of Corollary 3.79 hold. There exists an output feedback $u = F(t)y + w$ such that the closed loop system*

$$E(t)\dot{x} = (A(t) + B(t)F(t)C(t))x + B(t)w + f(t) + B(t)F(t)g(t) \tag{3.130}$$

is regular (as a free system) if and only if $\hat{v} = 0$, $\hat{d} + \hat{a} = n$, and $\phi = \omega$.

Proof. As in the proof of Corollary 3.81, we are allowed to study (3.128) instead of (3.119) when requiring the closed loop system to be regular and strangeness-free. If $\hat{v} = 0$, $\hat{d} + \hat{a} = n$, and $\phi = \omega$, then the unknown x_4 does not appear in (3.128) and u_1 and y_1 have the same size. The feedback

$$u_1 = y_1 + w_1, \quad u_2 = w_2$$

gives the closed loop system

$$\begin{aligned} \dot{x}_1 &= A_{13}(t)x_3 + B_{12}(t)w_2 + f_1(t), \\ 0 &= x_2 + B_{22}(t)w_2 + f_2(t), \\ 0 &= A_{31}(t)x_1 + x_3 + w_1 + f_3(t) + g_1(t), \end{aligned}$$

which is obviously regular and strangeness-free for $w = 0$. For the converse, again $\hat{v} = 0$ and $\hat{d} + \hat{a} = n$ are necessary. Moreover, for given

$$F = \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix},$$

partitioned conformly with (3.128), we obtain

$$\begin{bmatrix} 0 & 0 & A_{13} & A_{14} \\ 0 & I_{\hat{a}-\phi} & 0 & 0 \\ A_{31} & 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & B_{12} \\ 0 & B_{22} \\ I_\phi & 0 \end{bmatrix} \begin{bmatrix} F_{11} & F_{12} \\ F_{21} & F_{22} \end{bmatrix} \begin{bmatrix} 0 & 0 & I_\omega & 0 \\ C_{21} & C_{22} & 0 & 0 \end{bmatrix}$$
$$= \begin{bmatrix} B_{12}F_{22}C_{21} & B_{12}F_{22}C_{22} & A_{13} + B_{12}F_{21} & A_{14} \\ B_{22}F_{22}C_{21} & I_{\hat{a}-\phi} + B_{22}F_{22}C_{22} & B_{22}F_{21} & 0 \\ A_{31} + F_{12}C_{21} & F_{12}C_{22} & F_{11} & 0 \end{bmatrix}.$$

Thus, for the closed loop system to be regular and strangeness-free for $w = 0$, we must have that the square matrix function

$$\begin{bmatrix} I_{\hat{a}-\phi} + B_{22}F_{22}C_{22} & B_{22}F_{21} & 0 \\ F_{12}C_{22} & F_{11} & 0 \end{bmatrix}$$

is pointwise nonsingular. But this requires that the third block column is not present, hence $\phi = \omega$. □

Bibliographical remarks

The complete analysis for linear time varying systems has only recently been completed. It was observed in [95] that the concepts of linear constant coefficient systems cannot be directly generalized to variable coefficient systems. The analysis for the square and regular case was developed in different research groups using quite different concepts. An analysis based on matrix chains, leading to the *tractability index*, was given in the work of Griepentrog and März [100], [146], with a new and different formulation in [22], [148], [149], [150]. The concept of differentiation index and its generalizations based on derivative arrays were developed in the work of Campbell [44], [48], see also the monograph [29]. The concept of perturbation index was introduced in [105], see also [108]. The complete analysis, including the non-square case, as we have presented it in this chapter, was developed by the authors in [122], [123], [124], [126], [132].

The relationship between strangeness index, differentiation index, and perturbation index was derived in [126], [127]. A comparison between different index concepts and a discussion of many of the differences and difficulties with these concepts is given in [53], see also the preliminary work in [92]. Recent developments concerning the relationship between the tractability index and the strangeness index are given in [198].

The least squares approach was presented in [127] and the canonical form for control problems was derived in [183], [184] and extended in [132], see also [200]. The extension of this theory to generalized solutions is based on [97], [96], [123], [124] and was given in [179], [180].

Smooth factorizations of matrix functions as we have discussed them in Theorem 3.9 have been studied in [157], [191].

Exercises

1. For which inhomogeneities $f \in C(\mathbb{R}, \mathbb{C}^m)$ are the following differential-algebraic equations solvable with $x \in C^1(\mathbb{R}, \mathbb{C}^n)$?

   (a) $m = 1, n = 1, t\dot{x} = f(t)$,

   (b) $m = 1, n = 1, 0 = tx + f(t)$,

   (c) $m = 2, n = 2, \dot{x}_1 = f_1(t), 0 = tx_1 + f_2(t)$.

2. Prove Lemma 3.6.

3. Compute $\tilde{E} = PEQ$ and $\tilde{A} = PAQ - PE\dot{Q}$ for

$$E(t) = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad A(t) = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad P(t) = \begin{bmatrix} t & 1 \\ 1 & 0 \end{bmatrix}, \quad Q(t) = \begin{bmatrix} -1 & t \\ 0 & -1 \end{bmatrix}$$

and for

$$E(t) = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad A(t) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad P(t) = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}, \quad Q(t) = \begin{bmatrix} 0 & -1 \\ 1 & -t \end{bmatrix}.$$

Compare the so obtained pairs with those of the examples accompanying Section 3.1. Interpret the observations.

4. Let $(E, A) \sim (\tilde{E}, \tilde{A})$ with respect to global equivalence and assume that the corresponding matrix function Q is twice continuously differentiable. Show that then

$$\left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & I \\ -A & E \end{bmatrix} \right) \sim \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & I \\ -\tilde{A} & \tilde{E} \end{bmatrix} \right).$$

5. Repeat the computation of the (local) characteristic values r, a and s of Example 3.12 and Example 3.13 but use instead of a basis of cokernel E as columns of T' an arbitrary matrix T' for which $[\, T' \;\; T \,]$ is nonsingular. Verify that s is indeed independent of the choice of T' as proved in Remark 3.8.

6. Determine the (local) characteristic quantities (r, a, s) of

$$(E(t), A(t)) = \left(\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix}, \begin{bmatrix} -1 & -\eta t \\ 0 & -(1+\eta) \end{bmatrix} \right)$$

for every $t \in \mathbb{R}$ and $\eta \in \mathbb{R}$.

7. Determine the (local) characteristic quantities (r, a, s) of

$$(E(t), A(t)) = \left(\begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)$$

for every $t \in \mathbb{R}$.

8. Solve the inhomogeneous differential-algebraic equation

$$t\dot{x}_2 = x_1 + f_1(t), \quad 0 = x_2 + f_2(t)$$

belonging to the pair of matrix functions of Exercise 7. How smooth must f be if we require continuous differentiability only for x_2 as the structure suggests? What happens if we rewrite this system (using $\frac{d}{dt}(tx_2) = t\dot{x}_2 + x_2$) as

$$\frac{d}{dt}(tx_2) = x_1 + x_2 + f_1(t), \quad 0 = x_2 + f_2(t)?$$

9. For $E \in C(\mathbb{I}, \mathbb{C}^{n,n})$ we define the Drazin inverse E^D pointwise by $E^D(t) = E(t)^D$. Determine E^D for the matrix functions E of Exercise 6 and Exercise 7. What do you observe?

10. Determine the (global) characteristic quantities (r_i, a_i, s_i), $i = 0, \ldots, \mu$, of the pair of matrix functions given in Exercise 6 for every $\eta \in \mathbb{R}$.

11. Determine the (global) characteristic quantities (r_i, a_i, s_i), $i = 0, \ldots, \mu$, of the pairs of matrix functions belonging to the different blocks of the Kronecker canonical form given in Theorem 2.3.
12. Prove Lemma 3.20.
13. Let
$$(E, A) = \left(\begin{bmatrix} 0 & G_2 & E_{13} \\ 0 & 0 & G_1 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix} \right)$$
be sufficiently smooth with pointwise nonsingular matrix functions G_1 and G_2. Show that (E, A) is globally equivalent to a pair of constant matrix functions. Try to generalize this result.
14. Work out the details of the proof of Theorem 3.30.
15. Determine the (local) characteristic quantities $(\tilde{r}_\ell, \tilde{a}_\ell, \tilde{s}_\ell)$, $\ell = 0, \ldots, \mu$ of the inflated pairs (M_ℓ, N_ℓ) belonging to the pairs (E, A) of Example 3.1 and Example 3.2. From these, compute then the (global) characteristic quantities (r_i, a_i, s_i), $i = 0, \ldots, \mu$.
16. Let $A \in \mathbb{C}^{m,k}$ and $B \in \mathbb{C}^{k,n}$ have full row rank. Prove that then AB has full row rank.
17. Show that G from (3.23) with $c_i \neq 0$, $i = 0, \ldots, \mu$, and $\mu \geq 1$ satisfies $G(t)^\mu \neq 0$ and $G(t)^{\mu+1} = 0$ for all $t \in \mathbb{I}$.
18. Determine a possible pair $(\hat{E}, \hat{A})$ for the pair given in Exercise 6.
19. Show that the differentiation index is well defined for the pair given in Exercise 7 and determine a possible $(\hat{E}, \hat{A})$ as defined in (3.41) and (3.61).
20. Let (E, A) be given by
$$E(t) = \begin{bmatrix} 0 & 1 \\ 0 & \alpha t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \mathbb{I} = [-1, 1]$$
with real parameter α.
    (a) Determine a suitable decomposition of the form (3.26) and the corresponding (global) characteristic quantities.
    (b) Show that for $\alpha > 0$ the corresponding inflated matrix function M_2 is pointwise 1-full but not smoothly 1-full.
    (c) Analyze the 1-fullness of M_2 for arbitrary $\alpha \in \mathbb{R}$ and $t \in \mathbb{I}$.
21. Consider the optimal control problem
$$\frac{1}{2} \int_{\underline{t}}^{\overline{t}} (x(t)^H x(t) + u(t)^H u(t))\, dt = \min! \quad \text{s. t.}$$
$$\dot{x}(t) = A(t)x(t) + B(t)u(t) + f(t), \quad x(\underline{t}) = 0$$
in the notation of Theorem 3.61. Derive a complex version of the method of Lagrangian multipliers by rewriting the complex system of ordinary differential equations as a real system of doubled size and applying the real version of the method of Lagrangian multipliers.

22. Let (E, A) be given as in Example 7. Determine all $x_0 \in \mathbb{C}^2$ that are weakly consistent with a given $f \in \mathcal{C}^2_{\mathrm{imp}}$ satisfying $f_- = 0$. Compare the result (recalling that this example satisfies Hypothesis 3.48 due to Exercise 19) with the statement of Theorem 3.74.

23. Determine all solutions of

$$t\dot{x} = 0, \quad x_- = 0$$

in $\mathcal{C}_{\mathrm{imp}}$. Interpret the result.

Chapter 4
Nonlinear differential-algebraic equations

In this chapter, we study general nonlinear systems of differential-algebraic equations, i.e., equations of the form

$$F(t, x, \dot{x}) = 0. \tag{4.1}$$

For convenience, in this section we switch to real-valued problems. To obtain the results for complex-valued problems, we just have to analyze the real and imaginary part of the equation and the unknown separately. Moreover, we will first study the case $m = n$, i.e., the case where the number of equations equals the number of unknowns. Thus, we consider $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^n)$ with $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open. Again, we may have an initial condition

$$x(t_0) = x_0 \tag{4.2}$$

together with (4.1).

Note that in the following, for convenience of notation, we shall identify vectors $[x_1^T \ \cdots \ x_k^T]^T \in \mathbb{R}^n$ and tuples $(x_1, \ldots, x_k) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_k}$, $n_1 + \cdots + n_k = n$. In turn, we also write tuples as columns where it seems to be appropriate.

4.1 Existence and uniqueness of solutions

A typical approach to analyze nonlinear problems is to use the implicit function theorem in order to show that a (given) solution is locally unique. To apply the implicit function theorem, one must require that for the given solution the Fréchet derivative of the underlying nonlinear operator is *regular* (i.e., has a continuous inverse). Using the Fréchet derivative can be interpreted as *linearization* of the nonlinear problem. Of course, we expect that in the case of (4.1) this linearization will in general lead to linear differential-algebraic equations with variable coefficients. In view of the results of the previous chapter, we also expect that we must deal with inflated systems that are obtained by successive differentiation of the original equation with respect to time. We are then concerned with the question how these differentiated equations and their linearizations look like. In particular, we must investigate whether these two processes (differentiation and linearization) commute, i.e., whether it makes a difference first to differentiate and then to linearize or vice versa.

To answer this question, we assume that F in (4.1) is sufficiently smooth and that it induces an operator

$$\mathcal{F}\colon \mathbb{D} \to C^{\ell}(\mathbb{I}, \mathbb{R}^n)$$

with $\mathbb{D} \subseteq C^{\ell+1}(\mathbb{I}, \mathbb{R}^n)$ by

$$\mathcal{F}(x)(t) = F(t, x(t), \dot{x}(t))$$

for $t \in \mathbb{I}$, where $\mathbb{I}$ is a compact interval. Let $\mathcal{F}$ be Fréchet-differentiable. By definition, the Fréchet derivative

$$D\mathcal{F}(x)\colon C^{\ell+1}(\mathbb{I}, \mathbb{R}^n) \to C^{\ell}(\mathbb{I}, \mathbb{R}^n)$$

of $\mathcal{F}$ at $x \in \mathbb{D}$ is a continuous linear operator and satisfies

$$\mathcal{F}(x + \Delta x) = \mathcal{F}(x) + D\mathcal{F}(x)\Delta x + R(x, \Delta x) \tag{4.3}$$

for all $\Delta x \in C^{\ell+1}(\mathbb{I}, \mathbb{R}^n)$ in a neighborhood of the origin, where the remainder term has the property that

$$\frac{\|R(x, \Delta x)\|_{\ell}}{\|\Delta x\|_{\ell+1}} \to 0 \quad \text{for } \|\Delta x\|_{\ell+1} \to 0. \tag{4.4}$$

Here, $\|\cdot\|_{\ell}$ denotes the norm belonging to $C^{\ell}(\mathbb{I}, \mathbb{R}^n)$ as a Banach space, e.g.,

$$\|x\|_{\ell} = \sum_{i=0}^{\ell} \|x^{(i)}\|_0, \qquad \|x\|_0 = \sup_{t\in\mathbb{I}} \|x(t)\|_{\infty}. \tag{4.5}$$

Our aim then is to show that

$$D\mathcal{F}(x)\Delta x(t) = E(t)\Delta\dot{x}(t) - A(t)\Delta x(t) \tag{4.6}$$

holds, where

$$E(t) = F_{\dot{x}}(t, x(t), \dot{x}(t)), \qquad A(t) = -F_x(t, x(t), \dot{x}(t)). \tag{4.7}$$

For sufficiently smooth F, by Taylor expansion we have that

$$\begin{aligned} R(x, \Delta x)(t) &= \mathcal{F}(x + \Delta x)(t) - \mathcal{F}(x)(t) - D\mathcal{F}(x)\Delta x(t) \\ &= F(t, x(t) + \Delta x(t), \dot{x}(t) + \Delta\dot{x}(t)) - F(t, x(t), \dot{x}(t)) \\ &\quad - F_{\dot{x}}(t, x(t), \dot{x}(t))\Delta\dot{x}(t) - F_x(t, x(t), \dot{x}(t))\Delta x(t) \\ &= o(\|\Delta x(t)\|_{\infty} + \|\Delta\dot{x}(t)\|_{\infty}). \end{aligned}$$

Differentiating this relation then yields

$$\left(\tfrac{d}{dt}\right)^i R(x, \Delta x)(t) = o(\|\Delta x(t)\|_{\infty} + \|\Delta\dot{x}(t)\|_{\infty} + \cdots + \|\Delta x^{(i+1)}(t)\|_{\infty}), \quad i = 0, \ldots, \ell, \tag{4.8}$$

which immediately implies (4.4). Thus, (4.6) indeed defines the Fréchet derivative $D\mathcal{F}(x)$ of $\mathcal{F}$ at x. Moreover, defining

$$\left(\tfrac{d}{dt}\right)^i \mathcal{F} : \mathbb{D} \to C^{\ell-i}(\mathbb{I}, \mathbb{R}^n), \quad i = 0, \ldots, \ell \tag{4.9}$$

by

$$\left(\tfrac{d}{dt}\right)^i \mathcal{F}(x)(t) = \left(\tfrac{d}{dt}\right)^i F(t, x(t), \dot{x}(t)) \tag{4.10}$$

and $\left(\frac{d}{dt}\right)^i D\mathcal{F}$ accordingly, differentiation of (4.3) gives

$$\left(\tfrac{d}{dt}\right)^i \mathcal{F}(x + \Delta x) = \left(\tfrac{d}{dt}\right)^i \mathcal{F}(x) + \left(\tfrac{d}{dt}\right)^i D\mathcal{F}(x)\Delta x + \left(\tfrac{d}{dt}\right)^i R(x, \Delta x),$$

with

$$\left\| \left(\tfrac{d}{dt}\right)^i R(x, \Delta x) \right\|_\ell = o(\|\Delta x\|_{\ell+1}).$$

Thus, the Fréchet derivative of the differentiated operator is nothing else than the differentiated Fréchet derivative.

As for linear differential-algebraic equations, we gather the original equation and its derivatives up to order $\ell \in \mathbb{N}_0$ into an inflated system

$$F_\ell(t, x, \dot{x}, \ldots, x^{(\ell+1)}) = 0, \tag{4.11}$$

where F_ℓ has the form

$$\begin{aligned} F_\ell(t, x, \dot{x}, \ldots, x^{(\ell+1)}) &= \begin{bmatrix} F(t, x, \dot{x}) \\ \frac{d}{dt} F(t, x, \dot{x}) \\ \vdots \\ \left(\frac{d}{dt}\right)^\ell F(t, x, \dot{x}) \end{bmatrix} \\ &= \begin{bmatrix} F(t, x, \dot{x}) \\ F_t(t, x, \dot{x}) + F_x(t, x, \dot{x})\dot{x} + F_{\dot{x}}(t, x, \dot{x})\ddot{x} \\ \vdots \\ \vdots \end{bmatrix}, \end{aligned}$$

and define the Jacobians

$$\begin{aligned} M_\ell(t, x, \dot{x}, \ldots, x^{(\ell+1)}) &= F_{\ell;\dot{x},\ldots,x^{(\ell+1)}}(t, x, \dot{x}, \ldots, x^{(\ell+1)}), \\ N_\ell(t, x, \dot{x}, \ldots, x^{(\ell+1)}) &= -(F_{\ell;x}(t, x, \dot{x}, \ldots, x^{(\ell+1)}), 0, \ldots, 0) \end{aligned} \tag{4.12}$$

according to (3.28). In order to examine the structure of these Jacobians, let $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\ell+1)}) \in \mathbb{R}^{(\ell+2)n+1}$ be given with $(t_0, x_0, \dot{x}_0) \in \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}$. Defining a polynomial x of degree at most $\ell + 1$ by

$$x(t) = \sum_{i=0}^{\ell+1} \frac{x_0^{(i)}}{i!}(t - t_0)^i, \tag{4.13}$$

we obviously have that

$$x^{(i)}(t_0) = x_0^{(i)}, \quad i = 0, \dots, \ell + 1.$$

Moreover, there exists a small (relative) neighborhood $\mathbb{I}_\varepsilon$ of t_0 such that

$$(t, x(t), \dot{x}(t)) \in \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}$$

for all $t \in \mathbb{I}_\varepsilon$. By the preceding discussion, it follows that M_ℓ and N_ℓ have precisely the structure given by the formulas (3.29) when we use (4.7) with x given by (4.13).

In the linear case, we have shown that Hypothesis 3.48 is the correct way to define a *regular* differential-algebraic equation. Regularity here is to be understood as follows. A linear problem satisfying Hypothesis 3.48 fixes a strangeness-free differential-algebraic equation (3.60). The underlying differential-algebraic operator D, as considered in Section 3.4 with appropriately chosen spaces, is invertible and the inverse is continuous. The question then is how we can transfer these properties to the nonlinear case. A first idea could be to require that all possible linearizations of (4.1) satisfy Hypothesis 3.48 with the same characteristic values $\hat{\mu}$, $\hat{a}$ and $\hat{d}$. See [53] for such an approach on the basis of the differentiation index and 1-fullness. But, as the following example shows, this property is not invariant under simple transformations of the original problem.

Example 4.1. The differential-algebraic equation

$$\begin{bmatrix} 1 & 0 \\ \dot{x}_1 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_2 - x_1 \\ x_2 \end{bmatrix} = 0$$

has the unique solution $(x_1, x_2) = (0, 0)$. This system may be seen as a regularly transformed linear differential-algebraic equation satisfying Hypothesis 3.48 with $\hat{\mu} = 1$, $\hat{a} = 2$, and $\hat{d} = 0$. Its linearization is given by

$$\begin{bmatrix} 0 & 1 \\ \dot{x}_2 - x_1 & \dot{x}_1 \end{bmatrix} \begin{bmatrix} \Delta\dot{x}_1 \\ \Delta\dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \dot{x}_1 & -1 \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta x_2 \end{bmatrix} + \begin{bmatrix} x_1 - \dot{x}_2 \\ -\dot{x}_1\dot{x}_2 + x_1\dot{x}_1 - x_2 \end{bmatrix}.$$

For the linearization along the exact solution, we find that M_ℓ is always rank deficient, whereas for the linearization along the perturbed solution $(x_1, x_2) = (\varepsilon, 0)$ with $\varepsilon \neq 0$ it always has full rank. Thus, Hypothesis 3.48 does not hold uniformly for all linearizations in a neighborhood of the exact solution.

In general, we must therefore expect that away from the solution the constant rank assumptions as required in Hypothesis 3.48 may not hold for the linearization. One may argue that the transformation in Example 4.1 has no physical meaning. But from the mathematical point of view, the assumptions that we require should be invariant under all reversible (smooth) transformations which carry over the

given problem into a problem of the same structure. Of course, the class of such transformations is larger than in the linear case.

We here propose the following generalization of Hypothesis 3.48 to nonlinear problems. Recall also the interpretation of Hypothesis 3.48 given in Remark 3.55, which in principle also applies to the nonlinear case.

Hypothesis 4.2. *There exist integers μ, a, and d such that the set*

$$\mathbb{L}_\mu = \{(t, x, \dot{x}, \ldots, x^{(\mu+1)}) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = 0\} \quad (4.14)$$

associated with F is nonempty and such that for every $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$, there exists a (sufficiently small) neighborhood in which the following properties hold:

1. *We have* rank $M_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = (\mu+1)n - a$ *on $\mathbb{L}_\mu$ such that there exists a smooth matrix function Z_2 of size $(\mu+1)n \times a$ and pointwise maximal rank, satisfying $Z_2^T M_\mu = 0$ on $\mathbb{L}_\mu$.*

2. *We have* rank $\hat{A}_2(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = a$, *where $\hat{A}_2 = Z_2^T N_\mu [I_n \; 0 \; \cdots \; 0]^T$ such that there exists a smooth matrix function T_2 of size $n \times d$, $d = n - a$, and pointwise maximal rank, satisfying $\hat{A}_2 T_2 = 0$.*

3. *We have* rank $F_{\dot{x}}(t, x, \dot{x}) T_2(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = d$ *such that there exists a smooth matrix function Z_1 of size $n \times d$ and pointwise maximal rank, satisfying* rank $\hat{E}_1 T_2 = d$, *where $\hat{E}_1 = Z_1^T F_{\dot{x}}$.*

The local existence of the matrix functions Z_2, T_2, and Z_1 follows from the following (local) nonlinear version of Theorem 3.9.

Theorem 4.3. *Let $E \in C^\ell(\mathbb{D}, \mathbb{R}^{m,n})$, $\ell \in \mathbb{N}_0 \cup \{\infty\}$, with* rank $E(x) = r$ *for all $x \in \mathbb{M} \subseteq \mathbb{D}$, $\mathbb{D} \subseteq \mathbb{R}^k$ open. For every $\hat{x} \in \mathbb{M}$ there exists a sufficiently small neighborhood $\mathbb{V} \subseteq \mathbb{D}$ of $\hat{x}$ and matrix functions $T \in C^\ell(\mathbb{V}, \mathbb{R}^{n,n-r})$, $Z \in C^\ell(\mathbb{V}, \mathbb{R}^{m,m-r})$, with pointwise orthonormal columns such that*

$$ET = 0, \quad Z^T E = 0 \quad (4.15)$$

on $\mathbb{M}$.

Proof. For $\hat{x} \in \mathbb{M}$, using the singular value decomposition, there exist orthogonal matrices $\hat{U} \in \mathbb{R}^{m,m}$, $\hat{V} \in \mathbb{R}^{n,n}$ with

$$\hat{U}^T E(\hat{x}) \hat{V} = \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & 0 \end{bmatrix}$$

and $\hat{\Sigma} \in \mathbb{R}^{r,r}$ nonsingular.

Splitting $\hat{V} = [\ \hat{T}'\ \ \hat{T}\ ]$ according to the above block structure, we consider the linear system of equations

$$\begin{bmatrix} \hat{Z}'^T E(x) \\ \hat{T}^T \end{bmatrix} T = \begin{bmatrix} 0 \\ I_{n-r} \end{bmatrix}. \tag{4.16}$$

Since

$$\begin{bmatrix} \hat{Z}'^T E(\hat{x}) \\ \hat{T}^T \end{bmatrix} [\ \hat{T}'\ \ \hat{T}\ ] = \begin{bmatrix} \hat{\Sigma} & 0 \\ 0 & I_{n-r} \end{bmatrix},$$

there exists a neighborhood of $\hat{x}$, where the coefficient matrix in (4.16) is invertible. Thus, (4.16) locally defines a matrix function T via

$$T(x) = \begin{bmatrix} \hat{Z}'^T E(x) \\ \hat{T}^T \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ I_{n-r} \end{bmatrix},$$

which obviously has full column rank. Moreover, by construction, $[\ \hat{T}'\ \ T(x)\ ]$ is nonsingular. By the definition of T, it follows that

$$\hat{U}^T E(x)[\ \hat{T}'\ \ T(x)\ ] = \begin{bmatrix} \hat{Z}'^T E(x)\hat{T}' & 0 \\ \hat{Z}^T E(x)\hat{T}' & \hat{Z}^T E(x)T(x) \end{bmatrix}$$

such that

$$\operatorname{rank} E(x) = \operatorname{rank} \hat{Z}'^T E(x)\hat{T}' + \operatorname{rank} \hat{Z}^T E(x)T(x) = r + \operatorname{rank} \hat{Z}^T E(x)T(x).$$

If $x \in \mathbb{M}$, then we have $\operatorname{rank} E(x) = r$ implying that $\hat{Z}^T E(x)T(x) = 0$. Together with $\hat{Z}'^T E(x)T(x) = 0$, this gives $E(x)T(x) = 0$. Orthonormality of the columns of $T(x)$ can be obtained by the smooth Gram–Schmidt orthonormalization process.

The corresponding result for Z follows by considering the pointwise transpose E^T. □

Definition 4.4. Given a differential-algebraic equation as in (4.1), the smallest value of μ such that F satisfies Hypothesis 4.2 is called the *strangeness index* of (4.1). If $\mu = 0$, then the differential-algebraic equation is called strangeness-free.

Recalling Section 3.3, this definition of the strangeness index for a nonlinear differential-algebraic equation is a straightforward generalization of the strangeness index for a linear differential-algebraic equation. Note that although the quantities μ, a, and d of Hypothesis 4.2 correspond to $\hat{\mu}$, $\hat{a}$, and $\hat{d}$ of Hypothesis 3.48, for convenience of notation, we omit the hats in this chapter.

Example 4.5. For the nonlinear problem from Example 4.1 written in the form

$$\dot{x}_2 - x_1 = 0, \quad \dot{x}_1\dot{x}_2 - \dot{x}_1 x_1 + x_2 = 0,$$

differentiation yields

$$\ddot{x}_2 - \dot{x}_1 = 0, \quad \ddot{x}_1\dot{x}_2 + \dot{x}_1\ddot{x}_2 - \ddot{x}_1 x_1 - \dot{x}_1^2 + \dot{x}_2 = 0.$$

Thus, for $\mu = 1$ we find that

$$[\,-N_1[I_2\ 0]^T\ \ M_1\,](t,x,\dot{x},\ddot{x}) = \left[\begin{array}{cc|cc|cc} -1 & 0 & 0 & 1 & 0 & 0 \\ -\dot{x}_1 & 1 & \dot{x}_2 - x_1 & \dot{x}_1 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & 0 & 1 \\ -\ddot{x}_1 & 0 & \ddot{x}_2 - 2\dot{x}_1 & \ddot{x}_1 + 1 & \dot{x}_2 - x_1 & \dot{x}_1 \end{array}\right]$$

and

$$\begin{aligned}\mathbb{L}_1 &= \{(t,x_1,x_2,\dot{x}_1,\dot{x}_2,\ddot{x}_1,\ddot{x}_2) \mid x_1 = \dot{x}_2,\ x_2 = \dot{x}_1 x_1 - \dot{x}_1\dot{x}_2,\\ &\qquad\qquad \dot{x}_1 = \ddot{x}_2,\ \dot{x}_2 = \ddot{x}_1 x_1 + \dot{x}_1^2 - \ddot{x}_1\dot{x}_2 - \dot{x}_1\ddot{x}_2\}\\ &= \{(t,x_1,x_2,\dot{x}_1,\dot{x}_2,\ddot{x}_1,\ddot{x}_2) \mid x_1 = 0,\ x_2 = 0,\ \dot{x}_1 = \ddot{x}_2,\ \dot{x}_2 = 0\}.\end{aligned}$$

Hence, we have rank $M_1 = 2$ and rank$[\,-N_1[\,I_2\ \ 0\,]^T\ \ M_1\,] = 4$ on $\mathbb{L}_1$ and the problem satisfies Hypothesis 4.2 with $\mu = 1$, $a = 2$, and $d = 0$. Note that $\mathbb{L}_1$ can be parameterized by three scalars, e.g., by $(t, \ddot{x}_1, \ddot{x}_2)$.

Before we discuss implications of Hypothesis 4.2, we show that Hypothesis 4.2 is indeed invariant under a large class of equivalence transformations. These results will also motivate why we require uniform characteristic values μ, a and d only on the set $\mathbb{L}_\mu$. We begin with the nonlinear analogues of changes of bases.

Lemma 4.6. *Let F as in (4.1) satisfy Hypothesis 4.2 with characteristic values μ, a and d, and let $\tilde{F}$ be given by*

$$\tilde{F}(t,\tilde{x},\dot{\tilde{x}}) = F(t,x,\dot{x}), \quad x = Q(t,\tilde{x}), \quad \dot{x} = Q_t(t,\tilde{x}) + Q_{\tilde{x}}(t,\tilde{x})\dot{\tilde{x}}, \qquad (4.17)$$

with sufficiently smooth $Q \in C(\mathbb{I}\times\mathbb{R}^n,\mathbb{R}^n)$, where $Q(t,\,\cdot\,)$ is bijective for every $t \in \mathbb{I}$ and the Jacobian $Q_{\tilde{x}}(t,\tilde{x})$ is nonsingular for every $(t,\tilde{x}) \in \mathbb{I}\times\mathbb{R}^n$. Then, $\tilde{F}$ satisfies Hypothesis 4.2 with characteristic values μ, a and d.

Proof. Let $\mathbb{L}_\mu$ and $\tilde{\mathbb{L}}_\mu$ be the corresponding sets as defined by Hypothesis 4.2. Since $Q(t,\,\cdot\,)$ is bijective and smooth, we have

$$\tilde{z} = (t,\tilde{x},\dot{\tilde{x}},\ldots,\tilde{x}^{(\mu+1)}) \in \tilde{\mathbb{L}}_\mu \iff z = (t,x,\dot{x},\ldots,x^{(\mu+1)}) \in \mathbb{L}_\mu.$$

Setting

$$\tilde{E}(t,\tilde{x},\dot{\tilde{x}}) = \tilde{F}_{\dot{\tilde{x}}}(t,\tilde{x},\dot{\tilde{x}}), \quad \tilde{A}(t,\tilde{x},\dot{\tilde{x}}) = -\tilde{F}_{\tilde{x}}(t,\tilde{x},\dot{\tilde{x}})$$

and using (4.17), we get

$$\tilde{E}(t,\tilde{x},\dot{\tilde{x}}) = F_{\dot{x}}(t,x,\dot{x})Q_{\tilde{x}}(t,\tilde{x})$$

and

$$\tilde{A}(t,\tilde{x},\dot{\tilde{x}}) = -F_x(t,x,\dot{x})Q_{\tilde{x}}(t,\tilde{x}) - F_{\dot{x}}(t,x,\dot{x})(Q_{t\tilde{x}}(t,\tilde{x}) + Q_{\tilde{x}\tilde{x}}(t,\tilde{x})\dot{\tilde{x}}).$$

Together with (4.7), we can write this as

$$[\ \tilde{E}(t,\tilde{x},\dot{\tilde{x}})\ \ \tilde{A}(t,\tilde{x},\dot{\tilde{x}})\ ] = [\ E(t,x,\dot{x})\ \ A(t,x,\dot{x})\ ] \begin{bmatrix} Q_{\tilde{x}}(t,\tilde{x}) & -\frac{d}{dt}Q_{\tilde{x}}(t,\tilde{x}) \\ 0 & Q_{\tilde{x}}(t,\tilde{x}) \end{bmatrix}.$$

This relation has exactly the form of global equivalence (3.3). Since the corresponding inflated pairs (M_μ, N_μ) and $(\tilde{M}_\mu, \tilde{N}_\mu)$ are built according to (3.29), we get

$$[\ \tilde{M}_\mu(\tilde{z})\ \ \tilde{N}_\mu(\tilde{z})\ ] = [\ M_\mu(z)\ \ N_\mu(z)\ ] \begin{bmatrix} \Theta_\mu(\tilde{z}) & -\Psi_\mu(\tilde{z}) \\ 0 & \Theta_\mu(\tilde{z}) \end{bmatrix}$$

according to (3.35), where we only must replace Q by $Q_{\tilde{x}}(t,\tilde{x})$ in (3.34). The invariance of Hypothesis 4.2 follows then by the same proof as the invariance of Hypothesis 3.48. □

Lemma 4.7. *Let F as in (4.1) satisfy Hypothesis 4.2 with characteristic values μ, a and d, and let $\tilde{F}$ be given by*

$$\tilde{F}(t,x,\dot{x}) = P(t,x,\dot{x},F(t,x,\dot{x})), \tag{4.18}$$

with sufficiently smooth $P \in C(\mathbb{I}\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n,\mathbb{R}^n)$, where $P(t,x,\dot{x},\cdot)$ is bijective with $P(t,x,\dot{x},0) = 0$ and $P_w(t,x,\dot{x},\cdot)$ nonsingular (where P_w denotes the derivative of P with respect to the fourth argument) for every $(t,x,\dot{x}) \in \mathbb{I}\times\mathbb{R}^n\times\mathbb{R}^n$. Then, $\tilde{F}$ satisfies Hypothesis 4.2 with characteristic values μ, a and d.

Proof. By induction, it follows that $\tilde{\mathbb{L}}_\mu = \mathbb{L}_\mu$ for the corresponding sets as defined by Hypothesis 4.2. Setting

$$\tilde{E}(t,x,\dot{x}) = \tilde{F}_{\dot{x}}(t,x,\dot{x}), \quad \tilde{A}(t,x,\dot{x}) = -\tilde{F}_x(t,x,\dot{x}),$$

we get

$$\tilde{E}(t,x,\dot{x}) = P_{\dot{x}}(t,x,\dot{x},F(t,x,\dot{x})) + P_w(t,x,\dot{x},F(t,x,\dot{x}))F_{\dot{x}}(t,x,\dot{x})$$

and

$$\tilde{A}(t,x,\dot{x}) = -P_x(t,x,\dot{x},F(t,x,\dot{x})) - P_w(t,x,\dot{x},F(t,x,\dot{x}))F_x(t,x,\dot{x}).$$

In contrast to the proof of Lemma 4.6, we do not get a relation in the form of a global equivalence. But if we restrict our considerations to the set $\mathbb{L}_\mu$, then we obtain

$$[\ \tilde{E}(t,x,\dot{x})\ \ \tilde{A}(t,x,\dot{x})\ ] = P_w(t,x,\dot{x},0)[\ E(t,x,\dot{x})\ \ A(t,x,\dot{x})\ ]$$

on $\mathbb{L}_\mu$. By induction with respect to successive differentiation according to (3.35), we obtain

$$[\ \tilde{M}_\mu(z)\ \ \tilde{N}_\mu(z)\ ] = \Pi_\mu(z)[\ M_\mu(z)\ \ N_\mu(z)\ ],$$

with $z = (t, x, \dot{x}, \ldots, x^{(\mu+1)}) \in \mathbb{L}_\mu$ for the corresponding inflated pairs. In (3.34), we only must replace P by $P_w(t, x, \dot{x}, 0)$. Again, the invariance of Hypothesis 4.2 follows by the same proof as the invariance of Hypothesis 3.48. □

For general nonlinear differential-algebraic equations, there is a third type of transformations which does not alter the solution behavior, namely making the problem autonomous.

Lemma 4.8. *Let F as in (4.1) satisfy Hypothesis 4.2 with characteristic values μ, a and d, and let $\tilde{F}$ be given by*

$$\tilde{F}(\tilde{x}, \dot{\tilde{x}}) = \begin{bmatrix} F(t, x, \dot{x}) \\ \dot{t} - 1 \end{bmatrix}, \quad \tilde{x} = \begin{bmatrix} t \\ x \end{bmatrix}, \quad \dot{\tilde{x}} = \begin{bmatrix} \dot{t} \\ \dot{x} \end{bmatrix}. \tag{4.19}$$

Then, $\tilde{F}$ satisfies Hypothesis 4.2 with characteristic values μ, a and $d+1$.

Proof. Let $\mathbb{L}_\mu$ and $\tilde{\mathbb{L}}_\mu$ be the corresponding sets as defined by Hypothesis 4.2 and let (M_μ, N_μ) and $(\tilde{M}_\mu, \tilde{N}_\mu)$ be the corresponding inflated pairs. Since $\dot{t} = 1$ and $t^{(i)} = 0$ for $i > 1$, it follows by induction that

$$\tilde{z} = (t, x, \dot{t}, \dot{x}, \ldots, t^{(\mu+1)}, x^{(\mu+1)}) \in \tilde{\mathbb{L}}_\mu \iff z = (t, x, \dot{x}, \ldots, x^{(\mu+1)}) \in \mathbb{L}_\mu.$$

Moreover, the rows of $\tilde{M}_\mu$ belonging to $\dot{t} = 1$ and $t^{(i)} = 0$ for $i > 1$ cannot lead to a rank deficiency. Removing the corresponding rows and columns from $\tilde{M}_\mu$ just gives M_μ. Hence, the corresponding matrix functions $\hat{A}_2$ and $\tilde{A}_2$ defined in the second part of Hypothesis 4.2 differ only in an additional column w of $\tilde{A}_2$ belonging to t, i.e.,

$$\tilde{A}_2 = [\ w\ \ \hat{A}_2\ ].$$

Since $\hat{A}_2$ has full row rank a, the same holds for $\tilde{A}_2$. Using the Moore–Penrose pseudoinverse $\hat{A}_2^+$ of $\hat{A}_2$ and observing that $\hat{A}_2\hat{A}_2^+ = I$ due to the full row rank of $\hat{A}_2$, we get

$$\tilde{T}_2 = \begin{bmatrix} 0 & -1 \\ T_2 & \hat{A}_2^+ w \end{bmatrix}$$

and thus

$$\begin{aligned} \operatorname{rank}(\tilde{F}_{\dot{t},\dot{x}} \tilde{T}_2) &= \operatorname{rank} \left(\begin{bmatrix} 0 & F_{\dot{x}} \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & -1 \\ T_2 & \hat{A}_2^+ w \end{bmatrix} \right) \\ &= \operatorname{rank} \begin{bmatrix} F_{\dot{x}} T_2 & F_{\dot{x}} \hat{A}_2^+ w \\ 0 & -1 \end{bmatrix} = d + 1. \end{aligned}$$

□

We now consider the implications of Hypothesis 4.2 for the solvability of (4.1). In the linear case, Hypothesis 3.48 allowed for an index reduction in one step to a *regular* differential-algebraic equation, i.e., to a problem whose underlying linear operator has a continuous inverse. The following discussion shows that a similar procedure is possible even in the nonlinear case. The first immediate consequence of Hypothesis 4.2 is the following.

Lemma 4.9. *If F as in (4.1) satisfies Hypothesis 4.2 with characteristic values μ, a and d, then*

$$\operatorname{rank} F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}}(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = (\mu + 1)n \tag{4.20}$$

for all $(t, x, \dot{x}, \ldots, x^{(\mu+1)}) \in \mathbb{L}_\mu$.

Proof. The claim follows directly from

$$F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}} = [\, -N_\mu [I_n\ 0\ \cdots\ 0]^T \ \ M_\mu\,],$$

since the second assumption of Hypothesis 4.2 requires the right hand side to have full row rank. □

Let now $z_{\mu,0} = (t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ be fixed. Recall that we have $F_\mu(z_{\mu,0}) = 0$ by definition. Since $\mathbb{L}_\mu$ is contained in $\mathbb{R}^{(\mu+2)n+1}$, the Jacobian $F_{\mu;t,x,\dot{x},\ldots,x^{(\mu+1)}}(z_{\mu,0})$ has the size $(\mu + 1)n \times (\mu + 2)n + 1$ and has full row rank due to Lemma 4.9. Hence, we can select $n + 1$ columns such that removing them from the Jacobian does not lead to a rank deficiency. Again by Lemma 4.9, the column belonging to t can always be removed. Since

$$\operatorname{corank} M_\mu(z_{\mu,0}) = a, \quad \operatorname{rank} Z_2(z_{\mu,0})^T N_\mu(z_{\mu,0})[I_n\ 0\ \cdots\ 0]^T = a,$$

we can (without loss of generality) partition x as $x = (x_1, x_2)$, where x_1 has $d = n - a$ variables and x_2 has a variables, such that discarding the columns of $F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}}$ belonging to x_1 does not lead to a rank drop. It follows then from (4.20) that $Z_2^T F_{\mu;x_2}$ is nonsingular. The remaining a variables, say p, associated with columns of $F_{\mu;t,x,\dot{x},\ldots,x^{(\mu+1)}}(z_{\mu,0})$ that we can remove without having a rank drop, must then be chosen from $(\dot{x}, \ldots, x^{(\mu+1)})$.

Let $(t_0, x_{1,0}, p_0)$ be that part of $z_{\mu,0}$ that corresponds to the variables (t, x_1, p). The implicit function theorem then implies that there exists a neighborhood $\mathbb{V} \subseteq \mathbb{R}^{n+1}$ of $(t_0, x_{1,0}, p_0)$, without loss of generality an open ball with radius $\varepsilon > 0$ and center $(t_0, x_{1,0}, p_0)$, and a neighborhood $\tilde{\mathbb{U}}$ of $z_{\mu,0}$ such that

$$\mathbb{U} = \mathbb{L}_\mu \cap \tilde{\mathbb{U}} = \{\theta(t, x_1, p) \mid (t, x_1, p) \in \mathbb{V}\},$$

where $\theta : \mathbb{V} \to \mathbb{U}$ is a diffeomorphism, i.e., a differentiable homeomorphism.

Hence, the set $\mathbb{L}_\mu$ is locally diffeomorphic to an open ball in $\mathbb{R}^{n+1}$. In particular, it can be (locally) parameterized by $n+1$ scalars. We therefore call $\mathbb{L}_\mu$ a *manifold* and $n+1$ its *dimension*. In the present context, this more intuitive definition of a manifold will be sufficient. A more general definition will follow in Section 4.5 when we discuss differential-algebraic equations on manifolds.

With these preparations, we can state the above observation on the local properties of $\mathbb{L}_\mu$ as follows.

Corollary 4.10. *If F as in (4.1) satisfies Hypothesis 4.2 with characteristic values μ, a and d, then the set $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)n+1}$ forms a (smooth) manifold of dimension $n+1$.*

It follows from Corollary 4.10 that the equation

$$F_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = 0$$

can be locally solved according to

$$(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = \theta(t, x_1, p).$$

In particular, there exist locally defined functions $\mathcal{G}$ and $\mathcal{H}$ such that

$$F_\mu(t, x_1, \mathcal{G}(t, x_1, p), \mathcal{H}(t, x_1, p)) = 0 \tag{4.21}$$

for all $(t, x_1, p) \in \mathbb{V}$. Setting $y = (\dot{x}, \ldots, x^{(\mu+1)})$, it follows with Z_2 as defined by Hypothesis 4.2 that (omitting arguments)

$$\tfrac{d}{dp}(Z_2^T F_\mu) = (Z_{2;x_2}^T F_\mu + Z_2^T F_{\mu;x_2})\mathcal{G}_p + (Z_{2;y}^T F_\mu + Z_2^T F_{\mu;y})\mathcal{H}_p = Z_2^T F_{\mu;x_2}\mathcal{G}_p = 0$$

on $\mathbb{V}$, since $F_\mu = 0$ and $Z_2^T F_{\mu;y} = Z_2^T M_\mu = 0$ on $\mathbb{V}$. By construction, the variables in x_2 were selected such that $Z_2^T F_{\mu;x_2}$ is nonsingular. Hence,

$$\mathcal{G}_p(t, x_1, p) = 0$$

for all $(t, x_1, p) \in \mathbb{V}$, implying the existence of a function $\mathcal{R}$ such that

$$x_2 = \mathcal{G}(t, x_1, p) = \mathcal{G}(t, x_1, p_0) = \mathcal{R}(t, x_1),$$

and

$$F_\mu(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p)) = 0$$

for all $(t, x_1, p) \in \mathbb{V}$. We then get that (omitting arguments)

$$\begin{aligned}\tfrac{d}{dx_1}(Z_2^T F_\mu) &= (Z_{2;x_1}^T F_\mu + Z_2^T F_{\mu;x_1}) + (Z_{2;x_2}^T F_\mu + Z_2^T F_{\mu;x_2})\mathcal{R}_{x_1} \\ &\quad + (Z_{2;y}^T F_\mu + Z_2^T F_{\mu;y})\mathcal{H}_{x_1} \\ &= Z_2^T F_{\mu;x_1} + Z_2^T F_{\mu;x_2}\mathcal{R}_{x_1} = -Z_2^T N_\mu [I_n\, 0\, \cdots\, 0]^T \begin{bmatrix} I \\ \mathcal{R}_{x_1} \end{bmatrix} = 0\end{aligned}$$

on $\mathbb{V}$. Following Hypothesis 4.2, we can therefore choose

$$T_2(t, x_1) = \begin{bmatrix} I \\ \mathcal{R}_{x_1}(t, x_1) \end{bmatrix}$$

with the effect that Z_1 depends only on $(t, x, \dot{x})$. In fact, due to the requirement of a maximal rank and the local considerations here, we can even choose Z_1 to be constant.

Setting

$$\begin{aligned} \hat{F}_1(t, x_1, x_2, \dot{x}_1, \dot{x}_2) &= Z_1^T F(t, x_1, x_2, \dot{x}_1, \dot{x}_2), \\ \hat{F}_2(t, x_1, x_2) &= Z_2^T F_\mu(t, x_1, x_2, \mathcal{H}(t, x_1, p_0)), \end{aligned} \tag{4.22}$$

we then consider the *reduced differential-algebraic equation*

$$\hat{F}(t, x, \dot{x}) = \begin{bmatrix} \hat{F}_1(t, x, \dot{x}) \\ \hat{F}_2(t, x) \end{bmatrix} = 0. \tag{4.23}$$

By construction, we have $\hat{F}(t_0, x_0, \dot{x}_0) = 0$. Moreover, for all $(t, x, \dot{x})$ satisfying $\hat{F}(t, x, \dot{x}) = 0$ it follows that

$$\hat{F}_{\dot{x}}(t, x, \dot{x}) = \begin{bmatrix} Z_1^T F_{\dot{x}}(t, x_1, x_2, \dot{x}_1, \dot{x}_2) \\ 0 \end{bmatrix}$$

and

$$\hat{F}_x(t, x, \dot{x}) = \begin{bmatrix} Z_1^T F_x(t, x_1, x_2, \dot{x}_1, \dot{x}_2) \\ Z_2^T F_{\mu;x}(t, x_1, x_2, \mathcal{H}(t, x_1, p_0)) \end{bmatrix}.$$

Since

$$\hat{F}_2(t, x_1, \mathcal{R}(t, x_1)) = Z_2^T F_\mu(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p_0)) = 0$$

and since

$$\tfrac{d}{dx_2} \hat{F}_2(t, x_1, \mathcal{R}(t, x_1)) = Z_2^T F_{\mu;x_2}(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p_0))$$

is nonsingular, the implicit function theorem implies that $\hat{F}_2(t, x_1, x_2) = 0$ holds if and only if $x_2 = \mathcal{R}(t, x_1)$ holds. Hence,

$$\hat{F}_x(t, x, \dot{x}) = \begin{bmatrix} Z_1^T F_x(t, x_1, x_2, \dot{x}_1, \dot{x}_2) \\ Z_2^T F_{\mu;x}(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p_0)) \end{bmatrix}$$

provided that $\hat{F}(t, x, \dot{x}) = 0$, and the kernel of the second block row is given by the span of the columns of $T_2(t, x_1)$. Because of

$$\hat{F}_{\dot{x}} T_2(t, x, \dot{x}) = \begin{bmatrix} Z_1^T F_{\dot{x}} T_2(t, x, \dot{x}) \\ 0 \end{bmatrix}$$

and, since $Z_1^T F_{\dot{x}} T_2(t, x, \dot{x})$ is nonsingular by Hypothesis 4.2, the reduced differential-algebraic equation (4.23) satisfies Hypothesis 4.2 with characteristic values $\mu = 0$, a and d. Thus, (4.23) is strangeness-free.

Differentiating $x_2 = \mathcal{R}(t, x_1)$ and eliminating x_2 and $\dot{x}_2$ in the equation $\hat{F}_1(t, x, \dot{x}) = 0$ yields the relation

$$Z_1^T F(t, x_1, \mathcal{R}(t, x_1), \dot{x}_1, \mathcal{R}_t(t, x_1) + \mathcal{R}_{x_1}(t, x_1)\dot{x}_1) = 0. \tag{4.24}$$

If $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ solves (4.23) in its domain of definition, then $(t_0, x_1^*(t_0), \dot{x}_1^*(t_0))$ solves (4.24). Since

$$\tfrac{d}{dx_1}\hat{F}_1 = Z_1^T F_{\dot{x}_1} + Z_1^T F_{\dot{x}_2}\mathcal{R}_{x_1} = Z_1^T F_{\dot{x}} T_2$$

is nonsingular due to Hypothesis 4.2, we can locally solve (4.24) for $\dot{x}_1$. Hence, the reduced problem (4.23) yields a decoupled differential-algebraic equation of the form

$$\dot{x}_1 = \mathcal{L}(t, x_1), \quad x_2 = \mathcal{R}(t, x_1). \tag{4.25}$$

Note that solutions of (4.23) that are close to x^* (in the metric of $C^1(\mathbb{I}, \mathbb{R}^n)$) must also solve (4.25).

In summary, we have shown that, starting from a point $z_{\mu,0} \in \mathbb{L}_\mu$, the equation $F_\mu = 0$ locally implies a reduced strangeness-free differential-algebraic equation of the form (4.23). If we assume solvability of (4.23), then we can transform (4.23) into a strangeness-free differential-algebraic equation of the special form (4.25). In particular, the following theorem holds.

Theorem 4.11. *Let F as in (4.1) be sufficiently smooth and satisfy Hypothesis 4.2 with characteristic values μ, a, and d. Then every sufficiently smooth solution of (4.1) also solves the reduced problems (4.23) and (4.25) consisting of d differential and a algebraic equations.*

Proof. If x^* is a sufficiently smooth solution of (4.1), then it must also solve the reduced differential-algebraic equations (4.23) and (4.25), since

$$(t, x^*(t), \dot{x}^*(t), \ldots, \left(\tfrac{d}{dt}\right)^{\mu+1} x^*(t)) \in \mathbb{L}_\mu \tag{4.26}$$

for every $t \in \mathbb{I}$. Since (4.25) fixes a unique solution when we prescribe an initial value for x_1, locally there can be only one solution of (4.1) satisfying (4.2). □

For sufficiently small intervals $\mathbb{I}$, we can also state Theorem 4.11 in a Banach space setting according to Section 3.4. For this, without restriction, we may assume a homogeneous initial value, since we can shift the function x as it has been done in Section 3.4 in the linear case.

Theorem 4.12. *Let F as in (4.1) be sufficiently smooth and satisfy Hypothesis 4.2. Let $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ be a sufficiently smooth solution of (4.1). Let the (nonlinear) operator $\hat{\mathcal{F}} : \mathbb{D} \to \mathbb{Y}$, $\mathbb{D} \subseteq \mathbb{X}$ open, be defined by*

$$\hat{\mathcal{F}}(x)(t) = \begin{bmatrix} \dot{x}_1(t) - \mathcal{L}(t, x_1(t)) \\ x_2(t) - \mathcal{R}(t, x_1(t)) \end{bmatrix}, \tag{4.27}$$

with the Banach spaces

$$\mathbb{X} = \{x \in C(\mathbb{I}, \mathbb{R}^n) \mid x_1 \in C^1(\mathbb{I}, \mathbb{R}^d),\ x_1(t_0) = 0\}, \quad \mathbb{Y} = C(\mathbb{I}, \mathbb{R}^n), \tag{4.28}$$

according to the construction preceding (4.25). Then, x^* is a regular solution *of the (strangeness-free) problem*

$$\hat{\mathcal{F}}(x) = 0 \tag{4.29}$$

in the following sense. There exists a neighborhood $\mathbb{U} \subseteq \mathbb{X}$ of x^*, and a neighborhood $\mathbb{V} \subseteq \mathbb{Y}$ of the origin such that for every $f \in \mathbb{V}$ the equation $\hat{\mathcal{F}}(x) = f$ has a unique solution $x \in \mathbb{U}$ that depends continuously on f. In particular, x^* is the unique solution in $\mathbb{U}$ belonging to $f = 0$.

Proof. The Fréchet derivative $D\hat{\mathcal{F}}(x) : \mathbb{X} \to \mathbb{Y}$ of $\hat{\mathcal{F}}$ at some $x \in \mathbb{D}$, given by

$$\begin{aligned} D\hat{\mathcal{F}}(x)(\Delta x)(t) &= \begin{bmatrix} \Delta\dot{x}_1(t) - \mathcal{L}_{x_1}(t, x_1(t))\Delta x_1(t) \\ \Delta x_2(t) - \mathcal{R}_{x_1}(t, x_1(t))\Delta x_1(t) \end{bmatrix} \\ &= \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \Delta\dot{x}(t) - \begin{bmatrix} \mathcal{L}_{x_1}(t, x_1(t)) & 0 \\ \mathcal{R}_{x_1}(t, x_1(t)) & -I_a \end{bmatrix} \Delta x(t), \end{aligned}$$

is a linear homeomorphism, i.e., it is invertible and has a continuous inverse, cp. Section 3.4. The claim then follows by the implicit function theorem in Banach spaces or, more constructive, by means of Newton's method in Banach spaces, see, e.g., [68, Theorems 15.2 and 15.6]. □

If we do not start with a solution x^* of the original problem but only with a point $z_{\mu,0} \in \mathbb{L}_\mu$, then the same procedure that lead to (4.23) can be applied and we still obtain a reduced differential-algebraic equation (4.23). This reduced differential-algebraic equation, however, may not be solvable at all, cp. Exercise 11. Moreover, even if this reduced problem possesses a solution, then it is not clear whether this solution also solves the original differential-algebraic equation (4.1). Recall the corresponding results in Section 3.3 concerning the linear case. To show that the reduced system reflects (at least locally) the properties of the original system concerning solvability and structure of the solution set, we need the converse direction of Theorem 4.11. The following theorem gives sufficient conditions for this converse to hold.

Theorem 4.13. *Let F as in (4.1) be sufficiently smooth and satisfy Hypothesis 4.2 with characteristic values μ, a, d, and with characteristic values $\mu+1$ (replacing μ), a, d. Then, for every $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$, the reduced problem (4.23) has a unique solution satisfying the initial value given by $z_{\mu+1,0}$. Moreover, this solution locally solves the original problem (4.1).*

Proof. By assumption, there exists (locally with respect to $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$) a parameterization (t, x_1, p), where p is chosen out of $(\dot{x}, \dots, x^{(\mu+2)})$, with

$$F_{\mu+1}(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p)) \equiv 0.$$

This includes the equation

$$F_\mu(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p)) \equiv 0, \tag{4.30}$$

with trivial dependence on $x^{(\mu+2)}$, as well as the equation

$$\tfrac{d}{dt} F_\mu(t, x_1, \mathcal{R}(t, x_1), \mathcal{H}(t, x_1, p)) \equiv 0. \tag{4.31}$$

Equation (4.30) implies that (omitting arguments)

$$F_{\mu;t} + F_{\mu;x_2}\mathcal{R}_t + F_{\mu;\dot{x},\dots,x^{(\mu+2)}}\mathcal{H}_t \equiv 0, \tag{4.32a}$$
$$F_{\mu;x_1} + F_{\mu;x_2}\mathcal{R}_{x_1} + F_{\mu;\dot{x},\dots,x^{(\mu+2)}}\mathcal{H}_{x_1} \equiv 0, \tag{4.32b}$$
$$F_{\mu;\dot{x},\dots,x^{(\mu+2)}}\mathcal{H}_p \equiv 0. \tag{4.32c}$$

The relation $\frac{d}{dt}F_\mu = 0$ has the form

$$F_{\mu;t} + F_{\mu;x_1}\dot{x}_1 + F_{\mu;x_2}\dot{x}_2 + F_{\mu;\dot{x},\dots,x^{(\mu+1)}} \begin{bmatrix} \ddot{x} \\ \vdots \\ x^{(\mu+2)} \end{bmatrix} = 0.$$

Inserting the parameterization (t, x_1, p) yields that (4.31) can be written as

$$F_{\mu;t} + F_{\mu;x_1}\mathcal{H}_1 + F_{\mu;x_2}\mathcal{H}_2 + F_{\mu;\dot{x},\dots,x^{(\mu+1)}}\mathcal{H}_3 \equiv 0,$$

where $\mathcal{H}_i$, $i = 1, \dots, 3$, are the parts of $\mathcal{H}$ corresponding to $\dot{x}_1$, $\dot{x}_2$, and the remaining variables, respectively. Multiplication with Z_2^T (corresponding to Hypothesis 4.2 with μ, a, d) gives

$$Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x_1}\mathcal{H}_1 + Z_2^T F_{\mu;x_2}\mathcal{H}_2 \equiv 0.$$

Inserting the relations of (4.32) and observing that $Z_2^T F_{\mu;x_2}$ is nonsingular, we find that

$$Z_2^T F_{\mu;x_2}(\mathcal{H}_2 - \mathcal{R}_t - \mathcal{R}_{x_1}\mathcal{H}_1) \equiv 0,$$

or

$$\mathcal{H}_2 = \mathcal{R}_t + \mathcal{R}_{x_1}\mathcal{H}_1,$$

i.e.,

$$\dot{x}_2 = \mathcal{R}_t + \mathcal{R}_{x_1}\dot{x}_1.$$

In summary, the derivative array equation $F_{\mu+1} = 0$ implies that

$$Z_1^T F(t, x_1, x_2, \dot{x}_1, \dot{x}_2) = 0, \tag{4.33a}$$
$$x_2 = \mathcal{R}(t, x_1), \tag{4.33b}$$
$$\dot{x}_2 = \mathcal{R}_t(t, x_1) + \mathcal{R}_{x_1}(t, x_1)\dot{x}_1. \tag{4.33c}$$

Elimination of x_2 and $\dot{x}_2$ from (4.33a) gives an ordinary differential equation

$$\dot{x}_1 = \mathcal{L}(t, x_1).$$

In particular, it follows that $\dot{x}_1$ and $\dot{x}_2$ are not part of the parameterization. Therefore, the following construction is possible. Let $p = p(t)$ be arbitrary but smooth and consistent to the initial value $z_{\mu+1,0}$ and let $x_1 = x_1(t)$ and $x_2 = x_2(t)$ be the solution of the initial value problem

$$\dot{x}_1 = \mathcal{L}(t, x_1), \quad x_1(t_0) = x_{1,0},$$
$$x_2 = \mathcal{R}(t, x_1).$$

Although $\dot{x}_1$ and $\dot{x}_2$ are not part of the parameterization, we automatically get $\dot{x}_1 = \dot{x}_1(t)$ and $\dot{x}_2 = \dot{x}_2(t)$. Thus, we have

$$F_{\mu+1}(t, x_1(t), x_2(t), \dot{x}_1(t), \dot{x}_2(t), \mathcal{H}_3(t, x_1(t), p(t)) \equiv 0,$$

in a neighborhood of t_0, or

$$F(t, x_1(t), x_2(t), \dot{x}_1(t), \dot{x}_2(t)) \equiv 0$$

for the first block row of the derivative array. From the construction of (4.23), it then follows that also

$$\hat{F}(t, x_1(t), x_2(t), \dot{x}_1(t), \dot{x}_2(t)) \equiv 0.$$

Finally, uniqueness follows by Theorem 4.11. □

Remark 4.14. Although all the results that we have obtained in this section are of a local nature, they can be globalized as it can be done in the case of ordinary differential equations (see, e.g., [106, Th. I.7.4]). Like there, we can continue the process (under the assumption of sufficient smoothness) until we reach the boundary of $\mathbb{L}_\mu$ or $\mathbb{L}_{\mu+1}$, respectively. Note that this may happen in finite time.

Remark 4.15. The proof of Theorem 4.13, together with Remark 4.14, shows that under the stated assumptions, for all solutions $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (4.1) there exists locally a function $\mathcal{P} \in C(\mathbb{I}, \mathbb{R}^{(\mu+1)n})$ with $\mathcal{P}(t)[\, I_n \;\; 0 \;\; \cdots \;\; 0\,]^T = \dot{x}^*(t)$ such that $F_\mu(t, x^*(t), \mathcal{P}(t)) \equiv 0$. By the compactness of $\mathbb{I}$, there is only a finite number of intervals, which must be taken into account. It is therefore possible to compose the corresponding parameterizations to a globally continuous parameterization such that the existence of a continuous $\mathcal{P}$ is guaranteed globally. We then can define solutions of (4.1) as those which can be extended to a continuous path $(t, x^*(t), \mathcal{P}(t))$ in $\mathbb{L}_\mu$, where $\mathcal{P}(t)[\, I_n \; 0 \; \cdots \; 0\,]^T = \dot{x}^*(t)$. In this way, we can drastically reduce the smoothness requirements in the above constructions. In particular, we can replace (4.26) with the condition that $(t, x^*(t), \mathcal{P}(t)) \in \mathbb{L}_\mu$ for every $t \in \mathbb{I}$.

Example 4.16. Consider the differential-algebraic equation

$$F(t, x, \dot{x}) = \dot{x}^2 - 1 = 0,$$

with $\mathbb{I} = [0, 1]$, $\mathbb{D}_x = \mathbb{R}$, and $\mathbb{D}_{\dot{x}} = \mathbb{R} \setminus \{0\}$. We find that

$$\mathbb{L}_0 = \left\{(t, x, \dot{x}) \;\middle|\; t \in \mathbb{I},\ x \in \mathbb{R},\ \dot{x} \in \{-1, 1\}\right\}$$

and that $M_0(t, x, \dot{x}) = 2\dot{x}$ has full rank on $\mathbb{L}_0$. Differentiating once yields $2\dot{x}\ddot{x} = 0$, such that

$$\mathbb{L}_1 = \left\{(t, x, \dot{x}, \ddot{x}) \;\middle|\; t \in \mathbb{I},\ x \in \mathbb{R},\ \dot{x} \in \{-1, 1\},\ \ddot{x} = 0\right\}$$

and

$$M_1(t, x, \dot{x}, \ddot{x}) = \begin{bmatrix} 2\dot{x} & 0 \\ 2\ddot{x} & 2\dot{x} \end{bmatrix}$$

has full rank on $\mathbb{L}_1$. Thus, this problem satisfies Hypothesis 4.2 with $\mu = 0$, $a = 0$, $d = 1$ as well as with $\mu = 1$, $a = 0$, $d = 1$. All $(t_0, x_0) \in \mathbb{I} \times \mathbb{R}$ are consistent, since $(t_0, x_0, \pm 1) \in \mathbb{L}_0$. With the two possibilities for the value of $\dot{x}$, we obtain the local solutions $x^*(t) = \pm(t - t_0) + x_0$, which both can be extended (as smooth solutions) to the whole interval $\mathbb{I}$.

4.2 Structured problems

In many applications modeled by differential-algebraic equations, the arising systems exhibit special structures. A typical example of such a special structure are the mathematical models of mechanical multibody systems such as the physical pendulum in Example 1.3. Making use of this structure usually leads to a simplified analysis.

In this section, we want to study such classes of structured differential-algebraic equations in the context of Hypothesis 4.2. Historically, the notion index in the naming of the various structures refers to a counting in the spirit of the differentiation index. The following discussion will show, on the basis of Section 3.3, that it coincides with the differentiation index in the case of linear problems. For convenience, we consider only autonomous equations, although all results also hold in the non-autonomous case.

All structured problems that we will discuss in the sequel are semi-explicit, i.e., of the form

$$\dot{x}_1 = f(x_1, x_2), \quad 0 = g(x_1, x_2), \tag{4.34}$$

with different assumptions on the functions f and g. We start our exposition with *semi-explicit differential-algebraic equations of index* $\nu = 1$, which have the form (4.34), where the Jacobian $g_{x_2}(x_1, x_2)$ is nonsingular for all relevant points $(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$.

Example 4.17. Example 1.4 with

$$x_1 = \begin{bmatrix} c \\ T \end{bmatrix}, \quad x_2 = R$$

and

$$f(x_1, x_2) = \begin{bmatrix} k_1(c_0 - c) - R \\ k_1(T_0 - T) + k_2 R - k_3(T - T_C(t)) \end{bmatrix},$$
$$g(x_1, x_2) = R - k_3 \exp\left(-\tfrac{k_4}{T} c\right)$$

is a semi-explicit differential-algebraic equation of index $\nu = 1$, since the Jacobian g_{x_2} is the identity.

We assume in the following that $\mathbb{L}_0 \neq \emptyset$, where

$$\mathbb{L}_0 = \left\{(t, x_1, x_2, \dot{x}_1, \dot{x}_2) \mid \dot{x}_1 = f(x_1, x_2),\ 0 = g(x_1, x_2)\right\}.$$

In particular, we assume that the constraint $g(x_1, x_2) = 0$ can be satisfied. For the corresponding Jacobians of the derivative array, we get (omitting arguments)

$$M_0 = \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix}, \quad N_0 = \begin{bmatrix} f_{x_1} & f_{x_2} \\ g_{x_1} & g_{x_2} \end{bmatrix}.$$

Following Hypothesis 4.2, we obtain that rank $M_0 = n_1$. With

$$Z_2^T = [\, 0 \ \ I_{n_2} \,],$$

we then get $a = \operatorname{rank} Z_2^T N_0 = \operatorname{rank}[\, g_{x_1} \ \ g_{x_2} \,] = n_2$. Setting

$$T_2 = \begin{bmatrix} I_{n_1} \\ g_{x_2}^{-1} g_{x_1} \end{bmatrix},$$

we finally see that rank $F_{\dot{x}} T_2 = n_1 = n_1 + n_2 - a = d$. Hence, under the stated assumptions, (4.34) possesses the strangeness index $\mu = 0$, and we have proved the following theorem.

Theorem 4.18. *Semi-explicit differential-algebraic equations of index $\nu = 1$ satisfy Hypothesis 4.2 with characteristic values $\mu = 0$, $a = n_2$, and $d = n - a$, provided that $\mathbb{L}_1 \neq \emptyset$.*

The next type of systems that we want to study are *semi-explicit differential-algebraic equations of index $\nu = 2$*. These have the form (4.34), where $[g_{x_1}(x_1, x_2) g_{x_2}(x_1, x_2)]$ has full row rank for all relevant points $(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and the differential-algebraic equation

$$g_{x_1}(x_1(t), x_2)\dot{x}_1(t) + g_{x_2}(x_1(t), x_2)\dot{x}_2 = 0, \tag{4.35}$$

obtained by differentiating the constraint, satisfies Hypothesis 4.2 with $\mu = 0$ for all relevant functions $x_1 \in C^1(\mathbb{I}, \mathbb{R}^{n_1})$.

Example 4.19. Example 1.5 of the Stokes equation discretized in space has the form (4.36) with

$$x_1 = u_h, \quad x_2 = p_h, \quad f(x_1, x_2) = Au_h + Bp_h, \quad g(x_1) = B^T u_h.$$

If the nonuniqueness of a free constant in the pressure is also fixed by the discretization method, then B has full column rank. Thus, $[\, g_{x_1} \;\; g_{x_2} \,] = [\, B^T \;\; 0 \,]$ has full row rank. Differentiating the constraint, we get

$$0 = B^T \dot{x}_1 = B^T A x_1 + B^T B x_2.$$

For given x_1, this is a regular strangeness-free differential-algebraic equation for x_2 due to the invertibility of $B^T B$. Hence, the given problem is a semi-explicit differential-algebraic equation of index $\nu = 2$.

In the following, we assume that $\mathbb{L}_1 \neq \emptyset$, where

$$\begin{aligned} \mathbb{L}_1 = \big\{ (t, x_1, x_2, \dot{x}_1, \dot{x}_2, \ddot{x}_1, \ddot{x}_2) \mid \dot{x}_1 &= f(x_1, x_2),\ 0 = g(x_1), \\ \ddot{x} &= f_{x_1}(x_1, x_2)\dot{x}_1 + f_{x_2}(x_1, x_2)\dot{x}_2, \\ 0 &= g_{x_1}(x_1, x_2) f(x_1(t), x_2) + g_{x_2}(x_1, x_2)\dot{x}_2 \big\}, \end{aligned}$$

and that $\tilde{\mathbb{L}}_0 \neq \emptyset$, where

$$\tilde{\mathbb{L}}_0 = \big\{ (t, x_2, \dot{x}_2) \mid g_{x_1}(x_1(t), x_2) f(x_1(t), x_2) + g_{x_2}(x_1(t), x_2)\dot{x}_2 \big\}.$$

Let (omitting arguments)

$$\tilde{M}_0 = [\, g_{x_2} \,], \quad \tilde{N}_0 = -[\, g_{x_1 x_2} f + g_{x_1} f_{x_2} + g_{x_2 x_2}\dot{x}_2 \,].$$

Due to the assumptions, Hypothesis 4.2 yields that there exist smooth matrix functions $\tilde{Z}_2$, $\tilde{T}_2$, and $\tilde{Z}_1$ of appropriate size and pointwise full column rank such that

$$\begin{aligned} &\operatorname{rank} \tilde{M}_0 = \tilde{d}, && \tilde{Z}_2^T \tilde{M}_0 = 0, \\ &\operatorname{rank} \tilde{Z}_2^T \tilde{N}_0 = \tilde{a}, && \tilde{Z}_2^T \tilde{N}_0 \tilde{T}_2 = 0, \\ &\operatorname{rank} g_{x_2} T_2 = \tilde{d}, && \operatorname{rank} \tilde{Z}_1^T g_{x_2} T_2 = \tilde{d}. \end{aligned}$$

The corresponding characteristic values $\tilde{a}$ and $\tilde{d}$ satisfy $\tilde{a} + \tilde{d} = n_2$. Moreover, we have that $[\ \tilde{Z}_1\ \ \tilde{Z}_2\ ]$ is a square matrix function with pointwise full rank n_2.

In order to investigate (4.34) under the given assumptions, we consider the corresponding Jacobians of the derivative array given by

$$M_1 = \left[\begin{array}{cc|cc} I_{n_1} & 0 & & \\ 0 & 0 & & \\ \hline -f_{x_1} & -f_{x_2} & I_{n_1} & 0 \\ -g_{x_1} & -g_{x_2} & 0 & 0 \end{array}\right],$$

$$N_1 = \left[\begin{array}{cc|cc} & f_{x_1} & f_{x_2} & & 0 & 0 \\ & g_{x_1} & g_{x_2} & & 0 & 0 \\ \hline f_{x_1x_1}\dot{x}_1 + f_{x_1x_2}\dot{x}_2 & & f_{x_1x_2}\dot{x}_1 + f_{x_2x_2}\dot{x}_2 & & 0 & 0 \\ g_{x_1x_1}\dot{x}_1 + g_{x_1x_2}\dot{x}_2 & & g_{x_1x_2}\dot{x}_1 + g_{x_2x_2}\dot{x}_2 & & 0 & 0 \end{array}\right].$$

Following Hypothesis 4.2, we first compute

$$\operatorname{rank} M_1 = 2 \operatorname{rank} I_{n_1} + \operatorname{rank} g_{x_2} = 2n_1 + \tilde{d}.$$

A possible choice for Z_2 is given by

$$Z_2^T = \left[\begin{array}{cc|cc} 0 & I_{n_2} & 0 & 0 \\ \tilde{Z}_2^T g_{x_1} & 0 & 0 & \tilde{Z}_2^T \end{array}\right].$$

Since

$$g_{x_1x_2}\dot{x}_1 + g_{x_2x_2} = g_{x_1x_2} f + g_{x_2x_2}\dot{x}_2 = -\tilde{N}_0$$

on $\mathbb{L}_1$, we then see that

$$Z_2^T N_1 [\ I_n\ \ 0\ ]^T = \begin{bmatrix} g_{x_1} & g_{x_2} \\ * & -\tilde{Z}_2^T \tilde{N}_0 \end{bmatrix}.$$

Choosing $\tilde{T}_2'$ such that $[\ \tilde{T}_2'\ \ \tilde{T}_2\ ]$ is nonsingular and recalling that $[\ \tilde{Z}_1\ \ \tilde{Z}_2\ ]$ is nonsingular as well, elementary block row and column operations yield

$$\begin{bmatrix} g_{x_1} & g_{x_2} \\ * & -\tilde{Z}_2^T \tilde{N}_0 \end{bmatrix} \to \begin{bmatrix} \tilde{Z}_1^T g_{x_1} & \tilde{Z}_1^T \tilde{M}_0 \tilde{T}_2' & \tilde{Z}_1^T \tilde{M}_0 \tilde{T}_2 \\ \tilde{Z}_2^T g_{x_1} & 0 & 0 \\ * & \tilde{Z}_2^T \tilde{N}_0 \tilde{T}_2' & 0 \end{bmatrix}.$$

Full row rank of $[g_{x_1}\, g_{x_2}]$ implies full row rank of $\tilde{Z}_2^T g_{x_1}$. Moreover, by construction $\tilde{Z}_1^T \tilde{M}_0 \tilde{T}_2$ and $\tilde{Z}_2^T \tilde{N}_0 \tilde{T}_2'$ are nonsingular. Hence, $Z_2^T N_1 [\, I_n \;\; 0\,]^T$ has full row rank and a possible choice for T_2 due to Hypothesis 4.2 is given by

$$T_2 = \begin{bmatrix} K \\ * \end{bmatrix},$$

where K is of size $n_1 \times (n_1 - \tilde{a})$ and has full column rank. This finally implies that

$$F_{\dot{x}} T_2 = \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} K \\ * \end{bmatrix} = \begin{bmatrix} K \\ 0 \end{bmatrix}$$

has full column rank. This shows that Hypothesis 4.2 holds with characteristic values $\mu = 1$, $a = n_2 + \tilde{a}$, and $d = n_1 - \tilde{a}$. In particular, we have proved the following theorem.

Theorem 4.20. *Consider a semi-explicit differential-algebraic equations of index $\nu = 2$ given by (4.34), together with the additional properties that the matrix $[g_{x_1}(x_1, x_2) g_{x_2}(x_1, x_2)]$ has full row rank for all relevant points $(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$ and that (4.35) satisfies Hypothesis 4.2 with $\mu = 0$ for all relevant functions $x_1 \in C^1(\mathbb{I}, \mathbb{R}^{n_1})$. Then, this differential-algebraic equation satisfies Hypothesis 4.2 with characteristic values $\mu = 1$, $a = n_2 + \tilde{a}$, and $d = n_1 - \tilde{a}$, provided that $\mathbb{L}_1 \neq \emptyset$ and $\tilde{\mathbb{L}}_0 \neq \emptyset$.*

A special case of semi-explicit systems of index $\nu = 2$ are differential-algebraic equations of the form

$$\dot{x}_1 = f(x_1, x_2), \quad 0 = g(x_1), \tag{4.36}$$

with $g_{x_1}(x_1) f_{x_2}(x_1, x_2)$ nonsingular for all relevant points $(x_1, x_2) \in \mathbb{R}^{n_1} \times \mathbb{R}^{n_2}$. To see this, we differentiate the constraint to obtain

$$0 = g_{x_1}(x_1)\dot{x}_1 = g_{x_1}(x_1) f(x_1, x_2),$$

which for given x_1 can be solved for x_2 with the help of the implicit function theorem. Hence, it satisfies Hypothesis 4.2 with $\mu = 0$.

Systems of the form (4.36) with the required property are called differential-algebraic equations in *Hessenberg form of index $\nu = 2$*.

Example 4.21. The discretized Stokes equation of Example 4.19 is a differential-algebraic equation in Hessenberg form of index $\nu = 2$, since $g_{x_1}(x_1) f_{x_2}(x_1, x_2) = B^T B$ is nonsingular.

Differential-algebraic equations in Hessenberg form of index $\nu = 2$ belong to a more general class of structured problems, so-called differential-algebraic equations in *Hessenberg form of index* ν. These have the form

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \ldots, x_{\nu-1}, x_\nu), \\ \dot{x}_2 &= f_2(x_1, \ldots, x_{\nu-1}), \\ \dot{x}_3 &= f_3(x_2, \ldots, x_{\nu-1}), \\ &\vdots \\ \dot{x}_{\nu-1} &= f_{\nu-1}(x_{\nu-2}, x_{\nu-1}), \\ 0 &= f_\nu(x_{\nu-1}), \end{aligned} \tag{4.37}$$

with

$$\frac{\partial f_\nu}{\partial x_{\nu-1}} \cdot \frac{\partial f_{\nu-1}}{\partial x_{\nu-2}} \cdots \frac{\partial f_2}{\partial x_1} \cdot \frac{\partial f_1}{\partial x_\nu} \quad \text{nonsingular} \tag{4.38}$$

for all relevant points $(x_1, \ldots, x_\nu) \in \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_\nu}$. Note that it only makes sense to consider $\nu \geq 2$.

Example 4.22. An important class of applications, where modeling leads to differential-algebraic equations in Hessenberg form of higher index, are constrained multibody systems such as the physical pendulum of Example 1.3. A typical form of the arising equations is given by

$$\begin{aligned} \dot{p} &= v, \\ M(p)\dot{v} &= f(p, v) - g_p(p)^T \lambda, \\ g(p) &= 0, \end{aligned} \tag{4.39}$$

where p denotes the generalized positions, v the generalized velocities, and λ the Lagrangian multiplier belonging to the constraint $g(p) = 0$. Since the constraint is on the position only, (4.39) is also called the *formulation on position level*. The standard assumptions, supported by the application, are that the mass matrix $M(p)$ is symmetric and positive definite and that the constraints are (locally) independent in the sense that the Jacobian $g_p(p)$ has full row rank. Setting $x_1 = v$, $x_2 = p$, and $x_3 = \lambda$, as well as

$$\begin{aligned} f_1(x_1, x_2, x_3) &= M(x_2)^{-1}(f(x_2, x_1) - g_p(x_2)^T x_3), \\ f_2(x_1, x_2) &= x_1, \\ f_3(x_2) &= g(x_2) \end{aligned}$$

and observing that

$$\frac{\partial f_3}{\partial x_2} \cdot \frac{\partial f_2}{\partial x_1} \cdot \frac{\partial f_1}{\partial x_3}(x_1, x_2, x_3) = -g_p(x_2)M(x_2)^{-1}g_p(x_2)^T$$

is nonsingular, we see that (4.39) is a differential-algebraic equation in Hessenberg form of index $\nu = 3$.

For the analysis of differential-algebraic equations in Hessenberg form, we assume that $\mathbb{L}_\mu \neq \emptyset$ with $\mu = \nu - 1$, i.e., that all constraints which arise by successive differentiation of the last equation in (4.37) and elimination of the arising derivatives with the help of the other equations can be satisfied. Our aim then is to show that (4.37) satisfies Hypothesis 4.2. We start with the observation that

$$F_{\dot x} = \begin{bmatrix} I_{n_1} & & & & \\ & I_{n_2} & & & \\ & & \ddots & & \\ & & & I_{n_{\nu-1}} & \\ & & & & 0 \end{bmatrix}, \quad F_x = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,\nu-1} & A_{1,\nu} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,\nu-1} & 0 \\ & \ddots & \ddots & \vdots & \vdots \\ & & \ddots & A_{\nu-1,\nu-1} & 0 \\ & & & A_{\nu,\nu-1} & 0 \end{bmatrix},$$

where $A_{i,j} = \frac{\partial f_i}{\partial x_j}$. Note the Hessenberg-like structure of F_x which gave (4.37) its name. To utilize the structure of $F_{\dot x}$ and F_x, we write these in the form

$$F_{\dot x} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad F_x = \begin{bmatrix} H & U \\ V^T & 0 \end{bmatrix},$$

with

$$H = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,\nu-1} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,\nu-1} \\ & \ddots & \ddots & \vdots \\ & & A_{\nu-1,\nu-2} & A_{\nu-1,\nu-1} \end{bmatrix}, \quad U = \begin{bmatrix} A_{1,\nu} \\ 0 \\ \vdots \\ 0 \end{bmatrix},$$

$$V^T = \begin{bmatrix} 0 & \cdots & 0 & A_{\nu,\nu-1} \end{bmatrix}.$$

Thus,

$$M_\mu = \left[\begin{array}{cc|cc|cc|c} I & 0 & & & & & \\ 0 & 0 & & & & & \\ \hline -H & -U & I & 0 & & & \\ -V^T & 0 & 0 & 0 & & & \\ \hline * & * & -H & -U & I & 0 & \\ * & 0 & -V^T & 0 & 0 & 0 & \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array}\right]$$

and

$$N_\mu = \left[\begin{array}{cc|c|c|c} H & U & & & \\ V^T & 0 & & & \\ \hline \dot H & \dot U & & & \\ \dot V^T & 0 & & & \\ \hline * & * & & & \\ * & 0 & & & \\ \hline \vdots & \vdots & & & \end{array}\right].$$

Reordering (via block permutations) the block rows and columns such that all identity blocks in the diagonal of M_μ are moved to the upper left corner yields

$$\tilde{M}_\mu = \begin{bmatrix} I - \mathfrak{H} & -\mathfrak{U} \\ -\mathfrak{V}^T & 0 \end{bmatrix},$$

with

$$\mathfrak{H} = \begin{bmatrix} 0 & & & & \\ H & 0 & & & \\ * & H & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ * & \cdots & * & H & 0 \end{bmatrix}$$

and

$$\mathfrak{U} = \begin{bmatrix} 0 & & & & \\ U & 0 & & & \\ * & U & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ * & \cdots & * & U & 0 \end{bmatrix}, \quad \mathfrak{V}^T = \begin{bmatrix} 0 & & & & \\ V^T & 0 & & & \\ * & V^T & 0 & & \\ \vdots & \ddots & \ddots & \ddots & \\ * & \cdots & * & V^T & 0 \end{bmatrix}.$$

Accordingly, we obtain the permuted $\tilde{N}_\mu$. Observe that all nontrivial entries in $\mathfrak{H}$, $\mathfrak{U}$, and $\mathfrak{V}^T$ have the same block structure given by H, U and V^T, respectively. In particular, they have the block structure of a nilpotent matrix with nilpotency index ν such that

$$\begin{aligned} &\text{(a)} \quad \mathfrak{H}^\ell = 0 && \text{for } \ell \geq \nu, \\ &\text{(b)} \quad \mathfrak{V}^T \mathfrak{H}^\ell = 0 && \text{for } \ell \geq \nu - 1, \\ &\text{(c)} \quad \mathfrak{V}^T \mathfrak{H}^\ell \mathfrak{U} = 0 && \text{for } \ell \geq \nu - 2. \end{aligned}$$

Furthermore, we can utilize the Hessenberg form of H and its derivatives, noticing that for an ℓ-fold product of Hessenberg matrices we obtain

$$\begin{bmatrix} 0 & \cdots & 0 & * \end{bmatrix} \underbrace{\begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix} \cdots \begin{bmatrix} * & \cdots & \cdots & * \\ * & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & * & * \end{bmatrix}}_{\ell\text{-fold product}} \begin{bmatrix} * \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

when $\ell \leq \nu - 3$. These properties imply that

$$\mathfrak{V}^T (I - \mathfrak{H})^{-1} \mathfrak{U} = \sum_{\ell=0}^{\nu-3} \mathfrak{V}^T \mathfrak{H}^\ell \mathfrak{U} = 0,$$

such that $Z_2^T \tilde{M}_\mu = 0$ for the choice

$$Z_2^T = [\, \mathfrak{V}^T (I - \mathfrak{H})^{-1} \;\; I \,]$$

following Hypothesis 4.2. Because of

$$\begin{aligned}
\mathfrak{V}^T (I - \mathfrak{H})^{-1} &= \sum_{\ell=0}^{\nu-2} \mathfrak{V}^T \mathfrak{H}^\ell \\
&= \begin{bmatrix} 0 & & & \\ V^T & \ddots & & \\ \dot{V}^T & \ddots & \ddots & \\ & \ddots & \ddots & \ddots \\ * & & \dot{V}^T & V^T \;\; 0 \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ 0 & \ddots & & & \\ V^T H & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ * & & V^T H & 0 & 0 \end{bmatrix} \\
&\quad + \cdots + \begin{bmatrix} 0 & & & & \\ 0 & \ddots & & & \\ 0 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ V^T H^{\nu-2} & & 0 & 0 & 0 \end{bmatrix},
\end{aligned}$$

we then get that

$$\begin{aligned}
Z_2^T \tilde{N}_\mu [\, I \;\; 0 \;\; \cdots \;\; 0 \,]^T &= \mathfrak{V}^T (I - \mathfrak{H})^{-1} \begin{bmatrix} H & U \\ \dot{H} & \dot{U} \\ \vdots & \vdots \\ H^{(\mu)} & U^{(\mu)} \end{bmatrix} - \begin{bmatrix} V^T & 0 \\ \dot{V}^T & 0 \\ \vdots & \vdots \\ V^{(\mu)T} & 0 \end{bmatrix} \\
&= \begin{bmatrix} -V^T & 0 \\ V^T H & 0 \\ V^T H^2 & 0 \\ \vdots & \vdots \\ V^T H^{\nu-2} & 0 \\ V^T H^{\nu-1} & V^T H^{\nu-2} U \end{bmatrix} + \begin{bmatrix} 0 & 0 \\ \dot{V}^T & 0 \\ * & 0 \\ \vdots & \vdots \\ * & 0 \\ * & 0 \end{bmatrix}.
\end{aligned}$$

Setting $W_i = A_{\nu,\nu-1} \cdots A_{i+1,i}$ for $i = 1, \ldots, \nu - 1$, we have

$$\begin{aligned}
V^T &= [\, 0 \;\; \cdots \;\; 0 \;\; 0 \;\; 0 \;\; W_{\nu-1} \,], \\
V^T H &= [\, 0 \;\; \cdots \;\; 0 \;\; 0 \;\; W_{\nu-2} \;\; * \,], \\
V^T H^2 &= [\, 0 \;\; \cdots \;\; 0 \;\; W_{\nu-3} \;\; * \;\; * \,],
\end{aligned}$$

and so on. The corresponding entries in the other summand in the representation of $Z_2^T \tilde{N}_\mu [\, I \;\; 0 \;\; \cdots \;\; 0 \,]^T$ are sums of products, where the first factor is V^T or one of its derivatives and the other factors are H or one of its derivatives, but one factor less compared with $V^T H^\ell$. Hence, they do not perturb the zero entries and the entries $W_{\nu-\ell}$ in $V^T H^\ell$. This implies that

$$Z_2^T \tilde{N}_\mu [\, I \;\; 0 \;\; \cdots \;\; 0 \,]^T = \begin{bmatrix} \mathfrak{W} & 0 \\ * & W_1 A_{1,\nu} \end{bmatrix},$$

where

$$\mathfrak{W} = \begin{bmatrix} 0 & \cdots & 0 & -W_{\nu-1} & 0 \\ \vdots & \iddots & W_{\nu-2} & * & \vdots \\ 0 & \iddots & \iddots & \vdots & \vdots \\ W_1 & * & \cdots & * & 0 \end{bmatrix}.$$

Since $A_{\nu,\nu-1} \cdot A_{\nu-1,\nu-2} \cdots A_{2,1} \cdot A_{1,\nu}$ is nonsingular due to (4.38), the quantities $W_1, \ldots, W_{\nu-1}$ have full row rank and $W_1 A_{1,\nu}$ is nonsingular. Therefore, $Z_2^T \tilde{N}_\mu [\, I \;\; 0 \;\; \cdots \;\; 0 \,]^T$ has full row rank $a = \nu n_\nu$ and its kernel is given by

$$T_2 = \begin{bmatrix} \mathfrak{K} \\ * \end{bmatrix},$$

where the columns of $\mathfrak{K}$ span kernel $\mathfrak{W}$. We then end up with

$$F_{\dot{x}} T_2 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \mathfrak{K} \\ * \end{bmatrix} = \begin{bmatrix} \mathfrak{K} \\ 0 \end{bmatrix}$$

and rank $F_{\dot{x}} T_2 = \operatorname{rank} \mathfrak{K} = d = n - a$. Thus, we have proved the following theorem.

Theorem 4.23. *Differential-algebraic equations in Hessenberg form of index ν given by (4.37) satisfy Hypothesis 4.2 with characteristic values $\mu = \nu - 1$, $a = \nu n_\nu$, and $d = n - a$, provided that $\mathbb{L}_\mu \neq \emptyset$.*

There is an intuitive way to argue that Theorem 4.23 holds and to get an idea which conditions imply $\mathbb{L}_\mu \neq \emptyset$. In (4.37), we have differential equations for $x_1, \ldots, x_{\nu-1}$, but not for x_ν. In order to derive a strangeness-free differential-algebraic equation in the variables $x_1, \ldots, x_\nu$, we need a relation which is solvable at least for x_ν. Since x_ν is only present in the left hand side of the first differential equation, we apparently can only proceed in the following way. We differentiate the constraint and eliminate the occurring derivative $\dot{x}_{\nu-1}$ with the help of the corresponding differential equation to obtain a further constraint $0 = f_{\nu;x_{\nu-1}}(x_{\nu-1}) f_{\nu-1}(x_{\nu-2}, x_{\nu-1})$, which must be satisfied by any solution of (4.37). We then differentiate this new constraint and so on, until the newest

constraint contains x_ν. This occurs after differentiating $\nu - 1$ times. The condition $\mathbb{L}_\mu \neq \emptyset$ then guarantees that all obtained constraints can be simultaneously fulfilled. Moreover, (4.38) guarantees that we can locally solve for νn_ν variables out of $(x_1, \ldots, x_\nu)$ which must include x_ν.

Remark 4.24. Since the structured problems (4.34), (4.36), and (4.37) are all explicit in the derivatives, the corresponding reduced differential-algebraic equations (4.23) can be chosen in such a way that the differential part $Z_1^T F(t, x_1, x_2, \dot{x}_1, \dot{x}_2) = 0$ is explicit with respect to x_1 and that no derivative $\dot{x}_2$ occurs. Hence, (4.24) can be trivially solved for $\dot{x}_1$ and we do not need any additional assumptions to obtain (4.25).

In the context of numerical methods, in particular for the structured problems in this section, it is common practice to modify the given differential-algebraic equation in such a way that the index of the system is decreased or increased. Decreasing the index is typically performed to allow for the application of numerical methods which require that the index does not exceed a certain number, cp. Chapter 6. Increasing the index on the other hand may help to obtain a system which exhibits more structure. The reason for this may again be to allow for the application of specific numerical methods that are well suited for such structures. In both cases one must be aware of possible changes in the structure of the solution space of the systems.

Consider the case of a differential-algebraic equation in Hessenberg form of index $\nu = 2$. Differentiating the constraint in (4.36), and eliminating $\dot{x}_1$ with the help of the differential equation gives the *hidden constraint* $0 = g_{x_1}(x_1) f(x_1, x_2)$. If we replace the old constraint with the so obtained new one, we get the modified problem

$$\dot{x}_1 = f(x_1, x_2), \quad 0 = g_{x_1}(x_1) f(x_1, x_2). \tag{4.40}$$

To distinguish the quantities of Hypothesis 4.2 already determined for (4.36) from those for (4.40), we use tildes in the latter case. Assuming that $\mathbb{L}_1 \neq \emptyset$ for (4.36), we immediately get that

$$\tilde{\mathbb{L}}_0 = \left\{(t, x_1, x_2) \mid \dot{x}_1 = f(x_1, x_2),\ 0 = g_{x_1}(x_1) f(x_1, x_2)\right\}$$

is also nonempty. With the Jacobians of the derivative array

$$\tilde{M}_0 = \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix}, \quad \tilde{N}_0 = \begin{bmatrix} f_{x_1} & f_{x_2} \\ * & g_{x_1} f_{x_2} \end{bmatrix},$$

we find that

$$\tilde{Z}_2^T = [\ 0\ \ I_{n_2}\ ], \quad \tilde{Z}_2^T \tilde{N}_0 = [\ *\ \ g_{x_1} f_{x_2}\ ], \quad \tilde{T}_2 = \begin{bmatrix} I_{n_1} \\ * \end{bmatrix},$$

such that (4.40) satisfies Hypothesis 4.2 with $\tilde{\mu} = 0$, $\tilde{a} = n_2$, and $\tilde{d} = n_1$. Observe that passing from (4.36) to (4.40), we loose the original constraint $0 = g(x_1)$. This is indicated by a smaller value of $\tilde{a}$ compared with a. Thus, we have shown the following theorem.

Theorem 4.25. *For a semi-explicit differential-algebraic equation* (4.36) *in Hessenberg form of index* $\nu = 2$*, the related differential-algebraic equation* (4.40) *satisfies Hypothesis* 4.2 *with* $\mu = 0$*.*

A similar result also holds in the case of general Hessenberg systems. Replacing (4.37) by

$$\begin{aligned} \dot{x}_1 &= f_1(x_1, \dots, x_{\nu-1}, x_\nu), \\ \dot{x}_2 &= f_2(x_1, \dots, x_{\nu-1}), \\ \dot{x}_3 &= f_3(x_2, \dots, x_{\nu-1}), \\ &\vdots \\ \dot{x}_{\nu-1} &= f_{\nu-1}(x_{\nu-2}, x_{\nu-1}), \\ 0 &= f_{\nu;x_{\nu-1}}(x_{\nu-1}) f_{\nu-1}(x_{\nu-2}, x_{\nu-1}), \end{aligned} \tag{4.41}$$

we can define

$$\tilde{x}_1 = x_1, \ \dots, \ \tilde{x}_{\nu-3} = x_{\nu-3}, \quad \tilde{x}_{\nu-2} = (x_{\nu-2}, x_{\nu-1}), \quad \tilde{x}_{\nu-1} = x_\nu,$$

and

$$\begin{aligned} \tilde{f}_1(\tilde{x}_1, \dots, \tilde{x}_{\nu-2}, \tilde{x}_{\nu-1}) &= f_1(x_1, \dots, x_{\nu-1}, x_\nu), \\ \tilde{f}_2(\tilde{x}_1, \dots, \tilde{x}_{\nu-2}) &= f_2(x_1, \dots, x_{\nu-1}), \\ &\vdots \\ \tilde{f}_{\nu-3}(\tilde{x}_{\nu-4}, \dots, \tilde{x}_{\nu-2}) &= f_{\nu-3}(x_{\nu-4}, \dots, x_{\nu-1}), \\ \tilde{f}_{\nu-2}(\tilde{x}_{\nu-3}, \tilde{x}_{\nu-2}) &= \begin{bmatrix} f_{\nu-2}(x_{\nu-3}, x_{\nu-2}, x_{\nu-1}) \\ f_{\nu-1}(x_{\nu-2}, x_{\nu-1}) \end{bmatrix}, \\ \tilde{f}_{\nu-1}(\tilde{x}_{\nu-2}) &= f_{\nu;x_{\nu-1}}(x_{\nu-1}) f_{\nu-1}(x_{\nu-2}, x_{\nu-1}). \end{aligned}$$

It is then obvious that (4.41) has again Hessenberg form but with an index reduced by one. Moreover,

$$\begin{aligned} &\frac{\partial \tilde{f}_{\nu-1}}{\partial \tilde{x}_{\nu-2}} \cdot \frac{\partial \tilde{f}_{\nu-2}}{\partial \tilde{x}_{\nu-1}} \cdots \frac{\partial \tilde{f}_2}{\partial \tilde{x}_1} \cdot \frac{\partial \tilde{f}_1}{\partial \tilde{x}_{\nu-1}} \\ &\quad = \begin{bmatrix} \frac{\partial f_\nu}{\partial x_{\nu-1}} \cdot \frac{\partial f_{\nu-1}}{\partial x_{\nu-2}} & * \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial f_{\nu-2}}{\partial x_{\nu-3}} \\ 0 \end{bmatrix} \cdot \frac{\partial f_{\nu-3}}{\partial x_{\nu-4}} \cdots \frac{\partial f_2}{\partial x_1} \cdot \frac{\partial f_1}{\partial x_\nu} \\ &\quad = \frac{\partial f_\nu}{\partial x_{\nu-1}} \cdot \frac{\partial f_{\nu-1}}{\partial x_{\nu-2}} \cdots \frac{\partial f_2}{\partial x_1} \cdot \frac{\partial f_1}{\partial x_\nu}, \end{aligned}$$

and we have the following result.

Theorem 4.26. *For a semi-explicit differential-algebraic equation* (4.37) *in Hessenberg form of index ν, the related differential-algebraic equation* (4.41) *satisfies Hypothesis* 4.2 *with $\mu = \nu - 2$.*

Example 4.27. Consider again the multibody system from Example 4.22. Due to the above discussion, we can lower the index by one by replacing the constraint by its derivative. We obtain the new system

$$\begin{aligned} \dot{p} &= v, \\ M(p)\dot{v} &= f(p,v) - g_p(p)^T\lambda, \\ g_p(p)v &= 0, \end{aligned} \tag{4.42}$$

the so-called *formulation on velocity level*, which is a differential-algebraic equation in Hessenberg form with index $\nu = 2$. We can perform the same reduction step once more to arrive at

$$\begin{aligned} \dot{p} &= v, \\ M(p)\dot{v} &= f(p,v) - g_p(p)^T\lambda, \\ g_{pp}(p)(v,v) &+ g_p(p)M(p)^{-1}(f(p,v) - g_p(p)^T\lambda) = 0, \end{aligned} \tag{4.43}$$

the so-called *formulation on acceleration level*, which is now strangeness-free. Note, however, that (4.43) is not a reduced differential-algebraic equation belonging to (4.39) in the sense of Section 4.1, since (4.39) does not have the same number of algebraic and differential components as (4.43).

A reduced differential-algebraic equation in the sense of Section 4.1 can (locally) be obtained by dividing p into (p_1, p_2) such that $g(p_1, p_2) = 0$ can be solved for p_2. With (v_1, v_2) as corresponding velocities, we then consider

$$\begin{aligned} \dot{p}_1 &= v_1, \\ \dot{v}_1 &= [\, I \;\; 0\,]M(p)^{-1}(f(p,v) - g_p(p)^T\lambda), \\ g(p) &= 0, \\ g_p(p)v &= 0, \\ g_{pp}(p)(v,v) &+ g_p(p)M(p)^{-1}(f(p,v) - g_p(p)^T\lambda) = 0, \end{aligned} \tag{4.44}$$

which contains all constraints imposed by (4.39). Following Hypothesis 4.2, we have the Jacobians of the derivative array

$$M_0 = \begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad N_0 = \begin{bmatrix} 0 & 0 & -I & 0 & 0 \\ * & * & * & * & * \\ G_1 & G_2 & 0 & 0 & 0 \\ * & * & G_1 & G_2 & 0 \\ * & * & * & * & W \end{bmatrix},$$

with nonsingular blocks $G_2 = g_{p_2}$ and $W = -g_p M^{-1} g_p^T$ such that we can choose

$$T_2 = \begin{bmatrix} G_2 & 0 \\ -G_1 & 0 \\ 0 & G_2 \\ * & -G_1 \\ * & * \end{bmatrix}.$$

Hence, (4.44) satisfies Hypothesis 4.2 with $\mu = 0$, provided that the constraints can be fulfilled.

According to Section 4.1, a solution of (4.1) also solves the reduced differential-algebraic equation (4.23) provided that (4.1) satisfies Hypothesis 4.2. Thus, in order to compute this solution, we can work with (4.23). Unfortunately, some numerical schemes, such as Runge–Kutta based methods, require the problem to be semi-explicit, see [108]. This can be achieved by introducing a new variable $y = \dot{x}$ and transforming (4.23) to

$$\dot{x} = y, \quad \hat{F}_1(x, y) = 0, \quad \hat{F}_2(x) = 0, \tag{4.45}$$

still assuming for simplicity that the given problem is autonomous. If we require as in Section 4.1 that (4.23) implies a system of the from (4.23), then the set

$$\tilde{\mathbb{L}}_1 = \big\{(t, x, y, \dot{x}, \dot{y}, \ddot{x}, \ddot{y}) \mid \dot{x} = y,\ \hat{F}_1(x, y) = 0,\ \hat{F}_2(x) = 0,\\ \ddot{x} = \dot{y},\ \hat{F}_{1;x}(x, y)\dot{x} + F_{1;\dot{x}}(x, y)\dot{y} = 0,\ \hat{F}_{2;x}(x)\dot{x} = 0\big\}$$

can be locally parameterized according to

$$\begin{aligned} x_2 &= \mathcal{R}(x_1), \quad & \dot{x}_1 &= \mathcal{L}(x_1), \quad & \dot{x}_2 &= \mathcal{R}_{x_1}(x_1)\mathcal{L}(x_1), \\ y_1 &= \dot{x}_1, & y_2 &= \dot{x}_2, & \ddot{x} &= \dot{y} = \tilde{\mathcal{H}}(x_1, p), \end{aligned}$$

where $x = (x_1, x_2)$ is as in Section 4.1 and $y = (y_1, y_2)$ is split accordingly. The parameters p can be chosen out of $\dot{y}$ and the function $\tilde{\mathcal{H}}$ is defined implicitly by the linear problem

$$\hat{F}_{1;x}(x, y)\dot{x} + \hat{F}_{1;\dot{x}}(x, y)\dot{y} = 0,$$

due to the full row rank of $\hat{F}_{1;\dot{x}}(x, y)$. In particular, $\tilde{\mathbb{L}}_1$ is nonempty. The relevant

matrix functions in Hypothesis 4.2 for (4.45) are then given by (omitting arguments)

$$\tilde{M}_1 = \left[\begin{array}{cccc|cccc} I_d & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_a & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -I_d & 0 & I_d & 0 & 0 & 0 \\ 0 & 0 & 0 & -I_a & 0 & I_a & 0 & 0 \\ \hat{F}_{1;x_1} & \hat{F}_{1;x_2} & \hat{F}_{1;\dot{x}_1} & \hat{F}_{1;\dot{x}_2} & 0 & 0 & 0 & 0 \\ \hat{F}_{2;x_1} & \hat{F}_{2;x_2} & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right],$$

$$\tilde{N}_1 = \left[\begin{array}{cccc|cccc} 0 & 0 & I_d & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_a & 0 & 0 & 0 & 0 \\ -\hat{F}_{1;x_1} & -\hat{F}_{1;x_2} & -\hat{F}_{1;\dot{x}_1} & -\hat{F}_{1;\dot{x}_2} & 0 & 0 & 0 & 0 \\ -\hat{F}_{2;x_1} & -\hat{F}_{2;x_2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ * & * & * & * & 0 & 0 & 0 & 0 \\ * & * & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Hence, rank $\tilde{M}_1 = 2n + d$, where $n = a + d$ is the original system size, and

$$\tilde{Z}_2^T = \left[\begin{array}{cccc|cccc} 0 & 0 & I_d & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_a & 0 & 0 & 0 & 0 \\ -\hat{F}_{2;x_1} & -\hat{F}_{2;x_2} & 0 & 0 & 0 & 0 & 0 & I_a \end{array}\right].$$

We then get that

$$\tilde{Z}_2^T \tilde{N}_1 [\, I \;\; 0 \,]^T = \begin{bmatrix} -\hat{F}_{1;x_1} & -\hat{F}_{1;x_2} & -\hat{F}_{1;\dot{x}_1} & -\hat{F}_{1;\dot{x}_2} \\ -\hat{F}_{2;x_1} & -\hat{F}_{2;x_2} & 0 & 0 \\ * & * & -\hat{F}_{2;x_1} & -\hat{F}_{2;x_2} \end{bmatrix},$$

which has full row rank $n + a$ due to the properties of (4.23). Finally,

$$F_{\dot{x}} \tilde{T}_2 = \begin{bmatrix} I_d & 0 & 0 & 0 \\ 0 & I_a & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} I_d \\ * \\ * \\ * \end{bmatrix} = \begin{bmatrix} I_d \\ * \\ 0 \\ 0 \end{bmatrix}$$

shows that (4.45) satisfies Hypothesis 4.2 with $\tilde{\mu} = 1$, $\tilde{a} = n + a$, and $\tilde{d} = d$. Choosing $\tilde{Z}_1^T = [\, I_d \;\; 0 \;\; 0 \;\; 0 \,]^T$ yields the reduced differential-algebraic equation

$$\dot{x}_1 = y_1, \quad x_2 = \mathcal{R}(x_1), \quad y_1 = \mathcal{L}(x_1), \quad y_2 = \mathcal{R}_{x_1}(x_1)\mathcal{L}(x_1).$$

In particular, this system implies (4.23) such that there are no problems arising from the hidden constraints that are introduced by increasing the index. By this analysis, we have proved the following theorem.

Theorem 4.28. *Consider a strangeness-free differential-algebraic equation in the form* (4.23) *with a sufficiently smooth solution. Then* (4.45) *satisfies Hypothesis* 4.2 *with characteristic values* $\tilde{\mu} = 1$, $\tilde{a} = n + a$, *and* $\tilde{d} = d$.

Remark 4.29. The differentiation index, defined in Definition 3.37 for linear differential-algebraic equations, can be generalized to nonlinear problems (4.1), see [53], [54]. Taking the definition of [54], the differentiation index then typically coincides with the notion of index as we have used it in this section in the context of semi-explicit differential-algebraic equations. This definition, however, has not the same invariance properties as Hypothesis 4.2. In particular, a result similar to that of Lemma 4.7 does not hold.

4.3 Over- and underdetermined problems

So far in this chapter, we have restricted the analysis to regular systems with $m = n$. In this section, we generalize the results of Section 4.1 to possibly over- or underdetermined problems

$$F(t, x, \dot{x}) = 0, \tag{4.46}$$

i.e., with $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^m)$, $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open. In order to apply the same techniques as before, we must explicitly require that $\mathbb{L}_\mu = F_\mu^{-1}(\{0\})$ is a manifold. We therefore consider the following generalization of Hypothesis 4.2. For convenience, we omit the function arguments.

Hypothesis 4.30. *There exist integers* μ, r, a, d, *and* v *such that the set*

$$\mathbb{L}_\mu = \left\{(t, x, \dot{x}, \ldots, x^{(\mu+1)}) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = 0\right\}, \tag{4.47}$$

associated with F *is nonempty and such that for every* $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ *there exists a* (*sufficiently small*) *neighborhood in which the following properties hold:*

1. *The set* $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)n+1}$ *forms a manifold of dimension* $(\mu + 2)n + 1 - r$.
2. *We have* $\operatorname{rank} F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}} = r$ *on* $\mathbb{L}_\mu$.
3. *We have* $\operatorname{corank} F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}} - \operatorname{corank} F_{\mu-1;x,\dot{x},\ldots,x^{(\mu)}} = v$ *on* $\mathbb{L}_\mu$, *with the convention that* $\operatorname{corank} F_{-1;x} = 0$.

4. *We have* rank $M_\mu = r - a$ *on* $\mathbb{L}_\mu$ *such that there exist smooth matrix functions* Z_2 *and* T_2 *of size* $(\mu+1)m \times a$ *and* $n \times (n-a)$*, respectively, and pointwise maximal rank, satisfying* $Z_2^T M_\mu = 0$ *on* $\mathbb{L}_\mu$ *as well as* rank $Z_2^T F_{\mu;x} = a$ *and* $Z_2^T F_{\mu;x} T_2 = 0$.

5. *We have* rank $F_{\dot{x}} T_2 = d = m - a - v$ *such that there exists a smooth matrix function* Z_1 *of size* $n \times d$ *and pointwise maximal rank, satisfying* rank $Z_1^T F_{\dot{x}} T_2 = d$.

For square systems without redundancies, i.e., $m = n$ and $v = 0$, Hypothesis 4.30 reduces to Hypothesis 4.2. According to Definition 4.4, we again call the smallest possible μ in Hypothesis 4.30 the *strangeness-index* of (4.46). Systems with vanishing strangeness index are still called *strangeness-free*.

To derive the implications of Hypothesis 4.30 and to motivate the various assumptions, we proceed as follows. Compare with the more special approach of Section 4.1.

Let $z_{\mu,0} = (t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ be fixed. Since by assumption, $\mathbb{L}_\mu$ is a manifold of dimension $(\mu+2)n + 1 - r$, we can locally parameterize it by $(\mu+2)n + 1 - r$ parameters. These can be chosen from $(t, x, \dot{x}, \ldots, x^{(\mu+1)})$ in such a way that discarding the associated columns from

$$F_{\mu;t,x,\dot{x},\ldots,x^{(\mu+1)}}(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)})$$

does not lead to a rank drop. Because of Part 2 of Hypothesis 4.30, already $F_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}}$ has maximal rank. Hence, we can always choose t as a parameter.

Because of Part 4 of Hypothesis 4.30, we can choose $n - a$ parameters out of x. Without restriction we can write x as (x_1, x_2, x_3) with $x_1 \in \mathbb{R}^d$, $x_2 \in \mathbb{R}^{n-a-d}$, $x_3 \in \mathbb{R}^a$, and choose (x_1, x_2) as further parameters. In particular, the matrix $Z_2^T F_{\mu;x_3}$ is then nonsingular. The remaining parameters $p \in \mathbb{R}^{(\mu+1)n+a-r}$ can be chosen out of $(\dot{x}, \ldots, x^{(\mu+1)})$.

Therefore, Hypothesis 4.30 implies that there exists a neighborhood $\mathbb{V} \subseteq \mathbb{R}^{(\mu+2)n+1-r}$ of $(t_0, x_{1,0}, x_{2,0}, p_0)$ as part of $z_{\mu,0}$, corresponding to the selected parameters (t, x_1, x_2, p), and a neighborhood $\tilde{\mathbb{U}} \subseteq \mathbb{R}^{(\mu+2)n+1}$ of $z_{\mu,0}$ such that

$$\mathbb{U} = \mathbb{L} \cap \tilde{\mathbb{U}} = \{\theta(t, x_1, x_2, p) \mid (t, x_1, x_2, p) \in \mathbb{V}\},$$

where $\theta : \mathbb{V} \to \mathbb{U}$ is a diffeomorphism. Again we may assume that $\mathbb{V}$ is an open ball with radius $\varepsilon > 0$ and center $(t_0, x_{1,0}, x_{2,0}, p_0)$.

In this way, we have obtained that $F_\mu(z_\mu) = 0$ holds locally if and only if $z_\mu = \theta(t, x_1, x_2, p)$ for some $(t, x_1, x_2, p) \in \mathbb{U}$. In particular, there exist functions $\mathcal{G}$, corresponding to x_3, and $\mathcal{H}$, corresponding to $(\dot{x}, \ldots, x^{(\mu+1)})$ such that

$$F_\mu(t, x_1, x_2, \mathcal{G}(t, x_1, x_2, p), \mathcal{H}(t, x_1, x_2, p)) = 0 \tag{4.48}$$

on $\mathbb{V}$. As in Section 4.1, it follows that there exists a function $\mathcal{R}$ such that

$$x_3 = \mathcal{G}(t, x_1, x_2, p) = \mathcal{G}(t, x_1, x_2, p_0) = \mathcal{R}(t, x_1, x_2)$$

and

$$F_\mu(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \mathcal{H}(t, x_1, x_2, p)) = 0$$

on $\mathbb{V}$. Similarly, we can choose T_2 of Hypothesis 4.30 as

$$T_2(t, x_1, x_2) = \begin{bmatrix} I \\ \mathcal{R}_{x_1,x_2}(t, x_1, x_2) \end{bmatrix}.$$

Thus, Part 5 of Hypothesis 4.30 yields a matrix function Z_1 which only depends on the original variables $(t, x, \dot{x})$. Again, due to the full rank assumption, we can choose the neighborhood $\mathbb{V}$ so small that we can take a constant Z_1. The corresponding *reduced differential-algebraic equation* therefore reads

$$\hat{F}(t, x, \dot{x}) = \begin{bmatrix} \hat{F}_1(t, x, \dot{x}) \\ \hat{F}_2(t, x) \end{bmatrix} = 0, \tag{4.49}$$

with

$$\begin{aligned} \hat{F}_1(t, x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) &= Z_1^T F(t, x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3), \\ \hat{F}_2(t, x_1, x_2, x_3) &= Z_2^T F_\mu(t, x_1, x_2, x_3, \mathcal{H}(t, x_1, x_2, p_0)). \end{aligned} \tag{4.50}$$

Analogous to the construction in Section 4.1, it then follows that (4.49) satisfies Hypothesis 4.30 with characteristic values $\mu = 0, r = a+d, a, d$, and v. Similarly, it follows that $\hat{F}_2(t, x_1, x_2, x_3) = 0$ is locally equivalent to $x_3 = \mathcal{R}(t, x_1, x_2)$. Differentiating the latter relation, we can eliminate x_3 and $\dot{x}_3$ in the first equation of (4.49) to obtain

$$\begin{aligned} \hat{F}_1(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \dot{x}_1, \dot{x}_2, \mathcal{R}_t(t, x_1, x_2) \\ + \mathcal{R}_{x_1}(t, x_1, x_2)\dot{x}_1 + \mathcal{R}_{x_2}(t, x_1, x_2)\dot{x}_2) = 0. \end{aligned} \tag{4.51}$$

If the function $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ solves (4.49) in its domain of definition, then the point $(t_0, x_1^*(t_0), x_2^*(t_0), \dot{x}_1^*(t_0), \dot{x}_2^*(t_0))$ solves (4.51). By Part 5 of Hypothesis 4.30, it then follows that this system can be solved locally for $\dot{x}_1$. In this way, we obtain a decoupled differential-algebraic equation of the form

$$\dot{x}_1 = \mathcal{L}(t, x_1, x_2, \dot{x}_2), \quad x_3 = \mathcal{R}(t, x_1, x_2). \tag{4.52}$$

Obviously, in this system $x_2 \in C^1(\mathbb{I}, \mathbb{R}^{n-a-d})$ can be chosen arbitrarily (at least when staying in the domain of definition of $\mathcal{R}$ and $\mathcal{L}$), while the resulting system has locally a unique solution for x_1 and x_3 provided that an initial condition is given that satisfies the algebraic constraint. In this way, we have proved the following theorem.

Theorem 4.31. *Let F as in (4.46) be sufficiently smooth and satisfy Hypothesis 4.30 with characteristic values μ, r, a, d, and v. Then every sufficiently smooth solution of (4.46) also solves the reduced differential-algebraic equations (4.49) and (4.52) consisting of d differential and a algebraic equations.*

Proof. The proof is analogous to that of Theorem 4.11. □

So far, we have not discussed the quantity v. This quantity measures the number of equations in the original system that give rise to trivial equations $0 = 0$, i.e., it counts the number of redundancies in the system. Together with a and d it gives a complete classification of the m equations into d differential equations, a algebraic equations and v trivial equations. Of course, trivial equations can be simply removed without altering the solution set. Omitting Part 3 of Hypothesis 4.30, however, would mean that a given problem may satisfy the modified hypothesis for different characteristic values of a and d.

Example 4.32. Consider the differential-algebraic equation

$$F(t, x, \dot{x}) = \begin{bmatrix} \dot{x}_2 \\ \log x_2 \end{bmatrix} = 0, \tag{4.53}$$

with $m = 2$ equations and $n = 2$ unknowns x_1, x_2. To check Hypothesis 4.30 for $\mu = 0$, we consider the set

$$\mathbb{L}_0 = \big\{(t, x_1, x_2, \dot{x}_1, \dot{x}_2) \mid x_2 = 1,\ \dot{x}_2 = 0\big\}.$$

Obviously, $\mathbb{L}_0$ is a manifold parameterized by $(t, x_1, \dot{x}_1)$. Furthermore, we have

$$F_{0;\dot{x}} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad F_{0;x} = \begin{bmatrix} 0 & 0 \\ 0 & x_2^{-1} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

on $\mathbb{L}_0$. Thus,

$$\operatorname{rank} F_{0;x,\dot{x}} = 2, \quad \operatorname{corank} F_{0;x,\dot{x}} = 0, \quad \operatorname{rank} F_{0;\dot{x}} = 1.$$

With $Z_2^T = [\,0\ \ 1\,]$, we then obtain

$$\operatorname{rank} Z_2^T F_{0;x} = \operatorname{rank}[\,0\ \ 1\,] = 1,$$

and, with $T_2^T = [\,1\ \ 0\,]$, finally

$$\operatorname{rank} F_{\dot{x}} T_2 = 0.$$

Hence, we get the quantities $r = 2$, $v = 0$, $a = 1$, and $d = 0$. Hypothesis 4.30 is not satisfied, since $d \neq m - a - v = 1$. If we would drop Part 3 of Hypothesis 4.30, then there would be no condition on v and we could simply choose $v = 1$ to satisfy

all remaining requirements. To check Hypothesis 4.30 for $\mu = 1$, we must consider the next level of the derivative array $F_1 = 0$, which consists of the equations

$$\dot{x}_2 = 0, \quad \log x_2 = 0, \quad \ddot{x}_2 = 0, \quad \frac{\dot{x}_2}{x_2} = 0.$$

The set

$$\mathbb{L}_1 = \left\{(t, x_1, x_2, \dot{x}_1, \dot{x}_2, \ddot{x}_1, \ddot{x}_2) \mid x_2 = 1,\ \dot{x}_2 = 0,\ \ddot{x}_2 = 0\right\}$$

is a manifold parameterized by $(t, x_1, \dot{x}_1, \ddot{x}_1)$. Furthermore, we have

$$F_{1;\dot{x},\ddot{x}} = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & x_2^{-1} & 0 & 0 \end{array}\right] = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{array}\right]$$

and

$$F_{1;x} = \left[\begin{array}{cc} 0 & 0 \\ 0 & x_2^{-1} \\ \hline 0 & 0 \\ 0 & -x_2^{-2}\dot{x}_2 \end{array}\right] = \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 \\ \hline 0 & 0 \\ 0 & 0 \end{array}\right]$$

on $\mathbb{L}_1$. Thus,

$$\operatorname{rank} F_{1;x,\dot{x},\ddot{x}} = 3, \quad \operatorname{corank} F_{1;x,\dot{x},\ddot{x}} = 1, \quad \operatorname{rank} F_{1;\dot{x},\ddot{x}} = 2.$$

Proceeding as above, we compute

$$Z_2^T = \left[\begin{array}{cc|cc} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -1 \end{array}\right], \quad T_2 = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

and

$$\operatorname{rank} Z_2^T F_{0;x} = \operatorname{rank} \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = 1, \quad \operatorname{rank} F_{\dot{x}} T_2 = \operatorname{rank} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0.$$

Hence, Hypothesis 4.30 is satisfied with $\mu = 1$, $r = 3$, $v = 1$, $a = 1$, and $d = 0$.

Remark 4.33. It should be noted that in practical systems, due to modeling simplifications, measurement errors for the coefficients or round-off errors in the rank computations, these redundant equations of the form $0 = 0$ may get perturbed to equations of the form $0 = \varepsilon$ with small $|\varepsilon|$. In general, this creates large difficulties in the numerical methods, in particular, in the process of determining the characteristic values μ, a, d, v and in the solution of the nonlinear systems, see Chapter 6.

To show that the reduced systems (4.49) and (4.52) reflect (at least locally) the properties of the original system concerning solvability and structure of the solution set, we need the following theorem, which generalizes Theorem 4.13. Since the part x_2 in (4.48) requires a different treatment than the parts in Theorem 4.13, we include a detailed proof.

Theorem 4.34. *Let F as in (4.46) be sufficiently smooth and satisfy Hypothesis 4.30 with characteristic values μ, a, d, v and with characteristic values $\mu+1$ (replacing μ), a, d, v. Let $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$ be given and let the parameterization p in (4.48) for $F_{\mu+1}$ include $\dot{x}_2$. Then, for every function $x_2 \in C^1(\mathbb{I}, \mathbb{R}^{n-a-d})$ with $x_2(t_0) = x_{2,0}$, $\dot{x}_2(t_0) = \dot{x}_{2,0}$, the reduced differential-algebraic equations (4.49) and (4.52) have unique solutions x_1 and x_3 satisfying $x_1(t_0) = x_{1,0}$. Moreover, the so obtained function $x = (x_1, x_2, x_3)$ locally solves the original problem.*

Proof. By assumption, there exists (locally with respect to $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$) a parameterization (t, x_1, x_2, p), where p is chosen out of $(\dot{x}, \ldots, x^{(\mu+2)})$, with

$$F_{\mu+1}(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \mathcal{H}(t, x_1, x_2, p)) \equiv 0.$$

This includes the equation

$$F_\mu(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \mathcal{H}(t, x_1, x_2, p)) \equiv 0, \tag{4.54}$$

with trivial dependence on $x^{(\mu+2)}$, as well as the equation

$$\tfrac{d}{dt} F_\mu(t, x_1, x_2, \mathcal{R}(t, x_1, x_2), \mathcal{H}(t, x_1, x_2, p)) \equiv 0. \tag{4.55}$$

Equation (4.54) implies that (omitting arguments)

$$F_{\mu;t} + F_{\mu;x_3}\mathcal{R}_t + F_{\mu;\dot{x},\ldots,x^{(\mu+2)}}\mathcal{H}_t \equiv 0, \tag{4.56a}$$
$$F_{\mu;x_1,x_2} + F_{\mu;x_3}\mathcal{R}_{x_1,x_2} + F_{\mu;\dot{x},\ldots,x^{(\mu+2)}}\mathcal{H}_{x_1,x_2} \equiv 0, \tag{4.56b}$$
$$F_{\mu;\dot{x},\ldots,x^{(\mu+2)}}\mathcal{H}_p \equiv 0. \tag{4.56c}$$

The relation $\frac{d}{dt}F_\mu = 0$ has the form

$$F_{\mu;t} + F_{\mu;x_1}\dot{x}_1 + F_{\mu;x_2}\dot{x}_2 + F_{\mu;x_3}\dot{x}_3 + F_{\mu;\dot{x},\ldots,x^{(\mu+1)}} \begin{bmatrix} \ddot{x} \\ \vdots \\ x^{(\mu+2)} \end{bmatrix} = 0.$$

Inserting the parameterization yields that (4.55) can be written as

$$F_{\mu;t} + F_{\mu;x_1}\mathcal{H}_1 + F_{\mu;x_2}\mathcal{H}_2 + F_{\mu;x_3}\mathcal{H}_3 + F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}\mathcal{H}_4 \equiv 0,$$

where $\mathcal{H}_i$, $i = 1, \ldots, 4$, are the parts of $\mathcal{H}$ corresponding to $\dot{x}_1$, $\dot{x}_2$, $\dot{x}_3$, and the remaining variables, respectively. Multiplication with Z_2^T (corresponding to Hypothesis 4.30 with characteristic values μ, a, d, v) gives

$$Z_2^T F_{\mu;t} + Z_2^T F_{\mu;x_1} \mathcal{H}_1 + Z_2^T F_{\mu;x_2} \mathcal{H}_2 + Z_2^T F_{\mu;x_3} \mathcal{H}_3 \equiv 0.$$

Inserting the relations (4.56) and observing that $Z_2^T F_{\mu;x_3}$ is nonsingular, we find that

$$Z_2^T F_{\mu;x_3} (\mathcal{H}_3 - \mathcal{R}_t - \mathcal{R}_{x_1} \mathcal{H}_1 - \mathcal{R}_{x_2} \mathcal{H}_2) \equiv 0,$$

or

$$\mathcal{H}_3 = \mathcal{R}_t + \mathcal{R}_{x_1} \mathcal{H}_1 + \mathcal{R}_{x_2} \mathcal{H}_2,$$

i.e.,

$$\dot{x}_3 = \mathcal{R}_t + \mathcal{R}_{x_1} \dot{x}_1 + \mathcal{R}_{x_2} \dot{x}_2.$$

In summary, the derivative array equation $F_{\mu+1} = 0$ implies that

$$Z_1^T F(t, x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) = 0, \tag{4.57a}$$

$$x_3 = \mathcal{R}(t, x_1, x_2), \tag{4.57b}$$

$$\dot{x}_3 = \mathcal{R}_t(t, x_1, x_2) + \mathcal{R}_{x_1} \dot{x}_1(t, x_1, x_2) + \mathcal{R}_{x_2}(t, x_1, x_2) \dot{x}_2. \tag{4.57c}$$

Elimination of x_3 and $\dot{x}_3$ from (4.57a) gives

$$\dot{x}_1 = \mathcal{L}(t, x_1, x_2, \dot{x}_2).$$

In particular, this shows that $\dot{x}_1$ and $\dot{x}_3$ are not part of the parameterization.

Since $\dot{x}_2$ is part of p, the following construction is possible. Let $x_2 = x_2(t)$ and $\dot{x}_2 = \dot{x}_2(t)$. Let $p = p(t)$ be arbitrary but consistent to the choice of $\dot{x}_2$ and to the initial value $z_{\mu+1,0}$. Finally, let $x_1 = x_1(t)$ and $x_3 = x_3(t)$ be the solution of the initial value problem

$$Z_1^T F(t, x_1, x_2(t), x_3, \dot{x}_1, \dot{x}_2(t), \dot{x}_3) = 0, \quad x_1(t_0) = x_{1,0}, \quad x_3 = \mathcal{R}(t, x_1, x_2(t)).$$

Although $\dot{x}_1$ and $\dot{x}_3$ are not part of the parameterization, we automatically get $\dot{x}_1 = \dot{x}_1(t)$ and $\dot{x}_3 = \dot{x}_3(t)$. Thus, we have

$$F_{\mu+1}(t, x_1(t), x_2(t), x_3(t), \dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t), \mathcal{H}_4(t, x_1(t), x_2(t), p(t)) \equiv 0$$

for all t in a neighborhood of t_0, or

$$F(t, x_1(t), x_2(t), x_3(t), \dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t)) \equiv 0$$

for the first block. □

Remark 4.35. Let the assumptions of Theorem 4.34 hold and let $x_{2,0}$ and $\dot{x}_{2,0}$ be the part of $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$ belonging to x_2 and $\dot{x}_2$. If $\tilde{x}_{2,0}$ and $\dot{\tilde{x}}_{2,0}$ are sufficiently close to $x_{2,0}$ and $\dot{x}_{2,0}$, then they are part of a $\tilde{z}_{\mu+1,0} \in \mathbb{L}_{\mu+1}$ close to $z_{\mu+1,0}$ and we can apply Theorem 4.34 with $z_{\mu+1,0}$ replaced by $\tilde{z}_{\mu+1,0}$.

Remark 4.36. Note that in Theorem 4.34 we can drop the assumption that $\dot{x}_2$ is part of the parameters if we know from the structure of the problem that $\mathcal{L}$ in (4.52) does not depend on $\dot{x}_2$. In particular, this is the case if we can choose the splitting (x_1, x_2, x_3) in such a way that the original problem does not depend on $\dot{x}_2$ and on components of $\dot{x}_3$ that depend on $\dot{x}_2$. An important consequence of this special case is that we do not need to require the initial condition $\dot{x}_2(t_0) = \dot{x}_{2,0}$. This also applies to Remark 4.35.

Remark 4.37. The reduced differential-algebraic equations (4.49) and (4.52) may already follow from $F_\ell = 0$ with $\ell < \mu$, although μ is chosen as small as possible. This occurs in cases when further differentiations only lead to trivial equations $0 = 0$ and consistency is guaranteed. To check the consistency of the model, however, it is still necessary to consider $F_\mu = 0$.

Example 4.38. Consider the problem of Example 4.32. The reduced differential-algebraic equation simply consists of $\log x_2 = 0$ and is already implied by $F_0 = 0$. The same holds for the slightly modified differential-algebraic equation

$$\dot{x}_2 = 1, \quad \log x_2 = 0.$$

Observe that the corresponding set $\mathbb{L}_0$ is nonempty. Differentiating once gives

$$\ddot{x}_2 = 0, \quad x_2^{-1}\dot{x}_2 = 0,$$

implying the contradiction $\dot{x}_2 = 0$. Thus, $\mathbb{L}_1$ is empty and the modified problem is not solvable.

4.4 Control problems

In the general nonlinear case, control problems have the form

$$F(t, x, u, \dot{x}) = 0, \tag{4.58a}$$
$$y - G(t, x) = 0, \tag{4.58b}$$

where $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_u \times \mathbb{D}_{\dot{x}}, \mathbb{R}^m)$ and $G \in C(\mathbb{I} \times \mathbb{D}_x, \mathbb{R}^p)$ with $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$, $\mathbb{D}_u \subseteq \mathbb{R}^l$ open. As usual, x represents the state, u the input, and y the output of the system.

In a first step, we omit the output equation. Using a behavior approach as in Section 3.6, i.e., setting

$$z = \begin{bmatrix} x \\ u \end{bmatrix},$$

we assume that F, rewritten with respect to the unknown z, satisfies Hypothesis 4.30. Note that we must replace n in Hypothesis 4.30 by $n+l$. According to the previous section, we locally get a reduced problem

$$\begin{aligned} \hat{F}_1(t, x, u, \dot{x}) &= 0, \\ \hat{F}_2(t, x, u) &= 0 \end{aligned} \tag{4.59}$$

corresponding to (4.49). To perform the next steps of the construction would require to split z into (z_1, z_2, z_3), where each part may consist of components of both x and u. To avoid such a splitting, which would mix input and state variables, we proceed as follows. Starting from (4.49) in the form

$$\begin{aligned} \hat{F}_1(t, z, \dot{z}) &= 0, \\ \hat{F}_2(t, z) &= 0, \end{aligned}$$

Hypothesis 4.30 yields (without arguments)

$$\hat{F}_{2;z} T_2 = 0, \quad \operatorname{rank} T_2 = n + l - a, \quad \operatorname{rank} \hat{F}_{1;\dot{z}} = d.$$

Choosing T_2' such that $[\, T_2' \;\; T_2 \,]$ is nonsingular, we find that

$$\operatorname{rank} \begin{bmatrix} \hat{F}_{1;\dot{z}} \\ \hat{F}_{2;z} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \hat{F}_{1;\dot{z}} T_2' & \hat{F}_{1;\dot{z}} T_2 \\ \hat{F}_{2;z} T_2' & 0 \end{bmatrix} = \operatorname{rank} \hat{F}_{1;\dot{z}} T_2 + \operatorname{rank} \hat{F}_{2;z} T_2' = d + a.$$

Thus, the given matrix function has pointwise full row rank. In the present context, this means that the matrix function

$$\begin{bmatrix} \hat{F}_{1;\dot{x}} & 0 \\ \hat{F}_{2;x} & \hat{F}_{2;u} \end{bmatrix} \tag{4.60}$$

of size $(d+a) \times (n+l)$ has full row rank. Observe that, in general, fixing a control u does not give a regular strangeness-free reduced problem, since

$$\begin{bmatrix} \hat{F}_{1;\dot{x}} \\ \hat{F}_{2;x} \end{bmatrix}$$

may be singular. An immediate question is whether it is possible to choose a control such that the resulting reduced problem is regular and strangeness-free.

Necessarily, for this we must have $d + a = n$. As in the linear case, we consider feedback controls. In the nonlinear case, a state feedback has the form

$$u = K(t, x), \tag{4.61}$$

leading to a closed loop reduced problem

$$\begin{aligned} \hat{F}_1(t, x, K(t, x), \dot{x}) &= 0, \\ \hat{F}_2(t, x, K(t, x)) &= 0. \end{aligned} \tag{4.62}$$

The condition for this system to be regular and strangeness-free reads

$$\begin{bmatrix} \hat{F}_{1;\dot{x}} \\ \hat{F}_{2;x} + \hat{F}_{2;u} K_x \end{bmatrix} \quad \text{nonsingular.}$$

Since the reduced system is only defined locally, it is sufficient to satisfy this condition only locally. Thus, we can restrict ourselves to linear feedbacks

$$u(t) = \tilde{K} x(t) + w(t), \tag{4.63}$$

such that $K_x = \tilde{K}$. Since (4.60) has full row rank, the existence of a suitable K_x follows from Corollary 3.81. The function w can be used to satisfy initial conditions of the form

$$u^{(\ell)}(t_0) = \tilde{K} x_0^{(\ell)} + w^{(\ell)}(t_0) = u_0^{(\ell)}. \tag{4.64}$$

Hence, we have proved the following theorem.

Theorem 4.39. *Suppose that the control problem* (4.58a) *in behavior form satisfies Hypothesis* 4.30 *with characteristic values μ, a, d, v and assume that $d + a = n$. Finally, let $z_{\mu,0} = (t_0, x_0, u_0, \dots, x_0^{(\mu+1)}, u_0^{(\mu+1)}) \in \mathbb{L}_\mu$. Then there (locally) exists a state feedback $u = K(t, x)$ satisfying $u_0 = K(t_0, x_0)$ and $\dot{u}_0 = K_t(t_0, x_0) + K_x(t_0, x_0)\dot{x}_0$ such that the closed loop reduced problem is regular and strangeness-free.*

Corollary 4.40. *Suppose that the control problem* (4.58a) *in behavior form satisfies Hypothesis* 4.30 *with μ, a, d, v and with $\mu + 1$ (replacing μ), a, d, v and assume that $d + a = n$. Furthermore, let u be a control in the sense that u and $\dot{u}$ can be chosen as part of the parameterization of $\mathbb{L}_{\mu+1}$ at $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$. Let $u = K(t, x)$ be a state feedback which satisfies the initial conditions $u_0 = K(t_0, x_0)$ and $\dot{u}_0 = K_t(t_0, x_0) + K_x(t_0, x_0)\dot{x}_0$ and yields a regular and strangeness-free closed loop reduced system. Then, the closed loop reduced problem has a unique solution satisfying the initial values given by $z_{\mu+1,0}$. Moreover, this solution locally solves the closed loop problem*

$$F(t, x, K(t, x), \dot{x}) = 0.$$

Proof. The proof can be carried out along the lines of the proof of Theorem 4.34, see Exercise 18. □

Example 4.41. Consider the control problem

$$F(t, x, u, \dot{x}) = \begin{bmatrix} \dot{x}_2 \\ \log x_2 + \sin u \end{bmatrix} = 0,$$

with $n = 2$ and $l = 1$. The corresponding behavior system reads

$$F(t, z, \dot{z}) = \begin{bmatrix} \dot{z}_2 \\ \log z_2 + \sin z_3 \end{bmatrix} = 0.$$

To check Hypothesis 4.30 for $\mu = 0$, we must consider

$$\mathbb{L}_0 = \big\{(t, x_1, x_2, u, \dot{x}_1, \dot{x}_2, \dot{u}) \mid x_2 = \exp(-\sin u),\ \dot{x}_2 = 0\big\}.$$

Obviously, $\mathbb{L}_0$ is a manifold parameterized by $(t, x_1, u, \dot{x}_1, \dot{u})$. Furthermore, we have

$$F_{0;\dot{z}} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad F_{0;z} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & x_2^{-1} & \cos u \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \exp(\sin u) & \cos u \end{bmatrix}$$

on $\mathbb{L}_0$. Thus,

$$\operatorname{rank} F_{0;z,\dot{z}} = 2, \quad \operatorname{corank} F_{0;z,\dot{z}} = 0, \quad \operatorname{rank} F_{0;\dot{z}} = 1.$$

With $Z_2^T = [\,0\ \ 1\,]$, we then obtain

$$\operatorname{rank} Z_2^T F_{0;z} = \operatorname{rank}[\,0\ \ \exp(\sin u)\ \ \cos u\,] = 1, \quad T_2 = \begin{bmatrix} 1 & 0 \\ 0 & -\cos u \\ 0 & \exp(\sin u) \end{bmatrix},$$

and finally

$$\operatorname{rank} F_{\dot{z}} T_2 = \operatorname{rank} \begin{bmatrix} 0 & -\cos u \\ 0 & 0 \end{bmatrix} = 1,$$

when we restrict u to a neighborhood of zero. Hence, Hypothesis 4.30 is satisfied with $\mu = 0$, $v = 0$, $a = 1$, and $d = 1$. For $z_{0,0} = (0, 0, 1, 0, 0, 0, 0)$ we can choose $Z_1^T = [\,1\ \ 0\,]$ to obtain the reduced problem

$$\dot{x}_2 = 0, \quad \log x_2 + \sin u = 0.$$

Note that the reduced problem here coincides with the original problem due to its special form (we have $\mu = 0$ and do not need to apply any transformations to separate the algebraic equations) and due to the special choice for Z_1. Fixing the

control u according to $u = 0$ gives a closed loop system that is not regular and strangeness-free. Indeed, it satisfies Hypothesis 4.30 only for $\mu = 1$ and it even includes a trivial equation due to a redundancy, cp. Example 4.32. To get a regular and strangeness-free closed-loop reduced problem, we look for a regularizing state feedback. Since

$$\begin{bmatrix} \hat{F}_{1;\dot{x}} & 0 \\ \hat{F}_{2;x} & \hat{F}_{2;u} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & x_2^{-1} & \cos u \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}$$

at $z_{0,0}$, we can choose $\tilde{K} = [\,1\ \ 0\,]$ or $u = x_1$ observing the initial values given by $z_{0,0}$. The corresponding closed loop reduced problem is given by

$$\dot{x}_2 = 0, \quad \log x_2 + \sin x_1 = 0.$$

By construction, it is regular and strangeness-free near the initial value given by $z_{0,0}$. For $x_1(0) = 0$, we particularly get the unique solution $x_1(t) = 0$, $x_2(t) = 1$.

We turn now to control problems that include the output equation (4.58b). In a behavior framework, we set

$$z = \begin{bmatrix} x \\ u \\ y \end{bmatrix}$$

and again apply the theory of the previous section. Due to the explicit form of the output equation, it is obvious that it becomes part of the algebraic constraints and does not affect the other constraints, cp. also the linear case of Section 3.6. Therefore, assuming that F satisfies Hypothesis 4.30, the reduced differential-algebraic equation has the form

$$\begin{aligned} &\hat{F}_1(t, x, u, \dot{x}) = 0, \\ &\hat{F}_2(t, x, u) = 0, \\ &y = G(t, x). \end{aligned} \tag{4.65}$$

If we consider output feedbacks of the form

$$u = K(t, y), \tag{4.66}$$

then the closed loop reduced problem has the form

$$\begin{aligned} &\hat{F}_1(t, x, K(t, G(t, x)), \dot{x}) = 0, \\ &\hat{F}_2(t, x, K(t, G(t, x))) = 0. \end{aligned} \tag{4.67}$$

The condition for this system to be regular and strangeness-free reads

$$\begin{bmatrix} \hat{F}_{1;\dot{x}} \\ \hat{F}_{2;x} + \hat{F}_{2;u} K_y G_x \end{bmatrix} = \begin{bmatrix} \hat{F}_{1;\dot{x}} & 0 \\ \hat{F}_{2;x} & \hat{F}_{2;u} \end{bmatrix} \begin{bmatrix} I \\ K_y G_x \end{bmatrix} \quad \text{nonsingular.} \tag{4.68}$$

Note that we get back the state feedback case if $y = x$. To look at (4.68) more closely, we may proceed as for (3.127) in the linear case. In particular, we can set $E_1 = \hat{F}_{1;\dot{x}}$, $A_2 = \hat{F}_{2;x}$, $B_2 = \hat{F}_{2;u}$ and $C = G_x$ and determine the quantities ϕ and ω at a point $(t_0, z_0, \dot{z}_0)$ given by $z_{\mu,0} \in \mathbb{L}_\mu$ as in the construction following (3.127). Recall that we have omitted the hats in the notation when we deal with nonlinear problems. In this way we get a nonlinear version of Corollary 3.82.

Corollary 4.42. *Suppose that the output control problem* (4.58) *in behavior form satisfies Hypothesis* 4.30 *with* μ, a, d, v *and assume that* $d+a = n$ *and* $\phi = \omega$. *Then there (locally) exists an output feedback* $u = K(t, y)$ *satisfying* $u_0 = K(t_0, y_0)$ *and* $\dot{u}_0 = K_t(t_0, y_0) + K_x(t_0, y_0)\dot{y}_0$ *such that the closed loop reduced problem is regular and strangeness-free.*

Proof. Under the given assumptions, the linear theory of Section 3.6 yields a suitable matrix $\tilde{K} = K_y$ such that (4.68) holds. The claim then follows for the linear output feedback

$$u(t) = \tilde{K}y(t) + w(t),$$

where the function w is used to satisfy the given initial conditions. □

Corollary 4.43. *Suppose that the output control problem* (4.58) *in behavior form satisfies Hypothesis* 4.30 *with* μ, a, d, v *and with* $\mu + 1$ *(replacing* μ*),* a, d, v, *and assume that* $d + a = n$ *and* $\phi = \omega$. *Furthermore, let* u *be a control in the sense that* u *and* $\dot{u}$ *can be chosen as part of the parameterization of* $\mathbb{L}_{\mu+1}$ *at* $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$. *Let* $u = K(t, y)$ *be an output feedback which satisfies the initial conditions* $u_0 = K(t_0, y_0)$ *and* $\dot{u}_0 = K_t(t_0, y_0) + K_x(t_0, y_0)\dot{y}_0$ *and yields a regular and strangeness-free closed loop reduced system. Then, the closed loop reduced problem has a unique solution satisfying the initial values given by* $z_{\mu+1,0}$. *Moreover, this solution locally solves the closed loop problem*

$$F(t, x, K(t, G(t, x)), \dot{x}) = 0.$$

Proof. The proof is analogous to that of Theorem 4.34. □

Remark 4.44. For the determination of a reduced differential-algebraic equation of the form (4.49), it is sufficient to consider F_μ in order to compute the desired regularizing state or output feedback and the solution of the closed loop system.

Remark 4.45. Suppose that for a given control problem (4.58) the variable x can be split into (x_1, x_2) in such a way that the reduced problem (4.59) can be transformed to

$$\dot{x}_1 = \mathcal{L}(t, x_1, u), \quad x_2 = \mathcal{R}(t, x_1, u)$$

according to (4.52). Then for every u with $u(t_0)$ sufficiently close to u_0 the closed loop reduced problem obviously is regular and strangeness-free. Due to the structure

of the problem (cp. Remark 4.36), we do not need to require that $\dot{u}$ is part of the parameters in order to get the results of Corollaries 4.40 and 4.42. Accordingly, we do not need to require that $\dot{u}(t_0) = \dot{u}_0$.

Example 4.46. A control problem for a multibody system has the form

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

since the control typically acts via external forces. Assuming that $g_p(p)$ has full row rank and that $M(p)$ is symmetric and positive definite, a possible reduced problem has the form

$$\begin{aligned} &\dot{p}_1 - v_1 = 0, \\ &\dot{v}_1 = [\, I \;\; 0 \,] M(p)^{-1} (f(p, v, u) - g_p(p)^T \lambda), \\ &g(p) = 0, \\ &g_p(p) v = 0, \\ &g_{pp}(v, v) + g_p(p) M(p)^{-1} (f(p, v, u) + g_p(p)^T \lambda) = 0, \end{aligned}$$

cp. (4.44). Moreover, this system is regular and strangeness-free for given u near the initial value, cp. Exercise 4.46. Comparing with (4.52), we have the splitting of variables

$$x_1 = (p_1, v_1), \quad x_2 = u, \quad x_3 = (p_2, v_2, \lambda).$$

The special structure of the reduced problem implies that from $\dot{x}_3$ only $\dot{\lambda}$ may depend on $\dot{u}$. Thus, Remark 4.36 and Remark 4.45 apply.

4.5 Differential equations on manifolds

The aim of this section is to show that regular, strangeness-free differential-algebraic equations are closely related to differential equations on manifolds. Although we have mainly worked in some (metric) space $\mathbb{R}^n$, in this section we consider the case of a general topological space $\mathbb{M}$. As it is common in the context of manifolds, we assume that $\mathbb{M}$ is a Hausdorff space (i.e., we can separate two different points by disjoint open neighborhoods) and that $\mathbb{M}$ possesses a countable (topological) basis. We follow here the presentation in [66], [68], see also [6], [174], [175]. For brevity we omit proofs concerning the basic theory of manifolds.

Before we can define a manifold properly, we need some preparations.

Definition 4.47. A homeomorphism (i.e., a continuous and open bijective map) $\varphi : \mathbb{U} \to \mathbb{V}$ with open sets $\mathbb{U} \subseteq \mathbb{M}$ and $\mathbb{V} \subseteq \mathbb{R}^d$ for some $d \in \mathbb{N}_0$ is called a *chart* of $\mathbb{M}$.

Definition 4.48. Two charts $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i$, $i = 1, 2$, are called *consistent* if either $\mathbb{U}_1 \cap \mathbb{U}_2 = \emptyset$ or

$$\varphi_2 \circ \varphi_1^{-1} : \varphi_1(\mathbb{U}_1 \cap \mathbb{U}_2) \to \varphi_2(\mathbb{U}_2 \cap \mathbb{U}_1)$$

and

$$\varphi_1 \circ \varphi_2^{-1} : \varphi_2(\mathbb{U}_2 \cap \mathbb{U}_1) \to \varphi_1(\mathbb{U}_1 \cap \mathbb{U}_2)$$

are homeomorphisms.

Definition 4.49. A collection $\{\varphi_i\}_{i\in\mathbb{J}}$, $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i$, of charts of $\mathbb{M}$ is called an *atlas* of $\mathbb{M}$ if every two charts are consistent and

$$\mathbb{M} = \bigcup_{i\in\mathbb{J}} \mathbb{U}_i.$$

Definition 4.50. Two atlases of $\mathbb{M}$ are called *equivalent* if their union is again an atlas of $\mathbb{M}$.

Definition 4.51. Let $\mathbb{M}$ be a Hausdorff space that possesses a countable (topological) basis. Then, $\mathbb{M}$ together with an equivalence class of atlases is called a *manifold*.

Note that the quantity d, i.e., the vector space dimension of the image space of a chart, may be different for different charts. Given a chart $\varphi : \mathbb{U} \to \mathbb{V}$, the use of an equivalence class of atlases allows us to add charts of the form $\varphi|_{\tilde{\mathbb{U}}} : \tilde{\mathbb{U}} \to \varphi(\tilde{\mathbb{U}})$ with $\tilde{\mathbb{U}} \subseteq \mathbb{U}$ open without violating the consistency of the charts and thus without changing the manifold. We can therefore assume that there exist charts defined on suitably small neighborhoods of a given point in $\mathbb{M}$.

The value of d belonging to a given chart can be seen as the local dimension of the manifold. Hence, we define the dimension of a manifold as follows.

Definition 4.52. Let $\{\varphi_i\}_{i\in\mathbb{J}}$, $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i$, be an atlas of the manifold $\mathbb{M}$. If there exists a $d \in \mathbb{N}_0$ such that $\mathbb{V}_i \subseteq \mathbb{R}^d$ for all $i \in \mathbb{J}$, then $\mathbb{M}$ is called a *manifold of dimension d*.

To characterize which manifolds actually possess a dimension, we have the following results.

Lemma 4.53. *Every manifold is a (disjoint) union of pathwise connected manifolds.*

Theorem 4.54. *Let $\mathbb{M}$ be a pathwise connected manifold. Then, $\mathbb{M}$ can be assigned a dimension in the sense of Definition 4.52.*

Thus, if we restrict ourselves to pathwise connected manifolds $\mathbb{M}$, then we can always speak of the dimension of the manifold. It is common to denote it by $\dim \mathbb{M}$.

Example 4.55. Let $\mathbb{M} \subseteq \mathbb{R}^n$ be open. Then, $\mathbb{M}$ becomes a manifold by the trivial atlas $\{\varphi\}$ with

$$\varphi \colon \mathbb{M} \to \mathbb{M}, \quad \varphi(x) = x.$$

Moreover, we have $\dim \mathbb{M} = n$.

Example 4.56. Let $\mathbb{M} = \{(x_1, x_2, x_3) \in \mathbb{R}^3 \mid x_1^2 + x_2^2 + x_3^2 = 0\}$ and set

$$\varphi_1 \colon \mathbb{U}_1 \to \mathbb{R}^2, \quad \mathbb{U}_1 = \mathbb{M} \setminus \{(0,0,1)\}, \qquad \varphi_1(x_1, x_2, x_3) = \frac{1}{1 - x_3}(x_1, x_2),$$

$$\varphi_2 \colon \mathbb{U}_2 \to \mathbb{R}^2, \quad \mathbb{U}_2 = \mathbb{M} \setminus \{(0,0,-1)\}, \quad \varphi_2(x_1, x_2, x_3) = \frac{1}{1 + x_3}(x_1, x_2),$$

recalling that we identify column vectors and tuples. The maps φ_1 and φ_2 are the so-called stereographic projections of the sphere $\mathbb{M}$ from the north and south pole onto the (x_1, x_2)-plane. The set $\{\varphi_1, \varphi_2\}$ forms an atlas of $\mathbb{M}$. Obviously, we have $\dim \mathbb{M} = 2$.

We want to mention already at this point that we use the symbol d to denote the dimension of a manifold, since it will turn out that in the interpretation of a regular strangeness-free differential-algebraic equation as a differential equation on a manifold, the dimension of this manifold will coincide with the size of the differential part of the differential-algebraic equation.

Up to now, we have only considered topological aspects of manifolds. We also speak of topological manifolds. For an analytical point of view, as for example differentiability of functions between manifolds, we need smooth atlases.

Definition 4.57. Let $\{\varphi_i\}_{i \in \mathbb{J}}$, $\varphi_i \colon \mathbb{U}_i \to \mathbb{V}_i$, be an atlas of the manifold $\mathbb{M}$. Suppose that there exists a $k \in \mathbb{N}_0 \cup \{\infty\}$ such that

$$\varphi_j \circ \varphi_i^{-1} \in C^k(\varphi_i(\mathbb{U}_i \cap \mathbb{U}_j), \varphi_j(\mathbb{U}_j \cap \mathbb{U}_i))$$

for all $i, j \in \mathbb{J}$ with $\mathbb{U}_i \cap \mathbb{U}_j \neq \emptyset$. Then, $\mathbb{M}$ is called a *manifold of class* C^k.

Definition 4.58. Let $\mathbb{M}$ and $\mathbb{L}$ be manifolds of class C^k with atlases $\{\varphi_i\}_{i \in \mathbb{J}}$ and $\{\psi_j\}_{j \in \mathbb{K}}$ according to Definition 4.57. A map $f \colon \mathbb{M} \to \mathbb{L}$ is called *ℓ-times continuously differentiable with $\ell \leq k$*, denoted by $f \in C^\ell(\mathbb{M}, \mathbb{L})$, if $\psi_j \circ f \circ \varphi_i^{-1}$ is ℓ-times continuously differentiable for all $i \in \mathbb{J}$, $j \in \mathbb{K}$, for which the composition is defined.

Given a manifold $\mathbb{M} \subseteq \mathbb{R}^n$, we have an intuitive imagination when $\mathbb{M}$ may be called a submanifold of $\mathbb{R}^n$. Note that if we only know that we have a manifold $\mathbb{M} \subseteq \mathbb{R}^n$, then this does not imply that $\mathbb{M}$ is a topological subspace of $\mathbb{R}^n$, since $\mathbb{M}$ may have a topology that is not the induced topology from $\mathbb{R}^n$. In the case that $\mathbb{M}$ is of class C^k, the property of being a submanifold $\mathbb{R}^n$ of class C^k should at least guarantee that functions defined on $\mathbb{R}^n$ of class C^k remain of class C^k if they are restricted to $\mathbb{M}$.

Definition 4.59. Let $\mathbb{M}$ and $\mathbb{X}$ be manifolds of dimensions d and n, respectively, and let $d \leq n$. A *topological embedding* of $\mathbb{M}$ in $\mathbb{X}$ is a continuous map $i_{\mathbb{M}} \colon \mathbb{M} \to \mathbb{X}$ such that $\mathbb{M}$ and $i_{\mathbb{M}}(\mathbb{M})$ are homeomorphic.

Definition 4.60. Let $\mathbb{X}$ be a manifold and let $\tilde{\mathbb{M}} \subseteq \mathbb{X}$ be a topological subspace of $\mathbb{X}$. We call $\tilde{\mathbb{M}}$ a *topological submanifold* of $\mathbb{X}$ if there exists a manifold $\mathbb{M}$ and an embedding $i_{\mathbb{M}} \colon \mathbb{M} \to \mathbb{X}$ such that $\tilde{\mathbb{M}} = i_{\mathbb{M}}(\mathbb{M})$.

Remark 4.61. Let $\mathbb{X} = \mathbb{R}^n$ be as in Definition 4.60 and let $\mathbb{M} \subseteq \mathbb{R}^n$ be a manifold. Then, $\mathbb{M}$ is a submanifold of $\mathbb{R}^n$ if

$$i_{\mathbb{M}} \colon \mathbb{M} \to \mathbb{R}^n, \quad i_{\mathbb{M}}(x) = x$$

is an embedding with $\tilde{\mathbb{M}} = i_{\mathbb{M}}(\mathbb{M})$, where $\tilde{\mathbb{M}} = \mathbb{M}$ as sets, but where $\tilde{\mathbb{M}}$ is considered as topological subspace of $\mathbb{R}^n$, i.e., with the topology induced from $\mathbb{R}^n$.

Since the manifolds $\mathbb{M}$ and $\tilde{\mathbb{M}}$ as in Remark 4.61 cannot be distinguished as topological spaces, it is common to see them as the same object. Moreover, one can show that every manifold is a submanifold of $\mathbb{R}^n$ for a sufficiently large n. Hence, we only need to consider manifolds that are submanifolds of some $\mathbb{R}^n$.

Example 4.62. In the following we consider $\mathbb{R}^d$ as a submanifold of $\mathbb{R}^n$, where $d \leq n$, by the so-called standard embedding given by

$$i_{\mathbb{R}^d} \colon \mathbb{R}^d \to \mathbb{R}^n, \quad i_{\mathbb{R}^d}(x_1, \ldots, x_d) = (x_1, \ldots, x_d, 0, \ldots, 0).$$

In particular, we identify $\mathbb{R}^d$ with $i_{\mathbb{R}^d}(\mathbb{R}^d)$.

Definition 4.63. A map $\varphi \colon \mathbb{U} \to \mathbb{V}$, $\mathbb{U}, \mathbb{V} \subseteq \mathbb{R}^n$ open, is called a *diffeomorphism* if it is bijective and both φ and φ^{-1} are of class C^1. It is called a *diffeomorphism of class* C^k, $k \in \mathbb{N} \cup \infty$, if in addition both φ and φ^{-1} are of class C^k.

Definition 4.64. A topological subspace $\mathbb{M} \subseteq \mathbb{R}^n$ is said to be a *submanifold of* $\mathbb{R}^n$ *of dimension* d *and of class* C^k, $k \in \mathbb{N} \cup \infty$, if for every $x \in \mathbb{M}$ there exists an open neighborhood $\tilde{\mathbb{U}} \subseteq \mathbb{R}^n$ of x and a diffeomorphism $\tilde{\varphi} \colon \tilde{\mathbb{U}} \to \tilde{\mathbb{V}}$ of class C^k with $\tilde{\mathbb{V}} \subseteq \mathbb{R}^n$ open and

$$\tilde{\varphi}(\mathbb{M} \cap \tilde{\mathbb{U}}) = \tilde{\mathbb{V}} \cap \mathbb{R}^d. \tag{4.69}$$

Of course, a submanifold $\mathbb{M} \subseteq \mathbb{R}^n$ of dimension d and of class C^k is itself a manifold of dimension d and of class C^k. First of all, a topological subspace of $\mathbb{R}^n$ is a Hausdorff space with a countable topological basis. Moreover, we can construct an atlas in the following way. For every $x \in \mathbb{M}$, we define a chart $\varphi \colon \mathbb{U} \to \mathbb{V}$ by taking a diffeomorphism $\tilde{\varphi}$ according to Definition 4.64 and setting

$$\varphi = \tilde{\varphi}|_{\mathbb{U}}, \quad \mathbb{U} = \mathbb{M} \cap \tilde{\mathbb{U}}, \quad \mathbb{V} = \tilde{\mathbb{V}} \cap \mathbb{R}^d, \tag{4.70}$$

the latter considered as an open subset of $\mathbb{R}^d$. For two charts $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i, i = 1, 2$, we then have that

$$\varphi_2 \circ \varphi_1^{-1} = \tilde{\varphi}_2 \circ \tilde{\varphi}_1^{-1} \circ i_{\mathbb{R}^d}$$

is a diffeomorphism of class C^k as a map from $\varphi_1(\mathbb{U}_1 \cap \mathbb{U}_2)$ to $\varphi_2(\mathbb{U}_2 \cap \mathbb{U}_1)$.

In our context, the most important kind of manifolds are those that are defined as the set of zeros of (smooth) nonlinear equations $H(x) = 0$, where the Jacobian $H_x(x)$ is assumed to have full row rank for all x at least on the set of zeros.

Theorem 4.65. *Let $H \in C^k(\mathbb{D}, \mathbb{R}^a)$, $\mathbb{D} \subseteq \mathbb{R}^n$ open, $k \in \mathbb{N} \cup \infty$, with $\mathbb{M} = H^{-1}(\{0\}) \neq \emptyset$ and suppose that* rank $H_x(x) = a \leq n$ *for all $x \in \mathbb{M}$. Then, $\mathbb{M}$ is a submanifold of $\mathbb{R}^n$ of dimension $d = n - a$ and of class C^k.*

Proof. Let $x_0 \in \mathbb{M}$. Since rank $H_x(x_0) = a$, we can split x according to $x = (x_1, x_2)$ such that $H_{x_2}(x_{1,0}, x_{2,0})$ is nonsingular. Applying the implicit function theorem, there exists an open neighborhood $\mathbb{V}$ of $x_{1,0} \in \mathbb{R}^d$ and a function $G \in C^k(\mathbb{V}, \mathbb{R}^a)$ with $G(x_{1,0}) = x_{2,0}$ and

$$H(x_1, G(x_1)) = 0 \quad \text{for all } x_1 \in \mathbb{V}.$$

Now we define $\tilde{\varphi}$ by $\tilde{\varphi}(x_1, x_2) = (x_1, H(x_1, x_2))$. Since

$$\tilde{\varphi}'(x_{1,0}, x_{2,0}) = \begin{bmatrix} I_d & 0 \\ H_{x_1}(x_{1,0}, x_{2,0}) & H_{x_2}(x_{1,0}, x_{2,0}) \end{bmatrix},$$

the inverse function theorem yields that there exist neighborhoods $\tilde{\mathbb{U}} \subseteq \mathbb{R}^n$ of $(x_{1,0}, x_{2,0})$ and $\tilde{\mathbb{V}} \subseteq \mathbb{R}^n$ of $(x_{1,0}, 0)$ such that $\tilde{\varphi} : \tilde{\mathbb{U}} \to \tilde{\mathbb{V}}$ is a diffeomorphism of class C^k. Moreover, by construction $\tilde{\varphi}(x_1, x_2) = (x_1, 0)$ for $(x_1, x_2) \in \mathbb{M} \cap \tilde{\mathbb{U}}$ which implies (4.70). Thus, $\mathbb{M}$ is a submanifold of $\mathbb{R}^n$ of dimension $d = n - a$ and of class C^k by Theorem 4.65. In particular, the corresponding chart φ of $\mathbb{M}$ from (4.69) satisfies

$$\varphi(x_1, x_2) = x_1, \quad \varphi^{-1}(x_1) = (x_1, G(x_1)),$$

with possibly smaller neighborhoods $\mathbb{U}$ and $\mathbb{V}$ as from the implicit function theorem. □

In order to define differential equations on manifolds, we first need the notions of tangent spaces and vector fields. For this, let $\mathbb{M}$ in the following be a manifold of class C^1. Note also that in the following we will use a prime to denote the total derivative of a function.

Given $x_0 \in \mathbb{M}$ and a chart $\varphi : \mathbb{U} \to \mathbb{V} \subseteq \mathbb{R}^d$ of $\mathbb{M}$ with $x_0 \in \mathbb{U}$, we consider functions

$$\gamma \in C^1((-\varepsilon, \varepsilon), \mathbb{U}), \quad \gamma(0) = x_0,$$

and

$$\vartheta \in C^1(\mathbb{U}, \mathbb{R}), \quad \vartheta(x_0) = 0.$$

Since $\psi : (-\varepsilon, \varepsilon) \to (-\varepsilon, \varepsilon)$ with $\psi(t) = t$ is a chart for both the manifolds $(-\varepsilon, \varepsilon)$ and $\mathbb{R}$, we have by definition

$$\varphi \circ \gamma = \varphi \circ \gamma \circ \psi^{-1} \in C^1((-\varepsilon, \varepsilon), \mathbb{V})$$

and (at least for sufficiently small $\mathbb{V}$)

$$\vartheta \circ \varphi^{-1} = \psi \circ \vartheta \circ \varphi^{-1} \in C^1(\mathbb{V}, \mathbb{R}).$$

Hence,

$$\vartheta \circ \gamma = \vartheta \circ \varphi^{-1} \circ \varphi \circ \gamma \in C^1((-\varepsilon, \varepsilon), \mathbb{R})$$

and $(\vartheta \circ \gamma)'(0) \in \mathbb{R}$ is defined. Setting

$$\Gamma_{x_0} = \{\gamma \in C^1((-\varepsilon, \varepsilon), \mathbb{U}) \mid \gamma(0) = x_0,\ \varepsilon \text{ sufficiently small}\}$$

and

$$\Theta_{x_0} = \{\vartheta \in C^1(\mathbb{U}, \mathbb{R}) \mid \vartheta(x_0) = 0\},$$

we can therefore define the following.

Definition 4.66. Let $\gamma_1, \gamma_2 \in \Gamma_{x_0}$. We call γ_1 and γ_2 to be equivalent and write $\gamma_1 \sim \gamma_2$ if

$$(\vartheta \circ \gamma_1)'(0) = (\vartheta \circ \gamma_2)'(0) \quad \text{for all } \vartheta \in \Theta_{x_0}. \tag{4.71}$$

The corresponding equivalence class of a given $\gamma \in \Gamma_{x_0}$, denoted by $[\gamma]_{x_0}$, is called a *tangent vector to* $\mathbb{M}$ *at* x_0. The set

$$T_{x_0}(\mathbb{M}) = \{[\gamma]_{x_0} \mid \gamma \in \Gamma_{x_0}\} \tag{4.72}$$

is called the *tangent space of* $\mathbb{M}$ *at* x_0.

Let $\gamma_1, \gamma_2 \in \Gamma_{x_0}$ and $\gamma_1 \sim \gamma_2$. The defining relation (4.71) implies that

$$(\vartheta \circ \varphi^{-1} \circ \varphi \circ \gamma_1)'(0) = (\vartheta \circ \varphi^{-1} \circ \varphi \circ \gamma_2)'(0)$$

or

$$(\vartheta \circ \varphi^{-1})'(\varphi(x_0))(\varphi \circ \gamma_1)'(0) = (\vartheta \circ \varphi^{-1})'(\varphi(x_0))(\varphi \circ \gamma_2)'(0)$$

for all $\vartheta \in \Theta_{x_0}$. Choosing $\vartheta \in \Theta_{x_0}$ by

$$\vartheta(x) = e_k^T(\varphi(x) - \varphi(x_0)),$$

where e_k is the k-th canonical basis vector of $\mathbb{R}^d$, yields

$$(\vartheta \circ \varphi^{-1})'(\varphi(x_0)) = e_k^T.$$

Hence, $\gamma_1 \sim \gamma_2$ implies

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0)$$

and the function

$$\Phi_{x_0} \colon T_{x_0}(\mathbb{M}) \to \mathbb{R}^d, \quad \Phi_{x_0}([\gamma]_{x_0}) = (\varphi \circ \gamma)'(0) \tag{4.73}$$

is well defined. Choosing $\gamma \in \Gamma_{x_0}$ by

$$\gamma(t) = \varphi^{-1}(\varphi(x_0) + tz)$$

for a given $z \in \mathbb{R}^d$, we find that

$$(\phi \circ \gamma)(t) = \varphi(x_0) + tz,$$

hence $(\phi \circ \gamma)'(0) = z$, and it follows that Φ_{x_0} is surjective. Furthermore, if

$$(\varphi \circ \gamma_1)'(0) = (\varphi \circ \gamma_2)'(0),$$

then (4.71) holds, since

$$(\vartheta \circ \gamma)'(0) = (\vartheta \circ \varphi^{-1} \circ \varphi \circ \gamma)'(0) = (\vartheta \circ \varphi^{-1})'(\varphi(x_0))(\varphi \circ \gamma_1)'(0).$$

This implies that $\gamma_1 \sim \gamma_2$ or $[\gamma_1]_{x_0} = [\gamma_2]_{x_0}$, and thus Φ_{x_0} is injective.

Hence, we have shown that Φ_{x_0} is bijective and we can import the linear structure and topology from $\mathbb{R}^d$ into $T_{x_0}(\mathbb{M})$. In particular, $T_{x_0}(\mathbb{M})$ becomes a topological vector space homeomorphic to $\mathbb{R}^d$, justifying so the name tangent space.

As next step, we gather all tangent spaces that belong to a given manifold.

Definition 4.67. For a manifold $\mathbb{M}$, the (disjoint) union

$$T(\mathbb{M}) = \bigcup_{x \in \mathbb{M}} T_x(\mathbb{M}) \tag{4.74}$$

is called the *tangent bundle of* $\mathbb{M}$.

Let $\varphi \colon \mathbb{U} \to \mathbb{V}$ be a chart of $\mathbb{M}$. Setting

$$\mathbb{W} = \bigcup_{x \in \mathbb{U}} T_x(\mathbb{M}) \subseteq T(\mathbb{M})$$

and defining

$$\psi \colon \mathbb{W} \to \mathbb{V} \times \mathbb{R}^d, \quad \psi([\gamma]_x) = (\varphi(x), \Phi_x([\gamma]_x)), \tag{4.75}$$

we at once see that ψ is bijective. Taking the induced topology in $\mathbb{W}$, the map ψ becomes a homeomorphism. Let now $\{\varphi_i\}_{i \in \mathbb{J}}$, $\varphi \colon \mathbb{U}_i \to \mathbb{V}_i$, be an atlas of $\mathbb{M}$ and

consider $\psi_i : \mathbb{W}_i \to \mathbb{V}_i \times \mathbb{R}^d$ constructed as in (4.75). Defining $\mathbb{W} \subseteq T(\mathbb{M})$ to be open if and only if $\mathbb{W} \cap \mathbb{W}_i$ is open in $\mathbb{W}_i$ for all $i \in \mathbb{J}$ yields a topology in $T(\mathbb{M})$ such that it becomes a manifold. Moreover, if $\mathbb{M}$ has dimension d, then $T(\mathbb{M})$ has dimension $2d$. Observe that $\psi([\gamma]_x) = (u, z)$ with $(u, z) \in \mathbb{V} \times \mathbb{R}^d$ implies that $\varphi(x) = u$, $\Phi_x([\gamma]_x) = z$, and thus $x = \varphi^{-1}(u)$, $(\varphi \circ \gamma)'(0) = z$, $\gamma(0) = \varphi^{-1}(u)$. Therefore, the inverse of ψ_i is given by

$$\psi_i^{-1}(u, z) = [\gamma]_x, \quad x = \varphi_i^{-1}(u), \quad (\varphi_i \circ \gamma)'(0) = z, \quad \gamma(0) = \varphi_i^{-1}(u)$$

such that

$$\begin{aligned}
(\psi_j \circ \psi_i^{-1})(u, z) &= \psi_j([\gamma]_x) \\
&= (\varphi_j(x), \Phi_x([\gamma]_x)) \\
&= ((\varphi_j \circ \varphi_i^{-1})(u), (\varphi_j \circ \gamma)'(0)) \\
&= ((\varphi_j \circ \varphi_i^{-1})(u), (\varphi_j \circ \varphi_i^{-1} \circ \varphi_i \circ \gamma)'(0)) \\
&= ((\varphi_j \circ \varphi_i^{-1})(u), (\varphi_j \circ \varphi_i^{-1})'((\varphi_i \circ \gamma)(0))(\varphi_i \circ \gamma)'(0)) \\
&= ((\varphi_j \circ \varphi_i^{-1})(u), (\varphi_j \circ \varphi_i^{-1})'(u)z).
\end{aligned}$$

Hence, $\psi_j \circ \psi_i^{-1}$ is of class C^{k-1} if $\varphi_j \circ \varphi_i^{-1}$is of class C^k, implying that $T(\mathbb{M})$ is of class C^{k-1} if $\mathbb{M}$ is of class C^k.

Remark 4.68. Let $\mathbb{M} = H^{-1}(\{0\}) \subseteq \mathbb{R}^n$ according to Theorem 4.65, let $\varphi : \mathbb{U} \to \mathbb{V}$ be a chart of $\mathbb{M}$ and let $\gamma \in C^1((-\varepsilon, \varepsilon), \mathbb{U})$, $\gamma(0) = x_0$. With the notation used in the context of submanifolds, we define $\tilde{\gamma} \in C^1((-\varepsilon, \varepsilon), \tilde{\mathbb{U}})$ by

$$\tilde{\gamma} = \tilde{\varphi}^{-1} \circ i_{\mathbb{R}^d} \circ \varphi \circ \gamma.$$

Because of

$$(\tilde{\varphi}^{-1} \circ i_{\mathbb{R}^d} \circ \varphi)(x) = x \quad \text{for all } x \in \mathbb{U},$$

we have that

$$H(\tilde{\gamma}(t)) = H(\gamma(t)) = 0 \quad \text{for all } t \in (-\varepsilon, \varepsilon),$$

where $H \circ \tilde{\gamma}$ can now be differentiated with the help of the chain rule to obtain

$$H_x(x_0)\tilde{\gamma}'(0) = 0.$$

Observing that

$$(\tilde{\varphi}^{-1})'(i_{\mathbb{R}^d}(\varphi(x_0))) = (\tilde{\varphi}^{-1})'(\tilde{\varphi}(x_0)) = (\tilde{\varphi}(x_0))^{-1},$$

we get

$$\tilde{\gamma}'(0) = (\tilde{\varphi}(x_0))^{-1} \begin{bmatrix} I_d \\ 0 \end{bmatrix} (\varphi \circ \gamma)'(0).$$

Since we can choose γ such that $(\varphi \circ \gamma)'(0) \in \mathbb{R}^d$ is arbitrary, the possible values for $\tilde{\gamma}'(0)$ vary in a d-dimensional subspace of $\mathbb{R}^n$, which is nothing else than kernel $H_x(x_0)$. Hence, we can identify $T_{x_0}(\mathbb{M})$ with kernel $H_x(x_0)$ by the above relation. In the same way, we can identify $T(\mathbb{M})$ with the set $\{(x, \text{kernel } H_x(x)) \mid x \in \mathbb{M}\}$.

Definition 4.69. A *vector field* v on $\mathbb{M}$ is a continuous map $v\colon \mathbb{M} \to T(\mathbb{M})$ such that

$$v(x) \in T_x(\mathbb{M}) \quad \text{for all } x \in \mathbb{M}. \tag{4.76}$$

For given $x \in C^1((-\varepsilon, \varepsilon), \mathbb{M})$, we use the notation

$$\tfrac{d}{dt}x(t) = [\gamma]_{x(t)} \in T_{x(t)}(\mathbb{M}),$$

where $\gamma \in \Gamma_{x(t)}$ is defined by $\gamma(s) = x(s + t)$ for fixed $t \in (-\varepsilon, \varepsilon)$. With this, we now consider the problem of finding a function $x \in C^1((-\varepsilon, \varepsilon), \mathbb{M})$ that (pointwise) satisfies

$$\tfrac{d}{dt}x(t) = v(x(t)), \quad x(0) = x_0 \tag{4.77}$$

for given initial value $x_0 \in \mathbb{M}$ and given vector field v on $\mathbb{M}$. We call this an initial value problem for an ordinary differential equation on $\mathbb{M}$.

Let $\varphi\colon \mathbb{U} \to \mathbb{V}$ be a chart of $\mathbb{M}$ with $x_0 \in \mathbb{U}$ and $\psi\colon \mathbb{W} \to \mathbb{V} \times \mathbb{R}^d$ the corresponding chart of $T(\mathbb{M})$. Using the projections

$$\pi_1, \pi_2\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d, \quad \pi_1(u, z) = u, \quad \pi_2(u, z) = z,$$

we consider the initial value problem

$$y' = f(y), \quad y(0) = \varphi(x_0) \in \mathbb{V}, \tag{4.78}$$

with

$$f = \pi_2 \circ \psi \circ v \circ \varphi^{-1}\colon \mathbb{V} \to \mathbb{R}^d. \tag{4.79}$$

If we require f to be differentiable with Lipschitz continuous derivative (assuming $\mathbb{M}$ to be of class C^2 such that $T(\mathbb{M})$ is of class C^1, and v to be of class C^1 would guarantee this for a sufficiently small $\mathbb{V}$), then we get a unique local solution $y \in C^1((-\varepsilon, \varepsilon), \mathbb{V})$ with $\varepsilon > 0$ sufficiently small.

Setting

$$x = \varphi^{-1} \circ y\colon (-\varepsilon, \varepsilon) \to \mathbb{U}$$

and taking $\gamma \in \Gamma_{x(t)}$ with $\gamma(s) = x(s + t)$ for fixed $t \in (-\varepsilon, \varepsilon)$, we find that

$$(\varphi \circ \gamma)(s) = \varphi(\gamma(s)) = \varphi(x(s + t)) = (\varphi \circ x)(s + t) = y(s + t)$$

and thus

$$\begin{aligned} \tfrac{d}{dt}x(t) &= [\gamma]_{x(t)} = \psi^{-1}(\psi([\gamma]_{x(t)})) \\ &= \psi^{-1}(\varphi(x(t)), (\varphi \circ \gamma)'(0)) = \psi^{-1}(y(t), y'(t)), \end{aligned}$$

while

$$\begin{aligned} v(x(t)) &= \psi^{-1}((\psi \circ v \circ \varphi^{-1} \circ y)(t)) \\ &= \psi^{-1}((\pi_1 \circ \psi \circ v \circ \varphi^{-1} \circ y)(t), (\pi_2 \circ \psi \circ v \circ \varphi^{-1} \circ y)(t)). \end{aligned}$$

Because of

$$\begin{aligned} (\pi_1 \circ \psi \circ v \circ \varphi^{-1} \circ y)(t) &= (\pi_1 \circ \psi \circ v \circ x)(t) = (\pi_1 \circ \psi)(v(x(t))) \\ &= (\pi_1 \circ \psi)([\gamma]_{x(t)}) = \pi_1(\psi([\gamma]_{x(t)})) = \varphi(x(t)) = y(t) \end{aligned}$$

and

$$(\pi_2 \circ \psi \circ v \circ \varphi^{-1} \circ y)(t) = f(y(t)) = y'(t),$$

we also have

$$v(x(t)) = \psi^{-1}(y(t), y'(t)).$$

Hence, the so constructed $x \in C^1((-\varepsilon, \varepsilon), \mathbb{U})$ locally solves (4.77).

Theorem 4.70. *Under the stated smoothness assumptions, the initial value problem* (4.78) *locally possesses a unique solution* $x \in C^1((-\varepsilon, \varepsilon), \mathbb{U})$.

Proof. It remains to show that the local solution x constructed in the above way does not depend on the selected chart. For this, let $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i$, $i = 1, 2$, be two charts of $\mathbb{M}$ and $x_0 \in \mathbb{U}_1 \cap \mathbb{U}_2 \neq \emptyset$. Let $y_i \in C^1((-\varepsilon, \varepsilon), \mathbb{V}_1 \cap \mathbb{V}_2)$, $i = 1, 2$, be (local) solutions of

$$y_i' = f_i(y_i), \quad y_i(0) = \varphi_i(x_0),$$

with

$$f_i = \pi_2 \circ \psi_i \circ v \circ \varphi_i^{-1}.$$

We then must show that

$$\varphi_1^{-1} \circ y_1 = \varphi_2^{-1} \circ y_2.$$

Because of the unique solvability of initial value problems in $\mathbb{R}^d$, it is sufficient to show that $y_2 = \varphi_2 \circ \varphi_1^{-1} \circ y_1$ solves

$$y_2' = f_2(y_2), \quad y_2(0) = \varphi_2(x_0).$$

We first observe that

$$y_2(0) = (\varphi_2 \circ \varphi_1^{-1})(y_1(0)) = (\varphi_2 \circ \varphi_1^{-1})(\varphi_1(x_0)) = \varphi_2(x_0).$$

Because of

$$\begin{aligned} (\pi_2 \circ \psi_2)([\gamma]_x) &= (\varphi_2 \circ \gamma)'(0) = (\varphi_2 \circ \varphi_1^{-1} \circ \varphi_1 \circ \gamma)'(0) \\ &= (\varphi_2 \circ \varphi_1^{-1})'((\varphi_1 \circ \gamma)(0))(\varphi_1 \circ \gamma)'(0) \\ &= (\varphi_2 \circ \varphi_1^{-1})'(\varphi_1(x))(\varphi_1 \circ \gamma)'(0) \\ &= (\varphi_2 \circ \varphi_1^{-1})'(\varphi_1(x))(\pi_2 \circ \psi_1)([\gamma]_x), \end{aligned}$$

we find that

$$\begin{aligned} y_2'(t) &= (\varphi_2 \circ \varphi_1^{-1} \circ y_1)'(t) = (\varphi_2 \circ \varphi_1^{-1})(y_1(t)) y_1'(t) \\ &= (\varphi_2 \circ \varphi_1^{-1})(y_1(t)) f_1(y_1(t)) \\ &= (\varphi_2 \circ \varphi_1^{-1})(y_1(t))(\pi_2 \circ \psi_1 \circ v \circ \varphi_1^{-1})(y_1(t)) \\ &= (\pi_2 \circ \psi_2 \circ v \circ \varphi_1^{-1})(y_1(t)) = (\pi_2 \circ \psi_2 \circ v \circ \varphi_2^{-1})(y_2(t)) = f_2(y_2(t)), \end{aligned}$$

where we used that

$$(v \circ \varphi_1^{-1})(y_1(t)) \in T_{\varphi_1^{-1}(y_1(t))}(\mathbb{M}). \qquad \square$$

As usual, one can continue the so obtained local solution until the boundary of $\mathbb{M}$ is reached. Note that this does not imply that there exists a solution in $C^1([0, T], \mathbb{M})$ for a given $T > 0$. As for ordinary differential equations, the existence of a unique solution to the initial value problems allows for the definition of *flows*, i.e., of functions that map a given initial value on the final value of the corresponding solution after a given time interval. Starting with a solution $x^* \in C^1([0, T], \mathbb{M})$ of $\frac{d}{dt}x(t) = v(x(t))$, we can proceed as follows. There exists a grid

$$0 = t_0 < t_1 < \cdots < t_N = T$$

and charts $\varphi_i : \mathbb{U}_i \to \mathbb{V}_i$, $i = 0, \ldots, N-1$, of $\mathbb{M}$ such that $x^*(t) \in \mathbb{U}_i$ for all $i \in [t_i, t_{i+1}]$. Thus, for every chart φ_i, we have a solution $y_i \in C^1([t_i, t_{i+1}], \mathbb{V}_i)$ of

$$y' = f_i(y), \quad y(t_i) = y_i^*, \quad y_i^* = \varphi_i(x^*(t_i)) \in \mathbb{V}_i,$$

where $f_i = \pi_2 \circ \psi_i \circ v \circ \varphi_i^{-1}$ and where ψ_i is the chart of $T(\mathbb{M})$ belonging to φ_i. Let $\tilde{\mathbb{U}}_N \subseteq \mathbb{U}_{N-1}$ be a neighborhood of $x^*(t_N)$. Then there exists a sufficiently small neighborhood $\tilde{\mathbb{U}}_{N-1} \subseteq \mathbb{U}_{N-2} \cap \mathbb{U}_{N-1}$ of $x^*(t_{N-1})$ such that

$$y' = f_{N-1}(y), \quad y(t_{N-1}) = \varphi_{N-1}(x)$$

has a (unique) solution $y \in C^1([t_{N-1}, t_N], \mathbb{V}_{N-1})$ with $y(t_N) \in \varphi(\tilde{\mathbb{U}}_N)$ for all $x \in \tilde{\mathbb{U}}_{N-1}$. This defines a map

$$\phi_{N-1} : \tilde{\mathbb{U}}_{N-1} \to \tilde{\mathbb{U}}_N, \quad x \mapsto \varphi_{N-1}^{-1}(y(t_N)).$$

In this way, we inductively get neighborhoods $\tilde{\mathbb{U}}_i$ of $x^*(t_i)$ and maps

$$\phi_i : \tilde{\mathbb{U}}_i \to \tilde{\mathbb{U}}_{i+1}, \quad x \mapsto \varphi_i^{-1}(y(t_{i+1}))$$

for $i = 0, \ldots, N-1$, where $y \in C^1([t_i, t_{i+1}], \mathbb{V}_i)$ is the solution of

$$y' = f_i(y), \quad y(t_i) = \varphi_i(x),$$

with $y(t_{i+1}) \in \varphi_i(\tilde{\mathbb{U}}_{i+1})$ due to the construction of the sets $\tilde{\mathbb{U}}_0, \ldots, \tilde{\mathbb{U}}_N$. The composition

$$\phi = \phi_{N-1} \circ \cdots \circ \phi_1 \circ \phi_0$$

then is a map

$$\phi : \tilde{\mathbb{U}}_0 \to \tilde{\mathbb{U}}_N, \quad x_0 \mapsto x(T)$$

that maps a given initial value $x_0 \in \tilde{\mathbb{V}}_0$ to the final value $x(T)$ of the solution $x \in C^1([0, T], \mathbb{M})$ of the initial value problem (4.77). In particular, $\phi(x^*(0)) = x^*(T)$. In this construction, the assumption on the existence of a solution x^* can be dropped if T is sufficiently small, since solutions always exist at least locally. In this way, we can define maps

$$\phi_t : \mathbb{U} \to \mathbb{M}, \quad x_0 \mapsto x(t)$$

for a given $t \in \mathbb{R}$ being sufficiently small in modulus and an appropriately chosen open set $\mathbb{U} \in \mathbb{M}$. Obviously, ϕ_t can be inverted on $\phi_t(\mathbb{U})$ just by solving the differential equation backwards from t to 0. Since the differential equation is autonomous, this is the same as solving it from 0 to $-t$. Hence, there exists

$$\phi_t^{-1} : \phi_t(\mathbb{U}) \to \mathbb{U},$$

with

$$\phi_t^{-1} = \phi_{-t}.$$

In particular, we have that $\phi_0 = \mathrm{id}$. Let now

$$\phi_{t_i} : \mathbb{U}_i \to \mathbb{M}, \quad i = 1, 2, 3,$$

with $t_3 = t_1 + t_2$ be given. Setting $\mathbb{U} = \phi_{t_1}^{-1}(\mathbb{U}_2) \cap \mathbb{U}_3$ and restricting ϕ_{t_1} and ϕ_{t_3} to $\mathbb{U}$, we get

$$\phi_{t_2} \circ \phi_{t_1} = \phi_{t_3} = \phi_{t_1+t_2},$$

due to the unique solvability of the initial value problems. The functions ϕ_t are called *flows*.

In the following, we want to show that a solvable regular differential-algebraic equation with strangeness index $\mu = 0$ can be locally (near a given solution) interpreted as a differential equation on a manifold and vice versa. Due to Section 4.1, we are allowed to assume without loss of generality that the given differential-algebraic equation is autonomous and in the reduced form (4.23). We therefore consider

$$\hat{F}_1(x, \dot{x}) = 0, \quad \hat{F}_2(x) = 0, \tag{4.80}$$

with sufficiently smooth functions $\hat{F}_1$ and $\hat{F}_2$ and use the notation of Section 4.1.

Starting with a solution $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (4.80), the set

$$\mathbb{M} = \hat{F}_2^{-1}(\{0\}) \tag{4.81}$$

is nonempty by assumption. Because of

$$\operatorname{rank} \hat{F}_{2;x}(x) = a \quad \text{for all } x \in \mathbb{M},$$

the set $\mathbb{M} \subseteq \mathbb{R}^n$ forms a manifold of dimension $d = n - a$. Note that $\mathbb{M}$ actually is a submanifold of $\mathbb{R}^n$. Suppose that $\mathbb{M}$ is of class C^2. Given a chart $\varphi : \mathbb{U} \to \mathbb{V}$, say with $x^*(t_0) \in \mathbb{U}$ for some $t_0 \in \mathbb{I}$, we may assume due to Theorem 4.65 that $x = (x_1, x_2)$, $\varphi(x) = x_1$, and $\varphi^{-1}(x_1) = (x_1, \mathcal{R}(x_1))$, in particular

$$x_2 = \mathcal{R}(x_1),$$

with $\mathcal{R} : \mathbb{V} \to \mathbb{R}^a$. Since $\hat{F}_2(x_1, \mathcal{R}(x_1)) = 0$ for all $x_1 \in \mathbb{V}$, we have that

$$\mathcal{R}_{x_1}(x_1) = -\hat{F}_{2;x_2}(x_1, \mathcal{R}(x_1))^{-1} \hat{F}_{2;x_1}(x_1, \mathcal{R}(x_1)).$$

By assumption,

$$\hat{F}_1(x_1, \mathcal{R}(x_1), \dot{x}_1, \mathcal{R}_{x_1}(x_1)\dot{x}_1) = 0$$

is solved by $(x_1^*(t_0), \dot{x}_1^*(t_0))$ and the Jacobian with respect to x_1 is nonsingular. Hence, we can solve locally for x_1 according to

$$\dot{x}_1 = \mathcal{L}(x_1),$$

where we may assume that $\mathcal{L} : \mathbb{V} \to \mathbb{R}^d$. Note that

$$\begin{bmatrix} I_d \\ \mathcal{R}_{x_1}(x_1) \end{bmatrix} \mathcal{L}(x_1) \in \operatorname{kernel} \hat{F}_{2;x}(x_1, \mathcal{R}(x_1)).$$

Following Remark 4.68, this vector of the kernel coincides with $\tilde{\gamma}'(0)$, where $\tilde{\gamma}$ belongs to some $\gamma \in C^1((-\varepsilon, \varepsilon), \mathbb{U})$. In particular, it follows from Remark 4.68 that $(\varphi \circ \gamma)'(0) = \mathcal{L}(x_1)$ holds. Hence,

$$v(x) = \psi^{-1}(\varphi(x), (\mathcal{L} \circ \varphi)(x))$$

defines a vector field v on $\mathbb{U}$. The definition of v does not depend on the selected chart, since $\hat{F}_1(x, \dot{x}) = 0$ locally has a unique solution $\dot{x} \in \operatorname{kernel} \hat{F}_{2;x}$ for given $x \in \mathbb{M}$. Because of $y = \varphi(x) = x_1$ and

$$f(y) = (\pi_2 \circ \psi \circ v \circ \varphi^{-1})(y) = \pi_2 \circ \psi \circ \psi^{-1}(y, \mathcal{L}(y)) = \mathcal{L}(y),$$

the differential equation in $\mathbb{V}$ reads $\dot{x}_1 = \mathcal{L}(x_1)$. Thus, in order to solve the constructed ordinary differential equation on $\mathbb{M}$, we must solve

$$\dot{x}_1 = \mathcal{L}(x_1), \quad x_2 = \mathcal{R}(x_1),$$

which is nothing else than the local version of the given differential-algebraic equation.

Conversely, let $\mathbb{M} \subseteq \mathbb{R}^n$ be a manifold of dimension d with a vector field v and consider an initial value problem

$$\tfrac{d}{dt}x(t) = v(x(t)), \quad x(0) = x_0. \tag{4.82}$$

We assume that $\mathbb{M}$ is of class C^k for sufficiently large k and that the inclusion $i_{\mathbb{M}} : \mathbb{M} \to \mathbb{R}^n$ defined by $i_{\mathbb{M}}(x) = x$ for all $x \in \mathbb{M}$ is also of class C^k. Let $\varphi : \mathbb{U} \to \mathbb{V}$ be a chart of $\mathbb{M}$ with $x_0 \in \mathbb{U}$. Since $\mathbb{R}^n$ is a manifold with a trivial chart, we have

$$i_{\mathbb{M}} \circ \varphi^{-1} \in C^k(\mathbb{V}, \mathbb{R}^n).$$

Defining $\tilde{F}_2 : \mathbb{R}^n \times \mathbb{V} \to \mathbb{R}^n$ by

$$\tilde{F}_2(x, y) = x - (i_{\mathbb{M}} \circ \varphi^{-1})(y),$$

we have $\tilde{F}_2(x, \varphi(x)) = 0$ for all $x \in \mathbb{U}$. Moreover, since $\tilde{F}_2 \in C^k(\mathbb{R}^n \times \mathbb{V}, \mathbb{R}^n)$, we are allowed to differentiate, and obtain

$$\tilde{F}_{2;x,y}(x, y) = [\, I_n \;\; -(i_{\mathbb{M}} \circ \varphi^{-1})'(y) \,].$$

Thus, $\tilde{F}_2^{-1}(\{0\}) \subseteq \mathbb{R}^{n+d}$ forms a manifold $\tilde{\mathbb{U}}$ of dimension d with the only chart $\tilde{\varphi} : \tilde{\mathbb{U}} \to \mathbb{V}$, $\tilde{\varphi}(x, y) = y$, $\tilde{\varphi}^{-1}(y) = ((i_{\mathbb{M}} \circ \varphi^{-1})(y), y)$. The manifolds $\mathbb{U}$ and $\tilde{\mathbb{U}}$ are homeomorphic via

$$\tilde{\varphi}^{-1} \circ \varphi : \mathbb{U} \to \tilde{\mathbb{U}}, \quad (\tilde{\varphi}^{-1} \circ \varphi)(x) = ((i_{\mathbb{M}} \circ \varphi^{-1})(\varphi(x)), \varphi(x)) = (x, \varphi(x)).$$

Taking the chart $\psi : \mathbb{W} \to \mathbb{U} \times \mathbb{R}^d$ of $T(\mathbb{M})$ that corresponds to φ, the above differential equation reads

$$y' = f(y), \quad y = \varphi(x), \quad f = \pi_2 \circ \psi \circ v \circ \varphi^{-1}$$

in local coordinates. Defining $\tilde{F}_1 : \mathbb{R}^n \times \mathbb{V} \times \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^d$ by

$$\tilde{F}_1(x, y, x', y') = y' - f(y),$$

the so obtained (strangeness-free) differential-algebraic equation

$$y' = f(y), \quad x = (i_{\mathbb{M}} \circ \varphi^{-1})(y)$$

together with $y(0) = y_0$, $y_0 = \varphi(x_0)$, is locally equivalent to the initial value problem (4.82) on $\mathbb{M}$. Note that the resulting differential-algebraic equation is formulated in an unknown function (x, y) with values in $\mathbb{R}^{n+d}$. We cannot expect that we can reduce it to a differential-algebraic equation only for x without further assumptions, since the reverse construction would yield a submanifold of $\mathbb{R}^n$. Assuming thus that $\mathbb{M}$ indeed is a submanifold of $\mathbb{R}^n$, we can proceed as follows. Due

to Definition 4.64, there exists a diffeomorphism $\tilde{\varphi} : \tilde{\mathbb{U}} \to \tilde{\mathbb{V}}$, with $x_0 \in \tilde{\mathbb{U}} \subseteq \mathbb{R}^n$ and $\tilde{\mathbb{V}} \subseteq \mathbb{R}^n$ such that (4.70) defines a chart of $\mathbb{M}$. Because of

$$i_{\mathbb{M}} \circ \varphi^{-1} = \tilde{\varphi}^{-1} \circ i_{\mathbb{R}^d},$$

we are now allowed to differentiate according to

$$(i_{\mathbb{M}} \circ \varphi^{-1})'(y) = (\tilde{\varphi}^{-1} \circ i_{\mathbb{R}^d})'(y) = (\tilde{\varphi}^{-1})'(i_{\mathbb{R}^d}(y)) \begin{bmatrix} I_d \\ 0 \end{bmatrix}.$$

Since $(\tilde{\varphi}^{-1})'(i_{\mathbb{R}^d}(y)) \in \mathbb{R}^{n,n}$ is nonsingular, it follows that $(i_{\mathbb{M}} \circ \varphi^{-1})'(y) \in \mathbb{R}^{n,d}$ has full column rank. In particular, this holds for $U = (i_{\mathbb{M}} \circ \varphi^{-1})'(y_0)$. Hence, there exists an orthogonal matrix $Z \in \mathbb{R}^{n,n}$, $Z = [\, Z_1 \;\; Z_2 \,]$ with $Z_2 \in \mathbb{R}^{n,a}$ and $Z_2^T U = 0$ such that

$$Z_1^T \tilde{F}_2(x, y) = Z_1^T (x - (i_{\mathbb{M}} \circ \varphi^{-1})(y)) = 0$$

can be solved locally for $y = \mathcal{S}(x)$. In this way, we get a locally equivalent differential-algebraic equation of the form

$$\hat{F}_1(x, x') = 0, \quad \hat{F}_2(x) = 0$$

with

$$\begin{aligned} \hat{F}_1(x, x') &= \mathcal{S}_x(x)x' - f(\mathcal{S}(x)), \\ \hat{F}_2(x) &= Z_2^T (x - (i_{\mathbb{M}} \circ \varphi^{-1})(\mathcal{S}(x))). \end{aligned}$$

Differentiating the identity

$$Z_1^T (x - (i_{\mathbb{M}} \circ \varphi^{-1})(\mathcal{S}(x))) \equiv 0,$$

we obtain

$$Z_1^T - Z_1^T U \mathcal{S}_x(x_0) = 0.$$

Thus, for (x_0, x_0') with some $x_0' \in \mathbb{R}^n$, we find that

$$\begin{aligned} \hat{F}_{1;x'}(x_0, x_0') &= \mathcal{S}_x(x_0) = (Z_1^T U)^{-1} Z_1^T, \\ \hat{F}_{2;x}(x_0) &= Z_2^T - Z_2^T U \mathcal{S}_x(x_0) = Z_2^T \end{aligned}$$

and the so constructed differential-algebraic equation is strangeness-free.

Remark 4.71. Note that a standard ordinary differential equation $y' = f(y)$ trivially is a differential equation on the manifold $\mathbb{R}^n$. This observation is another reason not to distinguish between ordinary differential equations and strangeness-free differential-algebraic equations with $a \neq 0$ in the definition of an index.

Bibliographical Remarks

The theory for general nonlinear differential-algebraic equations has only recently been studied in detail and is still in a state of active research with many open problems. Early results come from the work of Griepentrog and März [100]. The key idea for the current state of research came through the work of Campbell on derivative arrays [44], [46], [48], [52]. The solvability theory for square systems was developed by Campbell and Griepentrog in [54] and the general theory for over- and underdetermined systems was given by the authors in [129].

The generalization of the differentiation index for nonlinear differential-algebraic systems was a complicated process starting with the work of Gear [92], [93]. It culminated in a detailed analysis of a multitude of different index concepts [53]. The generalization of the strangeness index to nonlinear systems was given in [128], [129].

The geometric theory to study differential-algebraic systems as differential equations on manifolds was developed mainly in the work by Rheinboldt [190], [192], Rabier and Rheinboldt [174], [175], and Reich [185], [186], [187]. The analysis of singular points of nonlinear differential-algebraic systems is still an open problem and only very few results in this direction have been obtained, mainly for specially structured systems, see [151], [152], [173], [176], [177], [215].

For structured nonlinear differential-algebraic systems like those arising in multibody dynamics or circuit simulation, a more detailed theory is available, see for example [29], [79], [103], [104], [181], [214].

Exercises

1. Determine all solutions $x \in C^1(\mathbb{R}, \mathbb{R})$ of the differential-algebraic equation
$$x^2 - t^2 = 0.$$
Do the same, allowing for solutions $x \in C^0(\mathbb{R}, \mathbb{R})$.
2. Determine all solutions $x \in C^1(\mathbb{R}, \mathbb{R})$ of the differential-algebraic equation
$$x\dot{x} = 0.$$
3. Determine all solutions $x \in C^1(\mathbb{R}, \mathbb{R})$ of the differential-algebraic equation
$$x(\dot{x} - 2t) = 0.$$
4. Sketch the solution behavior of the initial value problem
$$\dot{x}_1 = 1, \quad x_1 - x_2(x_2^2 - 1) = 0, \quad x_1(0) = 0, \quad x_2(0) = -1$$
for $t \geq 0$. What happens?

5. Let $x^* \in C^1(\mathbb{I}, \mathbb{R}^2)$ with $\mathbb{I} = [0, \frac{2}{9}\sqrt{3})$ be the solution of the initial value problem of Exercise 4. Linearize the differential-algebraic equation at x^* and determine the local characteristic quantities of the linearized equation as function of $t \in \mathbb{I}$. What happens for $t \to \frac{2}{9}\sqrt{3}$?

6. Consider a sufficiently smooth function $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}}, \mathbb{R}^n)$. Let the matrices $M_2(t, x(t), \dot{x}(t), \ddot{x}(t))$ and $N_2(t, x(t), \dot{x}(t), \ddot{x}(t))$ be defined according to (4.12) for some appropriate function $x \in C^2(\mathbb{I}, \mathbb{R}^n)$. Verify that these matrices coincide with $M_2(t)$ and $N_2(t)$ defined according to (3.29) setting $E(t) = F_{\dot{x}}(t, x(t), \dot{x}(t))$ and $A(t) = -F_x(t, x(t), \dot{x}(t))$.

7. Show that a (real) linear differential-algebraic equation with variable coefficients that satisfies Hypothesis 3.48 also satisfies Hypothesis 4.2 and vice versa.

8. Consider

$$F(t, x, \dot{x}) = \begin{bmatrix} \dot{x}_2 - x_1 \\ x_2 \end{bmatrix}$$

together with

$$\tilde{F}(t, x, \dot{x}) = P(t, x, \dot{x}, F(t, x, \dot{x})), \quad P(t, x, \dot{x}, w) = \begin{bmatrix} 1 & 0 \\ \dot{x}_1 & 1 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix},$$

cp. Example 4.1. Verify the relations in the proof of Lemma 4.7. In particular, show that $\dot{x}_2 - x_1 = 0$ is sufficient to get a relation between (E, A) and $(\tilde{E}, \tilde{A})$ that has the form of a global equivalence.

9. Let $f : \mathbb{D} \to \mathbb{R}$, $\mathbb{D} = S((x_0, y_0), \varepsilon) \subseteq \mathbb{R}^m \times \mathbb{R}^n$, be continuously differentiable, where $S((x_0, y_0), \varepsilon)$ denotes the open ball of radius ε around (x_0, y_0). Show that $f_y(x, y) = 0$ for all $(x, y) \in \mathbb{D}$ implies that

$$f(x, y) = f(x, y_0)$$

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

10. For the problems of Exercises 1, 2, and 3, determine the set $\mathbb{L}_0 = F^{-1}(\{0\})$ for the corresponding function $F : \mathbb{R} \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. Check whether it is possible to restrict the domain of F in such a way that the restricted problem satisfies Hypothesis 4.2. If this is the case, determine the corresponding reduced differential-algebraic equation and discuss its solvability.

11. Let the differential-algebraic equation

$$\dot{x}_1^2 = \dot{x}_2, \quad x_2 = -t,$$

with $\mathbb{I} = \mathbb{R}$, $\mathbb{D}_x = \mathbb{R}^2$, and $\mathbb{D}_{\dot{x}} = \{(\dot{x}_1, \dot{x}_2) \in \mathbb{R}^2 \mid \dot{x}_1, \dot{x}_2 > 0\}$ be given. Show that it satisfies Hypothesis 4.2 with $\mu = 0$, $a = 1$, and $d = 1$. Determine a corresponding reduced differential-algebraic equation (4.23). Why is there no corresponding system of the form (4.25)?

12. Discuss the properties of the differential-algebraic equation

$$x_n \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \vdots \\ \vdots \\ \dot{x}_n \end{bmatrix} = \begin{bmatrix} 1 & & \\ & \ddots & \\ & & \ddots \\ & & & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} f_1(t) \\ \vdots \\ \vdots \\ f_n(t) \end{bmatrix}$$

for $n \geq 2$.

13. Work out a simplified proof (compared with those given in Section 4.2) for the claim that a differential-algebraic equation in Hessenberg form of index $\nu = 2$ satisfies Hypothesis 4.2 provided that the relevant constraints can be satisfied.

14. Discuss the properties of the overdetermined problem (4.46) given by

$$\dot{x} = f_1(t), \quad x = f_2(t)$$

with $m = 2$ and $n = 1$.

15. Consider the overdetermined problem

$$\begin{aligned} &\dot{p} = v, \\ &M(p)\dot{v} = f(p, v) - g_p(p)^T \lambda, \\ &g(p) = 0, \\ &g_p(p)v = 0, \\ &g_{pp}(p)(v, v) + g_p(p)M(p)^{-1}(f(p, v) - g_p(p)^T \lambda) = 0, \end{aligned}$$

obtained by just adding the hidden constraints of a multibody system to the original differential-algebraic equation. Check whether this problem satisfies Hypothesis 4.30 under the usual assumptions on M and g. If this is the case, determine its characteristic values.

16. Discuss the properties of the control problem (4.58) given by

$$x^2 - t^2 + u = 0$$

with $m = 1$, $n = 1$, and $l = 1$.

17. Discuss the properties of the control problem (4.58) given by

$$\dot{x}_1 = 0, \quad x_2 u = 0$$

with $m = 2$, $n = 2$, and $l = 1$. In particular, consider suitable restrictions of the domain of the associated function F.

18. Give the proof of Corollary 4.40.

19. Show that the control problem associated with the multibody system in Example 4.46 is regular and strangeness-free for given u near the initial value.

20. Show that the differential-algebraic equation

$$\begin{aligned}
&\dot{x}_1 = x_4, \quad \dot{x}_4 = -2x_1x_7,\\
&\dot{x}_2 = x_5, \quad \dot{x}_5 = -2x_2x_7,\\
&0 = x_1^2 + x_2^2 - x_3,\\
&0 = 2x_1x_4 + 2x_2x_5 - x_6,\\
&0 = 2x_4^2 - 4x_1^2x_7 + 2x_5^2 - 4x_2^2x_7 - x_7 + 1 = 0
\end{aligned}$$

satisfies Hypothesis 4.2 with characteristic values $\mu = 0$, $a = 3$, and $d = 4$. Reformulate this problem as a differential equation on a manifold.

Part II

Numerical solution of differential-algebraic equations

Chapter 5
Numerical methods for strangeness-free problems

In the first part of this textbook, we have given a detailed analysis of the existence and uniqueness of solutions, the consistency of initial conditions, and the theory of generalized solutions and control problems. The second part now deals with the numerical solution of differential-algebraic equations. In principle, one could try to apply standard discretization schemes for ordinary differential equations directly to differential-algebraic equations, by replacing for example derivatives by finite differences. But it was observed immediately that, in contrast to the numerical solution of ordinary differential equations, many difficulties arise for differential-algebraic equations. These difficulties are due to the algebraic constraints, in particular to the hidden constraints, i.e., to those algebraic constraints that are not explicitly given in the system. In view of the discussion in the first part, these arise in problems with a strangeness index larger than zero. First of all, it may happen that, although the problem has a unique solution, the solution of the discretized equation is not unique, or vice versa. We will present such examples below. A second problem is that explicit methods cannot be used directly as one can already see from the linear constant coefficient problem (2.1), since an explicit method would require the solution of a linear system with the (typically) singular matrix E. A third effect is that, due to discretization errors, the numerical solution may drift off from the analytical solution if the constraints are not explicitly forced during the integration. In order to avoid this effect, the solution has to be forced to lie on the constraint manifold. But, to be able to do this, a parameterization of this manifold has to be known which is often difficult in higher index problems. A fourth observation is that many differential-algebraic systems behave in some respect like stiff differential equations which forces one to use methods with good stability properties.

In view of the described difficulties, different approaches may be considered. For problems with a particular structure, like for example mechanical multibody systems or circuit simulation problems, one can use the structure to derive and analyze classical discretization schemes. We will describe some of these approaches below, see also [29], [79], [105], [108].

An alternative to a direct discretization of a higher index differential-algebraic equation is to discretize an equivalent formulation of the problem with strangeness index zero, as we have obtained it in (3.60) for linear systems with variable coefficients and in (4.23) for nonlinear systems. In these equivalent strangeness-free formulations, the solution set is the same as that of the original equation and parame-

terizations of the constraint manifold are explicitly available. Hence, the numerical solution can be forced to lie on this manifold.

In this chapter, we will discuss the two main classes of discretization methods, namely one-step methods (concentrating on Runge–Kutta methods) and (linear) multi-step methods, and their generalization to differential-algebraic equations. In both cases, we will start with the treatment of linear differential-algebraic equations with constant coefficients and discuss why similar results cannot hold for linear problems with variable coefficients. We then present methods that are suited for the treatment of nonlinear semi-explicit systems of index one and of nonlinear systems in the form (4.23).

Many of the results in this chapter are based on the well-known analysis for the treatment of ordinary differential equations, see, e.g., [106], [108], [210]. We assume that the reader is familiar with the basic concepts of this area. Nevertheless, we sketch some of the fundamental results from the treatment of ordinary differential equations placed in the context of general discretization methods. We complement these results by some topics that become relevant in the investigation of the presented numerical methods.

5.1 Preparations

In general, we study the numerical solution of initial value problems for differential-algebraic systems of the form

$$F(t, x, \dot{x}) = 0, \quad x(t_0) = x_0 \tag{5.1}$$

in the interval $\mathbb{I} = [t_0, T] \subset \mathbb{R}$. We denote by $t_0 < t_1 < t_2 < \cdots < t_N = T$ gridpoints in the interval $\mathbb{I}$ and by x_i approximations to the solution $x(t_i)$. We concentrate on a fixed stepsize, i.e., we use $t_i = t_0 + ih$, $i = 0, \ldots, N$, and $T - t_0 = Nh$. Note that this notation is in conflict with the notation of the nilpotent part in the Weierstraß canonical form (2.7), when we treat linear differential-algebraic equations with constant coefficients. But there will be no problem to distinguish the two meanings of N from the context.

A *discretization method* for the solution of (5.1) is given by an iteration

$$\mathfrak{X}_{i+1} = \mathfrak{F}(t_i, \mathfrak{X}_i; h), \tag{5.2}$$

where the $\mathfrak{X}_i$ are elements in some $\mathbb{R}^{\mathrm{n}}$, together with quantities $\mathfrak{X}(t_i) \in \mathbb{R}^{\mathrm{n}}$ representing the actual solution at t_i. We are then interested in conditions that guarantee convergence of the methods in the sense that $\mathfrak{X}_N$ tends to $\mathfrak{X}(t_N)$ when h tends to zero. Throughout this chapter we consider only real problems and we assume, for convenience, that all functions are defined on a compact set and are sufficiently

smooth, and that all numerical approximations stay in the domain of all relevant functions.

Definition 5.1. The discretization method (5.2) is said to be *consistent of order p* if

$$\|\mathfrak{X}(t_{i+1}) - \mathfrak{F}(t_i, \mathfrak{X}(t_i); h)\| \le Ch^{p+1}, \tag{5.3}$$

with a constant C independent of h.

Definition 5.2. The discretization method (5.2) is said to be *stable* if there exists a vector norm $\|\cdot\|$ such that

$$\|\mathfrak{F}(t_i, \mathfrak{X}(t_i); h) - \mathfrak{F}(t_i, \mathfrak{X}_i; h)\| \le (1 + hK)\|\mathfrak{X}(t_i) - \mathfrak{X}_i\| \tag{5.4}$$

in this vector norm, with a constant K independent of h.

Definition 5.3. The discretization method (5.2) is said to be *convergent of order p* if

$$\|\mathfrak{X}(t_N) - \mathfrak{X}_N\| \le Ch^p, \tag{5.5}$$

with a constant C independent of h, provided that

$$\|\mathfrak{X}(t_0) - \mathfrak{X}_0\| \le \tilde{C}h^p, \tag{5.6}$$

with a constant $\tilde{C}$ independent of h.

Note that consistency and convergence do not depend on the selected vector norm, since in $\mathbb{R}^n$ all vector norms are equivalent. Thus, if stability is involved, then we always work with the vector norm selected for (5.4).

Theorem 5.4. *If the discretization method (5.2) is stable and consistent of order p, then it is convergent of order p.*

Proof. From

$$\begin{aligned}\|\mathfrak{X}(t_{i+1}) - \mathfrak{X}_{i+1}\| &= \|\mathfrak{X}(t_{i+1}) - \mathfrak{F}(t_i, \mathfrak{X}(t_i); h) + \mathfrak{F}(t_i, \mathfrak{X}(t_i); h) - \mathfrak{X}_{i+1}\| \\ &\le Ch^{p+1} + (1 + hK)\|\mathfrak{X}(t_i) - \mathfrak{X}_i\|,\end{aligned}$$

it follows that

$$\begin{aligned}\|\mathfrak{X}(t_N) - \mathfrak{X}_N\| &\le Ch^{p+1} + (1 + hK)\|\mathfrak{X}(t_{N-1}) - \mathfrak{X}_{N-1}\| \\ &\le Ch^{p+1} + (1 + hK)Ch^{p+1} + (1 + hK)^2\|\mathfrak{X}(t_{N-2}) - \mathfrak{X}_{N-2}\| \\ &\le Ch^{p+1}(1 + (1 + hK) + \cdots + (1 + hK)^{N-1}) \\ &\quad + (1 + hK)^N\|\mathfrak{X}(t_0) - \mathfrak{X}_0\| \\ &\le Ch^{p+1}\frac{(1 + hK)^N - 1}{(1 + hK) - 1} + (1 + hK)^N\tilde{C}h^p \\ &\le (C/K + \tilde{C})\exp(K(T - t_0))h^p.\end{aligned}$$

□

When studying Runge–Kutta methods or linear multi-step methods, it is often convenient to describe the structure of some matrices via the Kronecker product of two matrices. The Kronecker product of $R = [r_{ij}] \in \mathbb{C}^{k,l}$ with $S \in \mathbb{C}^{m,n}$ is defined as the block matrix $R \otimes S = [r_{ij}S] \in \mathbb{C}^{km,ln}$. Its main properties are given in the following lemma.

Lemma 5.5. *The Kronecker product has the following properties:*

1. *Let matrices U, V and R, S be given such that the products UR, VS exist. Then,*

$$(U \otimes V)(R \otimes S) = UR \otimes VS. \tag{5.7}$$

2. *Let $R \in \mathbb{C}^{k,l}$ and $S \in \mathbb{C}^{m,n}$. Considering $R \otimes S$ as a block matrix consisting of blocks of size $m \times n$, we define the so-called perfect shuffle matrices Π_1 and Π_2 with respect to the rows and columns, respectively, by the following process: Take the first row/column of the first block, then the first row/column of the second block, and so on until the first row/column of the last block, continue in the same way with the second row/column of every block, until the last row/column of every block. With the so obtained permutation matrices Π_1 and Π_2, we get*

$$\Pi_1^T (R \otimes S)\Pi_2 = S \otimes R. \tag{5.8}$$

If $k = l$ and $m = n$, then $\Pi_1 = \Pi_2 = \Pi$ and $S \otimes R$ is similar to $R \otimes S$.

Proof. The proof is left as an exercise, cp. Exercise 1. □

When we discuss the numerical solution of strangeness-free nonlinear differential-algebraic equations of the form (4.23), we will always assume that a unique solution $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of the corresponding initial value problem exists. Linearizing (4.23) along x^* yields the matrix functions

$$\begin{aligned} \hat{E}_1(t) = \hat{F}_{1;\dot{x}}(t, x^*(t), \dot{x}^*(t)), \quad \hat{A}_1(t) &= -\hat{F}_{1;x}(t, x^*(t), \dot{x}^*(t)), \\ \hat{A}_2(t) &= -\hat{F}_{2;x}(t, x^*(t)). \end{aligned} \tag{5.9}$$

The central property of this linearization is given by the following lemma.

Lemma 5.6. *Consider a pair $(\hat{E}, \hat{A})$ of continuous matrix functions of the form*

$$\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \quad \hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} \tag{5.10}$$

according to (5.9). Then there exist (smooth) pointwise nonsingular matrix functions $P \in C(\mathbb{I}, \mathbb{R}^{n,n})$ and $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$P\hat{E}Q = \begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}, \quad P\hat{A}Q - P\hat{E}\dot{Q} = \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix}. \tag{5.11}$$

In particular, P has the special form

$$P = \begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix}$$

matching the block structure of $\hat{E}$ and $\hat{A}$.

Proof. Recalling that due to the construction of (4.23) in Section 4.1 the pair $(\hat{E}, \hat{A})$ of matrix functions is regular and strangeness-free, the claim is a special version of Theorem 3.32. The structure of P follows from the fact that in the sequence of equivalence relations in the proof of Theorem 3.32 the first block row never contributes to the second block row. □

For the systems of nonlinear equations that arise by discretization, we must show the existence of solutions near the corresponding value of the true solution. We will always do that on the basis of the following convergence result for a Newton-like method. Note that this Newton-like iteration is a tool for the analysis, it should not be used in the actual computation of numerical approximations. Observe also that for simplicity we use an adapted notation in this theorem and a related corollary.

Theorem 5.7. *Let $F \in C^1(\mathbb{D}, \mathbb{R}^n)$ with an open and convex set $\mathbb{D}$. Let $x^0 \in \mathbb{D}$ and let $\hat{x} \in \mathbb{D}$ be such that $F'(\hat{x})$ is invertible, where $F'(\hat{x})$ denotes the Fréchet derivative of F evaluated at $\hat{x}$. Furthermore, let constants α, β, γ be given such that for some vector norm and the associated matrix norm*

$$\|F'(\hat{x})^{-1} F(x^0)\| \le \alpha, \tag{5.12a}$$

$$\|F'(\hat{x})^{-1}\| \le \beta, \tag{5.12b}$$

$$\|F'(x) - F'(y)\| \le \gamma \|x - y\| \quad \text{for all } x, y \in \mathbb{D}, \gamma \neq 0, \tag{5.12c}$$

$$\|x^0 - \hat{x}\| < \frac{1}{\beta\gamma}, \tag{5.12d}$$

$$2\alpha\beta\gamma \le (1 + \beta\gamma\hat{\tau})^2 \quad \text{with } \hat{\tau} = -\|x^0 - \hat{x}\| \tag{5.12e}$$

$$\overline{S}(x^0, \rho_-) \subseteq \mathbb{D}, \quad \rho_\pm = \frac{1}{\beta\gamma}(1 + \beta\gamma\hat{\tau} \pm \sqrt{(1 + \beta\gamma\hat{\tau})^2 - 2\alpha\beta\gamma}), \tag{5.12f}$$

where $\overline{S}(x^0, \rho_-)$ denotes the closure of the open ball $S(x^0, \rho_-)$ of radius ρ_- around x^0. Then,

$$x^{m+1} = x^m - F'(\hat{x})^{-1} F(x^m) \tag{5.13}$$

defines a sequence $\{x^m\}$ of points in $\overline{S}(x^0, \rho_-)$ which converges to a point x^* in $\overline{S}(x^0, \rho_-)$ that satisfies $F(x^*) = 0$. There is no other solution of $F(x) = 0$ in

$$\overline{S}(x^0, \rho_-) \cup (S(x^0, \rho_+) \cap \mathbb{D}).$$

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

Proof. If $x^0, \ldots, x^m \in \mathbb{D}$, then we have

$$\begin{aligned}
\|x^{m+1} - x^m\| &= \|F'(\hat{x})^{-1} F(x^m)\| \\
&= \|F'(\hat{x})^{-1}[F(x^m) - F(x^{m-1}) - F'(\hat{x})(x^m - x^{m-1})]\| \\
&= \|F'(\hat{x})^{-1}[F(x^{m-1} + s(x^m - x^{m-1}))|_{s=0}^{s=1} - F'(\hat{x})(x^m - x^{m-1})]\| \\
&= \Big\| F'(\hat{x})^{-1} \int_0^1 \big[F'(x^{m-1} + s(x^m - x^{m-1})) - F'(\hat{x})\big](x^m - x^{m-1}) ds \Big\| \\
&\leq \beta\gamma \|x^m - x^{m-1}\| \int_0^1 \|x^{m-1} + s(x^m - x^{m-1}) - \hat{x}\| ds \\
&= \beta\gamma \|x^m - x^{m-1}\| \int_0^1 \|s(x^m - \hat{x}) + (1-s)(x^{m-1} - \hat{x})\| ds \\
&\leq \frac{1}{2}\beta\gamma \|x^m - x^{m-1}\| (\|x^m - \hat{x}\| + \|x^{m-1} - \hat{x}\|).
\end{aligned}$$

Considering the quadratic polynomial $\varphi : \mathbb{R} \to \mathbb{R}$ given by

$$\varphi(\tau) = \tfrac{1}{2}\beta\gamma\tau^2 - (\beta\gamma\hat{\tau} + 1)\tau + \alpha$$

and observing that $\dot{\varphi}(\hat{\tau}) = -1$, we define a sequence $\{\tau_m\}$ via

$$\tau_{m+1} = \tau_m - \frac{\varphi(\tau_m)}{\dot{\varphi}(\hat{\tau})} = \tau_m + \varphi(\tau_m), \quad \tau_0 = 0.$$

Using $\hat{\tau} = -\|x^0 - \hat{x}\|$, we obtain that

$$\|x^1 - x^0\| \leq \alpha = \tau_1 - \tau_0.$$

By induction, it follows that

$$\begin{aligned}
\|x^m - \hat{x}\| &\leq \|x^m - x^{m-1}\| + \cdots + \|x^1 - x^0\| + \|x^0 - \hat{x}\| \\
&\leq (\tau_m - \tau_{m-1}) + \cdots + (\tau_1 - \tau_0) + (\tau_0 - \hat{\tau}) = \tau_m - \hat{\tau}
\end{aligned}$$

and

$$\begin{aligned}
\|x^{m+1} - x^m\| &\leq \frac{1}{2}\beta\gamma(\tau_m - \tau_{m-1})((\tau_m - \hat{\tau}) + (\tau_{m-1} - \hat{\tau})) \\
&= \frac{1}{2}\beta\gamma(\tau_m - \tau_{m-1})(\tau_m + \tau_{m-1} - 2\hat{\tau}) \\
&= \frac{1}{2}\beta\gamma\tau_m^2 - \frac{1}{2}\beta\gamma\tau_{m-1}^2 - \beta\gamma\hat{\tau}(\tau_m - \tau_{m-1}) \\
&= \frac{1}{2}\beta\gamma\tau_m^2 - (\tau_m - \tau_{m-1} + (\beta\gamma\hat{\tau} + 1)\tau_{m-1} - \alpha) - \beta\gamma\hat{\tau}(\tau_m - \tau_{m-1}) \\
&= \varphi(\tau_m) = \tau_{m+1} - \tau_m.
\end{aligned}$$

Since $\rho_\pm$ are the two zeros of φ, it is obvious that the increasing sequence $\{\tau_m\}$ converges with $\tau_m \to \rho_-$. Hence, we have

$$\|x^m - x^0\| \le \tau_m - \tau_0 = \tau_m < \rho_-,$$

i.e., the sequence $\{x^m\}$ is well defined and stays in $S(x^0, \rho_-)$. Since $\{\tau_m\}$ is a converging majorizing sequence, $\{x^m\}$ converges to some $x^* \in \overline{S}(x^0, \rho_-)$. By the continuity of F, the limit x^* satisfies $F(x^*) = 0$.

To show the claimed uniqueness, let $x^{**} \in \overline{S}(x^0, \rho_-) \cup (S(x^0, \rho_+) \cap \mathbb{D})$ with $F(x^{**}) = 0$ be another solution. Then we have the estimate

$$\begin{aligned}
\|x^{m+1} - x^{**}\| &= \|x^m - F'(\hat{x})^{-1}F(x^m) - x^{**}\| \\
&= \|F'(\hat{x})^{-1}[F(x^{**}) - F(x^m) - F'(\hat{x})(x^{**} - x^m)]\| \\
&= \|F'(\hat{x})^{-1}[F(x^m + s(x^{**} - x^m))|_{s=0}^{s=1} - F'(\hat{x})(x^{**} - x^m)]\| \\
&= \Big\|F'(\hat{x})^{-1}\int_0^1 \big[F'(x^m + s(x^{**} - x^m)) - F'(\hat{x})\big](x^{**} - x^m)ds\Big\| \\
&\le \beta\gamma\|x^m - x^{**}\| \int_0^1 \|x^m + s(x^{**} - x^m) - \hat{x}\|ds \\
&\le \frac{1}{2}\beta\gamma\|x^m - x^{**}\|(\|x^m - \hat{x}\| + \|x^{**} - \hat{x}\|).
\end{aligned}$$

Defining a sequence $\{\sigma_m\}$ via

$$\sigma_{m+1} = \sigma_m - \frac{\varphi(\sigma_m)}{\dot{\varphi}(\hat{\tau})} = \sigma_m + \varphi(\sigma_m), \quad \sigma_0 = \|x^0 - x^{**}\|,$$

it is obvious that the decreasing sequence $\{\sigma_m\}$ converges with $\sigma_m \to \rho_-$, because of $\sigma_0 < \rho_+$. Starting with $\|x^0 - x^{**}\| \le \sigma_0 - \tau_0$, we have by induction that

$$\begin{aligned}
\|x^{m+1} - x^{**}\| &\le \frac{1}{2}\beta\gamma(\sigma_m - \tau_m)((\tau_m - \hat{\tau}) + (\sigma_m - \tau_m) + (\tau_m - \hat{\tau})) \\
&= \frac{1}{2}\beta\gamma(\sigma_m - \tau_m)(\sigma_m + \tau_m - 2\hat{\tau}) \\
&= \frac{1}{2}\beta\gamma\sigma_m^2 - \frac{1}{2}\beta\gamma\tau_m^2 - \beta\gamma\hat{\tau}(\sigma_m - \tau_m) \\
&= (\varphi(\sigma_m) + \sigma_m - \alpha) - (\varphi(\tau_m) + \tau_m - \alpha) \\
&= \sigma_{m+1} - \tau_{m+1},
\end{aligned}$$

and therefore $\|x^m - x^{**}\| \to 0$ or $x^m \to x^{**}$. Hence, $x^{**} = x^*$ due to the uniqueness of limits. Finally, if $\rho_- < \rho_+$, then x^* lies in the interior of $\overline{S}(x^0, \rho_-) \cup (S(x^0, \rho_+) \cap \mathbb{D})$ and is thus locally unique. □

Corollary 5.8. *If, in addition to the assumptions of Theorem 5.7,*

$$\|x^0 - \hat{x}\| \leq \tfrac{1}{2\beta\gamma}$$

holds, then

$$\|x^* - x^0\| \leq 4\alpha.$$

Proof. Theorem 5.7 yields

$$\begin{aligned}\|x^* - x^0\| \leq \rho_- &= \frac{2\alpha}{1 - \beta\gamma\|x^0 - \hat{x}\| + \sqrt{(1 - \beta\gamma\|x^0 - \hat{x}\|)^2 - 2\alpha\beta\gamma}} \\ &\leq \frac{2\alpha}{1 - \beta\gamma\|x^0 - \hat{x}\|} \\ &\leq 4\alpha.\end{aligned}$$

□

After these preparations, we begin our analysis of numerical discretization methods for differential-algebraic equations.

5.2 One-step methods

A one-step method for the computation of numerical approximations x_i to the values $x(t_i)$ of a solution x of an ordinary differential equation $\dot{x} = f(t, x)$ has the form

$$x_{i+1} = x_i + h\Phi(t_i, x_i; h), \tag{5.14}$$

where Φ is the so-called *increment function*. In the context of ordinary differential equations, a one-step method is called *consistent* of order p if, under the assumption that $x_i = x(t_i)$, the *local discretization error* $x_{i+1} - x(t_{i+1})$ satisfies

$$\|x(t_{i+1}) - x_{i+1}\| \leq Ch^{p+1}, \tag{5.15}$$

with a constant C that is independent of h. Using (5.14), this is equivalent to

$$\|x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i); h)\| \leq Ch^{p+1}. \tag{5.16}$$

Setting $\mathfrak{X}_i = x_i$, $\mathfrak{X}(t_i) = x(t_i)$, and $\mathfrak{F}(t_i, \mathfrak{X}_i; h) = x_i + h\Phi(t_i, x_i; h)$, the one-step method (5.14) can be seen as a general discretization method as introduced in Section 5.1. Since

$$\begin{aligned}&\|\mathfrak{F}(t_i, \mathfrak{X}(t_i); h) - \mathfrak{F}(t_i, \mathfrak{X}_i; h)\| \\ &\quad = \|(\mathfrak{X}(t_i) + h\Phi(t_i, \mathfrak{X}(t_i); h)) - (\mathfrak{X}_i + h\Phi(t_i, \mathfrak{X}_i; h))\| \\ &\quad \leq (1 + hK)\|\mathfrak{X}(t_i) - \mathfrak{X}_i\|,\end{aligned}$$

where K is the Lipschitz constant of Φ with respect to its second argument, these methods are stable without any further assumptions. Hence, consistency implies convergence of the one-step method.

As mentioned in the introduction of this chapter, in our discussion of one-step methods we will concentrate on Runge–Kutta methods. The general form of an s-stage Runge–Kutta method for the solution of $\dot{x} = f(t, x)$, $x(t_0) = x_0$ is given by

$$x_{i+1} = x_i + h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j}, \tag{5.17}$$

where

$$\dot{X}_{i,j} = f(t_i + \gamma_j h, X_{i,j}), \quad j = 1, \ldots, s, \tag{5.18}$$

and the so-called *internal stages* $X_{i,j}$ are given by

$$X_{i,j} = x_i + h \sum_{l=1}^{s} \alpha_{jl} \dot{X}_{i,l}, \quad j = 1, \ldots, s. \tag{5.19}$$

The coefficients α_{jl}, β_j, and γ_j determine the particular method and are conveniently displayed in a so-called *Butcher tableau*

$$\begin{array}{c|c} \gamma & \mathcal{A} \\ \hline & \beta^T \end{array}, \tag{5.20}$$

with $\mathcal{A} = [\alpha_{jl}]$, $\beta = [\beta_j]$, and $\gamma = [\gamma_j]$.

The coefficients are assumed to satisfy the condition

$$\gamma_j = \sum_{l=1}^{s} \alpha_{jl}, \quad j = 1, \ldots, s, \tag{5.21}$$

which implies that the Runge–Kutta method yields the same approximations to the solution x of $\dot{x} = f(t, x)$, $x(t_0) = x_0$, when we transform it to an equivalent autonomous problem by adding the trivial equation $\dot{t} = 1$, $t(t_0) = t_0$ to the system of ordinary differential equations. The remaining freedom in the coefficients is used to obtain a certain order of consistency. The following result is due to Butcher [37], see, e.g., [106, p. 208].

Theorem 5.9. *If the coefficients α_{jl}, β_j and γ_j of the Runge–Kutta method given by (5.17), (5.18), and (5.19) satisfy the conditions*

$$\begin{aligned}
B(p)\colon &\quad \sum_{j=1}^{s} \beta_j \gamma_j^{k-1} = \tfrac{1}{k}, && k = 1, \ldots, p,\\
C(q)\colon &\quad \sum_{l=1}^{s} \alpha_{jl} \gamma_l^{k-1} = \tfrac{1}{k}\gamma_j^{k}, && j = 1, \ldots, s,\ k = 1, \ldots, q, \qquad (5.22)\\
D(r)\colon &\quad \sum_{j=1}^{s} \beta_j \gamma_j^{k-1} \alpha_{jl} = \tfrac{1}{k}\beta_l(1-\gamma_l^{k}), && l = 1, \ldots, s,\ k = 1, \ldots, r,
\end{aligned}$$

with $p \le q + r + 1$ and $p \le 2q + 2$, then the method is consistent and hence convergent of order p.

There are many possibilities to determine appropriate coefficients for Runge–Kutta methods. See [38], [106], [108], [210] for details on the construction of Runge–Kutta methods. Some methods which are important in the numerical solution of differential-algebraic equations can be obtained as follows. Requiring $B(2s)$, $C(s)$, and $D(s)$ of (5.22) leads to a unique method of order $p = 2s$ for every stage number s. The simplest of these so-called *Gauß methods* can be found in Table 5.1.

Table 5.1. The simplest Gauß methods

$$\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{1}{2} - \frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4} - \frac{\sqrt{3}}{6} \\ \frac{1}{2} + \frac{\sqrt{3}}{6} & \frac{1}{4} + \frac{\sqrt{3}}{6} & \frac{1}{4} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

$$\begin{array}{c|ccc} \frac{1}{2} - \frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9} - \frac{\sqrt{15}}{15} & \frac{5}{36} - \frac{\sqrt{15}}{30} \\ \frac{1}{2} & \frac{5}{36} + \frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36} - \frac{\sqrt{15}}{24} \\ \frac{1}{2} + \frac{\sqrt{15}}{10} & \frac{5}{36} + \frac{\sqrt{15}}{30} & \frac{2}{9} + \frac{\sqrt{15}}{15} & \frac{5}{36} \\ \hline & \frac{5}{18} & \frac{4}{9} & \frac{5}{18} \end{array}$$

Requiring $B(2s-1)$, $C(s)$, and $D(s-1)$ together with $\gamma_s = 1$ leads to a unique method of order $p = 2s - 1$ for every stage number s. The simplest of these so-called *Radau IIA methods* can be found in Table 5.2.

The given formulation of Runge–Kutta methods immediately suggests a generalization to differential-algebraic equations of the form (5.1) by defining x_{i+1} as

Table 5.2. The simplest Radau IIA methods

$$\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

$$\begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-196\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{297+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

the solution of (5.17) and (5.19) together with

$$F(t_i + \gamma_j h, X_{i,j}, \dot{X}_{i,j}) = 0, \quad j = 1, \ldots, s. \tag{5.23}$$

Of course, these relations only define a method if we can show that they define a unique x_{i+1} in the vicinity of x_i at least for sufficiently small stepsize h. If this is the case, then we can analyze the convergence properties of the resulting methods. We will do this by considering these methods as general discretization methods.

We start our investigations by considering the case of linear differential-algebraic equations with constant coefficients $E\dot{x} = Ax + f(t)$, $x(t_0) = x_0$. In this case, the Runge–Kutta method has the form (5.17), with $\dot{X}_{i,l}$ obtained via the solution of the linear system

$$\begin{bmatrix} E - h\alpha_{1,1}A & -h\alpha_{1,2}A & \cdots & -h\alpha_{1,s}A \\ -h\alpha_{2,1}A & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & -h\alpha_{s-1,s}A \\ -h\alpha_{s,1}A & \cdots & -h\alpha_{s,s-1}A & E - h\alpha_{s,s}A \end{bmatrix} \dot{X}_i = Z_i, \tag{5.24}$$

where

$$\dot{X}_i = \begin{bmatrix} \dot{X}_{i,1} \\ \dot{X}_{i,2} \\ \vdots \\ \dot{X}_{i,s} \end{bmatrix}, \quad Z_i = \begin{bmatrix} Ax_i + f(t_i + \gamma_1 h) \\ Ax_i + f(t_i + \gamma_2 h) \\ \vdots \\ Ax_i + f(t_i + \gamma_s h) \end{bmatrix}.$$

Using the Kronecker product as introduced in Section 5.1, we can rewrite (5.24) as

$$(I_s \otimes E - h\mathcal{A} \otimes A)\dot{X}_i = Z_i. \tag{5.25}$$

It is clear that for nonsquare coefficient matrices E, A this system is not uniquely solvable for arbitrary right hand sides. But even in the square case, if the pair (E, A) is not regular, then the coefficient matrix in (5.25) is singular, see Exercise 3. Therefore, in order to obtain a well-defined method, we must require that the pair (E, A) is regular. Since for (nonsingular) matrices P, Q of appropriate dimensions

$$(I_s \otimes P)(I_s \otimes E - h\mathcal{A} \otimes A)(I_s \otimes Q) = (I_s \otimes PEQ - h\mathcal{A} \otimes PAQ),$$

we may assume for the analysis that the pair (E, A) is in Weierstraß canonical form (2.7), i.e.,

$$(E, A) = \left(\begin{bmatrix} I_d & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I_a \end{bmatrix} \right),$$

with J, N in Jordan canonical form and N nilpotent. Obviously, the system decouples and it suffices to investigate the subsystems as given in (2.8) and (2.9). The behavior of Runge–Kutta methods applied to ordinary differential equations is well studied, see, e.g., [106], [108]. For this reason, it suffices to consider the nilpotent part, i.e., a system of the form

$$N\dot{x} = x + f(t). \tag{5.26}$$

Furthermore, since the matrix N is in Jordan canonical form, the system decouples further and we can treat each Jordan block in N separately. Thus, for the analysis, we may assume that N in (5.26) consists of a single nilpotent Jordan block of size ν. In Lemma 2.8, we have shown that the solution of (5.26) is $x = -\sum_{j=0}^{\nu-1} N^j f^{(j)}$, independent of any initial values.

In this case the linear system (5.25) has the form

$$(I_s \otimes N - h\mathcal{A} \otimes I_\nu)\dot{X}_i = Z_i. \tag{5.27}$$

Using the perfect shuffle matrix Π of Lemma 5.5, we find that

$$\begin{aligned} \Pi^T (I_s \otimes N - h\mathcal{A} \otimes I_\nu)\Pi &= N \otimes I_s - I_\nu \otimes h\mathcal{A} \\ &= \begin{bmatrix} -h\mathcal{A} & I_s & & \\ & \ddots & \ddots & \\ & & \ddots & I_s \\ & & & -h\mathcal{A} \end{bmatrix}. \end{aligned} \tag{5.28}$$

Hence, to obtain a reasonable method, it is necessary that $\mathcal{A}$ is nonsingular, which implies that we are restricted to *implicit Runge–Kutta methods*. We then have the following result on the order of the local error.

Theorem 5.10. *Consider the differential-algebraic equation* (5.26) *with* $\nu = \mathrm{ind}(N, I)$. *Apply a Runge–Kutta method with coefficients* $\mathcal{A}$, β, *and* γ, *and assume that* $\mathcal{A}$ *is invertible. If* $\kappa_j \in \mathbb{N}$, $j = 1, \ldots, \nu$, *exist such that*

$$\begin{aligned} \beta^T \mathcal{A}^{-k} e &= \beta^T \mathcal{A}^{-j} \gamma^{j-k}/(j-k)!, && k = 1, 2, \ldots, j-1, \\ \beta^T \mathcal{A}^{-j} \gamma^k &= k!/(k-j+1)!, && k = j, j+1, \ldots, \kappa_j, \end{aligned} \tag{5.29}$$

where $e = [\,1 \;\cdots\; 1\,]^T$ *of appropriate size and* $\gamma^j = [\,\gamma_1^j \;\cdots\; \gamma_s^j\,]^T$, *then the local error satisfies*

$$x(t_{i+1}) - x_{i+1} = \mathcal{O}(h^{\kappa_\nu - \nu + 2}) + \mathcal{O}(h^{\kappa_{\nu-1} - \nu + 3}) + \cdots + \mathcal{O}(h^{\kappa_1 + 1}). \tag{5.30}$$

Proof. To derive an estimate for the local error, we study one step of the Runge–Kutta method from t_i to t_{i+1} and assume that we start with exact initial values, i.e., that $x_i = x(t_i)$. In addition, we may assume without loss of generality that N consists only of one Jordan block of size ν. From (5.28) we see that (5.27) is uniquely solvable for $\dot{X}_i$, with

$$\dot{X}_i = -\Pi \begin{bmatrix} (h\mathcal{A})^{-1} & (h\mathcal{A})^{-2} & \cdots & (h\mathcal{A})^{-\nu} \\ & \ddots & \ddots & \vdots \\ & & \ddots & (h\mathcal{A})^{-2} \\ & & & (h\mathcal{A})^{-1} \end{bmatrix} \Pi^T Z_i.$$

Moreover,

$$\Pi^T Z_i = \begin{bmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,\nu} \end{bmatrix}, \quad Z_{i,l} = \begin{bmatrix} x_{i,l} + f_l(t_i + \gamma_1 h) \\ x_{i,l} + f_l(t_i + \gamma_2 h) \\ \vdots \\ x_{i,l} + f_l(t_i + \gamma_s h) \end{bmatrix}, \quad f(t) = \begin{bmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_\nu(t) \end{bmatrix}.$$

Using (formal) Taylor expansion of $x(t_{i+1})$ and the representation (5.17) of the numerical solution, we obtain

$$x(t_{i+1}) - x_{i+1} = -h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j} + \sum_{k \geq 1} \frac{h^k}{k!} x^{(k)}(t_i).$$

It is sufficient to look at the first component τ_1 of the local error $x(t_{i+1}) - x_{i+1}$, since the j-th component becomes the first component if we consider problem (5.26) with index $\nu + 1 - j$.

Since $(\beta^T \otimes I_\nu)\Pi = I_\nu \otimes \beta^T$, we get that

$$\tau_1 = h\beta^T \sum_{j=1}^{\nu} (h\mathcal{A})^{-j} Z_{i,j} + \sum_{k \geq 1} \frac{h^k}{k!} x_1^{(k)}(t_i).$$

The representation $x(t) = -\sum_{j=0}^{\nu-1} N^j f^{(j)}$ of the solution yields for the different components

$$x_l(t) = -\sum_{j=l}^{\nu} f_j^{(j-l)}(t).$$

Together with the Taylor expansion

$$Z_{i,l} = x_l(t_i)e + \sum_{k\geq 0} \frac{h^k}{k!} f_l^{(k)}(t_i)\gamma^k,$$

we then obtain that

$$\begin{aligned}
\tau_1 &= \sum_{j=1}^{\nu} h^{-j+1}\beta^T \mathcal{A}^{-j}\Big(-\sum_{k=j}^{\nu} f_k^{(k-j)}(t_i)e + \sum_{k\geq 0}\frac{h^k}{k!} f_j^{(k)}(t_i)\gamma^k\Big) \\
&\quad + \sum_{k\geq 1}\frac{h^k}{k!}\Big(-\sum_{j=1}^{\nu} f_j^{(k+j-1)}(t_i)\Big) \\
&= -\sum_{k=1}^{\nu}\sum_{j=1}^{k} h^{-j+1}\beta^T \mathcal{A}^{-j} e f_k^{(k-j)}(t_i) \\
&\quad + \sum_{k\geq 0}\sum_{j=1}^{\nu}\frac{h^{k-j+1}}{k!}\beta^T \mathcal{A}^{-j}\gamma^k f_j^{(k)}(t_i) - \sum_{k\geq 1}\sum_{j=1}^{\nu}\frac{h^k}{k!} f_j^{(k+j-1)}(t_i) \\
&= \sum_{j=1}^{\nu}\Big(-\sum_{k=1}^{j} h^{-k+1}\beta^T \mathcal{A}^{-k} e f_j^{(j-k)}(t_i) \\
&\qquad + \sum_{k\geq 0}\frac{h^{k-j+1}}{k!}\beta^T \mathcal{A}^{-j}\gamma^k f_j^{(k)}(t_i) - \sum_{k\geq 1}\frac{h^k}{k!} f_j^{(k+j-1)}(t_i)\Big).
\end{aligned}$$

Reordering with respect to powers of h gives

$$\begin{aligned}
\tau_1 = \sum_{j=1}^{\nu}\Big(&-\sum_{k=1}^{j} h^{-k+1}\beta^T \mathcal{A}^{-k} e f_j^{(j-k)}(t_i) + \sum_{k=0}^{j-1}\frac{h^{k-j+1}}{k!}\beta^T \mathcal{A}^{-j}\gamma^k f_j^{(k)}(t_i) \\
&+ \sum_{k\geq j}\frac{h^{k-j+1}}{k!}\beta^T \mathcal{A}^{-j}\gamma^k f_j^{(k)}(t_i) - \sum_{k\geq 1}\frac{h^k}{k!} f_j^{(k+j-1)}(t_i)\Big)
\end{aligned}$$

$$= \sum_{j=1}^{\nu} \Big(- \sum_{k=1}^{j} h^{-k+1} \beta^T \mathcal{A}^{-k} e f_j^{(j-k)}(t_i) + \sum_{k=1}^{j} \frac{h^{-k+1}}{(j-k)!} \beta^T \mathcal{A}^{-j} \gamma^{j-k} f_j^{(j-k)}(t_i)$$
$$+ \sum_{k \geq j} \frac{h^{k-j+1}}{k!} \beta^T \mathcal{A}^{-j} \gamma^k f_j^{(k)}(t_i) - \sum_{k \geq j} \frac{h^{k-j+1}}{(k-j+1)!} f_j^{(k)}(t_i) \Big).$$

and the order conditions are obvious. □

An important class of Runge–Kutta methods for differential-algebraic equations are the so-called *stiffly accurate Runge–Kutta methods*. These are defined to satisfy $\beta_j = \alpha_{sj}$ for all $j = 1, \ldots, s$, see [108]. Writing this as $\beta^T = e_s^T \mathcal{A}$ with $e_s^T = [\, 0 \ \cdots \ 0 \ 1 \,]^T$ of appropriate size, we then obtain from (5.21) that $\gamma_s = e_s^T \mathcal{A} e = \beta^T e$. Since $\beta^T e = 1$ for consistent Runge–Kutta methods, it follows that $\gamma_s = 1$. Moreover, we get that $\beta^T \mathcal{A}^{-1} e = e_s^T e = 1$, implying that κ_1 in (5.29) is infinite. Hence, these methods show the same order of consistency for regular linear differential-algebraic equations with constants coefficients as for ordinary differential equations. For differential-algebraic equations of higher index or for methods which are not stiffly accurate, however, we may have a reduction of the order. Note that the Radau IIA methods are stiffly accurate by construction.

Example 5.11. For the Gauß method with $s = 2$, see Table 5.1, we have that $\kappa_1 = 2$ and $\kappa_2 = 2$, cp. Exercise 4. For the Radau IIA method with $s = 2$, see Table 5.2, we have that $\kappa_1 = \infty$ and $\kappa_2 = 2$, cp. Exercise 5. In particular, for both methods the local error is $\mathcal{O}(h^2)$ when we apply them to (5.26) with $\mathrm{ind}(N, I) = 2$.

For the global error, we then have the following order result.

Theorem 5.12. *Consider a Runge–Kutta method consisting of* (5.17), (5.19), *and* (5.24) *with invertible* $\mathcal{A}$ *applied to a linear differential-algebraic equation with constant coefficients of the form* $E\dot{x} = Ax + f(t)$, $x(t_0) = x_0$ *with a regular pair* (E, A) *and* $\nu = \mathrm{ind}(E, A)$. *Furthermore, let* $\kappa_j \geq j$, $j = 1, \ldots, \nu$, *according to Theorem* 5.10 *and let*

$$|1 - \beta^T \mathcal{A}^{-1} e| < 1. \tag{5.31}$$

Then the Runge–Kutta method is convergent of order

$$\min_{1 \leq j \leq \nu} \{p, \kappa_j - \nu + 2\}, \tag{5.32}$$

where p *is the order of the method when applied to ordinary differential equations.*

Proof. As in the proof of Theorem 5.10, it is sufficient to consider a single nilpotent Jordan block (N, I_ν). But in contrast to that proof, for the analysis of the global

error we cannot assume that $x_i = x(t_i)$. We are therefore in the following situation. The numerical approximations satisfy

$$x_{i+1} = x_i + h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j},$$

with

$$(I_s \otimes N - h\mathcal{A} \otimes I_\nu)\dot{X}_i = Z_i,$$

where

$$\dot{X}_i = \begin{bmatrix} \dot{X}_{i,1} \\ \dot{X}_{i,2} \\ \vdots \\ \dot{X}_{i,s} \end{bmatrix}, \quad Z_i = \begin{bmatrix} x_i + f(t_i + \gamma_1 h) \\ x_i + f(t_i + \gamma_2 h) \\ \vdots \\ x_i + f(t_i + \gamma_s h) \end{bmatrix}.$$

Due to Theorem 5.10, the actual solution satisfies

$$x(t_{i+1}) = x(t_i) + h \sum_{j=1}^{s} \beta_j \dot{\tilde{X}}_{i,j} + \delta_i,$$

with

$$(I_s \otimes N - h\mathcal{A} \otimes I_\nu)\dot{\tilde{X}}_i = \tilde{Z}_i,$$

where

$$\dot{\tilde{X}}_i = \begin{bmatrix} \dot{\tilde{X}}_{i,1} \\ \dot{\tilde{X}}_{i,2} \\ \vdots \\ \dot{\tilde{X}}_{i,s} \end{bmatrix}, \quad \tilde{Z}_i = \begin{bmatrix} x(t_i) + f(t_i + \gamma_1 h) \\ x(t_i) + f(t_i + \gamma_2 h) \\ \vdots \\ x(t_i) + f(t_i + \gamma_s h) \end{bmatrix},$$

The quantity δ_i is the local error committed in the i-th step and is bounded according to (5.30). Hence, the propagation of the global error $\varepsilon_i = x(t_i) - x_i$ is described by the recursion

$$\varepsilon_{i+1} = \varepsilon_i + h \sum_{j=1}^{s} \beta_j (\dot{\tilde{X}}_{i,j} - \dot{X}_{i,j}) + \delta_i.$$

Using again the perfect shuffle matrix as in the proof of Theorem 5.10, we may rewrite

$$(I_s \otimes N - h\mathcal{A} \otimes I_\nu)(\dot{\tilde{X}}_i - \dot{X}_i) = \tilde{Z}_i - Z_i = e \otimes \varepsilon_i$$

as

$$\begin{aligned}
&\sum_{j=1}^{s} \beta_j (\dot{\tilde{X}}_{i,j} - \dot{X}_{i,j}) \\
&\quad = -\begin{bmatrix} \beta^T & & \\ & \beta^T & \\ & & \ddots \\ & & & \beta^T \end{bmatrix} \begin{bmatrix} (h\mathcal{A})^{-1} & (h\mathcal{A})^{-2} & \cdots & (h\mathcal{A})^{-\nu} \\ & \ddots & \ddots & \vdots \\ & & \ddots & (h\mathcal{A})^{-2} \\ & & & (h\mathcal{A})^{-1} \end{bmatrix} \begin{bmatrix} \varepsilon_{i,1} e \\ \varepsilon_{i,2} e \\ \vdots \\ \varepsilon_{i,\nu} e \end{bmatrix} \\
&\quad = -\begin{bmatrix} \beta^T (h\mathcal{A})^{-1} e \varepsilon_{i,1} + \cdots + \beta^T (h\mathcal{A})^{-\nu} e \varepsilon_{i,\nu} \\ \vdots \\ \beta^T (h\mathcal{A})^{-1} e \varepsilon_{i,\nu} \end{bmatrix} \\
&\quad = -\begin{bmatrix} \beta^T (h\mathcal{A})^{-1} e & \cdots & \beta^T (h\mathcal{A})^{-\nu} e \\ & \ddots & \vdots \\ & & \beta^T (h\mathcal{A})^{-1} e \end{bmatrix} \varepsilon_i .
\end{aligned}$$

Thus, the error recursion takes the form

$$\varepsilon_{i+1} = \varepsilon_i - h \begin{bmatrix} \beta^T (h\mathcal{A})^{-1} e & \cdots & \beta^T (h\mathcal{A})^{-\nu} e \\ & \ddots & \vdots \\ & & \beta^T (h\mathcal{A})^{-1} e \end{bmatrix} \varepsilon_i + \delta_i = M \varepsilon_i + \delta_i ,$$

with

$$M = \begin{bmatrix} 1 - \beta^T \mathcal{A}^{-1} e & -h^{-1} \beta^T \mathcal{A}^{-2} e & \cdots & -h^{-\nu+1} \beta^T \mathcal{A}^{-\nu} e \\ & \ddots & \ddots & \vdots \\ & & \ddots & -h^{-1} \beta^T \mathcal{A}^{-2} e \\ & & & 1 - \beta^T \mathcal{A}^{-1} e \end{bmatrix} .$$

Note that the matrix M is the same in each step of the recursion, since the stepsize h is constant. Hence, with $\varepsilon_0 = 0$, we get that

$$\varepsilon_N = \sum_{i=0}^{N-1} M^{N-1-i} \delta_i .$$

Taking componentwise worst case bounds $|\delta_i| \le \delta$ for $i = 0, \ldots, N-1$, with $\delta \ge 0$, we obtain the componentwise estimate

$$|\varepsilon_N| \le \sum_{i=0}^{N-1} |M|^{N-1-i} \delta .$$

By assumption, there exists a constant $\sigma > 0$ independent of h such that $|1 - \beta^T \mathcal{A}^{-1} e| < \sigma < 1$. Thus, we have a componentwise upper bound $\tilde{M}$ for $|M|$ of the form

$$\tilde{M} = \begin{bmatrix} \sigma & \sigma(C/h) & \cdots & \sigma(C/h)^{\nu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \sigma(C/h) \\ & & & \sigma \end{bmatrix}$$

with an appropriate constant $C > 0$ independent of h, and we can express $\tilde{M}$ as

$$\tilde{M} = \sigma(I_\nu + \tilde{N} + \tilde{N}^2 + \cdots + \tilde{N}^{\nu-1}), \quad \tilde{N} = (C/h)N.$$

Defining the polynomials P_j, $j = 1, \ldots, \nu - 1$, by

$$P_j(i) = \binom{i+j-1}{i-1},$$

we will prove by induction that

$$\tilde{M}^i = \sigma^i(I_\nu + P_1(i)\tilde{N} + P_2(i)\tilde{N}^2 + \ldots + P_{\nu-1}(i)\tilde{N}^{\nu-1}).$$

Obviously, the claim holds for $i = 1$. Assuming that the claim holds for some $i \geq 1$, we compute

$$\tilde{M}^{i+1} = \tilde{M}^i \tilde{M} = \sigma^{i+1}(I_\nu + \tilde{P}_1\tilde{N} + \tilde{P}_2\tilde{N}^2 + \cdots + \tilde{P}_{\nu-1}\tilde{N}^{\nu-1}),$$

with

$$\tilde{P}_j = P_j(i) + \cdots + P_2(i) + P_1(i) + 1$$

for $j = 1, \ldots, \nu - 1$. Inserting the definition of the polynomials P_j, we get that

$$\begin{aligned} \tilde{P}_j &= \tbinom{i+j-1}{i-1} + \cdots + \tbinom{i+1}{i-1} + \tbinom{i}{i-1} + \tbinom{i}{i} = \tbinom{i+j-1}{i-1} + \cdots + \tbinom{i+1}{i-1} + \tbinom{i+1}{i} \\ &= \tbinom{i+j-1}{i-1} + \cdots + \tbinom{i+2}{i} = \cdots = \tbinom{i+j}{i} = P_j(i+1). \end{aligned}$$

Thus, it follows that

$$\begin{aligned} \sum_{i=0}^{N-1} \tilde{M}^{N-1-i} &= \sum_{i=0}^{N-1} \tilde{M}^i \\ &= \sum_{i=0}^{N-1} \sigma^i I_\nu + \sum_{i=0}^{N-1} \sigma^i P_1(i)\tilde{N} + \cdots + \sum_{i=0}^{N-1} \sigma^i P_{\nu-1}(i)\tilde{N}^{\nu-1} \\ &\leq \frac{1}{1-\sigma} I_\nu + \tilde{C}\tilde{N} + \cdots + \tilde{C}\tilde{N}^{\nu-1}, \end{aligned}$$

with a constant $\tilde{C} > 0$ independent of h which is defined as the maximum value of the converging infinite sums $\sum_{i=0}^{\infty} \sigma^i P_j(i)$, $j = 1, \ldots, \nu - 1$. We therefore end up with

$$|\varepsilon_N| \le \begin{bmatrix} \mathcal{O}(1) & \mathcal{O}(h^{-1}) & \cdots & \mathcal{O}(h^{-\nu+1}) \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mathcal{O}(h^{-1}) \\ & & & \mathcal{O}(1) \end{bmatrix} \begin{bmatrix} \mathcal{O}(h^{\kappa_\nu-\nu+2}) \\ \mathcal{O}(h^{\kappa_{\nu-1}-\nu+3}) \\ \vdots \\ \mathcal{O}(h^{\kappa_1+1}) \end{bmatrix},$$

which implies that

$$|\varepsilon_N| \le \begin{bmatrix} \mathcal{O}(h^{\kappa_\nu-\nu+2}) + \mathcal{O}(h^{\kappa_{\nu-1}-\nu+2}) + \cdots + \mathcal{O}(h^{\kappa_1-\nu+2}) \\ \mathcal{O}(h^{\kappa_{\nu-1}-\nu+3}) + \cdots + \mathcal{O}(h^{\kappa_1-\nu+3}) \\ \vdots \\ \mathcal{O}(h^{\kappa_1+1}) \end{bmatrix}.$$

□

This result shows that, even if the given Runge–Kutta method satisfies (5.31), it is not stable in the sense of Definition 5.2 for higher index problems, i.e., for problems with $\nu > 1$. In particular, we observe that in these cases the order of convergence may be lower than the order of consistency. It even may happen that we loose convergence.

Example 5.13. Consider again the differential-algebraic equation (5.26) with $\mathrm{ind}(N, I) = 2$. For this problem, the Gauß method with $s = 2$ given in Table 5.1 is not convergent at all while the Radau IIA method with $s = 2$ given in Table 5.2 is convergent of order two, see Exercises 6 and 7.

In contrast to the case of constant coefficients, where for regular pairs in each step the next iterate and the stage values are uniquely determined at least for sufficiently small h, difficulties already arise in the case of variable coefficients. Similar to (5.24), the linear system for the stage variables has the form $(\tilde{E} - h\tilde{A})\dot{X}_i = Z_i$, with

$$\tilde{E} = \begin{bmatrix} E(t_i + \gamma_1 h) & & \\ & \ddots & \\ & & E(t_i + \gamma_s h) \end{bmatrix},$$

$$\tilde{A} = \begin{bmatrix} A(t_i + \gamma_1 h)\alpha_{11} & \cdots & A(t_i + \gamma_1 h)\alpha_{1s} \\ \vdots & & \vdots \\ A(t_i + \gamma_s h)\alpha_{s1} & \cdots & A(t_l + \gamma_s h)\alpha_{ss} \end{bmatrix}$$

and

$$Z_i = \begin{bmatrix} A(t_i + \gamma_1 h)x_i + f(t_i + \gamma_1 h) \\ \vdots \\ A(t_i + \gamma_s h)x_i + f(t_i + \gamma_s h) \end{bmatrix}.$$

If we use for example the implicit Euler method, then $\tilde{E} = E(t_i)$ and $\tilde{A} = A(t_i)$. Thus, in order to solve for the stage variables, we need that $E(t_i) - hA(t_i)$ is invertible, which requires the pair $(E(t_i), A(t_i))$ to be regular. But we have seen in Chapter 3 that it may happen that the pair $(E(t), A(t))$ is singular for all t even though there exists a unique solution of the differential-algebraic equation, cp. Example 3.2. Replacing the implicit Euler method in such a case by some other Runge–Kutta methods, the pairs $(\tilde{E}, \tilde{A})$ may be invertible, but typically then the linear systems that have to be solved for the stage variables are ill-conditioned.

Example 5.14. If we apply the Radau IIA method with $s = 2$ given in Table 5.2 to the problem of Example 3.2 with $\mathbb{I} = [0, 1]$, then the coefficient matrix

$$\tilde{E} - h\tilde{A} = \left[\begin{array}{cc|cc} \frac{5}{12}h & -\frac{5}{12}h(t_i + \frac{1}{3}h) & -\frac{1}{12}h & \frac{1}{12}h(t_i + \frac{1}{3}h) \\ 1 & -(t_i + \frac{1}{3}h) & 0 & 0 \\ \hline \frac{3}{4}h & -\frac{3}{4}h(t_i + h) & \frac{1}{4}h & -\frac{1}{4}h(t_i + h) \\ 0 & 0 & 1 & -(t_i + h) \end{array}\right]$$

is invertible for all $h \neq 0$, but the condition number is $\mathcal{O}(h^{-2})$, cp. Exercise 9.

A related effect is that the Runge–Kutta method may lead to an unstable recursion for the numerical solutions x_i.

Example 5.15. Consider the linear differential-algebraic equation

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & -\eta t \\ 0 & -(1+\eta) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix} \tag{5.33}$$

with a parameter $\eta \in \mathbb{R}$. Note that for $\eta = -1$ we get the problem of Example 3.2. Performing a change of basis via

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} -\eta t & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix},$$

we obtain the equivalent constant coefficient system

$$\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} + \begin{bmatrix} f_1(t) \\ f_2(t) \end{bmatrix}. \tag{5.34}$$

Obviously, (5.34) and thus (5.33) has strangeness index $\mu = 1$ independent of η. A short computation shows that (5.33) has a unique solution given by

$$\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \begin{bmatrix} f_1(t) - \eta t(f_2(t) - \dot{f}_1(t)) \\ f_2(t) - \dot{f}_1(t) \end{bmatrix}$$

without specifying initial values.

Applying the implicit Euler method gives

$$0 = -x_{1,i+1} - \eta t_{i+1} x_{2,i+1} + f_1(t_{i+1}),$$
$$\tfrac{1}{h}(x_{1,i+1} - x_{1,i}) + \eta t_{i+1} \tfrac{1}{h}(x_{2,i+1} - x_{2,i}) = -(1+\eta) x_{2,i+1} + f_2(t_{i+1}).$$

Solving the first equation in two consecutive steps for $x_{1,i}$, $x_{1,i+1}$ and inserting in the second equation yields a recursion for $x_{2,i}$ of the form

$$x_{2,i+1} = \tfrac{\eta}{1+\eta} x_{2,i} + \tfrac{1}{1+\eta}(f_2(t_{i+1}) - \tfrac{1}{h}(f_1(t_{i+1}) - f_1(t_i))),$$

which is obviously divergent for $\eta < -\frac{1}{2}$.

Finally, it is also possible that the Runge–Kutta method determines a unique numerical solution, although the given problem is not uniquely solvable at all, cp. Exercise 11.

In order to obtain convergence results for classes of Runge–Kutta methods that include the implicit Euler method, we are therefore forced to restrict the class of problems. In view of the theoretical results, which say that we can transform higher index problems to strangeness-free problems with the same solution, it is natural to consider problems of the form (3.60) in the linear case or of the form (4.23) in the nonlinear case. Recall that, under the assumption that this problem possesses a unique solution, we can make use of the equivalent formulation (4.25).

In the following, we consider two cases. In the first case, we assume (4.23) to be a semi-explicit differential-algebraic equation of index $\nu = 1$, see Section 4.2, and develop convergence results for some larger classes of Runge–Kutta methods. In the second case, we drop the additional assumption on the structure of (4.23) but treat only a restricted class of Runge–Kutta methods.

Consider first a uniquely solvable initial value problem (5.1) with a semi-explicit differential-algebraic equation of the form

$$\dot{x} = f(t, x, y), \quad 0 = g(t, x, y). \tag{5.35}$$

If we assume that along the given solution (x, y) the Jacobian $g_y(t, x(t), y(t))$ is nonsingular, then we can solve the second equation in (5.35) for y. Following the notation in Section 4.1, we write $y = \mathcal{R}(t, x)$. In order to generalize a given Runge–Kutta method for solving an ordinary differential equation to (5.35), we introduce a small parameter ε to get

$$\dot{x} = f(t, x, y), \quad \varepsilon \dot{y} = g(t, x, y). \tag{5.36}$$

Applying a given Runge–Kutta method, we get

$$x_{i+1} = x_i + h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j}, \quad y_{i+1} = y_i + h \sum_{j=1}^{s} \beta_j \dot{Y}_{i,j}, \tag{5.37}$$

together with

$$\dot{X}_{i,j} = f(t_i + \gamma_j h, X_{i,j}, Y_{i,j}), \quad \varepsilon \dot{Y}_{i,j} = g(t_i + \gamma_j h, X_{i,j}, Y_{i,j}), \tag{5.38}$$

and

$$X_{i,j} = x_i + h \sum_{l=1}^{s} \alpha_{jl} \dot{X}_{i,l}, \quad Y_{i,j} = y_i + h \sum_{l=1}^{s} \alpha_{jl} \dot{Y}_{i,l}. \tag{5.39}$$

Setting then $\varepsilon = 0$ in (5.38) yields

$$\dot{X}_{i,j} = f(t_i + \gamma_j h, X_{i,j}, Y_{i,j}), \quad 0 = g(t_i + \gamma_j h, X_{i,j}, Y_{i,j}) \tag{5.40}$$

which is just (5.23) for (5.35). Thus, the method defined by (5.37), (5.39), and (5.40) is nothing else than the general method consisting of (5.17), (5.19), and (5.23) applied to (5.35). According to [105], this kind of constructing a method for a semi-explicit differential-algebraic equation is called *direct approach*.

In the so-called *indirect approach*, the second equation in (5.37) and (5.40) is replaced by

$$0 = g(t_{i+1}, x_{i+1}, y_{i+1}). \tag{5.41}$$

It then follows that

$$Y_{i,j} = \mathcal{R}(t_i + \gamma_j h, X_{i,j}), \quad y_{i+1} = \mathcal{R}(t_{i+1}, x_{i+1}) \tag{5.42}$$

such that we can eliminate all quantities which are related to the part y of (5.35). The remaining equations for the variables associated with x then read

$$x_{i+1} = x_i + \sum_{j=1}^{s} \beta_j f(t_i + \gamma_j h, X_{i,j}, \mathcal{R}(t_i + \gamma_j h, X_{i,j})) \tag{5.43}$$

and

$$X_{i,j} = x_i + \sum_{l=1}^{s} \alpha_{jl} f(t_i + \gamma_l h, X_{i,l}, \mathcal{R}(t_i + \gamma_l h, X_{i,l})), \tag{5.44}$$

which is just the given Runge–Kutta method applied to the ordinary differential equation

$$\dot{x} = f(t, x, \mathcal{R}(t, x)). \tag{5.45}$$

From this observation, it is obvious that in the indirect approach all convergence results for a given Runge–Kutta method when applied to an ordinary differential equation carry over to semi-explicit differential-algebraic equations of index $\nu = 1$. In particular, we have convergence with the same order.

Thus, it remains to perform the analysis for the direct approach. For convenience, we restrict ourselves to autonomous problems by dropping the argument t of f and g in (5.35). The general case then follows by the usual approach to turn a non-autonomous problem into an autonomous problem.

Theorem 5.16. *Consider an autonomous differential-algebraic system* (5.35) *of index* $\nu = 1$ *together with consistent initial values* (x_0, y_0) *due to* $0 = g(x_0, y_0)$. *Apply a Runge–Kutta method given by* (5.37), (5.39), *and* (5.40) *with invertible coefficient matrix* $\mathcal{A}$. *Assume that it has order* p *for ordinary differential equations and that it satisfies condition* $C(q)$ *of* (5.22) *with* $p \geq q+1$, *and let* $\varrho = 1-\beta^T \mathcal{A}^{-1} e$. *If* $|\varrho| \leq 1$, *then the global error satisfies*

$$\|x(t_N) - x_N\| = \mathcal{O}(h^k),$$

where k *is given as follows:*

1. *If* $\varrho = 0$, *then* $k = p$.
2. *If* $-1 \leq \varrho < 1$, *then* $k = \min\{p, q+1\}$.
3. *If* $\varrho = 1$, *then* $k = \min\{p-1, q\}$.

If $|\varrho| > 1$, *then the method is not convergent.*

Proof. Since the system is autonomous, the stage values $X_{i,j}$, $Y_{i,j}$ satisfy

$$\dot{X}_{i,j} = f(X_{i,j}, Y_{i,j}), \quad 0 = g(X_{i,j}, Y_{i,j}).$$

By the implicit function theorem applied to the constraint equation, we always have consistency according to $Y_{i,j} = \mathcal{R}(X_{i,j})$, $j = 1, \ldots, s$. Inserting this into (5.37) and (5.39), we obtain

$$x_{i+1} = x_i + h \sum_{j=1}^{s} \beta_j f(X_{i,j}, \mathcal{R}(X_{i,j})),$$

$$X_{i,j} = x_i + h \sum_{l=1}^{s} \alpha_{jl} f(X_{i,l}, \mathcal{R}(X_{i,l})),$$

and the values x_i are the numerical approximations obtained when we apply the given Runge–Kutta method to (5.45) in the autonomous form $\dot{x} = f(x, \mathcal{R}(x))$. In particular, we have

$$x(t_i) - x_i = \mathcal{O}(h^p).$$

For the other variables, (5.37) and (5.39) imply that

$$y_{i+1} = y_i + h(\beta^T \otimes I_s)\dot{Y}_i \tag{5.46}$$

and

$$Y_i = e \otimes y_i + h(\mathcal{A} \otimes I_s)\dot{Y}_i, \tag{5.47}$$

where

$$Y_i = \begin{bmatrix} Y_{i,1} \\ \vdots \\ Y_{i,s} \end{bmatrix}, \quad \dot{Y}_i = \begin{bmatrix} \dot{Y}_{i,1} \\ \vdots \\ \dot{Y}_{i,s} \end{bmatrix},$$

using a notation that is similar to that in the proof of Theorem 5.10. Solving in (5.47) for $\dot{Y}_i$ and eliminating it in (5.46) yields

$$\begin{aligned} y_{i+1} &= y_i + (\beta^T \otimes I_s)(\mathcal{A}^{-1} \otimes I_s)(Y_i - e \otimes y_i) \\ &= (1 - \beta^T \mathcal{A}^{-1} e) y_i + (\beta^T \mathcal{A}^{-1} \otimes I_s) Y_i. \end{aligned}$$

Let us first consider the case that $\varrho = 0$. We then obtain that

$$y_{i+1} = Y_{i,s} = \mathcal{R}(X_{i,s}) = \mathcal{R}(x_{i+1}).$$

In particular, it follows that $g(x_{i+1}, y_{i+1}) = 0$ such that direct and indirect approach coincide. With

$$\|y(t_i) - y_i\| = \|\mathcal{R}(x(t_i)) - \mathcal{R}(x_i)\| \leq L \|x(t_i) - x_i\|,$$

where L is the Lipschitz constant associated with $\mathcal{R}$, convergence for the y-part of the solution follows with the same order p as for the x-part.

In order to develop an error recursion for the y-part of the solution in the other cases, we observe that the conditions $B(p)$ and $C(q)$ of (5.22) require that the quadrature rules given by the coefficients $(\beta_j, \gamma_j)_{j=1,\ldots,s}$ for intervals $[t_i, t_{i+1}]$ and by the coefficients $(\alpha_{jl}, \gamma_l)_{l=1,\ldots,s}$ for intervals $[t_i, t_i + \gamma_j h]$, $j = 1, \ldots, s$, are of order p and q, respectively. Hence, if the Runge–Kutta method has order p then $B(p)$ must hold. From the conditions $B(p)$ and $C(q)$, we then obtain that

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + h \sum_{j=1}^{s} \beta_j \dot{y}(t_i + \gamma_j h) + \mathcal{O}(h^{p+1}), \\ y(t_i + \gamma_j h) &= y(t_i) + h \sum_{l=1}^{s} \alpha_{jl} \dot{y}(t_i + \gamma_l h) + \mathcal{O}(h^{q+1}), \quad j = 1, \ldots, s. \end{aligned} \tag{5.48}$$

Introducing the vectors

$$\tilde{Y}_i = \begin{bmatrix} y(t_i + \gamma_1 h) \\ \vdots \\ y(t_i + \gamma_s h) \end{bmatrix}, \quad \dot{\tilde{Y}}_i = \begin{bmatrix} \dot{y}(t_i + \gamma_1 h) \\ \vdots \\ \dot{y}(t_i + \gamma_s h) \end{bmatrix},$$

we can write (5.48) as

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + h(\beta^T \otimes I_s) \dot{\tilde{Y}}_i + \mathcal{O}(h^{p+1}), \\ \tilde{Y}_i &= e \otimes y(t_i) + h(\mathcal{A} \otimes I_s) \dot{\tilde{Y}}_i + \mathcal{O}(h^{q+1}). \end{aligned}$$

Solving the second relation for $\dot{\tilde{Y}}_i$ and eliminating it in the first relation gives

$$\begin{aligned} y(t_{i+1}) &= y(t_i) + (\beta^T \mathcal{A}^{-1} \otimes I_s)(\tilde{Y}_i - (e \otimes y(t_i)) + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}) \\ &= (1 - \beta^T \mathcal{A}^{-1} e) y(t_i) + (\beta^T \mathcal{A}^{-1} \otimes I_s)\tilde{Y}_i + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}). \end{aligned}$$

Introducing $\varepsilon_i = y(t_i) - y_i$, we have derived the error recursion

$$\varepsilon_{i+1} = \varrho \varepsilon_i + \delta_i, \tag{5.49}$$

where

$$\delta_i = (\beta^T \mathcal{A}^{-1} \otimes I_s)(\tilde{Y}_i - Y_i) + \mathcal{O}(h^{p+1}) + \mathcal{O}(h^{q+1}).$$

From

$$\begin{aligned} x(t_i + \gamma_j h) &= x(t_i) + h \sum_{l=1}^{s} \alpha_{jl} f(x(t_i + \gamma_l h), \mathcal{R}(x(t_i + \gamma_l h))) + \mathcal{O}(h^{q+1}), \\ X_{i,j} &= x_i + h \sum_{l=1}^{s} \alpha_{jl} f(X_{i,l}, \mathcal{R}(X_{i,l})), \end{aligned}$$

we obtain that

$$\begin{aligned} x(t_i + \gamma_j h) - X_{i,j} &= x(t_i) - x_i \\ &+ h \sum_{l=1}^{s} \alpha_{jl} \big(f(x(t_i + \gamma_l h), \mathcal{R}(x(t_i + \gamma_l h))) - f(X_{i,l}, \mathcal{R}(X_{i,l})) \big) + \mathcal{O}(h^{q+1}), \end{aligned}$$

which implies that

$$x(t_i + \gamma_j h) - X_{i,j} = \mathcal{O}(h^p) + \mathcal{O}(h^{q+1}).$$

Hence,

$$Y_{i,j} - y(t_i + \gamma_j h) = \mathcal{R}(X_{i,j}) - \mathcal{R}(x(t_i + \gamma_j h)) = \mathcal{O}(h^p) + \mathcal{O}(h^{q+1})$$

and thus

$$\delta_i = \mathcal{O}(h^p) + \mathcal{O}(h^{q+1}) = \mathcal{O}(h^k), \quad k = \min\{p, q+1\}.$$

Together with $\varepsilon_0 = 0$, the error recursion (5.49) gives

$$\varepsilon_N = \sum_{i=0}^{N-1} \varrho^{N-1-i} \delta_i.$$

If $|\varrho| < 1$, then we can estimate

$$\|\varepsilon_N\| \leq Ch^k \sum_{i=0}^{N-1} \varrho^{N-1-i} \leq Ch^k/(1-\varrho),$$

with a constant C that is independent of h. If $\varrho = 1$, then we obtain the estimate

$$\|\varepsilon_N\| \leq CNh^k = C(T-t_0)h^{k-1}.$$

If $\varrho = -1$, then we make use of the fact that x_i as a numerical solution of an ordinary differential equation allows for an asymptotic expansion of the form

$$x(t_i) - x_i = \xi_p(t_i)h^p + \mathcal{O}(h^{p+1})$$

with a smooth function ξ_p satisfying $\xi(t_0) = 0$, see, e.g., [106]. It follows that the stage values $X_{i,j}$, which depend smoothly on x_i, as well as $Y_{i,j} = \mathcal{R}(X_{i,j})$ and thus also δ_i possess asymptotic expansions. The latter then has the form

$$\delta_i = \zeta_k(t_i)h^k + \mathcal{O}(h^{k+1}),$$

with a smooth function ζ_k and $k = \min\{p, q+1\}$. Since

$$\delta_{i+1} - \delta_i = (\zeta_k(t_{i+1}) - \zeta_k(t_i))h^k + \mathcal{O}(h^{k+1}) = \dot{\zeta}_k(t_i)h^{k+1} + \mathcal{O}(h^{k+1}) = \mathcal{O}(h^{k+1}),$$

we obtain the estimate

$$\|\varepsilon_N\| \leq \|\delta_{N-1} - \delta_{N-2}\| + \cdots + \|\delta_1 - \delta_0\| \leq CNh^{k+1}/2 = C(T-t_0)h^k/2,$$

if N is even. A similar estimate also holds for N being odd.

Finally, if $|\varrho| > 1$, then the error recursion (5.49) shows that errors are amplified by a fixed factor. Hence, the method cannot be convergent. □

The most important part of Theorem 5.16 says that, if $\varrho = 0$, then the order of a Runge–Kutta method is not reduced when we apply it to semi-explicit differential-algebraic equations of index $\nu = 1$. Recall that a sufficient condition for $\varrho = 0$ is that the given method is stiffly accurate, i.e., that $\beta_j = \alpha_{sj}$ for $j = 1, \ldots, s$.

If we consider the general case of a regular, strangeness-free differential-algebraic equation (4.23), then it is no longer obvious that stiffly accurate Runge–Kutta methods behave in the same way, since in (4.23) we have to deal with the derivative $\dot{x}$, which in contrast to (5.35) includes all derivatives of the vector x of unknowns. Nevertheless, in the following we will show that for a special class of stiffly accurate Runge–Kutta methods a result similar to the corresponding part of Theorem 5.16 holds. It should also be noted that stiffly accurate Runge–Kutta methods applied to (4.23) automatically guarantee that the numerical approximations satisfy the algebraic constraints.

In order to solve regular strangeness-free differential-algebraic equations of the form

$$\hat{F}_1(t, x, \dot{x}) = 0, \quad \hat{F}_2(t, x) = 0, \tag{5.50}$$

it is convenient to work with so-called *collocation Runge–Kutta methods*. Starting with parameters γ_j, $j = 1, \ldots, s$, that satisfy

$$0 < \gamma_1 < \cdots < \gamma_s = 1, \tag{5.51}$$

and setting $\gamma_0 = 0$, we define the Lagrange interpolation polynomials

$$L_l(\xi) = \prod_{\substack{j=0 \\ j \neq l}}^{s} \frac{\xi - \gamma_j}{\gamma_l - \gamma_j}, \quad \tilde{L}_l(\xi) = \prod_{\substack{m=1 \\ m \neq l}}^{s} \frac{\xi - \gamma_m}{\gamma_l - \gamma_m} \tag{5.52}$$

and the coefficients

$$\alpha_{jl} = \int_0^{\gamma_j} \tilde{L}_l(\xi)\, d\xi, \quad \beta_j = \int_0^1 \tilde{L}_l(\xi)\, d\xi, \qquad j, l = 1, \ldots, s. \tag{5.53}$$

This fixes a Runge–Kutta method with $\beta_j = \alpha_{sj}$ for $j = 1, \ldots, s$. Hence the method is stiffly accurate.

If $\mathbb{P}_k$ denotes the space of polynomials of maximal degree $k - 1$, or, synonymously, of maximal order k, then the stage values $X_{i,l}$, $l = 1, \ldots, s$, together with $X_{i,0} = x_i$ define a polynomial $x_\pi \in \mathbb{P}_{s+1}$ via

$$x_\pi(t) = \sum_{l=0}^{s} X_{i,l} L_l\left(\frac{t - t_i}{h}\right). \tag{5.54}$$

The derivatives of x_π at the stages are then given by

$$\dot{X}_{i,j} = \dot{x}_\pi(t_i + \gamma_j h) = \frac{1}{h} \sum_{l=0}^{s} X_{i,l} \dot{L}_l(\gamma_j). \tag{5.55}$$

In order to fix the new approximation $x_{i+1} = x_\pi(t_{i+1}) = X_{i,s}$, we require that the polynomial x_π satisfies (5.50) at the so-called *collocation points* $t_{ij} = t_i + \gamma_j h$. This requirement leads to the nonlinear system

$$\hat{F}_1(t_i + \gamma_j h, X_{i,j}, \dot{X}_{i,j}) = 0, \quad \hat{F}_2(t_i + \gamma_j h, X_{i,j}) = 0. \tag{5.56}$$

Since $\tilde{L} \in \mathbb{P}_s$, the polynomials $P_l \in \mathbb{P}_{s+1}$ defined by

$$P_l(\sigma) = \int_0^{\sigma} \tilde{L}_l(\xi)\, d\xi$$

possess the representation

$$P_l(\sigma) = \sum_{j=0}^{s} P_l(\gamma_j) L_j(\sigma).$$

Differentiating P_l and using $\dot{P}_l = \tilde{L}_l$ then gives

$$\tilde{L}_l(\sigma) = \sum_{j=0}^{s} \left(\int_0^{\gamma_j} \tilde{L}_l(\xi)\, d\xi \right) \dot{L}_j(\sigma)$$

and thus

$$\delta_{lm} = \tilde{L}_l(\gamma_m) = \sum_{j=1}^{s} \alpha_{jl} v_{mj} = \begin{cases} 1 & \text{for } l = m, \\ 0 & \text{otherwise,} \end{cases} \tag{5.57}$$

with $v_{mj} = \dot{L}_j(\gamma_m)$. Hence, with $V = [v_{mj}]_{m,j=1,\ldots,s}$ we have $V = \mathcal{A}^{-1}$. Moreover, with $v_{m0} = \dot{L}_0(\gamma_m)$ and the fact that the polynomials L_j, $j = 0, \ldots, s$, sum up to one, we obtain that

$$\sum_{j=0}^{s} \dot{L}_j(\gamma_m) = \sum_{j=0}^{s} v_{mj} = 0,$$

such that

$$v_{m0} = -e_m^T V e,$$

where as before e_m is the m-th unit vector and e the vector of all ones of appropriate size.

With these relations, we can rewrite (5.55) as

$$h\dot{X}_m = v_{m0} x_i + \sum_{j=1}^{s} v_{mj} X_j.$$

Thus,

$$\begin{aligned} h \sum_{m=1}^{s} \alpha_{lm} \dot{X}_{i,m} &= \sum_{m=1}^{s} \alpha_{lm} v_{m0} x_i + \sum_{j,m=1}^{s} \alpha_{lm} v_{mj} X_{i,j} \\ &= -e_l^T \mathcal{A} V e x_i + \sum_{j=1}^{s} e_l^T \mathcal{A} V e_j X_{i,j} = -x_i + X_{i,l}. \end{aligned}$$

Comparing with (5.19), (5.17) and recalling that the coefficients α_{jl} and β_j fix a stiffly accurate Runge–Kutta method due to (5.53), we see that collocation just leads to a special class of Runge–Kutta methods.

For the analysis of consistency of the method, we now drop the index i for convenience and set $\hat{t}_j = t_{ij} = t_i + \gamma_j h$. Assuming that x^* is the unique solution of the given regular strangeness-free differential-algebraic equation, we must first show that (5.56) in the form

$$\hat{F}_1\Big(\hat{t}_j, X_j, \frac{1}{h}\sum_{l=0}^{s} v_{jl} X_l\Big) = 0, \quad \hat{F}_2(\hat{t}_j, X_j) = 0 \tag{5.58}$$

with $j = 1, \ldots, s$ is uniquely solvable for X_j, $j = 1, \ldots, s$. According to Section 4.1, the special properties of (5.50) near a solution allow us to assume that we can split x into (x_1, x_2) in such a way that $\hat{F}_2(t, x_1, x_2) = 0$ is locally solvable for $x_2 = \mathcal{R}(t, x_1)$. Note that this splitting is also applied to the numerical solutions x_i which are split accordingly as $(x_{i,1}, x_{i,2})$ so that there is no danger of mixing up the double meanings of x_1, x_2. With this splitting, we can make use of (4.25), which then follows from (5.50). Thus, splitting X_j accordingly into $(X_{j,1}, X_{j,2})$, the algebraic constraints imply

$$X_{j,2} = \mathcal{R}(\hat{t}_j, X_{j,1}) \tag{5.59}$$

such that we remain with the implicit difference equation

$$\hat{F}_1\Big(\hat{t}_j, X_{j,1}, \mathcal{R}(\hat{t}_j, X_{j,1}), \frac{1}{h}\sum_{l=0}^{s} v_{jl} X_{l,1}, \frac{1}{h}\sum_{l=0}^{s} v_{jl}\mathcal{R}(\hat{t}_l, X_{l,1})\Big) = 0. \tag{5.60}$$

It is therefore sufficient to show that (5.60) is solvable when we are sufficiently close to the solution x_1^* of the implicit ordinary differential equation

$$\hat{F}_1(t, x_1, \mathcal{R}(t, x_1), \dot{x}_1, \mathcal{R}_t(t, x_1) + \mathcal{R}_{x_1}(t, x_1)\dot{x}_1) = 0. \tag{5.61}$$

We will prove this by using Theorem 5.7 and interpreting the resulting method for (5.61) as a general discretization method. For this, we consider the linear problem

$$\hat{E}(t)\dot{y} - \hat{A}(t)y = \hat{E}(t)\dot{x}^*(t) - \hat{A}(t)x^*(t), \quad y(t_i) = x^*(t_i). \tag{5.62}$$

with $\hat{E}$, $\hat{A}$ as in (5.10). By construction, a solution is given by x^* and this solution is unique due to Lemma 5.6. Discretization of (5.62) according to (5.58) yields the linear system

$$\begin{aligned} \hat{E}_1(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl} X_l - \hat{A}_1(\hat{t}_j)X_j &= \hat{f}_1(\hat{t}_j), \\ -\hat{A}_2(\hat{t}_j)X_j &= \hat{f}_2(\hat{t}_j), \end{aligned} \tag{5.63}$$

with $\hat{f}(t) = \hat{E}(t)\dot{x}^*(t) - \hat{A}(t)x^*(t)$ split into $(\hat{f}_1(t), \hat{f}_2(t))$. Splitting the solution vector and its discretization as for (5.60), we obtain

$$\begin{aligned}\hat{E}_{11}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}X_{l,1} + \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}X_{l,2} \\ - \hat{A}_{11}(\hat{t}_j)X_{j,1} - \hat{A}_{12}(\hat{t}_j)X_{j,2} = \hat{f}_1(\hat{t}_j)\end{aligned} \tag{5.64}$$

and

$$-\hat{A}_{21}(\hat{t}_j)X_{j,1} - \hat{A}_{22}(\hat{t}_j)X_{j,2} = \hat{f}_2(\hat{t}_j), \tag{5.65}$$

with appropriately partitioned coefficient functions. With this partition and the assumption that the system is regular and strangeness-free, it follows that $\hat{A}_{22}$ is pointwise nonsingular and we can solve (5.65) for $X_{j,2}$. For convenience, we write

$$X_{j,2} = R(\hat{t}_j)X_{j,1} + g(\hat{t}_j) \tag{5.66}$$

with appropriately defined functions R and g. Elimination of $X_{j,2}$, $j = 1, \ldots, s$, in (5.64) yields a linear system for $X_{j,1}$, $j = 1, \ldots, s$, given by

$$\begin{aligned}\hat{E}_{11}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}X_{l,1} + \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}R(\hat{t}_l)X_{l,1} \\ - \hat{A}_{11}(\hat{t}_j)X_{j,1} - \hat{A}_{12}(\hat{t}_j)R(\hat{t}_j)X_{j,1} \\ = \hat{f}_1(\hat{t}_j) - \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}g(\hat{t}_l) + \hat{A}_{12}(\hat{t}_j)g(\hat{t}_j),\end{aligned} \tag{5.67}$$

with a coefficient matrix $B = [B_{jl}]_{j,l=1,\ldots,s}$ given by

$$B_{jl} = \frac{1}{h}(\hat{E}_{11}(\hat{t}_j) + \hat{E}_{12}(\hat{t}_j)R(\hat{t}_l))v_{jl} - \delta_{jl}(\hat{A}_{11}(\hat{t}_j) + \hat{A}_{12}(\hat{t}_j)R(\hat{t}_l)). \tag{5.68}$$

By assumption, the matrix valued function $\hat{E}_{11} + \hat{E}_{12}R$ is pointwise nonsingular. Since $V = [v_{jl}]_{j,l=1,\ldots,s}$ is nonsingular as well, it follows that the matrix B is nonsingular for sufficiently small h. In particular,

$$\|B\| \le Ch^{-1}, \quad \|B^{-1}\| \le Ch, \tag{5.69}$$

with some constant C independent of h. Therefore, the numerical solution for (5.62) is well defined at least for sufficiently small stepsizes h. In order to investigate the local error in more detail, we also split the exact solution x^* into (x_1^*, x_2^*) and get

$$\begin{aligned}\hat{E}_{11}(t)\dot{x}_1^*(t) + \hat{E}_{12}(t)\dot{x}_2^*(t) &= \hat{A}_{11}(t)x_1^*(t) + \hat{A}_{12}(t)x_2^*(t) + \hat{f}_1(t), \\ x_2^*(t) &= R(t)x_1^*(t) + g(t).\end{aligned} \tag{5.70}$$

We then write x_1^* in the form

$$x_1^*(t) = \sum_{l=0}^{s} x_1^*(\hat{t}_l) L_l\left(\frac{t-t_i}{h}\right) + \psi(t), \tag{5.71}$$

where the interpolation error ψ, given by

$$\psi(t) = \frac{x^{*(s+1)}(\theta(t))}{(s+1)!} \prod_{j=0}^{s} (t - \hat{t}_j), \quad \theta(t) \in (t_i, t_{i+1}), \tag{5.72}$$

is as smooth as x_1^*. In particular, we have that $\psi(t) = \mathcal{O}(h^{s+1})$. Differentiation of x_1^* then leads to

$$\dot{x}_1^*(t) = \frac{1}{h} \sum_{l=0}^{s} x_1^*(\hat{t}_l) \dot{L}_l\left(\frac{t-t_i}{h}\right) + \dot{\psi}(t), \tag{5.73}$$

with $\dot{\psi}(t) = \mathcal{O}(h^s)$. In the same way, we get that

$$R(t)x_1^*(t) + g(t) = \sum_{l=0}^{s} (R(\hat{t}_l)x_1^*(\hat{t}_l) + g(\hat{t}_l)) L_l\left(\frac{t-t_i}{h}\right) + \phi(t) \tag{5.74}$$

and

$$R(t)\dot{x}_1^*(t) + \dot{R}(t)x_1^*(t) + \dot{g}(t) = \frac{1}{h} \sum_{l=0}^{s} (R(\hat{t}_l)x_1^*(\hat{t}_l) + g(\hat{t}_l)) \dot{L}_l\left(\frac{t-t_i}{h}\right) + \dot{\phi}(t), \tag{5.75}$$

with $\phi(t) = \mathcal{O}(h^{s+1})$ and $\dot{\phi}(t) = \mathcal{O}(h^s)$. Eliminating x_2^* in the first relation of (5.70) with the help of the second relation and using the just derived representations, we obtain

$$\begin{aligned} \hat{E}_{11}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}x_1^*(\hat{t}_l) + \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}R(\hat{t}_l)x_1^*(\hat{t}_l) \\ - \hat{A}_{11}(\hat{t}_j)x_1^*(\hat{t}_j) - \hat{A}_{12}(\hat{t}_j)R(\hat{t}_j)x_1^*(\hat{t}_j) \\ = \hat{f}_1(\hat{t}_j) - \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}g(\hat{t}_l) + \hat{A}_{12}(\hat{t}_j)g(\hat{t}_j) + \mathcal{O}(h^s) \end{aligned} \tag{5.76}$$

for $j = 1, \ldots, s$. Hence,

$$B \begin{bmatrix} x_1^*(\hat{t}_1) - X_{1,1} \\ \vdots \\ x_1^*(\hat{t}_s) - X_{s,1} \end{bmatrix} = \mathcal{O}(h^s) \tag{5.77}$$

or, with (5.69),

$$\begin{bmatrix} x_1^*(\hat{t}_1) - X_{1,1} \\ \vdots \\ x_1^*(\hat{t}_s) - X_{s,1} \end{bmatrix} = \mathcal{O}(h^{s+1}). \tag{5.78}$$

In order to construct a solution of (5.60), we apply the Newton-like method (5.13). To start the iteration, we use the just obtained solution $X_{j,1}$, $j = 1, \ldots, s$, of (5.67) and set $X_{j,1}^0 = X_{j,1}$, $j = 1, \ldots, s$. The iteration (5.13) for (5.60) then reads

$$\begin{aligned} &\hat{E}_{11}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}(X_{l,1}^{m+1} - X_{l,1}^m) + \hat{E}_{12}(\hat{t}_j)\frac{1}{h}\sum_{l=0}^{s} v_{jl}R(\hat{t}_l)(X_{l,1}^{m+1} - X_{l,1}^m) \\ &\qquad - \hat{A}_{11}(\hat{t}_j)(X_{j,1}^{m+1} - X_{j,1}^m) - \hat{A}_{12}(\hat{t}_j)R(\hat{t}_j)(X_{j,1}^{m+1} - X_{j,1}^m) \\ &= -\hat{F}_1\Big(\hat{t}_j, X_{j,1}^m, \mathcal{R}(\hat{t}_j, X_{j,1}^m), \frac{1}{h}\sum_{l=0}^{s} v_{jl}X_{l,1}^m, \frac{1}{h}\sum_{l=0}^{s} v_{jl}\mathcal{R}(\hat{t}_l, X_{l,1}^m)\Big) \end{aligned} \tag{5.79}$$

for $j = 1, \ldots, s$, when we use the convention that $X_{0,1}^m = x_1^*(t_i)$. Observing that

$$\begin{aligned} &\hat{F}_1\Big(\hat{t}_j, X_{j,1}^0, \mathcal{R}(\hat{t}_j, X_{j,1}^0), \frac{1}{h}\sum_{l=0}^{s} v_{jl}X_{l,1}^0, \frac{1}{h}\sum_{l=0}^{s} v_{jl}\mathcal{R}(\hat{t}_l, X_{l,1}^0)\Big) \\ &\quad = \hat{F}_1\Big(\hat{t}_j, x_1^*(\hat{t}_j) + \mathcal{O}(h^{s+1}), x_2^*(\hat{t}_j) + \mathcal{O}(h^{s+1}), \\ &\qquad \frac{1}{h}\sum_{l=0}^{s} v_{jl}x_1^*(\hat{t}_l) + \mathcal{O}(h^s), \frac{1}{h}\sum_{l=0}^{s} v_{jl}x_1^*(\hat{t}_l) + \mathcal{O}(h^s)\Big) = \mathcal{O}(h^s) \end{aligned}$$

due to (5.78), we have that

$$B\begin{bmatrix} X_{1,1}^1 - X_{1,1}^0 \\ \vdots \\ X_{s,1}^1 - X_{s,1}^0 \end{bmatrix} = \mathcal{O}(h^s). \tag{5.80}$$

By (5.69), it follows that

$$\begin{bmatrix} X_{1,1}^1 - X_{1,1}^0 \\ \vdots \\ X_{s,1}^1 - X_{s,1}^0 \end{bmatrix} = \mathcal{O}(h^{s+1}). \tag{5.81}$$

Hence, the assumptions of Theorem 5.7 are satisfied with

$$\left\| \begin{bmatrix} X_{1,1}^1 - X_{1,1}^0 \\ \vdots \\ X_{s,1}^1 - X_{s,1}^0 \end{bmatrix} \right\| \le \alpha = Ch^{s+1}, \quad \|B^{-1}\| \le \beta = Ch, \quad \gamma = Ch^{-1}$$

for a suitable constant C independent of h. We now choose h so small that

$$\left\| \begin{bmatrix} x_1^*(\hat{t}_1) - X_{1,1}^0 \\ \vdots \\ x_1^*(\hat{t}_s) - X_{s,1}^0 \end{bmatrix} \right\| \leq \frac{1}{2\beta\gamma} = \frac{1}{2C^2}$$

and that

$$\left\| \begin{bmatrix} X_{1,1}^1 - X_{1,1}^0 \\ \vdots \\ X_{s,1}^1 - X_{s,1}^0 \end{bmatrix} \right\| \leq \frac{1}{9\beta\gamma} = \frac{1}{9C^2}.$$

Furthermore, we choose h so small that (5.12f) holds. Then, Theorem 5.7 implies that the quantities $X_{j,1}^m$ converge to locally unique quantities $X_{j,1}^*$, $j = 1, \ldots, s$. Moreover, by Corollary 5.8, it follows that

$$\begin{aligned} \left\| \begin{bmatrix} x_1^*(\hat{t}_1) - X_{1,1}^* \\ \vdots \\ x_1^*(\hat{t}_s) - X_{s,1}^* \end{bmatrix} \right\| &\leq \left\| \begin{bmatrix} x_1^*(\hat{t}_1) - X_{1,1}^0 \\ \vdots \\ x_1^*(\hat{t}_s) - X_{s,1}^0 \end{bmatrix} \right\| + \left\| \begin{bmatrix} X_{1,1}^* - X_{1,1}^0 \\ \vdots \\ X_{s,1}^* - X_{s,1}^0 \end{bmatrix} \right\| \\ &\leq Ch^{s+1} + 4Ch^{s+1} = 5Ch^{s+1}. \end{aligned} \tag{5.82}$$

This shows that (5.60) defines a numerical method for the solution of (5.45) that is consistent of order s.

In order to show stability, we proceed as follows. We have already shown that the nonlinear system consisting of

$$\begin{aligned} \hat{F}_1\Big(\hat{t}_j, X_{j,1}, \mathcal{R}(\hat{t}_j, X_{j,1}), \frac{1}{h} v_{j0} x_{1,i} + \sum_{l=1}^{s} v_{jl} X_{l,1}, \\ \frac{1}{h} v_{j0} \mathcal{R}(t_i, x_{1,i}) + \frac{1}{h} \sum_{l=1}^{s} v_{jl} \mathcal{R}(\hat{t}_l, X_{l,1})\Big) = 0, \end{aligned} \tag{5.83}$$

with $j = 1, \ldots, s$, is uniquely solvable in a neighborhood of $x_1^*(t_i)$ for the stage values $X_{j,1}$. Thus, we can see these $X_{j,1}$ as functions of $x_{1,i}$ according to

$$X_{j,1} = \mathcal{G}_j(x_{1,i}), \quad j = 1, \ldots, s, \tag{5.84}$$

where, by construction, the functions $\mathcal{G}_j$, $j = 1, \ldots, s$, satisfy

$$\begin{aligned} \hat{F}_1\Big(\hat{t}_j, \mathcal{G}_j(x_{1,i}), \mathcal{R}(\hat{t}_j, \mathcal{G}_j(x_{1,i})), \frac{1}{h} v_{j0} x_{1,i} + \sum_{l=1}^{s} v_{jl} \mathcal{G}_l(x_{1,i}), \\ \frac{1}{h} v_{j0} \mathcal{R}(t_i, x_{1,i}) + \frac{1}{h} \sum_{l=1}^{s} v_{jl} \mathcal{R}(\hat{t}_l, \mathcal{G}_l(x_{1,i}))\Big) \equiv 0. \end{aligned}$$

Differentiation with respect to $x_{1,i}$ yields (omitting arguments)

$$\begin{aligned}\hat{F}_{1;x_1}\mathcal{G}_{j;x_1} + \hat{F}_{1;x_2}\mathcal{R}_{x_1}\mathcal{G}_{j;x_1} + \hat{F}_{1;\dot{x}_1}\frac{1}{h}v_{j0} + \hat{F}_{1;\dot{x}_1}\frac{1}{h}\sum_{l=1}^{s} v_{jl}\mathcal{G}_{l;x_1} \\ + \hat{F}_{1;\dot{x}_2}\frac{1}{h}v_{j0}\mathcal{R}_{x_1} + \hat{F}_{1;\dot{x}_2}\frac{1}{h}\sum_{l=1}^{s} v_{jl}\mathcal{R}_{x_1}\mathcal{G}_{l;x_1} = 0,\end{aligned} \tag{5.85}$$

representing a linear system for the derivatives $\mathcal{G}_{j;x_1}$, $j = 1, \ldots, s$. Similar as for the matrix B from (5.68), it follows that the coefficient matrix is $\mathcal{O}(h^{-1})$ and invertible for sufficiently small h. Since the right hand side is $\mathcal{O}(h^{-1})$ as well, it follows that the derivatives $\mathcal{G}_{j;x_1}$ are $\mathcal{O}(1)$. Multiplying (5.85) with h, we then get

$$\hat{F}_{1;\dot{x}_1}\Big(v_{j0}I + \sum_{l=1}^{s} v_{jl}\mathcal{G}_{l;x_1}\Big) + \hat{F}_{1;\dot{x}_2}\Big(v_{j0}\mathcal{R}_{x_1} + \sum_{l=1}^{s} v_{jl}\mathcal{R}_{x_1}\mathcal{G}_{l;x_1}\Big) = \mathcal{O}(h).$$

Since the arguments of $\mathcal{R}_{x_1}$ only differ by terms of $\mathcal{O}(h)$, all arguments can be shifted to a common argument by Taylor expansion, yielding

$$(\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_2}\mathcal{R}_{x_1})\Big(v_{j0}I + \sum_{l=1}^{s} v_{jl}\mathcal{G}_{l;x_1}\Big) = \mathcal{O}(h),$$

with a nonsingular coefficient matrix $\hat{F}_{1;\dot{x}_1} + \hat{F}_{1;\dot{x}_2}\mathcal{R}_{x_1}$. Hence,

$$v_{j0}I + \sum_{l=1}^{s} v_{jl}\mathcal{G}_{l;x_1} = \mathcal{O}(h).$$

If we define the block column matrix $\mathcal{G}_{x_1}$ to consist of the blocks $\mathcal{G}_{j;x_1}$, $j = 1, \ldots, s$, we can write this relation as

$$(V \otimes I)\mathcal{G}_{x_1} = Ve \otimes I + \mathcal{O}(h).$$

This immediately gives $\mathcal{G}_{x_1} = e \otimes I + \mathcal{O}(h)$. In particular, $\mathcal{G}_{s;x_1} = I + \mathcal{O}(h)$ and $\mathcal{G}_{s;x_1}$ is Lipschitz continuous. Moreover, we can choose the norm such that the Lipschitz constant becomes $L = 1 + hK$. Using

$$x_{1,i+1} = \mathcal{G}_s(x_{1,i}), \quad x_1^*(t_{i+1}) = \mathcal{G}_s(x_1^*(t_i)) + \mathcal{O}(h^{s+1}),$$

the latter expressing the consistency of the method, we find that

$$\begin{aligned}\|x_1^*(t_{i+1}) - x_{1,i+1}\| &= \|\mathcal{G}_s(x_1^*(t_i)) - \mathcal{G}_s(x_{1,i}) + \mathcal{O}(h^{s+1})\| \\ &\le (1 + hK)\|x_1^*(t_i) - x_{1,i}\| + Ch^{s+1} \\ &\le 1 + h\tilde{K},\end{aligned} \tag{5.86}$$

which shows the stability of (5.60). Together with $x_{2,i+1} = \mathcal{R}(t_{i+1}, x_{1,i+1})$, we have shown that the resulting method (5.58) for the numerical solution of (5.50) is consistent and stable in the sense of Theorem 5.4, if we set $\mathfrak{X}_i = x_i = (x_{1,i}, x_{2,i})$ and $\mathfrak{X}(t_i) = x^*(t_i)$. Thus, we have the following theorem.

Theorem 5.17. *The collocation Runge–Kutta methods defined by* (5.58) *and* $x_{i+1} = X_{i,s}$ *with collocation points as in* (5.51) *are convergent of order* $p = s$.

A special class of Runge–Kutta methods that are covered by Theorem 5.17 are the already mentioned Radau IIA methods defined by $B(2s-1)$, $C(s)$, and $D(s-1)$ of (5.22) together with $\gamma_s = 1$. The order of these methods is given by $p = 2s-1$. That the order is higher than suggested by Theorem 5.17 is due to the special choice of the nodes. This effect is called *superconvergence*. Because of $B(2s-1)$, the underlying quadrature rule of the Radau IIA methods is required to be exact for all polynomials $\phi \in \mathbb{P}_{2s-1}$. Hence, we have that

$$\int_{t_i}^{t_{i+1}} \psi(r) \prod_{j=1}^{s} (r - \hat{t}_j)\, dr = 0 \quad \text{for all } \psi \in \mathbb{P}_{s-1}. \tag{5.87}$$

In order to show superconvergence for the Radau IIA methods, when applied to regular strangeness-free differential-algebraic equations, we first recall that the given solution x^* solves (5.62). The corresponding collocation solution is given by the stage values $X_1^0, \ldots, X_s^0$. Defining $x_\pi^0 \in \mathbb{P}_{s+1}$ by

$$x_\pi^0(t) = \sum_{l=0}^{s} X_l^0 L_l\left(\frac{t - \hat{t}_l}{h}\right), \tag{5.88}$$

where $X_0^0 = x^*(t_i)$, the polynomial x_π^0 solves the system

$$\hat{E}(t)\dot{y} - \hat{A}(t)y = \hat{E}(t)\dot{x}_\pi^0(t) - \hat{A}(t)x_\pi^0(t), \quad y(t_i) = x^*(t_i). \tag{5.89}$$

Transformation of (5.62) to the canonical form (5.11) yields

$$\begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \tfrac{d}{dt}(Q^{-1}y) = \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix} (Q^{-1}y) = P\hat{f}(t).$$

With

$$Q^{-1}y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad P\hat{f} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix} = \begin{bmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{bmatrix} \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix} = \begin{bmatrix} [\, I_d \;\; 0\,] P\hat{f} \\ P_{22}\hat{f}_2 \end{bmatrix},$$

it follows that

$$\dot{y}_1 = g_1(t), \quad 0 = y_2 + g_2(t).$$

Hence, setting

$$\begin{bmatrix} y_1^* \\ y_2^* \end{bmatrix} = Q^{-1}x^*, \quad \begin{bmatrix} y_{\pi,1}^0 \\ y_{\pi,2}^0 \end{bmatrix} = Q^{-1}x_\pi^0,$$

we get that

$$\begin{aligned} y_1^*(t) &= y_1^*(t_i) + \int_{t_i}^{t} [\, I_d \;\; 0 \,] P(r) \hat{f}(r)\, dr, \\ y_2^*(t) &= -P_{22}(t)\hat{f}_2(t), \end{aligned}$$

and, accordingly,

$$\begin{aligned} y_{\pi,1}^0(t) &= y_{\pi,1}^0(t_i) + \int_{t_i}^{t} [\, I_d \;\; 0 \,] P(r)(\hat{E}(r)\dot{x}_\pi^0(r) - \hat{A}(r)x_\pi^0(r))\, dr, \\ y_{\pi,2}^0(t) &= -P_{22}(t)(-\hat{A}_2(t)x_\pi^0(t)). \end{aligned}$$

Together, these relations imply that

$$x^*(t_{i+1}) - x_\pi^0(t_{i+1}) = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi(r)\, dr \\ 0 \end{bmatrix},$$

with

$$\phi(r) = [\, I_d \;\; 0 \,] P(r)(\hat{f}(r) + \hat{A}(r)x_\pi^0(r) - \hat{E}(r)\dot{x}_\pi^0(r)),$$

where we have used that $A_2(t_{i+1})X_j^0 + \hat{f}_2(t_{i+1}) = 0$ due to the construction of the numerical method. Since $X_1^0, \cdots, X_s^0$ represent the solution of (5.62), we find that

$$\phi(\hat{t}_j) = 0, \quad j = 1, \ldots, s,$$

and we can write ϕ in the form

$$\phi(r) = \omega(r) \prod_{j=1}^{s} (r - \hat{t}_j)$$

with a smooth function ω. Taylor expansion yields $\omega = \psi + \mathcal{O}(h^{s-1})$ with a polynomial $\psi \in \mathbb{P}_{s-1}$. Thus, the special choice of the nodes according to (5.87) leads to

$$\int_{t_i}^{t_{i+1}} \phi(r)\, dr = \int_{t_i}^{t_{i+1}} (\psi(r) \prod_{j=1}^{s} (r - \hat{t}_j) + \mathcal{O}(h^{2s-1}))\, dr = \mathcal{O}(h^{2s}).$$

We therefore end up with

$$x^*(t_{i+1}) - x_\pi^0(t_{i+1}) = \mathcal{O}(h^{2s}). \tag{5.90}$$

In particular, we observe consistency of order $p = 2s - 1$ at least in the linear case. In the following, we now transfer this result to the nonlinear case.

Denoting the stage values in the solution of (5.58) by $X_1^*, \ldots, X_s^*$ and defining $x_\pi^* \in \mathbb{P}_{s+1}$ by

$$x_\pi^*(t) = \sum_{l=0}^{s} X_l^* L_l\left(\frac{t - \hat{t}_l}{h}\right), \tag{5.91}$$

where $X_0^* = x^*(t_i)$, the derived consistency estimate, which also holds for the part of the variables associated with x_2 in the numerical solution, gives

$$x^*(\hat{t}_j) - x_\pi^*(\hat{t}_j) = \mathcal{O}(h^{s+1}), \quad \dot{x}^*(\hat{t}_j) - \dot{x}_\pi^*(\hat{t}_j) = \mathcal{O}(h^s).$$

Hence, using $\hat{F}(\hat{t}_j, x^*(\hat{t}_j), \dot{x}^*(\hat{t}_j)) = 0$, we obtain that

$$\begin{aligned} 0 &= \hat{F}(\hat{t}_j, x_\pi^*(\hat{t}_j), \dot{x}_\pi^*(\hat{t}_j)) \\ &= \hat{F}(\hat{t}_j, x^*(\hat{t}_j) + (x_\pi^*(\hat{t}_j) - x^*(\hat{t}_j)), \dot{x}^*(\hat{t}_j) + (\dot{x}_\pi^*(\hat{t}_j) - \dot{x}^*(\hat{t}_j))) \\ &= \hat{E}(\hat{t}_j)(\dot{x}_\pi^*(\hat{t}_j) - \dot{x}^*(\hat{t}_j)) - \hat{A}(\hat{t}_j)(x_\pi^*(\hat{t}_j) - x^*(\hat{t}_j)) + \mathcal{O}(h^{2s}). \end{aligned}$$

Together with

$$\hat{E}(\hat{t}_j)(\dot{x}_\pi^0(\hat{t}_j) - \dot{x}^*(\hat{t}_j)) - \hat{A}(\hat{t}_j)(x_\pi^0(\hat{t}_j) - x^*(\hat{t}_j)) = 0,$$

we get that

$$\hat{E}(\hat{t}_j)(\dot{x}_\pi^*(\hat{t}_j) - \dot{x}_\pi^0(\hat{t}_j)) - \hat{A}(\hat{t}_j)(x_\pi^*(\hat{t}_j) - x_\pi^0(\hat{t}_j)) = \mathcal{O}(h^{2s}).$$

Since

$$x_\pi^*(\hat{t}_j) - x_\pi^0(\hat{t}_j) = X_j^* - X_j^0, \quad \dot{x}_\pi^*(\hat{t}_j) - \dot{x}_\pi^0(\hat{t}_j) = \sum_{l=0}^{s} v_{jl}(X_l^* - X_l^0),$$

the first part of the quantities $X_j^* - X_j^0$, $j = 1, \ldots, s$, associated with x_1, solves a linear system with the coefficient matrix B defined in (5.68), implying that this part is $\mathcal{O}(h^{2s+1})$. The second part, associated with x_2, form the solution of a linear system with a coefficient matrix of $\mathcal{O}(1)$ and right hand side of $\mathcal{O}(h^{2s})$. Therefore, we get that $X_j^* - X_j^0 = \mathcal{O}(h^{2s})$, $j = 1, \ldots, s$. In particular,

$$x^*(t_{i+1}) - x_\pi^*(t_{i+1}) + x_\pi^0(t_{i+1}) - x^*(t_{i+1}) = \mathcal{O}(h^{2s}),$$

and (5.90) yields

$$x^*(t_{i+1}) - x_\pi^*(t_{i+1}) = \mathcal{O}(h^{2s}). \tag{5.92}$$

Since we have already shown stability of the methods, the consistency of order $p = 2s - 1$ implies convergence with the same order. Hence, we have the following theorem.

Theorem 5.18. *Choosing the nodes γ_j, $j = 1, \ldots, s$, in (5.51) such that (5.87) holds, the corresponding collocation Runge–Kutta methods defined by (5.58) and $x_{i+1} = X_{i,s}$ are convergent of order $p = 2s - 1$.*

One can show that the methods that are covered by Theorem 5.18 are just the Radau IIA methods that we have introduced in the beginning of this section, see [108]. In this way, we have also verified the claimed order $p = 2s - 1$.

Remark 5.19. All presented convergence results are based on the assumption that we use a constant stepsize. For real-life applications, however, a sophisticated strategy for the adaption of the stepsize during the numerical solution of the problem is indispensable. In the case of semi-explicit differential-algebraic equations of index $\nu = 1$ or regular strangeness-free problems (5.50), it is possible to use the same techniques as in the case of ordinary differential equations, see, e.g., [106]. However, when discretizing higher index systems, changes of the stepsize may lead to undesired effects, see Exercise 18.

In this section, we have only discussed a selected class of one-step methods that are applicable for differential-algebraic systems. Further classes of one-step methods include extrapolation methods, Rosenbrock methods, and methods that utilize further structure of the given problem, see [108] and the bibliographical remarks of this chapter.

5.3 Multi-step methods

It is the idea of multi-step methods to use several of the previous approximations $x_{i-1}, \ldots, x_{i-k}$ for the computation of the approximation x_i to the solution value $x(t_i)$. Given real coefficients α_l and β_l for $l = 0, \ldots, k$, a *linear multi-step method* for the numerical solution of an ordinary differential equation $\dot{x} = f(t, x)$ is given by

$$\sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = h \sum_{l=0}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}). \tag{5.93}$$

In order to fix x_i at least for sufficiently small stepsizes h, we must require that $\alpha_k \neq 0$. In addition, we assume that $\alpha_0^2 + \beta_0^2 \neq 0$ and speak of a *k-step method*. Of course, we must provide $x_0, \ldots, x_{k-1}$ to initialize the iteration. These are usually generated via appropriate one-step methods or within a combined order and stepsize control.

Since we can multiply (5.93) by any nonzero scalar without changing the numerical method, we are allowed to assume that either $\alpha_k = 1$ or, in the case of

implicit methods where $\beta_k \neq 0$, that $\beta_k = 1$. For the moment, we suppose that $\alpha_k = 1$.

The coefficients α_l and β_l define the so-called *characteristic polynomials*

$$\varrho(\lambda) = \sum_{l=0}^{k} \alpha_l \lambda^l, \quad \sigma(\lambda) = \sum_{l=0}^{k} \beta_l \lambda^l \tag{5.94}$$

of the multi-step method.

Setting

$$\mathfrak{X}_i = \begin{bmatrix} x_{i+k-1} \\ x_{i+k-2} \\ \vdots \\ x_i \end{bmatrix}, \quad \mathfrak{X}(t_i) = \begin{bmatrix} x(t_{i+k-1}) \\ x(t_{i+k-2}) \\ \vdots \\ x(t_i) \end{bmatrix},$$

we can treat multi-step methods in the context of general discretization methods as discussed in Section 5.1. The multi-step (5.93) is called *consistent of order* p if

$$\sum_{l=0}^{k} \alpha_{k-l} x(t_{i-l}) - h \sum_{l=0}^{k} \beta_{k-l} \dot{x}(t_{i-l}) = \mathcal{O}(h^{p+1}) \tag{5.95}$$

for all sufficiently smooth functions x with the constant involved in $\mathcal{O}(h^{p+1})$ being independent of h. By the implicit function theorem, we can solve (5.93) for x_i in the form

$$x_i = \mathcal{F}(t_i, x_{i-1}, \ldots, x_{i-k}; h)$$

with

$$\begin{aligned} &\mathcal{F}(t_i, x_{i-1}, \ldots, x_{i-k}; h) + \sum_{l=1}^{k} \alpha_{k-l} x_{i-l} \\ &\quad - h\beta_k f(t_i, \mathcal{F}(t_i, x_{i-1}, \ldots, x_{i-k}; h)) - h \sum_{l=1}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}) \equiv 0. \end{aligned} \tag{5.96}$$

If x solves the given ordinary differential equation, then we have that

$$\begin{aligned} &\mathcal{F}(t_i, x(t_{i-1}), \ldots, x(t_{i-k}); h) + \sum_{l=1}^{k} \alpha_{k-l} x(t_{i-l}) \\ &\quad - h\beta_k f(t_i, \mathcal{F}(t_i, x(t_{i-1}), \ldots, x(t_{i-k}); h)) - h \sum_{l=1}^{k} \beta_{k-l} f(t_{i-l}, x(t_{i-l})) = 0. \end{aligned}$$

Subtracting this from (5.95) gives

$$\begin{aligned} &x(t_i) - \mathcal{F}(t_i, x(t_{i-1}), \ldots, x(t_{i-k}); h) \\ &\quad = h\beta_k (f(t_i, x(t_i)) - f(t_i, \mathcal{F}(t_i, x(t_{i-1}), \ldots, x(t_{i-k}); h))) + \mathcal{O}(h^{p+1}) \end{aligned}$$

or, for sufficiently small h,

$$\|x(t_i) - \mathcal{S}(t_i, x(t_{i-1}), \ldots, x(t_{i-k}); h)\| \leq \frac{C}{1 - h|\beta_k|L} h^{p+1},$$

where L is the Lipschitz constant of f with respect to the second argument. Introducing

$$\mathcal{F}(t_i, \mathcal{X}_i; h) = \begin{bmatrix} \mathcal{S}(t_{i+k}, x_{i+k-1}, \ldots, x_i; h) \\ x_{i+k-1} \\ \vdots \\ x_{i+1} \end{bmatrix},$$

this estimate immediately leads to (5.3).

A multi-step method (5.93) is called *stable* if there exists a vector norm such that in the associated matrix norm

$$\|\mathcal{C}_\alpha \otimes I_n\| \leq 1, \tag{5.97}$$

where

$$\mathcal{C}_\alpha = \begin{bmatrix} -\alpha_{k-1} & \cdots & -\alpha_1 & -\alpha_0 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}$$

is the companion matrix of the characteristic polynomial ϱ given by the coefficients $\alpha_0, \ldots, \alpha_{k-1}, \alpha_k = 1$. In view of (5.4), we must investigate

$$\begin{aligned} &\mathcal{F}(t_i, \mathcal{X}(t_i); h) - \mathcal{F}(t_i, \mathcal{X}_i; h) \\ &\quad = \begin{bmatrix} \mathcal{S}(t_{i+k}, x(t_{i+k-1}), \ldots, x(t_i); h) - \mathcal{S}(t_{i+k}, x_{i+k-1}, \ldots, x_i; h) \\ x(t_{i+k-1}) - x_{i+k-1} \\ \vdots \\ x(t_{i+1}) - x_{i+1} \end{bmatrix}. \end{aligned}$$

Inserting (5.96) in the first block, we get

$$\begin{aligned} &\mathcal{S}(t_{i+k}, x(t_{i+k-1}), \ldots, x(t_i); h) - \mathcal{S}(t_{i+k}, x_{i+k-1}, \ldots, x_i; h) \\ &\quad = -\sum_{l=1}^{k} \alpha_{k-l}(x(t_{i+k-l}) - x_{i+k-l}) \\ &\qquad + h \sum_{l=1}^{k} \beta_{k-l}(f(t_{i+k-l}, x(t_{i+k-l})) - f(t_{i+k-l}, x_{i+k-l})) \\ &\qquad + h\beta_k(f(t_{i+k}, \mathcal{S}(t_{i+k}, x(t_{i+k-1}), \ldots, x(t_i); h)) \\ &\qquad - f(t_{i+k}, \mathcal{S}(t_{i+k}, x_{i+k-1}, \ldots, x_i; h))). \end{aligned}$$

It then follows that

$$\begin{aligned}\|\mathfrak{F}(t_i, \mathfrak{X}(t_i); h) - \mathfrak{F}(t_i, \mathfrak{X}_i; h)\| &\le \frac{1}{1 - h|\beta_k|L}(\|\mathcal{C}_\alpha \otimes I_n\| + h\tilde{K})\|\mathfrak{X}(t_i) - \mathfrak{X}_i\| \\ &\le (1 + hK)\|\mathfrak{X}(t_i) - \mathfrak{X}_i\|,\end{aligned}$$

when we use the specific norms associated with the stability (5.97). Hence, (5.4) holds and Theorem 5.4 yields convergence of multi-step methods that are consistent and stable.

To check whether a multi-step methods is consistent and stable in terms of the coefficients α_l and β_l, $l = 1, \dots, k$, we have the following results.

Theorem 5.20. *If the coefficients α_l and β_l, $l = 1, \dots, k$, of the multi-step method (5.93) satisfy the conditions*

$$\sum_{l=0}^{k} \alpha_l l^q = q \sum_{l=0}^{k} \beta_l l^{q-1}, \quad q = 0, \dots, p, \tag{5.98}$$

with the convention that the right hand side vanishes for $q = 0$ and that $0^0 = 1$, then the method is consistent of order p.

Proof. Taylor expansion of the left hand side of (5.95) yields

$$\begin{aligned}&\sum_{l=0}^{k}(\alpha_l x(t_l) - h\beta\dot{x}(t_l)) \\ &\quad= \sum_{l=0}^{k}\Big[\alpha_l \sum_{q=0}^{p} \frac{(t_l - t_0)^q}{q\,!}x^{(q)}(t_0) - h\beta_l \sum_{q=0}^{p-1} \frac{(t_l - t_0)^q}{q\,!}x^{(q+1)}(t_0)\Big] + \mathcal{O}(h^{p+1}) \\ &\quad= \sum_{l=0}^{k}\Big[\alpha_l \sum_{q=0}^{p} \frac{l^q h^q}{q\,!}x^{(q)}(t_0) - h\beta_l \sum_{q=1}^{p} \frac{l^{q-1}h^{q-1}}{(q-1)\,!}x^{(q)}(t_0)\Big] + \mathcal{O}(h^{p+1}) \\ &\quad= \sum_{q=0}^{p}\Big[\sum_{l=0}^{k} \alpha_l l^q - q \sum_{l=0}^{k} \beta_l l^{q-1}\Big]\frac{h^q}{q\,!}x^{(q)}(t_0) + \mathcal{O}(h^{p+1}).\end{aligned}$$

□

In order to get consistency of order at least $p = 1$, we must therefore require the characteristic polynomials (5.94) to satisfy

$$\varrho(1) = 0, \quad \dot{\varrho}(1) = \sigma(1). \tag{5.99}$$

For stability of the multi-step method, we have the following result.

Theorem 5.21. *Suppose that the characteristic polynomial ϱ of the multi-step method (5.93) satisfies the so-called root condition given by:*

1. *The roots of ϱ lie in the closed unit disk.*

2. *The roots of ϱ with modulus one are simple.*

Then the multi-step method is stable.

Proof. With the normalization $\alpha_k = 1$, the matrix $\mathcal{C}_\alpha$ from (5.97) is the companion matrix to the characteristic polynomial ϱ. In particular, the eigenvalues $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$ of $\mathcal{C}_\alpha$ are just the roots of ϱ. Due to the root condition, we may assume that

$$1 = |\lambda_1| = \cdots = |\lambda_m| > |\lambda_{m+1}| \geq \cdots \geq |\lambda_k|,$$

where the roots $\lambda_1, \ldots, \lambda_m$ are simple. Setting $\varepsilon = 1 - |\lambda_{m+1}|$, it follows that there exists a nonsingular matrix $T \in \mathbb{C}^{k,k}$ such that $\mathcal{C}_\alpha$ can be transformed to Jordan canonical form given by

$$J = T^{-1}\mathcal{C}_\alpha T = \left[\begin{array}{ccc|ccccc} \lambda_1 & & & & & & & \\ & \ddots & & & & & & \\ & & \lambda_m & & & & & \\ \hline & & & \lambda_{m+1} & \varepsilon_{m+1} & & & \\ & & & & \ddots & \ddots & & \\ & & & & & \ddots & \varepsilon_{k-1} & \\ & & & & & & \lambda_k & \end{array}\right],$$

with $|\varepsilon_j| \leq \varepsilon$ for $j = m+1, \ldots, k-1$. Hence, $\|J\|_\infty \leq 1$, and the relation

$$\|\mathfrak{X}\| = \|(T^{-1} \otimes I_n)\mathfrak{X}\|_\infty$$

defines a vector norm whose associated matrix norm satisfies

$$\|\mathcal{C}_\alpha \otimes I_n\| = \|(T^{-1} \otimes I_n)(\mathcal{C}_\alpha \otimes I_n)(T \otimes I_n)\|_\infty = \|J \otimes I_n\|_\infty \leq 1. \qquad \square$$

The general convergence result of Theorem 5.4 then yields the following sufficient conditions for a multi-step method to be convergent.

Theorem 5.22. *Suppose that the coefficients α_l and β_l, $l = 1, \ldots, k$, of the multi-step method (5.93) satisfy (5.98) and that the characteristic polynomial ϱ satisfies the root condition of Theorem 5.21. Then the multi-step method is convergent of order p.*

In the context of differential-algebraic equations, the most popular linear multi-step methods are the so-called *BDF methods*. The abbreviation BDF stands for *backward differentiation formulae*. These methods are obtained by setting

$$\beta_0 = \cdots = \beta_{k-1} = 0, \quad \beta_k = 1, \tag{5.100}$$

and choosing α_l, $l = 1, \ldots, k$, to satisfy (5.98) with p as large as possible. Since (5.98) constitutes a linear system for the coefficients α_l, with a Vandermonde matrix as coefficient matrix, we can achieve $p = k$. See Table 5.3 for the resulting

Table 5.3. The simplest BDF methods

α_{k-l}	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
$k=1$	1	-1					
$k=2$	$\frac{3}{2}$	-2	$\frac{1}{2}$				
$k=3$	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$			
$k=4$	$\frac{25}{12}$	-4	3	$-\frac{4}{3}$	$\frac{1}{4}$		
$k=5$	$\frac{137}{60}$	-5	5	$-\frac{10}{3}$	$\frac{5}{4}$	$-\frac{1}{5}$	
$k=6$	$\frac{147}{60}$	-6	$\frac{15}{2}$	$-\frac{20}{3}$	$\frac{15}{4}$	$-\frac{6}{5}$	$\frac{1}{6}$

coefficients α_l, $l = 1, \ldots, k$, of the simplest BDF methods. Recalling (5.95), the BDF methods satisfy

$$\dot{x}(t_k) - \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x(t_{i-l}) = \mathcal{O}(h^k). \tag{5.101}$$

Thus, the BDF methods can be interpreted as discretization methods for $\dot{x} = f(t, x)$ in the following way. We evaluate the ordinary differential equation at t_k obtaining $\dot{x}(t_k) = f(t_k, x(t_k))$ and simply replace $\dot{x}(t_k)$ by the backward differentiation formula given in (5.101) and $x(t_l)$ by x_l, $l = 0, \ldots, k$. Obviously, this kind of discretization can also be used in the same way for differential-algebraic equations. However, since the BDF methods are constructed only with a large consistency order in mind, we must still check their stability. Unfortunately, this leads to a restriction of the possible order.

Theorem 5.23. *The BDF methods are stable only for* $1 \leq k \leq 6$.

Proof. A nice proof can be found in the paper [107]. □

In the following, we discuss the properties of multi-step methods when employed for differential-algebraic equations. We will mainly consider the BDF methods,

which take the form

$$F(t_i, x_i, D_h x_i) = 0, \quad D_h x_i = \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l}. \tag{5.102}$$

Again, (5.102) only presents a reasonable numerical method if we can show that it defines a unique x_i in the vicinity of x_{i-1} at least for sufficiently small stepsize h.

As in Section 5.2, we begin the analysis of the convergence properties with the case of linear differential-algebraic equations with constant coefficients $E\dot{x} = Ax + f(t)$. Again, we assume that the pair (E, A) is regular such that the solution is unique. We then obtain the following convergence result.

Theorem 5.24. *Let (E, A) be regular with $\nu = \operatorname{ind}(E, A)$. Then the BDF methods (5.102) with $1 \le k \le 6$, applied to the system $E\dot{x} = Ax + f(t)$, $x(t_0) = x_0$, are convergent of order $p = k$.*

Proof. If we apply (5.102) to $E\dot{x} = Ax + f(t)$, then we obtain

$$E \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = A x_i + f(t_i). \tag{5.103}$$

Since (E, A) is regular, we may assume for the analysis that the pair (E, A) is in Weierstraß canonical form (2.7) such that the system decouples into two parts. The first part represents the BDF method applied to an ordinary differential equation for which Theorem 5.22 applies. We may therefore restrict our analysis without loss of generality to systems of the form

$$N \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = x_i + f(t_i). \tag{5.104}$$

To analyze this system, we again take the formal approach of Lemma 2.8 using the operator D_h of (5.102). Equation (5.104) then becomes

$$N D_h x_i = x_i + f(t_i).$$

Since D_h is a linear operator and commutes with N, we obtain

$$x_i = -(I - N D_h)^{-1} f(t_i) = -\sum_{j=0}^{\nu-1} N^j D_h^j f(t_i).$$

Using (5.101), we get the formal power series

$$D_h f(t_i) = \dot{f}(t_i) + \sum_{q \ge k} c_q \tfrac{h^q}{q!} f^{(q)}(t_i) = \dot{f}(t_i) + \mathcal{O}(h^k)$$

Applying D_h once more, we obtain

$$\begin{aligned} D_h^2 f(t_i) &= D_h \dot{f}(t_i) + \sum_{q \geq k} c_q \frac{h^q}{q!} D_h f^{(q)}(t_i) \\ &= \ddot{f}(t_i) + \sum_{q \geq k} \tilde{c}_q \frac{h^q}{q!} f^{(q+1)}(t_i) = \ddot{f}(t_i) + \mathcal{O}(h^k). \end{aligned}$$

Proceeding inductively, this gives

$$D_h^j f(t_i) - f^{(j)}(t_i) = \mathcal{O}(h^k), \quad j = 0, \ldots, \nu - 1$$

such that

$$x(t_i) - x_i = -\sum_{j=0}^{\nu-1} N^j (f^{(j)}(t_i) - D_h^j f(t_i)) = \mathcal{O}(h^k).$$

Hence, we have convergence of the BDF method of order $p = k$. □

Remark 5.25. If a BDF method with constant stepsize h is applied to a regular linear differential-algebraic equation $E\dot{x} = Ax + f(t)$ with constant coefficients and constant inhomogeneity and if $\nu = \mathrm{ind}(E, A) \geq 2$, then the numerical approximations x_i are consistent for $i \geq \nu - 1$, independent of the (possibly inconsistent) starting values.

As for Runge–Kutta methods, the situation becomes more complex in the case of linear differential-algebraic systems with variable coefficients. If we discretize (3.1) by a BDF method, then we obtain

$$E(t_i) \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = A(t_i) x_i + f(t_i),$$

or, equivalently, by collecting the terms in x_i,

$$(\alpha_k E(t_i) - hA(t_i)) x_i = hf(t_i) - E(t_i) \sum_{l=1}^{k} \alpha_{k-l} x_{i-l}. \tag{5.105}$$

It follows that a unique numerical solution x_i exists if and only if the matrix $\alpha_k E(t_i) - hA(t_i)$ is invertible. This is only possible if the matrix pair $(E(t_i), A(t_i))$ is regular. Thus, in order to guarantee that BDF methods can be used, we would need to assume that $(E(t), A(t))$ is regular for all $t \in \mathbb{I}$. But we have already seen in Chapter 3 that the pointwise regularity of (E, A) is not invariant under equivalence transformations. Moreover, it is completely independent from the solvability properties of the differential-algebraic equation.

If we apply for example a BDF method (5.105) to Example 3.1, then, since the pair $(E(t), A(t))$ is regular for all $t \in \mathbb{I}$, the BDF method determines a unique numerical solution, whereas the given differential-algebraic equation is not uniquely solvable at all. On the other hand, if we apply a BDF method (5.105) to Example 3.2, then, since the pair $(E(t), A(t))$ is singular for all $t \in \mathbb{I}$, the linear system in (5.105) is either not solvable or does not have a unique solution, whereas the given differential-algebraic equation is uniquely solvable.

These considerations lead to the conclusion that, as for one-step methods, we need to restrict the class of problems in such a way that BDF methods and other multi-step methods can be applied. It is clear that we must assume that the initial value problem for the differential-algebraic equation itself is uniquely solvable.

Following the lines of Section 5.2, we first consider the class of semi-explicit differential-algebraic equations of index $\nu = 1$ given by (5.35). Again, we may proceed in two ways to generalize multi-step methods for the solution of (5.35). In the *direct approach*, we apply (5.93) to (5.36). Linear multi-step methods then take the form

$$\begin{aligned} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} &= h \sum_{l=0}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}, y_{i-l}), \\ \varepsilon \sum_{l=0}^{k} \alpha_{k-l} y_{i-l} &= h \sum_{l=0}^{k} \beta_{k-l} g(t_{i-l}, x_{i-l}, y_{i-l}). \end{aligned} \tag{5.106}$$

Setting $\varepsilon = 0$ yields a method of the form

$$\begin{aligned} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} &= h \sum_{l=0}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}, y_{i-l}), \\ 0 &= h \sum_{l=0}^{k} \beta_{k-l} g(t_{i-l}, x_{i-l}, y_{i-l}). \end{aligned} \tag{5.107}$$

In the *indirect approach*, we simply replace the second part of (5.107) by $g(t_i, x_i, y_i) = 0$. Using the solvability of $g(t, x, y) = 0$ with respect to y, we have that $y_i = \mathcal{R}(t_i, x_i)$. If this also holds for $i = 0, \ldots, k-1$, i.e., if all the starting values are consistent, then the indirect approach is equivalent to the solution of the ordinary differential equation (5.45) by the given multi-step method and setting $y_i = \mathcal{R}(t_i, x_i)$. It is then clear that we obtain the same convergence properties as for ordinary differential equations. In particular, we may even apply explicit multi-step methods, i.e., methods of the form (5.93) with $\beta_k = 0$. Thus, for the analysis, it remains to look at the direct approach.

Theorem 5.26. *Consider an implicit multi-step method* (5.93) *of order* p *applied to* (5.35) *as in* (5.107). *Suppose that* ϱ *as well as* σ *satisfy the root condition. Then the method is convergent of order* p.

Proof. Since the multi-step method (5.93) is required to be implicit, we may assume that the coefficients are scaled so that $\beta_k = 1$. It then follows that there exists a vector norm such that in the associated matrix norm

$$\|\mathcal{C}_\beta \otimes I_n\| \leq 1, \tag{5.108}$$

with the companion matrix

$$\mathcal{C}_\beta = \begin{bmatrix} -\beta_{k-1} & \cdots & -\beta_1 & -\beta_0 \\ 1 & & & \\ & \ddots & & \\ & & 1 & \end{bmatrix}.$$

The second equation of (5.107) yields

$$g(t_i, x_i, y_i) = -\sum_{l=1}^{k} \beta_{k-l} g(t_{i-l}, x_{i-l}, y_{i-l}).$$

Setting $\eta_i = g(t_i, x_i, y_i)$ and

$$\mathcal{Y}_i = \begin{bmatrix} \eta_{i+k-1} \\ \vdots \\ \eta_i \end{bmatrix},$$

this can be written as

$$\mathcal{Y}_{i+1} = (\mathcal{C}_\beta \otimes I_n)\mathcal{Y}_i.$$

Using the vector norm selected in (5.108) and the bound (5.6) for the initial values, it follows that

$$\|\mathcal{Y}_{i+1}\| \leq \|\mathcal{Y}_i\| \leq \cdots \leq \|\mathcal{Y}_0\| \leq \tilde{C}h^p.$$

Hence, we can solve $\eta_i = g(t_i, x_i, y_i)$ for y_i whenever h is sufficiently small, and we get that

$$y_i = \mathcal{R}(t_i, x_i) + \mathcal{O}(h^p).$$

Inserting this into the first equation of (5.107), we obtain

$$\sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = h \sum_{l=0}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}, \mathcal{R}(t_{i-l}, x_{i-l})) + \mathcal{O}(h^{p+1}).$$

But this is a perturbed multi-step method applied to the ordinary differential equation (5.45). Comparing with (5.93), we see that it is stable and consistent of order p. Hence, it is convergent of order p by Theorem 5.4. From $x(t_i) - x_i = \mathcal{O}(h^p)$, we then get that

$$y(t_i) - y_i = \mathcal{R}(t_i, x(t_i)) - \mathcal{R}(t_i, x_i) + \mathcal{O}(h^p) = \mathcal{O}(h^p). \qquad \square$$

As a more general case, we discuss the application of BDF methods to regular strangeness-free differential-algebraic equations of the form (5.50), i.e., we consider the numerical methods given by

$$\hat{F}_1(t_i, x_i, D_h x_i) = 0, \quad \hat{F}_2(t_i, x_i) = 0. \tag{5.109}$$

If the starting values $x_0, \ldots, x_{k-1}$ satisfy the algebraic constraints, then we immediately see that all numerical approximations x_i satisfy the algebraic constraints as well. We can therefore split x_i into $(x_{i,1}, x_{i,2})$ and use

$$x_{i,2} = \mathcal{R}(t_i, x_{i,1}) \tag{5.110}$$

to eliminate $x_{i,2}$ in the first equation of (5.109). Inserting the definition of D_h, we obtain the nonlinear system

$$\hat{F}_1\Big(t_i, x_{i,1}, \mathcal{R}(t_i, x_{i,1}), \frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} x_{i-l,1}, \frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} \mathcal{R}(t_{i-l}, x_{i-l,1})\Big) = 0 \tag{5.111}$$

as discretization of the ordinary differential equation (5.61).

To show that this is a reasonable numerical method, we must first show that (5.111) uniquely determines the numerical approximation $x_{i,1}$ when we are close to the actual solution $x_1^*(t_i)$. To see this, we again discretize the linear problem (5.62) with the exact solution x^* by (5.109) to obtain

$$\hat{E}(t_i)\frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} x_{i-l} - \hat{A}(t_i) x_i = \hat{E}(t_i)\dot{x}^*(t_i) - \hat{A}(t_i) x^*(t_i). \tag{5.112}$$

Collecting the terms that include x_i gives the linear system

$$(\alpha_k \hat{E}(t_i) - h\hat{A}(t_i)) x_i = h(\hat{E}(t_i)\dot{x}^*(t_i) - \hat{A}(t_i)x^*(t_i)) - \hat{E}(t_i)\sum_{l=1}^{k} \alpha_{k-l} x_{i-l}. \tag{5.113}$$

Splitting (5.113) as (5.63) in the previous section and using the same notation, we get

$$\begin{aligned} &(\alpha_k \hat{E}_{11}(t_i) - h\hat{A}_{11}(t_i)) x_{i,1} + (\alpha_k \hat{E}_{12}(t_i) - h\hat{A}_{12}(t_i)) x_{i,2} \\ &\qquad = h\hat{f}_1(t_i) - \hat{E}_1(t_i)\sum_{l=1}^{k} \alpha_{k-l} x_{i-l} \end{aligned} \tag{5.114}$$

and

$$-h\hat{A}_{21}(t_i)x_{i,1} - h\hat{A}_{22}(t_i)x_{i,2} = h\hat{f}_2(t_i). \tag{5.115}$$

Similar to (5.66), we then have that

$$x_{i,2} = R(t_i)x_{i,1} + g(t_i). \tag{5.116}$$

Due to (5.6), we assume that $x_{i-l} = x^*(t_{i-l}) + \mathcal{O}(h^k)$, $l = 1, \ldots, k$. Using (5.101) for x^*, this leads to

$$\begin{aligned}\sum_{l=1}^{k} \alpha_{k-l} x_{i-l} &= \sum_{l=1}^{k} \alpha_{k-l} x^*_{i-l} + \mathcal{O}(h^k) \\ &= \sum_{l=0}^{k} \alpha_{k-l} x^*_{i-l} - \alpha_k x^*(t_i) + \mathcal{O}(h^k) \\ &= h\dot{x}^*(t_i) - \alpha_k x^*(t_i) + \mathcal{O}(h^k).\end{aligned}$$

Inserting this into (5.114) and eliminating $x_{i,2}$ with the help of (5.116) then gives

$$\begin{aligned}&(\alpha_k(\hat{E}_{11}(t_i) + \hat{E}_{12}(t_i)R(t_i)) - h(\hat{A}_{11}(t_i) + \hat{A}_{12}(t_i)R(t_i)))x_{i,1} \\ &\quad = h\hat{f}_1(t_i) - (\alpha_k\hat{E}_{12}(t_i) - h\hat{A}_{12}(t_i))g(t_i) - \hat{E}_1(t_i)\sum_{l=1}^{k}\alpha_{k-l}x_{i-l} \\ &\quad = h(\hat{E}_1(t_i)\dot{x}^*(t_i) - \hat{A}_1(t_i)x^*(t_i)) - (\alpha_k\hat{E}_{12}(t_i) - h\hat{A}_{12}(t_i))g(t_i) \\ &\qquad - \hat{E}_1(t_i)(h\dot{x}^*(t_i) - \alpha_k x^*(t_i)) + \mathcal{O}(h^k) \\ &\quad = (\alpha_k\hat{E}_{11}(t_i) - h\hat{A}_{11}(t_i))x_1^*(t_i) + (\alpha_k\hat{E}_{12}(t_i) - h\hat{A}_{12}(t_i))x_2^*(t_i) \\ &\qquad - (\alpha_k\hat{E}_{12}(t_i) - h\hat{A}_{12}(t_i))(x_2^*(t_i) - R(t_i)x_1^*(t_i)) + \mathcal{O}(h^k) \\ &\quad = (\alpha_k(\hat{E}_{11}(t_i) + \hat{E}_{12}(t_i)R(t_i)) - h(\hat{A}_{11}(t_i) + \hat{A}_{12}(t_i)R(t_i)))x_1^*(t_i) + \mathcal{O}(h^k).\end{aligned}$$

Since the coefficient matrix $\alpha_k(\hat{E}_{11}(t_i) + \hat{E}_{12}(t_i)R(t_i)) - h(\hat{A}_{11}(t_i) + \hat{A}_{12}(t_i)R(t_i))$ is bounded and boundedly invertible, we end up with $x_1^*(t_i) - x_{i,1} = \mathcal{O}(h^k)$ and hence

$$x^*(t_i) - x_i = \mathcal{O}(h^k). \tag{5.117}$$

In order to show that the nonlinear system (5.111) fixes a numerical solution $x_{i,1}$ at least for sufficiently small h, we apply the Newton-like method (5.13), which

here takes the form

$$\begin{aligned}\Big(\frac{\alpha}{h}(\hat{E}_{11}(t_i)+\hat{E}_{12}(t_i)R(t_i))-(\hat{A}_{11}(t_i)+\hat{A}_{12}(t_i)R(t_i))\Big)(x_{i,1}^{m+1}-x_{i,1}^{m})\\ =-\hat{F}_1\Big(t_i,x_{i,1}^{m},\mathcal{R}(t_i,x_{i,1}^{m}),\frac{1}{h}(\alpha_k x_{i,1}^{m}+\sum_{l=1}^{k}\alpha_{k-l}x_{i-l,1}),\\ \frac{1}{h}\Big(\alpha_k\mathcal{R}(t_i,x_{i,1}^{m})+\sum_{l=1}^{k}\alpha_{k-l}\mathcal{R}(t_{i-l},x_{i-l,1})\Big)\Big).\end{aligned}\tag{5.118}$$

The iteration is started with the first part of the just constructed solution of (5.112), which is therefore denoted by $x_{i,1}^{0}$. Then (5.117) implies that

$$x_1^*(t_i)-x_{i,1}^{0}=\mathcal{O}(h^k).\tag{5.119}$$

For the first correction in the Newton-like iteration, we then compute

$$\begin{aligned}&(\alpha(\hat{E}_{11}(t_i)+\hat{E}_{12}(t_i)R(t_i))-h(\hat{A}_{11}(t_i)+\hat{A}_{12}(t_i)R(t_i)))(x_{i,1}^{1}-x_{i,1}^{0})\\ &\quad=-h\hat{F}_1\Big(t_i,x_1^*(t_i)+\mathcal{O}(h^k),\mathcal{R}(t_i,x_1^*(t_i))+\mathcal{O}(h^k),\\ &\qquad\frac{1}{h}\sum_{l=0}^{k}\alpha_{k-l}x_1^*(t_{i-l})+\mathcal{O}(h^{k-1}),\frac{1}{h}\sum_{l=0}^{k}\alpha_{k-l}\mathcal{R}(t_{i-l},x_1^*(t_{i-l}))+\mathcal{O}(h^{k-1})\Big)\\ &\quad=-h\hat{F}_1\big(t_i,x_1^*(t_i),\mathcal{R}(t_i,x_1^*(t_i)),\\ &\qquad\dot{x}_1^*(t_i),\mathcal{R}_t(t_i,x_1^*(t_i))+\mathcal{R}_{x_1}(t_i,x_1^*(t_i))\dot{x}_1^*(t_i)\big)+\mathcal{O}(h^k)\\ &\quad=\mathcal{O}(h^k),\end{aligned}$$

since x_1^* solves (5.61). Thus, the assumptions of Theorem 5.7 are satisfied with

$$\|x_{i,1}^{1}-x_{i,1}^{0}\|\le\alpha=Ch^k,\quad\beta=Ch,\quad\gamma=Ch^{-1},$$

with a suitable constant C independent of h. The same analysis as in Section 5.2 shows that the Newton-like process generates iterates $x_{i,1}^{m}$ that converge to a solution $x_{i,1}^*$ of (5.111) with $x_1^*(t_i)-x_{i,1}^*=\mathcal{O}(h^k)$. Setting $x_{i,2}^*=\mathcal{R}(t_i,x_{i,1}^*)$ then yields a solution x_i^* of (5.109) satisfying

$$x^*(t_i)-x_i^*=\mathcal{O}(h^k).\tag{5.120}$$

In particular, we have shown that (5.111) locally defines a numerical solution $x_{i,1}^*$, provided that the iterates $x_{i-l,1}$, $l=1,\dots,k$, are close to the solution. Hence, writing (5.111) in a simplified form as

$$\tilde{F}_1(t_i,x_{i,1},x_{i-1,1},\dots,x_{i-k,1};h)=0,\tag{5.121}$$

this equation is locally solved by

$$x_{i,1} = \mathcal{I}(t_i, x_{i-1,1}, \ldots, x_{i-k,1}; h), \tag{5.122}$$

dropping the superscript $*$. By definition, the function $\mathcal{I}$ then satisfies

$$\tilde{F}_1(t_i, \mathcal{I}(t_i, x_{i-1,1}, \ldots, x_{i-k,1}; h), x_{i-1,1}, \ldots, x_{i-k,1}; h) \equiv 0. \tag{5.123}$$

Our task now is to show that (5.122) constitutes a convergent discretization method for the determination of the solution x_1^* of (5.111). To do so, we define

$$\mathfrak{X}_i = \begin{bmatrix} x_{i+k-1,1} \\ x_{i+k-2,1} \\ \vdots \\ x_{i,1} \end{bmatrix}, \quad \mathfrak{X}(t_i) = \begin{bmatrix} x_1^*(t_{i+k-1}) \\ x_1^*(t_{i+k-2}) \\ \vdots \\ x_1^*(t_i) \end{bmatrix},$$

together with

$$\mathfrak{F}(t_i, \mathfrak{X}_i; h) = \begin{bmatrix} \mathcal{I}(t_{i+k}, x_{i+k-1,1}, \ldots, x_{i,1}; h) \\ x_{i+k-1,1} \\ \vdots \\ x_{i+1,1} \end{bmatrix},$$

and proceed as in the beginning of this section.

For consistency, we must study $\mathfrak{X}(t_{i+1}) - \mathfrak{F}(t_i, \mathfrak{X}(t_i); h)$. Obviously, only the first block is relevant. We therefore consider

$$x_1^*(t_i) - \mathcal{I}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h).$$

Starting from

$$\begin{aligned} &\hat{F}_1\Big(t_i, x_1^*(t_i), \mathcal{R}(t_i, x_1^*(t_i)), \frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} x_1^*(t_{i-l}), \frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} \mathcal{R}(t_{i-l}, x_1^*(t_{i-l}))\Big) \\ &\quad = \hat{F}_1(t_i, x_1^*(t_i), \mathcal{R}(t_i, x_1^*(t_i)), \dot{x}_1^*(t_i) + \mathcal{O}(h^k), \\ &\qquad \mathcal{R}_t(t_i, x_1^*(t_i)) + \mathcal{R}_{x_1}(t_i, x_1^*(t_i))\dot{x}_1^*(t_i) + \mathcal{O}(h^k)) = \mathcal{O}(h^k), \end{aligned}$$

we consider (5.121) in the slightly more general form

$$\tilde{F}_1(t_i, x_{i,1}, x_{i-1,1}, \ldots, x_{i-k,1}; h) = \varepsilon.$$

For sufficiently small ε, this relation is still locally solvable for $x_{i,1}$ according to

$$x_{i,1} = \tilde{\mathcal{I}}(t_i, x_{i-1,1}, \ldots, x_{i-k,1}; h, \varepsilon)$$

such that

$$\tilde{F}_1(t_i, \tilde{\mathcal{G}}(t_i, x_{i-1,1}, \ldots, x_{i-k,1}; h, \varepsilon), x_{i-1,1}, \ldots, x_{i-k,1}; h, \varepsilon) \equiv 0. \tag{5.124}$$

Hence, we can write

$$x_1^*(t_i) = \tilde{\mathcal{G}}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h, \varepsilon),$$

with $\varepsilon = \mathcal{O}(h^k)$. It follows that

$$\begin{aligned} x_1^*(t_i) &- \mathcal{G}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h) \\ &= \tilde{\mathcal{G}}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h, \varepsilon) - \tilde{\mathcal{G}}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h, 0). \end{aligned}$$

In order to get a bound for this, we need a bound for the derivative $\tilde{\mathcal{G}}_\varepsilon$ of $\tilde{\mathcal{G}}$ with respect to ε. Differentiating (5.124) and using (5.111), we get (omitting arguments)

$$(\hat{F}_{1;x_1} + \hat{F}_{1;x_2}\mathcal{R}_{x_1} + \tfrac{\alpha_k}{h}\hat{F}_{1;\dot{x}_1} + \tfrac{\alpha_k}{h}\hat{F}_{1;\dot{x}_2}\mathcal{R}_{x_1})\tilde{\mathcal{G}}_\varepsilon = I.$$

Hence, $\tilde{\mathcal{G}}_\varepsilon = \mathcal{O}(h)$ and $\tilde{\mathcal{G}}$ is Lipschitz continuous with respect to ε with a Lipschitz constant $L_\varepsilon = \mathcal{O}(h)$. With this, we can estimate

$$\|x_1^*(t_i) - \mathcal{G}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h)\| \le L_\varepsilon \varepsilon = \mathcal{O}(h)\,\mathcal{O}(h^k) = \mathcal{O}(h^{k+1}),$$

showing that the discretization method is consistent of order k.

For stability, we must study $\mathfrak{F}(t_i, \mathfrak{X}(t_i); h) - \mathfrak{F}(t_i, \mathfrak{X}_i; h)$. Looking again only at the first block, we must consider

$$\mathcal{G}(t_i, x_1^*(t_{i-1}), \ldots, x_1^*(t_{i-k}); h) - \mathcal{G}(t_i, x_{i-1,1}, \ldots, x_{i-k,1}; h).$$

In this case, we need the derivatives $\mathcal{G}_{x_{i-l,1}}$ of $\mathcal{G}$ with respect to $x_{i-l,1}$ for $l = 1, \ldots, k$. Differentiating (5.123) and using (5.111), we get (omitting arguments)

$$(\hat{F}_{1;x_1} + \hat{F}_{1;x_2}\mathcal{R}_{x_1} + \tfrac{\alpha_k}{h}\hat{F}_{1;\dot{x}_1} + \tfrac{\alpha_k}{h}\hat{F}_{1;\dot{x}_2}\mathcal{R}_{x_1})\mathcal{G}_{x_{i-l,1}} + \tfrac{\alpha_{k-l}}{h}\hat{F}_{1;\dot{x}_1} + \tfrac{\alpha_{k-l}}{h}\hat{F}_{1;\dot{x}_2}\mathcal{R}_{x_1} = 0.$$

Hence, assuming again that the coefficients are scaled so that $\alpha_k = 1$, we have

$$\mathcal{G}_{x_{i-l,1}} = -\alpha_{k-l} I_n + \mathcal{O}(h).$$

In particular, we find

$$\begin{aligned}
&\mathcal{S}(t_i, x_1^*(t_{i-1}), \dots, x_1^*(t_{i-k}); h) - \mathcal{S}(t_i, x_{i-1,1}, \dots, x_{i-k,1}; h) \\
&\quad = \mathcal{S}(t_i, x_{i-1,1} + s(x_1^*(t_{i-1}) - x_{i-1,1}), \dots, \\
&\qquad\qquad x_{i-k,1} + s(x_1^*(t_{i-k}) - x_{i-k,1}); h)|_{s=0}^{s=1} \\
&\quad = \int_0^1 \sum_{l=1}^{k} \mathcal{S}_{x_{i-l,1}}(t_i, x_{i-1,1} + s(x_1^*(t_{i-1}) - x_{i-1,1}), \dots, \\
&\qquad\qquad x_{i-k,1} + s(x_1^*(t_{i-k}) - x_{i-k,1}); h)(x_1^*(t_{i-l}) - x_{i-l,1})\, ds \\
&\quad = \sum_{l=1}^{k} (-\alpha_{k-l} I_n + \mathcal{O}(h))(x_1^*(t_{i-l}) - x_{i-l,1})
\end{aligned}$$

such that

$$\begin{aligned}
&\mathcal{F}(t_i, \mathcal{X}(t_i); h) - \mathcal{F}(t_i, \mathcal{X}_i; h) \\
&\qquad = \begin{bmatrix} \sum_{l=1}^{k} (-\alpha_{k-l} I_n + \mathcal{O}(h))(x_1^*(t_{i+k-l}) - x_{i+k-l,1}) \\ x_1^*(t_{i+k-1}) - x_{i+k-1,1} \\ \vdots \\ x_1^*(t_{i+1}) - x_{i+1,1} \end{bmatrix}.
\end{aligned}$$

This then leads to the estimate

$$\|\mathcal{F}(t_i, \mathcal{X}(t_i); h) - \mathcal{F}(t_i, \mathcal{X}_i; h)\| \le (\|\mathcal{C}_\alpha \otimes I_n\| + Kh)\|\mathcal{X}(t_i) - \mathcal{X}_i\|.$$

Thus, the discretization method is stable if the underlying BDF method is stable.

Theorem 5.4 then yields convergence of order k of the discretization method and hence of (5.111) to the first part x_1^* of the solution $x^* = (x_1^*, x_2^*)$ of (5.50). Convergence of the second part associated with x_2^* follows, since $x_{i,2} = \mathcal{R}(t_i, x_{i,1})$ together with $x_2^*(t_i) = \mathcal{R}(t_i, x_1^*(t_i))$ gives

$$\|x_2^*(t_i) - x_{i,2}\| = \|\mathcal{R}(t_i, x_1^*(t_i)) - \mathcal{R}(t_i, x_{i,1})\| \le L\|x_1^*(t_i) - x_{i,1}\| = \mathcal{O}(h^k)$$

for all $i = 0, \dots, N$, where L denotes the Lipschitz constant of $\mathcal{R}$ with respect to the second argument. Thus, we have proved the following result.

Theorem 5.27. *The BDF discretization (5.109) of (5.50) is convergent of order $p = k$ for $1 \le k \le 6$ provided that the initial values $x_0, \dots, x_{k-1}$ are consistent.*

Remark 5.28. As for Runge–Kutta methods, there are no difficulties to supply BDF methods with a stepsize control if we restrict their application to semi-explicit differential-algebraic equations of index $\nu = 1$ or to regular strangeness-free problems (5.50). Of course, we can also combine a stepsize control with an order control, as it is typical in the context of multi-step methods. See [29] for more details. Compare also with Remark 5.19 and Exercise 18.

Bibliographical remarks

The analysis of Runge–Kutta methods for ordinary differential equations is a classical theme of numerical analysis. It can be found in most textbooks on this topic, see, e.g., [11], [35], [38], [72], [106], [108], [210]. The direct application of implicit Runge–Kutta methods to semi-explicit differential-algebraic equations was studied in [36], [73], [100], [165], [194]. The existing results are summarized in [105], [108]. Other one-step methods such as Rosenbrock methods or extrapolation methods for ordinary differential equations and their extensions to differential-algebraic equations are discussed in [38], [72], [106], [108], [210].

Multi-step methods for differential-algebraic equations were first studied by Gear [89]. Based on this work, several implementations of BDF methods were presented in [28], [30], [164], [195], see also [29], [108]. Other multi-step methods for semi-explicit or strangeness-free differential-algebraic equations are described for example in [114], [115], [140], [204] and for multibody systems in [79], [181].

Using these methods for higher index problems, it was observed that difficulties may arise, see, e.g., [126], [163], [203]. As we have shown, these difficulties can be avoided by transforming the differential-algebraic equation to an equivalent strangeness-free formulation, see also [125], [126], [127], [128], [132].

Exercises

1. Verify the claims of Lemma 5.5.
2. Show that $\|A \otimes I\|_\infty = \|A\|_\infty$ for arbitrary matrices $A \in \mathbb{C}^{m,n}$.
3. Assume that (E, A) with $E, A \in \mathbb{C}^{n,n}$ is a singular matrix pair. Prove with the help of Theorem 2.3 that then $I_n \otimes E - h\mathcal{A} \otimes A$ is singular as well, independent of the choice of $\mathcal{A} \in \mathbb{C}^{n,n}$.
4. Show that for the Gauß method with $s = 2$, see Table 5.1, we have that $\kappa_1 = 2$ and $\kappa_2 = 2$ as defined in Theorem 5.10.
5. Show that for the Radau IIA method with $s = 2$, see Table 5.2, we have that $\kappa_1 = \infty$ and $\kappa_2 = 2$ as defined in Theorem 5.10.
6. Implement the Gauß method with $s = 2$ from Table 5.1 using a constant stepsize. Use Newton's method and finite differences to approximate the Jacobians for the solution of the arising nonlinear systems. Apply your program to the problem (5.26), where N is a single nilpotent Jordan block with $\mathrm{ind}(N, I) = 2$ and $f(t) = [\,0 \;\; \exp(t)\,]^T$, and verify the corresponding claim of Theorem 5.10. Explain why this method is not covered by Theorem 5.12.
7. Implement the Radau IIA method with $s = 2$ from Table 5.2 using a constant stepsize. Use Newton's method and finite differences to approximate the Jacobians for the solution

of the arising nonlinear systems. Apply your program to the problem (5.26), where N is a single nilpotent Jordan block with $\mathrm{ind}(N, I) = 2$ and $f(t) = [\,0 \;\; \exp(t)\,]^T$, and verify the corresponding claims of Theorem 5.10 and Theorem 5.12.

8. Apply your program of Exercise 7 to the problem (5.26) where N is a single nilpotent Jordan block with $\mathrm{ind}(N, I) = \nu$, $\nu = 2, 3, 4$, and $f(t) = [\,0 \;\; \cdots \;\; 0 \;\; \exp(t)\,]^T$, and determine the resulting orders of convergence.

9. Show that the matrix $\tilde{E} - h\tilde{A}$ in Example 5.14 is invertible and that the condition number is
$$\|\tilde{E} - h\tilde{A}\|_\infty \|(\tilde{E} - h\tilde{A})^{-1}\|_\infty = \mathcal{O}(h^{-2}).$$

10. Apply the Radau IIA method with $s = 2$ given in Table 5.2 to (5.33) as well as to the transformed system (5.34) and to the associated reduced problem (3.60). Give a theoretical analysis and compare the numerical results.

11. Apply the programs developed in Exercises 6 and 7 to the problem of Example 3.1
$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$
Do the methods recognize that the solution is not unique?

12. Show that if the conditions $B(p)$ and $C(q)$ in (5.22) are satisfied, then the quadrature rules given by the coefficients $(\beta_j, \gamma_j)_{j=1,\ldots,s}$ for intervals $[t_i, t_{i+1}]$, $j = 1, \ldots, s$ and by the coefficients $(\alpha_{jl}, \gamma_l)_{l=1,\ldots,s}$ for intervals $[t_i, t_i + \gamma_j h]$, $j = 1, \ldots, s$, are of order p and q, respectively, i.e., show that (5.48) holds.

13. Show that the Gauß method with $s = 1$ given in Table 5.1 satisfies $B(p)$ and $C(q)$ of (5.22) with $p = 2$ and $q = 1$ as best choice. Determine the value ϱ of Theorem 5.16 and verify the claimed order by numerical experiments.

14. Show that the Gauß method with $s = 2$ given in Table 5.1 satisfies $B(p)$ and $C(q)$ of (5.22) with $p = 4$ and $q = 2$ as best choice. Determine the value ϱ of Theorem 5.16 and verify the claimed order by numerical experiments.

15. Let $x \in C^1(\mathbb{I}, \mathbb{R})$ with a compact interval $\mathbb{I}$ be sufficiently smooth. Show that $x(t) = \mathcal{O}(h^{k+1})$ for all $t \in \mathbb{I}$ implies that $\dot{x}(t) = \mathcal{O}(h^{k+1})$ for all $t \in \mathbb{I}$.

16. Let $\phi \in C(\mathbb{I}, \mathbb{R})$ with a compact interval $\mathbb{I}$ be sufficiently smooth. Furthermore, assume that ϕ has the pairwise distinct zeros $t_j \in \mathbb{I}$, $j = 0, \ldots, k$. Show that there exists a smooth function $\omega \in C(\mathbb{I}, \mathbb{R})$ such that ϕ has the representation
$$\phi(t) = \omega(t) \prod_{j=0}^{k} (t - t_j)$$
for all $t \in \mathbb{I}$.

17. Apply the program developed in Exercise 7 to the regular strangeness-free problem
$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} \exp(t) + \cos(t) \\ \exp(t) \end{bmatrix}, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$
Compare the results with the claim of Theorem 5.18.

18. Consider the differential-algebraic equation

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \\ u(t) \end{bmatrix}$$

with some sufficiently smooth $u \in C(\mathbb{I}, \mathbb{R})$. Let $t_i \in \mathbb{I}$, $i = 0, \ldots, N$, be given with $t_0 < t_1 < t_2 < \cdots < t_N$, $h_i = t_i - t_{i-1}$, $i = 1, \ldots, N$. Furthermore, let $[x_{1,i}\, x_{2,i}\, x_{3,i}]^T$, $i = 0, \ldots, N$, be the numerical approximations to the solution obtained by the implicit Euler method.

Show that the numerical approximations with $i \geq 2$ are independent of the choice of the initial values. Furthermore, show that for constant stepsize the implicit Euler method is convergent of order one. Finally, show that convergence is lost if the stepsize is changed.

19. Prove the claim of Remark 5.25.

20. Implement the BDF methods with $k = 1, \ldots, 6$, see Table 5.3, using a constant stepsize. Use Newton's method and finite differences to approximate the Jacobians for the solution of the arising nonlinear systems. Apply the program to the problem (5.26) where N is a single nilpotent Jordan block with $\operatorname{ind}(N, I) = \nu$, $\nu = 2, 3, 4$, and $f(t) = [\, 0 \ \cdots \ 0 \ \exp(t) \,]^T$, using exact initial values. Compare with the claim of Theorem 5.24.

21. Apply the program developed in Exercise 20 to the problem of Example 3.1

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Do the BDF methods recognize that the solution is not unique?

22. Apply the program developed in Exercise 20 to the regular strangeness-free problem

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} \exp(t) + \cos(t) \\ \exp(t) \end{bmatrix}, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

using exact initial values. Compare the results with the claim of Theorem 5.27.

23. Apply the programs developed in Exercise 7 and Exercise 20 to the regular strangeness-free problem

$$\begin{aligned} \dot{x}_1 &= x_4, & \dot{x}_4 &= -2x_1x_7, \\ \dot{x}_2 &= x_5, & \dot{x}_5 &= -2x_2x_7, \\ 0 &= x_1^2 + x_2^2 - x_3, \\ 0 &= 2x_1x_4 + 2x_2x_5 - x_6, \\ 0 &= 2x_4^2 - 4x_1^2x_7 + 2x_5^2 - 4x_2^2x_7 - x_7 + 1 = 0. \end{aligned}$$

Chapter 6
Numerical methods for index reduction

In Chapter 5, we have seen that stiffly accurate Runge–Kutta methods and BDF methods are well-suited for the numerical solution of semi-explicit differential-algebraic equations of index $\nu = 1$ or, more general, of regular strangeness-free differential-algebraic equations, while difficulties may arise for systems of higher index. An immediate idea for higher index problems is to use index reduction by differentiation, cp. Theorem 4.26, to turn the problem into one of index $\nu = 1$ or even into an ordinary differential equation. Although this seems very appealing and actually was the common approach until the early seventies, there are major disadvantages to this approach.

The algebraic equations typically represent constraints or conservation laws. Using differentiation as in Theorem 4.26 to reduce the index, these algebraic equations are not present any more during the numerical integration, and hence discretization and roundoff errors may lead to numerical results that violate the constraints. This was already observed in the context of mechanical multibody systems and led to several stabilization techniques, like in [24], [94]. These techniques are incorporated in many solution methods, see, e.g., [79], [108], [181], [201].

Another way out of this dilemma is to not replace the algebraic equations by their differentiated versions but to simply add these to the system, thus creating an overdetermined system that includes all the algebraic equations as well as all the necessary information from the derivatives. This approach was examined in [23], [85], [86] and for the derivative arrays (3.28) or (4.11) in [48], [49], [58]. The disadvantage of this approach is that it typically leads to a larger system and, in particular, that it requires special solution methods for the linear and nonlinear systems at every integration step to take care of the overdetermined systems.

An interesting variation of this approach was presented in [154] and modified in [130]. There new variables, so called *dummy derivatives*, are introduced to make these overdetermined systems regular and strangeness-free. The necessary information which variables have to be introduced is obtained in [154] from the calculation of the *structural* or *generic index* as defined in [161], [162]. Unfortunately, it was shown in [189] that the structural index may not give the correct information about the properties of the system, which makes this approach mathematically doubtful. In the case of structured problems like those discussed in Section 4.2 this information can be obtained differently, see Section 6.4 below.

There are also many other approaches for index reduction that we will not discuss in detail here. These include matrix chains as introduced in [100], projection

methods which repeatedly project the numerical solution onto the constraints, see, e.g., [7], [9], [108], [143], the differential-geometric approach presented in [169], and tangent space parameterization [169], [228].

In view of the discussion in Chapters 3 and 4, for general high index problems, however, it seems more promising to discretize the reduced differential-algebraic equations (3.60) in the linear case or (4.23) in the nonlinear case. Recall that these reduced equations possess the same solutions as the original differential-algebraic equation. Moreover, due to Chapter 5, a number of numerical discretization schemes are available to integrate such systems. The only problem that we are faced with is that the reduced equations are defined implicitly and are therefore not available in a numerical procedure. However, for the use within a discretization method, we only must be able to evaluate the reduced equations at distinct points. In the following, we will discuss how this can be done. We will treat the linear and the nonlinear case separately. For ease of presentation, we will restrict ourselves to the BDF discretization. For a discretization with Runge–Kutta type methods, see the results of Section 7.3, which in a similar way also hold for the Radau IIA methods, cp. [208]. In addition, we will address here the problem of the computation of consistent initial values.

6.1 Index reduction for linear problems

Let us first discuss linear differential-algebraic systems with variable coefficients that satisfy Hypothesis 3.48. In this case, the constructed strangeness-free system is given by (3.60) and the initial value problem that has to be solved has the form

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t), \quad x(t_0) = x_0, \tag{6.1}$$

where

$$(\hat{E}, \hat{A}) = \left(\begin{bmatrix} \hat{E}_1 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \end{bmatrix} \right), \quad \hat{f} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \end{bmatrix}. \tag{6.2}$$

The coefficients are obtained from the derivative array

$$M_\mu(t)\dot{z}_\mu = N_\mu(t)z_\mu + g_\mu(t), \tag{6.3}$$

where in contrast to (3.28) we have omitted the hat of $\hat{\mu}$, via

$$\begin{aligned} \hat{E}_1 = Z_1^T E, \quad \hat{A}_1 = Z_1^T A, & \qquad \hat{f}_1 = Z_1^T f, \\ \hat{A}_2 = Z_2^T N_\mu [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^T, & \qquad \hat{f}_2 = Z_2^T g_\mu. \end{aligned} \tag{6.4}$$

The (smooth) functions Z_1, Z_2 pointing to the differential and algebraic part of the given differential-algebraic equation, respectively, are defined in Hypothesis 3.48

as follows, again omitting hats. Since M_μ has constant rank, there exists a smooth matrix function Z_2 of size $(\mu+1)n \times a$ and pointwise maximal rank satisfying

$$Z_2^T M_\mu = 0, \quad \operatorname{rank}(Z_2^T N_\mu [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^T) = a. \tag{6.5}$$

Then, there exists a smooth matrix function T_2 of size $n \times d$, $d = n - a$, satisfying

$$Z_2^T N_\mu [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^T T_2 = 0, \quad \operatorname{rank}(ET_2) = d. \tag{6.6}$$

Finally, there exists a smooth matrix function Z_1 of size $n \times d$ and pointwise maximal rank satisfying

$$\operatorname{rank}(Z_1^T E T_2) = d. \tag{6.7}$$

For numerical purposes, we may even assume that Z_1, Z_2, and T_2 possess pointwise orthonormal columns, cp. Theorem 3.9.

When we discretize (6.1) by a BDF method, then the approximations x_i to the values $x(t_i)$ are the solutions of the discrete problem

$$\hat{E}(t_i) D_h x_i = \hat{A}(t_i) x_i + \hat{f}(t_i), \tag{6.8}$$

with $D_h x_i = \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l}$. Hence, in order to compute x_i, we must evaluate $\hat{E}$, $\hat{A}$, and $\hat{f}$ at t_i. For this purpose, we need $Z_1(t_i)$ and $Z_2(t_i)$.

Considering the definition of Z_1 and Z_2, we are able to compute a matrix $\tilde{Z}_2 \in \mathbb{R}^{(\mu+1)n,a}$ with orthonormal columns satisfying

$$\tilde{Z}_2^T M_\mu(t_i) = 0, \quad \operatorname{rank}(\tilde{Z}_2^T N_\mu(t_i) [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^T) = a, \tag{6.9}$$

say with the help of a singular value decomposition, a rank-revealing QR decomposition, or a URV decomposition of $M_\mu(t_i)$, see, e.g., [99], [206]. Using the same techniques, we can then determine a matrix $\tilde{T}_2 \in \mathbb{R}^{n,d}$ with orthonormal columns such that

$$\tilde{Z}_2^T N_\mu(t_i) [\, I_n \;\; 0 \;\; \cdots \;\; 0 \,]^T \tilde{T}_2 = 0, \quad \operatorname{rank}(E(t_i)\tilde{T}_2) = d. \tag{6.10}$$

Similarly, we finally get a matrix $\tilde{Z}_1 \in \mathbb{R}^{n,d}$, again possessing orthonormal columns, with

$$\operatorname{rank}(\tilde{Z}_1^T E(t_i) \tilde{T}_2) = d. \tag{6.11}$$

Since $\tilde{Z}_1$, $\tilde{Z}_2$, and $\tilde{T}_2$ have orthonormal columns which can be chosen such that these columns span the same subspaces as the columns of $Z_1(t_i)$, $Z_2(t_i)$, and $T_2(t_i)$, respectively, there exist unitary matrices U_1 and U_2 satisfying

$$\tilde{Z}_1 = Z_1(t_i) U_1, \quad \tilde{Z}_2 = Z_2(t_i) U_2. \tag{6.12}$$

Thus, in the numerical method we do not evaluate smooth functions Z_1 and Z_2, which we know to exist by theory, at a given point t_i, but only matrices $\tilde{Z}_1$ and $\tilde{Z}_2$ with (6.12). However, since (6.8), written in more detail as

$$\begin{aligned} Z_1(t_i)^T E(t_i) D_h x_i &= Z_1(t_i)^T A(t_i) x + Z_1(t_i)^T f_1(t_i), \\ 0 &= Z_2(t_i)^T N_\mu(t_i) x + Z_2(t_i)^T g_\mu(t_i), \end{aligned} \tag{6.13}$$

is equivalent to

$$\begin{aligned} \tilde{Z}_1^T E(t_i) D_h x_i &= \tilde{Z}_1^T A(t_i) x + \tilde{Z}_1^T f_1(t_i), \\ 0 &= \tilde{Z}_2^T N_\mu(t_i) x + \tilde{Z}_2^T g_\mu(t_i) \end{aligned} \tag{6.14}$$

as a system of linear equations that determines x_i, it is sufficient to solve (6.14) in order to compute x_i.

In principle, singular value decomposition and rank-revealing QR decompositions can be modified to evaluate smooth functions Z_1 and Z_2, see [31], [122], [227]. But during the integration procedure such computations would be too costly. As we have seen by the above discussion, this does not present any difficulties as long as the numerical integration method that we are using is invariant under transformations with unitary matrix functions form the left. All the methods that we have discussed in Chapter 5 fulfill this property.

Remark 6.1. During the integration process, the determination of Z_1 and Z_2 allows us simultaneously to check whether the quantities a and d stay constant. If this is not the case, then we must distinguish two cases. If $a+d=n$ still holds, then there is a change in the number of constraints. In such a case, it is important to locate the point where the change in the values a and d occurs. Moreover, if the value of a is increasing, then we must check the consistency of the actual solution value before we may continue the integration. If on the other hand $a+d<n$ holds, then the value μ may be too small. Hence, we must increase μ and check if then the corresponding values a and d stay constant or not.

Example 6.2. A typical situation, where the effects described in Remark 6.1 may occur, are *hybrid* (or *switched*) systems, see, e.g., [109], [110].

As a (nonlinear) example, consider a pendulum of length l and mass m under the influence of gravity $G=-mg$ and assume that the pendulum is tangentially accelerated by a (linearly) increasing force. Using the classical Euler–Lagrange formalism [79], [108] in Cartesian coordinates $[\,x_1\ x_2\,]^T=[\,x\ y\,]^T$ and velocities $[\,x_3\ x_4\,]^T=[\,\dot{x}\ \dot{y}\,]^T$ and the acceleration forces $f(t,x_1,x_2)=[\,f_1(t,x_1,x_2)\ \ f_2(t,x_1,x_2)\,]^T$, one obtains the differential-algebraic equation

$$\begin{aligned} \dot{x}_1 &= x_3, \\ \dot{x}_2 &= x_4, \end{aligned}$$

$$
\begin{aligned}
m\dot{x}_3 &= -2x_1\lambda + f_1(t, x_1, x_2),\\
m\dot{x}_4 &= -mg - 2x_2\lambda + f_2(t, x_1, x_2),\\
0 &= x_1^2 + x_2^2 - l^2.
\end{aligned}
$$

Now suppose that, when a certain centrifugal force is reached, the system changes from a pendulum to an flying mass point, i.e., the rope or rod is cut. In this case the system is not constrained anymore and the equations of motion are given by

$$
\begin{aligned}
\dot{x}_1 &= x_3,\\
\dot{x}_2 &= x_4,\\
m\dot{x}_3 &= 0,\\
m\dot{x}_4 &= -mg.
\end{aligned}
$$

If we consider the complete system, then it consists of two (operating) modes and it switches once between them. A typical task would be to determine the switching point, to simulate the movement of the mass point, or to control the switching and the successive flight. Since we have a change from a system with characteristic values $\mu = 2$, $a = 3$, and $d = 2$ to a system with characteristic values $\mu = 0$, $a = 0$, and $d = 5$, when we add the equation $\dot{\lambda} = 0$ for the no longer needed Lagrange multiplier, there are no problems with the consistency of the solution of the first system with respect to the second system at the switching point. Of course, we do not need to consider λ when we solve the second system numerically.

Remark 6.3. For linear differential-algebraic equations with constant coefficient systems the functions Z_1 and Z_2 can be chosen to be constant in the whole interval. In this case, we only need to compute $\tilde{Z}_1$ and $\tilde{Z}_2$ at the initial point t_0 and can use them during the whole integration process.

Remark 6.4. If we are in the case of an over- or underdetermined system with well-defined strangeness index as introduced in Section 3.1, then we can achieve the more general reduced differential-algebraic equation (3.44) with (3.41). As we have discussed in Section 3.1, it is not clear how the inhomogeneity $\hat{f}_3$ belonging to the third block of equations can be chosen in an invariant way.

From the point of view of numerical methods, this is not a severe problem, since a nonzero $\hat{f}_3$ just indicates that the given differential-algebraic equation has no solution due to an inconsistent inhomogeneity. In order to check this, we may simply omit that part of the system such that we are (after possibly fixing some free solution components) back in the case (6.2). Having then computed an x_i, we can test whether the residual $E(t_i)D_h x_i - A(t_i)x_i - f(t_i)$ in the original system is significantly larger than the expected discretization error. If this is the case, it indicates that the given differential-algebraic equation is not solvable, which usually means that the underlying mathematical model should be modified.

Having discussed how we can integrate a given differential-algebraic equation, there is still one problem that must be considered. Before we can start the integration, we must guarantee that the given initial value x_0 is consistent. In general, the computation of consistent initial values for higher index differential-algebraic equations is a difficult problem, see [82], [214]. In the case of a strangeness-free reduced differential-algebraic equation (6.1), however, where the algebraic equations are displayed directly, the computation of consistent initial values is straightforward. Since the condition for x_0 to be consistent is simply given by

$$0 = \hat{A}(t_0)x_0 + \hat{f}_2(t_0), \tag{6.15}$$

it is easy to check whether the given x_0 is consistent or not. In the case that a given $\tilde{x}_0$ is not consistent, we may use (6.15) to determine a related consistent x_0. One way to do this is to set $\tilde{x}_0 = x_0 + \delta$ and to determine the correction δ by solving the minimization problem

$$\|\delta\|_2 = \min!\,, \tag{6.16}$$

subject to the constraint

$$\|\hat{A}_2(t_0)\delta - \hat{f}_2(t_0) - \hat{A}_2(t_0)\tilde{x}_0\|_2 = \min!\,. \tag{6.17}$$

The solution of this least-squares problem is given by

$$\delta = \hat{A}_2(t_0)^+(\hat{A}_2(t_0)\tilde{x}_0 + \hat{f}_2(t_0)), \tag{6.18}$$

where $\hat{A}_2(t_0)^+$ is the Moore–Penrose pseudoinverse of $\hat{A}_2(t_0)$.

Since $\hat{A}_2(t_0)$ has full row rank a due to Hypothesis 3.48, it follows that $\hat{A}_2(t_0)\hat{A}_2(t_0)^+ = I_a$, hence

$$\begin{aligned}\hat{A}_2(t_0)x_0 + \hat{f}_2(t_0) &= \hat{A}_2(t_0)(\tilde{x}_0 - \delta) + \hat{f}_2(t_0)\\ &= \hat{A}_2(t_0)\tilde{x}_0 - (\hat{A}_2(t_0)\tilde{x}_0 + \hat{f}_2(t_0)) + \hat{f}_2(t_0) = 0.\end{aligned}$$

If we are interested in finding some consistent initial value x_0, then this least-squares approach is the easiest solution. But in some applications that lead to differential-algebraic systems, one wants to prescribe initial values for the differential variables, whereas initial values for the algebraic variables are not known. In such a situation, we may proceed as follows. Partition (6.15) as

$$0 = [\ \hat{A}_{21}(t_0)\ \ \hat{A}_{22}(t_0)\ ] \begin{bmatrix} x_{1,0} \\ x_{2,0} \end{bmatrix} + \hat{f}_2(t_0).$$

Suppose that $\hat{A}_{22}(t_0)$ has full row rank indicating that the quantities $x_{1,0}$ belong to the differential variables. If an estimate $[\tilde{x}_{1,0}^T, \tilde{x}_{2,0}^T]^T$ for a consistent initial value

is given and if we want $\tilde{x}_{1,0}$ to stay as it is, then we may determine a correction δ_2 according to $\tilde{x}_{2,0} = x_{2,0} + \delta_2$ by solving the minimization problem

$$\|\delta_2\|_2 = \min!\,, \tag{6.19}$$

subject to the constraint

$$\|\hat{A}_{22}(t_0)\delta_2 - \hat{A}_2(t_0)\tilde{x}_0 - \hat{f}_2(t_0)\|_2 = \min!, \tag{6.20}$$

i.e.,

$$\delta_2 = \hat{A}_{22}^{+}(t_0)(\hat{A}_2(t_0)\tilde{x}_0 + \hat{f}_2(t_0)). \tag{6.21}$$

The same argument as for the first case shows that setting $x_{1,0} = \tilde{x}_{1,0}$ and $x_{2,0} = \tilde{x}_{2,0} - \delta_2$ yields a consistent initial value.

6.2 Index reduction for nonlinear problems

For general nonlinear initial value problems of the form

$$F(t, x, \dot{x}) = 0, \quad x(t_0) = x_0, \tag{6.22}$$

we proceed with the index reduction as suggested in Chapter 4. In particular, we assume that Hypothesis 4.2 holds in a neighborhood of a path $(t, x^*(t), \mathcal{P}(t))$, $t \in \mathbb{I}$, belonging to the unique solution $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (6.22). Following Section 4.1, the function x^* also (locally) solves the reduced problem

$$\hat{F}(t, x, \dot{x}) = 0, \quad x(t_0) = x_0. \tag{6.23}$$

Recall that $\hat{F}$ has the special structure

$$\hat{F}(t, x, \dot{x}) = \begin{bmatrix} \hat{F}_1(t, x, \dot{x}) \\ \hat{F}_2(t, x) \end{bmatrix}, \tag{6.24}$$

where $\hat{F}_1(t, x, \dot{x}) = \tilde{Z}_1^T F(t, x, \dot{x})$ with an appropriate matrix $\tilde{Z}_1$ that possesses orthonormal columns and a function $\hat{F}_2$ defined by (4.22). In order to determine a numerical approximation x_i to $x^*(t_i)$, we apply a BDF method to (6.23) and obtain the nonlinear system

$$\hat{F}(t_i, x_i, D_h x_i) = 0. \tag{6.25}$$

Whereas an appropriate $\tilde{Z}_1$ can be obtained in the same way as in the linear case, we are here faced with the problem that $\hat{F}_2$ is only defined implicitly by means of

the implicit function theorem. In order to deal with $\hat{F}_2$ numerically, we therefore must go back to its defining equation. According to (4.21), we (locally) have

$$F_\mu(t, x, y) = 0 \implies y = \mathcal{H}(t, x_1, p), \tag{6.26}$$

where the variables x_1 out of x and p out of $(\dot{x}, \ldots, x^{(\mu+1)})$ denote the selected parameterization of $\mathbb{L}_\mu$. Accordingly to the splitting of $x = (x_1, x_2)$, we also split the numerical approximations as $x_i = (x_{1,i}, x_{2,i})$. As before, there is no danger of mixing up the double meanings of x_1, x_2. Utilizing the splitting, we observe that

$$F_\mu(t_i, x_i, y_i) = 0 \implies y_i = \mathcal{H}(t_i, x_{1,i}, p_i). \tag{6.27}$$

In the construction of Section 4.1, it is possible to fix the parameters p to be p_i instead of p_0. The corresponding definition of $\hat{F}_2$ then reads

$$\hat{F}_2(t, x_1, x_2) = Z_2^T F_\mu(t, x_1, x_2, \mathcal{H}(t, x_1, p_i)). \tag{6.28}$$

If we replace $\hat{F}_2$ in (6.23) by this, then we have

$$F_\mu(t_i, x_i, y_i) = 0 \implies \hat{F}_2(t_i, x_i) = 0. \tag{6.29}$$

Hence, instead of (6.25) we can consider

$$\tilde{Z}_1^T F(t_i, x_i, D_h x_i) = 0, \quad F_\mu(t_i, x_i, y_i) = 0, \tag{6.30}$$

which is an underdetermined nonlinear system in the unknowns (x_i, y_i). Any solution (x_i, y_i) yields an x_i that can be interpreted as the unique solution of (6.25) with an appropriately adapted part $\hat{F}_2$.

The standard method for the solution of underdetermined systems of nonlinear equations $\mathcal{F}(z) = 0$ as (6.30) is the Gauß–Newton method

$$z^{m+1} = z^m - \mathcal{F}_z^+(z^m)\mathcal{F}(z^m), \quad z^0 \text{ given}, \tag{6.31}$$

where $\mathcal{F}_z^+(z)$ is the Moore–Penrose pseudoinverse of $\mathcal{F}_z(z)$. In order to have (quadratic) convergence of the iterates z^m to a solution of $\mathcal{F}(z) = 0$, we must show that the Jacobians $\mathcal{F}_z(z)$ possess full row rank at a solution and thus in a whole neighborhood, see, e.g., [71], [160].

Theorem 6.5. *Let F of (6.22) satisfy Hypothesis 4.2. Then the Jacobian J of (6.30) at a solution (x_i, y_i) has full row rank for sufficiently small h, provided that $\tilde{Z}_1$ is a sufficiently good approximation to Z_1 of Hypothesis 4.2 corresponding to this solution.*

Proof. We have (without arguments)

$$J = \begin{bmatrix} \tilde{Z}_1^T \left(\frac{\alpha_k}{h} F_{\dot{x}} + F_x\right) & 0 \\ -N_\mu [I_n \; 0 \; \cdots \; 0]^T & M_\mu \end{bmatrix}.$$

Hypothesis 4.2 implies that

$$\begin{aligned}\operatorname{rank} J &= (\mu+1)n - a + \operatorname{rank}\begin{bmatrix}\tilde{Z}_1^T\left(\frac{\alpha_k}{h}F_{\dot{x}} + F_x\right)\\ \hat{A}_2\end{bmatrix}\\ &= (\mu+1)n + \operatorname{rank}\left(\tilde{Z}_1^T\left(\tfrac{\alpha_k}{h}F_{\dot{x}} + F_x\right)T_2\right).\end{aligned}$$

For sufficiently small h, we then obtain that

$$\operatorname{rank} J = (\mu+1)n + \operatorname{rank}(\tilde{Z}_1^T F_{\dot{x}} T_2).$$

For $\tilde{Z}_1 = Z_1$, this reduces to

$$\operatorname{rank} J = (\mu+1)n + d,$$

and, hence, $\tilde{Z}_1$ must approximate Z_1 in such a way that $\tilde{Z}_1^T F_{\dot{x}} T_2$ keeps full rank. □

Remark 6.6. If the matrix function Z_2 of Hypothesis 4.2 only depends on t, then one can determine $\tilde{Z}_2 = Z_2(t_i)$ in the same way as Z_2 in the linear case. The discrete system (6.30) can then be reduced to

$$\tilde{Z}_1^T F(t_i, x_i, D_h x_i) = 0, \tag{6.32a}$$

$$\tilde{Z}_2^T F_\mu(t_i, x_i, y_i) = 0. \tag{6.32b}$$

Due to Hypothesis 4.2, we have that $\tilde{Z}_2^T F_{\mu;y}(t_i, x, y) = 0$ independent of (x, y) such that (6.32b) does not depend on y_i. Hence, (6.32) can be seen as a nonlinear system for x_i only. In particular, this simplification applies to ordinary differential equations where $\tilde{Z}_2$ is an empty matrix and $\tilde{Z}_1 = I_n$ such that (6.32) becomes the standard BDF method applied to an ordinary differential equation. Moreover, this simplification also applies to linear systems of differential-algebraic equations, where (6.32) becomes (6.14) such that the linear case can be seen as a special form of the nonlinear case.

Remark 6.7. Looking at the presented index reduction procedure, it is clear that the problem sizes that can be handled by this approach is limited, since we must perform rank revealing factorizations of the full derivative array at every integration step. Furthermore, the index reduction procedure heavily relies on rank decisions that depend critically on the used method. To compute the rank of a matrix numerically is an intrinsically difficult mathematical problem, see, e.g., [99]. In this respect, the presented approach still needs a subtle error and perturbation analysis that is currently not available. Only for the special case of constant coefficient systems this analysis has been done partially in [153]. This analysis suggests a conservative strategy which, in the case of an unclear rank decision, takes the decision to rather use a higher value μ.

In order to compute consistent initial values, we recall that in Chapter 4 it has been shown that every (x_0, y_0) in a neighborhood of $(x^*(t_0), \mathcal{P}(t_0))$ can be locally extended to a solution of (6.22). Thus, consistency of an initial value x_0 at t_0 means that (t_0, x_0) is part of some $(t_0, x_0, y_0) \in \mathbb{L}_\mu$. To determine a consistent initial value or to check it for consistency, we must therefore solve the underdetermined system

$$F_\mu(t_0, x_0, y_0) = 0 \tag{6.33}$$

for (x_0, y_0). Again, the method of choice is the Gauß–Newton method, started with a sufficiently good guess $(\tilde{x}_0, \tilde{y}_0)$. As before, we can expect local quadratic convergence due to the following theorem.

Theorem 6.8. *Let F of (6.22) satisfy Hypothesis 4.2. Then the Jacobian J of (6.33) at a solution (x_0, y_0) has full row rank.*

Proof. Hypothesis 4.2 requires that $F_{\mu;x}$ and $F_{\mu;y}$ together have full row rank on $\mathbb{L}_\mu$. Hence, $J = [\, F_{\mu;x} \;\; F_{\mu;y} \,]$ has full row rank at a solution of (6.33). □

Remark 6.9. To determine the numerical solution of a differential-algebraic equation by means of the derivative array seems to be very expensive, especially if we look at the size of the function F_μ which has values in $\mathbb{R}^{(\mu+1)n}$. But there seems to be no way to avoid this when the given differential-algebraic equation has no special structure that we can utilize.

If, however, the differential-algebraic equation has further structure, as for example in the case of multibody systems, then often a part $F_{\mu,\mathrm{red}}$ of F_μ will be sufficient for the computation. We only need that

$$F_{\mu,\mathrm{red}}(t_i, x_i, y_i) = 0 \implies \hat{F}_2(t_i, x_i) = 0.$$

In the extreme case, $F_{\mu,\mathrm{red}}$ just contains all algebraic constraints, see, e.g., the formulation (4.44) for multibody systems. It is then obvious that we can just replace F_μ by $F_{\mu,\mathrm{red}}$ in (6.30) and (6.33) without changing the essential properties of these systems. On the other hand, we may expect that the computational work to solve the modified systems is drastically reduced. Of course, these observations also apply to the linear case. See Section 6.4 for the construction of reduced derivative arrays $F_{\mu,\mathrm{red}}$ in special applications.

Remark 6.10. In the context of nonlinear differential-algebraic equations, it is important to note that already for the computation of consistent initial values we must know at least the value of μ. In the linear case, μ can be computed even when we have no consistent value, because there the linearization does not depend on the trajectory along which we linearize. But in the nonlinear case we must take into consideration that we can compute μ (as well as a and d) only when we already have a consistent value. This happens due to the fact that the quantities in Hypothesis 4.2

may be different away from the manifold $\mathbb{L}_\mu$. In the case that the values of μ, a, and d are not known in advance, the only possibility is to try to fix them during the computation of a consistent initial value. In this situation, one may proceed as follows. Set $\mu = 0$, $a = 0$, and $d = n - a$. In each iteration step for solving (6.33), we numerically check the ranks (omitting arguments)

$$\operatorname{corank} F_{\mu;y} \leq a, \tag{6.34a}$$

$$\operatorname{rank} \tilde{Z}_2^T F_{\mu;x} = a, \tag{6.34b}$$

$$\operatorname{rank} F_{\dot{x}} \tilde{T}_2 = d, \tag{6.34c}$$

using say the singular value decomposition. Here $\tilde{Z}_2$ denotes a matrix whose orthonormal columns span the corange of $F_{\mu;y}$ perturbed to a matrix of corank a and $\tilde{T}_2$ denotes a matrix whose orthonormal columns span the kernel of $\tilde{Z}_2^T F_{\mu;x}$ perturbed to a matrix of rank a. These perturbations can be easily obtained from the singular value decomposition by neglecting the smallest (possibly nonzero) singular values, i.e., expecting a certain rank r proposed by the current setting of a and d according to (6.34), we simply set the $(r+1)$-st and all following singular values to zero. This corresponds to approximating the given matrix by the nearest matrix (with respect to the Frobenius norm) of the required (lower) rank r, see e.g., [99].

A violation of the numerical rank check in (6.34a) then indicates that there may be additional constraints and that the value of a should be increased by one (and d decreased by one). A violation of (6.34b) indicates that local uniqueness may not hold for the given problem and one should terminate. A violation of (6.34c) indicates that one is near a higher index problem and the value of μ should be increased by one.

The determination of the characteristic values μ, a, and d is successful if this process ends with accepting a consistent initial value under the rank checks (6.34). Note, however, that this procedure may not work if the given problem is not well scaled or the initial guess is poor. A detailed perturbation analysis for this procedure is a difficult open problem.

As in the linear case, it is possible to modify (6.33) when we want some components of the initial guess $\tilde{x}_0$ to be kept fixed in the determination of consistent initial values. We only must guarantee that the remaining columns of the Jacobian still have full row rank. This corresponds to the classification of a component of x to be a differential variable by the requirement that eliminating the associated column from the Jacobian does not lead to a rank deficiency.

Remark 6.11. In order to solve underdetermined nonlinear systems of equations by the Gauß–Newton method, the most important property that is needed is full row rank at the desired solution. This property then not only extends to a neighborhood of the solution set and guarantees local and quadratic convergence to some solution,

but also allows for some techniques known from Newton's method to improve the efficiency of the method, such as the simplified Gauß–Newton method (fixing the Jacobian) or the quasi-Gauß–Newton method (using Broyden rank one updates of the Jacobian), see, e.g., [71], [160].

6.3 Index reduction via feedback control

In the context of control problems

$$F(t, x, u, \dot{x}) = 0, \quad x(t_0) = x_0, \tag{6.35}$$

possibly together with an output equation

$$y = G(t, x), \tag{6.36}$$

see Section 4.4, different choices of the control u in (6.35) may lead to different properties of the resulting controlled problem

$$F(t, x, u(t), \dot{x}) = 0, \quad x(t_0) = x_0. \tag{6.37}$$

In particular, the free system corresponding to the choice $u = 0$ may not even satisfy Hypothesis 4.2. Section 4.4 gives several sufficient conditions when u can be chosen as state or output feedback in such a way that the closed loop system is regular and strangeness-free.

Let us first consider a system without the output equation (6.36). As we have shown in Section 4.4, the regularization can be obtained by a piecewise linear feedback of the form (4.63) with a suitable matrix $\tilde{K}$, cp. Theorem 4.39 and Corollary 4.40. Combining the analysis in Section 4.4 and the numerical procedures introduced in Section 6.2, discretization of (4.59) with a BDF method leads to

$$\hat{F}_1(t_i, x_i, u_i, D_h x_i) = 0, \quad \hat{F}_2(t_i, x_i, u_i) = 0. \tag{6.38}$$

The algebraic constraints can be replaced by

$$F_\mu(t_i, x_i, u_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)}) = 0. \tag{6.39}$$

Choosing a suitable feedback matrix $\tilde{K}$ which yields a regularizing state feedback, we set

$$u_i = \tilde{K} x_i + w_i. \tag{6.40}$$

The vector w_i should be chosen in such a way that u_i is near the old value u_{i-1}. We therefore take $w_i = u_{i-1} - \tilde{K} x_{i-1}$. Together with a suitable matrix $\tilde{Z}_1$, we thus

obtain the problem

$$\begin{aligned} \tilde{Z}_1^T F(t_i, x_i, \tilde{K}x_i + w_i, D_h x_i) &= 0 \\ F_\mu(t_i, x_i, \tilde{K}x_i + w_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)}) &= 0 \end{aligned} \tag{6.41}$$

for $(x_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)})$. By construction, the value x_i is the unique solution of

$$\hat{F}_1(t_i, x_i, \tilde{K}x_i + w_i, D_h x_i) = 0, \quad \hat{F}_2(t_i, x_i, \tilde{K}x_i + w_i) = 0. \tag{6.42}$$

Under the assumptions of Theorem 6.5, the Gauß–Newton method applied to (6.41) will again show quadratic convergence for sufficiently small h.

To obtain consistent initial values, we accordingly must solve

$$F_\mu(t_0, x_0, u_0, \dot{x}_0, \dot{u}_0, \ldots, x_0^{(\mu+1)}, u_0^{(\mu+1)}) = 0 \tag{6.43}$$

for $(x_0, u_0, \dot{x}_0, \dot{u}_0, \ldots, x_0^{(\mu+1)}, u_0^{(\mu+1)})$. Similar to Theorem 6.8, we expect quadratic convergence of the corresponding Gauß–Newton method.

If the output equation is included, then we must discretize (4.65). Applying a BDF method yields

$$\hat{F}_1(t_i, x_i, u_i, D_h x_i) = 0, \quad \hat{F}_2(t_i, x_i, u_i) = 0, \quad y_i = G(t_i, x_i). \tag{6.44}$$

The algebraic constraints (without the output relation) can again be replaced by (6.39). Choosing a suitable feedback matrix $\tilde{K}$ which yields a regularizing output feedback, we set

$$u_i = \tilde{K}y_i + w_i \tag{6.45}$$

and choose $w_i = u_{i-1} - \tilde{K}y_{i-1}$. Together with a suitable matrix $\tilde{Z}_1$, we obtain the problem

$$\begin{aligned} \tilde{Z}_1^T F(t_i, x_i, \tilde{K}y_i + w_i, D_h x_i) &= 0, \\ F_\mu(t_i, x_i, \tilde{K}y_i + w_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)}) &= 0, \\ y_i - G(t_i, x_i) &= 0 \end{aligned} \tag{6.46}$$

for $(x_i, y_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)})$. Because of the explicit form of the output equation, we can eliminate y_i. It then remains to solve

$$\begin{aligned} \tilde{Z}_1^T F(t_i, x_i, \tilde{K}G(t_i, x_i) + w_i, D_h x_i) &= 0, \\ F_\mu(t_i, x_i, \tilde{K}G(t_i, x_i) + w_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)}) &= 0 \end{aligned} \tag{6.47}$$

for $(x_i, \dot{x}_i, \dot{u}_i, \ldots, x_i^{(\mu+1)}, u_i^{(\mu+1)})$. By construction, the value x_i is the unique solution of

$$\hat{F}_1(t_i, x_i, \tilde{K}G(t_i, x_i) + w_i, D_h x_i) = 0, \quad \hat{F}_2(t_i, x_i, \tilde{K}G(t_i, x_i) + w_i) = 0. \tag{6.48}$$

Still, due to Theorem 6.5, we expect quadratic convergence of the Gauß–Newton method for sufficiently small h.

Concerning consistency, there is no difference to the case of state feedback, since we work with the same derivative array. Hence, we also solve (6.43) with quadratic convergence of the related Gauß–Newton method.

Having performed an integration step, we always end up with a new consistent value in $\mathbb{L}_\mu$, since in both cases the equation $F_\mu(z_\mu) = 0$ is part of the numerical procedure. Thus, we can iteratively proceed with the integration and obtain at least piecewise smooth regularizing controls and associated solutions.

6.4 Index reduction by minimal extension

As we have already discussed in Section 6.2, the general approach which works with full derivative arrays may be limited by memory requirements and by the fact that the quantities Z_1, Z_2, T_2 of Hypothesis 4.2 have to be computed in every integration step. However, in important applications like multibody systems ([79], [201]) or circuit simulation problems ([103], [104], [214]), the differential-algebraic equation has extra structure that can be used to determine the desired quantities without this large computational effort and memory requirement.

As we have presented the general method, the complete derivative array is used to determine an approximation $\tilde{Z}_1$ to the projection Z_1 at the actual point, and to compute the next integration step in (6.30). If, however, the structure of the problem allows to identify the equations that have to be differentiated, then we do not have to work with the complete derivative array but with a (possibly much) smaller system that replaces F_μ in (6.30). See Remark 6.9 on these so-called *reduced derivative arrays*. Using a reduced derivative array does not only reduce the computational effort in the solution of (6.30) but it also reduces the complexity of computing the projector Z_1 or an approximation to it, since we can replace the computation of Z_2 and T_2 from the Jacobian of F_μ by corresponding computations from the smaller Jacobian of $F_{\mu,\mathrm{red}}$.

An optimally small reduced derivative array is obtained when we are able to derive all hidden algebraic constraints of the given differential-algebraic equation analytically. We then just add these hidden constraints to the original differential-algebraic equation. While the whole derivative array consists of $(\mu + 1)n + d$ equations, the optimal reduced derivative array consists of at most $2n + d$ equations. If the number of algebraic constraints is small compared to the size of the original differential-algebraic equation, then this size is close to n.

But even with the reduction of computational work due to a reduced derivative array, the computation of the projectors Z_1, Z_2, T_2 may still be infeasible for large

scale systems. In these cases, the only hope is that the whole reduced differential-algebraic equation (6.23) can be obtained analytically.

In large scale applications, there may be another problem when dealing with the reduced differential-algebraic equation. The successful numerical treatment of large scale problems heavily relies on a utilizable structure of the arising linear subproblems. This structure typically is lost when we go over to the reduced problem. We are therefore interested in an index reduction method that allows to conserve certain structural properties of the given problem. A possible approach (at least for some important classes of applications) is based on the idea to achieve an index reduction by introducing some new variables, thus ending up with an extended system, i.e., with a system which contains more unknowns compared with the original and the reduced system. Such an approach was first suggested in [154] and modified to obtain a minimal extension in [130].

Since this approach must be adapted to the structure of the given problem, we restrict ourselves to the discussion of two classes of problems which are important in applications.

We first demonstrate this approach for mechanical multibody systems. The classical first order form of a multibody system, see, e.g., [79], [181], [196], is

$$\begin{aligned} \dot{p} &= v, \\ M(p)\dot{v} &= f(p,v) - g_p(p)^T\lambda, \\ g(p) &= 0, \end{aligned} \tag{6.49}$$

where p are the positions, v the velocities, $M(p)$ is the mass matrix, g describes the constraints and λ is the associated Lagrange multiplier. Under the usual assumptions that $M(p)$ is symmetric and positive definite and that the Jacobian $g_p(p)$ has full row rank, this system has strangeness index $\mu = 2$. In particular, it is a Hessenberg system with index $\nu = 3$, cp. Example 4.27.

A well-known index reduction technique is given by the Gear–Gupta–Leimkuhler stabilization [94], that couples the time derivative of the constraint equations via further Lagrange multipliers $\tilde{\lambda}$ into the dynamics according to

$$\begin{aligned} \dot{p} &= v - g_p(p)^T\tilde{\lambda}, \\ M(p)\dot{v} &= f(p,v) - g_p(p)^T\lambda, \\ 0 &= g(p), \\ 0 &= g_p(p)v. \end{aligned} \tag{6.50}$$

Thus, this stabilization introduces new variables and therefore yields an extended system, but this extended system is not strangeness-free. We have to perform one more differentiation of the constraint equations to obtain the optimal reduced

derivative array

$$\begin{aligned}
&\dot{p} = v,\\
&M(p)\dot{v} = f(p,v) - g_p(p)^T\lambda,\\
&0 = g(p),\\
&0 = g_p(p)v,\\
&0 = g_{pp}(p)(v,v) + g_p(p)\dot{v}.
\end{aligned} \tag{6.51}$$

To obtain a minimally extended strangeness-free system, we (locally) determine an orthogonal matrix U such that for the Jacobian matrix $g_p(p)$ we have

$$g_p(p)U = [\, G_1 \;\; G_2 \,], \tag{6.52}$$

with G_2 being square and nonsingular. We then partition

$$U^T p = \begin{bmatrix} p_1 \\ p_2 \end{bmatrix}, \quad U^T v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$$

conformably and replace every occurrence of $\dot{p}_2$ by a new variable $\hat{p}_2$ and every occurrence of $\dot{v}_2$ by a new variable $\hat{v}_2$. This gives the extended system

$$\dot{p}_1 = v_1, \tag{6.53a}$$
$$\hat{p}_2 = v_2, \tag{6.53b}$$
$$M(p)U \begin{bmatrix} \dot{v}_1 \\ \hat{v}_2 \end{bmatrix} = f(p,v) - g_p(p)^T\lambda, \tag{6.53c}$$
$$0 = g(p), \tag{6.53d}$$
$$0 = g_p(p)v, \tag{6.53e}$$
$$0 = g_{pp}(p)(v,v) + g_p(p)U \begin{bmatrix} \dot{v}_1 \\ \hat{v}_2 \end{bmatrix}. \tag{6.53f}$$

The following theorem shows that this system is strangeness-free.

Theorem 6.12. *Consider a multibody system of the form* (6.49) *with* $M(p)$ *symmetric and positive definite and suppose that* $g_p(p)$ *has full row rank. Then the extended system* (6.53) *is strangeness-free.*

Proof. Since G_2 in (6.52) is square nonsingular, we can (locally) solve (6.53d) by means of the implicit function theorem for p_2 in terms of p_1 and (6.53e) for v_2 in terms of p_1 and v_1. Since $M(p)$ is symmetric and positive definite, we can solve (6.53c) for $\dot{v}_1$ and $\hat{v}_2$. Moreover, $W(p) = g_p(p)M(p)^{-1}g_p(p)^T$ is symmetric and positive definite due to the full row rank of $g_p(p)$. Hence, we can eliminate $\dot{v}_1$ and $\hat{v}_2$ from (6.53f) and solve for λ according to

$$\lambda = W(p)^{-1}(g_{pp}(p)(v,v) + g_p(p)M(p)^{-1}f(p,v))$$

and we end up with an ordinary differential equation in the unknowns p_1 and v_1. Thus, the system has strangeness index $\mu = 0$. □

Example 6.13. Consider a multibody system describing the movement of a mass point restricted to a parabola under gravity from [190] given by

$$\begin{aligned} \dot{p}_1 &= v_1, & \dot{v}_1 &= 2\lambda p_1, \\ \dot{p}_2 &= v_2, & \dot{v}_2 &= 2\lambda p_2, \\ \dot{p}_3 &= v_3, & \dot{v}_3 &= -\lambda - 1, \\ 0 &= p_1^2 + p_2^2 - p_3. \end{aligned}$$

Here the coupling between p_3 and $\dot{p}_3$ causes a higher index. Differentiating the constraint once and eliminating the differentiated variables with the help of the other equations yields

$$0 = 2p_1v_1 + 2p_2v_2 - v_3.$$

The coupling between v_3 and $\dot{v}_3$ still causes a higher index. Differentiating once more and eliminating gives

$$0 = 2v_1^2 + 4\lambda p_1^2 + 2v_2^2 + 4\lambda p_2^2 + \lambda + 1.$$

A minimally extended strangeness-free system is obtained by adding the two derivatives of the constraint to the system and by replacing $\dot{p}_3$ and $\dot{v}_3$ say by $\hat{p}_3$ and $\hat{v}_3$, respectively. The system then reads

$$\begin{aligned} \dot{p}_1 &= v_1, & \dot{v}_1 &= 2\lambda p_1, \\ \dot{p}_2 &= v_2, & \dot{v}_2 &= 2\lambda p_2, \\ \hat{p}_3 &= v_3, & \hat{v}_3 &= -\lambda - 1, \\ 0 &= p_1^2 + p_2^2 - p_3, \\ 0 &= 2p_1v_1 + 2p_2v_2 - v_3, \\ 0 &= 2v_1^2 + 4\lambda p_1^2 + 2v_2^2 + 4\lambda p_2^2 + \lambda + 1. \end{aligned} \tag{6.54}$$

A reduced system is achieved by simply omitting the equations that involve the variables $\hat{p}_3$ and $\hat{v}_3$ which gives

$$\begin{aligned} \dot{p}_1 &= v_1, & \dot{v}_1 &= 2\lambda p_1, \\ \dot{p}_2 &= v_2, & \dot{v}_2 &= 2\lambda p_2, \\ 0 &= p_1^2 + p_2^2 - p_3, \\ 0 &= 2p_1v_1 + 2p_2v_2 - v_3, \\ 0 &= 2v_1^2 + 4\lambda p_1^2 + 2v_2^2 + 4\lambda p_2^2 + \lambda + 1. \end{aligned} \tag{6.55}$$

Numerical experiments show that working with the full derivative is about ten times slower than solving the analytically determined reduced problem (6.55) while solving the minimally extended system (6.54) is only about a factor two more expensive than solving (6.55). Note, however, that this is only a small example, where the utilization of the structure does not pay off so much.

A second class of structured problems, where the minimal extension approach can be employed successfully, is the simulation of electrical circuit equations. It is well-known which influence specific elements and their combination may have on the index, for a survey see [103], [104]. Furthermore, in [83], [84], [214] topological methods have been derived that analyze the network topology and show which equations are responsible for higher index. These methods also allow to derive (in a purely combinatorial way) projectors to filter out these equations from the system. We will briefly review these results, so that we can produce a minimal extension.

Let us denote by e the node potentials, by j_L and j_V the currents through inductances and voltage sources, respectively, by i and v the functions describing the current and voltage sources, respectively, by r the function describing the resistances, and finally by q_C and ϕ_L the functions describing the charges of the capacitances and the fluxes of the inductances, respectively. The modified nodal analysis leads to a quasi-linear system of differential-algebraic equations of the form

$$\begin{aligned} 0 &= A_C \tfrac{d}{dt} q_C(A_C^T e, t) + A_R r(A_R^T e, t) + A_L j_L \\ &\qquad + A_V j_V + A_I i(A^T e, \tfrac{d}{dt} q_C(A_C^T e, t), j_L, j_V, t), \\ 0 &= \tfrac{d}{dt} \phi_L(j_L, t) - A_L^T e, \\ 0 &= A_V^T e - v(A^T e, \tfrac{d}{dt} q_C(A_C^T e, t), j_L, j_V, t), \end{aligned} \tag{6.56}$$

where the matrix A that contains the information on the topology of the circuit is split as $[\, A_C \;\; A_L \;\; A_R \;\; A_V \;\; A_I \,]$, with A_C, A_L, A_R, A_V and A_I describing the branch-current relation for capacitive, inductive, resistive branches and branches for voltage sources and current sources, respectively.

In more detail, the matrix A is obtained in the following way. We first form the incidence matrix $\tilde{A} = [\tilde{a}_{k,l}]$ of the network graph, where the rows represent the nodes and the columns represent the branches of the network graph. There are exactly two nonvanishing entries in every column of $\tilde{A}$ namely $\tilde{a}_{k_1,l} = 1$ and $\tilde{a}_{k_2,l} = -1$ if the l-th branch connects node k_1 with node k_2. Thus, the network graph is a directed graph, since we must define a positive direction for the currents. The matrix A is then obtained by discarding the row of $\tilde{A}$ associated with the zero potential.

For the conventional modified nodal analysis, the vector of unknown variables consists of all node potentials e and all branch currents j_L, j_V of current-controlled

elements. Introducing new functions

$$C(u,t) = \tfrac{\partial}{\partial u} q_C(u,t), \quad L(j,t) = \tfrac{\partial}{\partial j}\phi_L(j,t),$$

and forming the partial derivatives

$$q_t(u,t) = \tfrac{\partial}{\partial t} q_C(u,t), \quad \phi_t(j,t) = \tfrac{\partial}{\partial t}\phi_L(j,t),$$

the system is reformulated as

$$\begin{aligned}
0 &= A_C C(A_C^T e,t) A_C^T \tfrac{d}{dt} e + A_C q_t(A_C^T e,t) \\
&\quad + A_R r(A_R^T e,t) + A_L j_L + A_V j_V \\
&\quad + A_I i(A^T e, C(A_C^T e,t) A_C^T \tfrac{d}{dt} e + A_C q_t(A_C^T e,t), j_L, j_V, t), \\
0 &= L(j_L,t)\tfrac{d}{dt} j_L + \phi_t(j_L,t) - A_L^T e, \\
0 &= A_V^T e - v(A^T e, C(A_C^T e,t) A_C^T \tfrac{d}{dt} e + A_C q_t(A_C^T e,t), j_L, j_V, t).
\end{aligned} \tag{6.57}$$

In the charge/flux oriented modified nodal analysis, the vector of unknowns is extended by the charges q of capacitances and the fluxes ϕ of inductances. Including the original voltage-charge and current-flux equations in the system yields the differential-algebraic equation

$$\begin{aligned}
0 &= A_C \tfrac{d}{dt} q + A_R r(A_R^T e,t) + A_L j_L \\
&\quad + A_V j_V + A_I i(A^T e, \tfrac{d}{dt} q, j_L, j_V, t), \\
0 &= \tfrac{d}{dt}\phi - A_L^T e, \\
0 &= A_V^T e - v(A^T e, \tfrac{d}{dt} q, j_L, j_V, t), \\
0 &= q - q_C(A_C^T e,t), \\
0 &= \phi - \phi_L(j_L,t).
\end{aligned} \tag{6.58}$$

Denote by the matrices Q_C, Q_{V-C}, Q_{R-CV}, and $\bar{Q}_{V-C}$ projections onto kernel A_C^T, kernel $Q_C^T A_V$, kernel $Q_{V-C}^T Q_C^T A_R$, and kernel $A_V^T Q_C$, respectively, and by Q_{CRV} a projection onto the intersection of the first three of these kernels. These constant projection matrices can be obtained very cheaply by purely topological analysis of the network. For the conventional modified nodal analysis (6.57), the equations that are responsible for a nonvanishing strangeness index are then given by the projected equations

$$\begin{aligned}
0 &= Q_{CRV}^T(A_L j_L + A_I i(\cdot)), \\
0 &= \bar{Q}_{V-C}^T(A_V^T e - v(\cdot)).
\end{aligned} \tag{6.59}$$

It follows that the equations in (6.57) together with the derivatives of (6.59)

$$\begin{aligned}
0 &= Q_{CRV}^T(A_L \tfrac{d}{dt} j_L + A_I \tfrac{d}{dt} i(\cdot)), \\
0 &= \bar{Q}_{V-C}^T(A_V^T \tfrac{d}{dt} e - \tfrac{d}{dt} v(\cdot)).
\end{aligned} \tag{6.60}$$

form a reduced derivative array.

To construct a minimal extension, we determine nonsingular matrices Π_j, Π_e such that

$$Q_{CRV}^T A_L \Pi_j^{-1} = [\, J_1 \;\; 0 \,], \quad \bar{Q}_{V-C}^T A_V^T \Pi_e^{-1} = [\, F_1 \;\; 0 \,]$$

with J_1, F_1 square nonsingular. Since $Q_{CRV}^T A_L$ and $\bar{Q}_{V-C}^T A_V^T$ are still only incidence-like matrices (containing topological information on the circuit in form of integers) the computation of Π_j, Π_e and their inverses is possible with very small computational effort and very accurately. We partition

$$\tilde{j}_L = \Pi_j j_L = \begin{bmatrix} \tilde{j}_{L_1} \\ \tilde{j}_{L_2} \end{bmatrix}, \quad \tilde{e}_L = \Pi_e e = \begin{bmatrix} \tilde{e}_1 \\ \tilde{e}_2 \end{bmatrix}$$

conformably and introduce new variables

$$\hat{e}_1 = \tfrac{d}{dt}\tilde{e}_1, \quad \hat{j}_1 = \tfrac{d}{dt}\tilde{j}_{L_1}. \tag{6.61}$$

It has been shown in [130] that the minimally extended strangeness-free system for the conventional modified nodal analysis is given by the system

$$\begin{aligned}
0 &= A_C C(A_C^T \Pi_e^{-1}\tilde{e}, t) A_C^T \Pi_e^{-1} \begin{bmatrix} \hat{e}_1 \\ \frac{d}{dt}\tilde{e}_2 \end{bmatrix} + A_C q_t(A_C^T \Pi_e^{-1}\tilde{e}, t) \\
&\quad + A_R r(A_R^T \Pi_e^{-1}\tilde{e}, t) + A_L \Pi_j^{-1}\tilde{j}_L + A_V j_V + A_I i(\cdot), \\
0 &= L(j_L, t)\Pi_j^{-1} \begin{bmatrix} \hat{j}_1 \\ \frac{d}{dt}\tilde{j}_{L_2} \end{bmatrix} + \phi_t(\Pi_j^{-1}\tilde{j}_L, t) - A_L^T \Pi_e^{-1}\tilde{e}, \\
0 &= A_V^T \Pi_e^{-1}\tilde{e} - v(\cdot), \\
0 &= Q_{CRV}^T (A_L \Pi_j^{-1} \begin{bmatrix} \hat{j}_1 \\ \frac{d}{dt}\tilde{j}_{L_2} \end{bmatrix} + \tfrac{d}{dt} i(\cdot)), \\
0 &= \bar{Q}_{V-C}^T (A_V^T \Pi_e^{-1} \begin{bmatrix} \hat{e}_1 \\ \frac{d}{dt}\tilde{e}_2 \end{bmatrix} - \tfrac{d}{dt} v(\cdot)).
\end{aligned} \tag{6.62}$$

Remark 6.14. If the original system has size n and there are l equations in (6.59) then the extended system has size $n + l$. Since typically l is much smaller than n, the extended system is only slightly larger than the original system.

For the charge/flux oriented modified nodal analysis (6.58), the equations that are responsible for a nonvanishing strangeness are given by the projected equations in (6.59) together with the last two equations in (6.58).

Using the replacements as in (6.61) and in addition

$$\hat{q} = \tfrac{d}{dt} q, \quad \hat{\phi} = \tfrac{d}{dt}\phi, \tag{6.63}$$

we obtain the following minimally extended strangeness-free system.

$$
\begin{aligned}
0 &= A_C\hat{q} + A_R r(A_R^T\Pi_e^{-1}\tilde{e}, t) + A_L\Pi_j^{-1}\tilde{j}_L + A_V j_V + A_I i(\cdot),\\
0 &= \hat{\phi} - A_L^T\Pi_e^{-1}\tilde{e},\\
0 &= A_V^T\Pi_e^{-1}\tilde{e} - v(\cdot),\\
0 &= q - q_C(A_C^T\Pi_e^{-1}\tilde{e}, t),\\
0 &= \phi - \phi_L(\Pi_j^{-1}\tilde{j}_L, t),\\
0 &= Q_{CRV}^T(A_L\Pi_j^{-1}\begin{bmatrix}\hat{j}_1\\ \frac{d}{dt}\tilde{j}_{L_2}\end{bmatrix} + \frac{d}{dt}i(\cdot)),\\
0 &= \bar{Q}_{V-C}^T(A_V^T\Pi_e^{-1}\begin{bmatrix}\hat{e}_1\\ \frac{d}{dt}\tilde{e}_2\end{bmatrix} - \frac{d}{dt}v(\cdot)),\\
0 &= \hat{q} - C(A_C^T\Pi_e^{-1}\tilde{e}, t)A_C^T\Pi_e^{-1}\begin{bmatrix}\hat{e}_1\\ \frac{d}{dt}\tilde{e}_2\end{bmatrix} + q_t(A_C^T\Pi_e^{-1}\tilde{e}, t),\\
0 &= \hat{\phi} - L(\Pi_j^{-1}\tilde{j}_L, t)\Pi_j^{-1}\begin{bmatrix}\hat{j}_1\\ \frac{d}{dt}\tilde{j}_{L_2}\end{bmatrix} + \phi_t(\Pi_j^{-1}\tilde{j}_L, t).
\end{aligned}
\tag{6.64}
$$

Obviously, we can use the last two relations to eliminate the just introduced variables $\hat{q}$ and $\hat{\phi}$ obtaining just the minimally extended system (6.62) for the conventional modified nodal analysis. Hence, system (6.64) is strangeness-free as well. Moreover, from a numerical point of view the reduced problems and minimally extended systems belonging to the conventional and charge oriented modified nodal analysis are the same or at least equivalent (in the sense that the common part of the numerical solution would be same when using the same stepsizes and ignoring roundoff errors). Concerning efficiency, however, it should be noted that in the charge oriented modified nodal analysis the minimally extended strangeness-free system is often significantly larger than the original system.

Remark 6.15. The presented minimal extensions allow to preserve certain symmetry properties of the original higher index formulation. This symmetry becomes important when linear subproblems in large scale problems must be solved iteratively. See [17] for details.

Remark 6.16. Circuit simulation packages do not work directly with the equations but rather with netlists that represent these equations. It has been shown in [18], [19] how the equations (6.60) that are added to the system can be obtained by replacing capacitances and inductances by new network elements, so that this process of index reduction can be incorporated directly into existing packages. Using this technique, it is furthermore possible to remove some equations and variables so that the resulting strangeness-free system even has the same number of variables and equations as the original system, see [18], [19] and the following example.

Example 6.17. Consider the circuit in Figure 6.1 which has a loop that consists of two capacitances and a voltage source. The equations of the modified nodal analysis for this circuit with values $R_1 = R_2 = 1$ and $C_1 = C_2 = 1$ are given by

$$0 = \begin{bmatrix} -1 & 1 & 0 \\ 1 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} e_1 \\ e_2 \\ j_V \end{bmatrix} + \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ j_V \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -v(t) \end{bmatrix}.$$

One can show that the strangeness index of this system is $\mu = 1$, cp. Exercise 15, and that the third equation is responsible for the higher index. Differentiating this

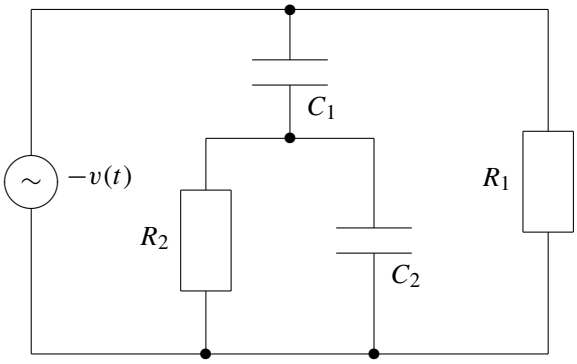


Figure 6.1. Circuit with strangeness-index $\mu = 1$

equation, adding it to the system and introducing the new variable $\hat{e}_1 = \frac{d}{dt} e_1$ yields the minimally extended system

$$0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} e_1 \\ e_2 \\ j_V \\ \hat{e}_1 \end{bmatrix} + \begin{bmatrix} -1 & 0 & -1 & -1 \\ 0 & -1 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ j_V \\ \hat{e}_1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ -v(t) \\ -\dot{v}(t) \end{bmatrix}.$$

Carrying out the procedure described in [18], [19], which consists essentially in eliminating again the newly introduced variable $\hat{e}_1$ and dropping the last equation, we obtain the system

$$0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 0 \end{bmatrix} \frac{d}{dt} \begin{bmatrix} e_1 \\ e_2 \\ j_V \end{bmatrix} + \begin{bmatrix} -1 & 0 & -1 \\ 0 & -1 & 0 \\ -1 & 0 & 0 \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ j_V \end{bmatrix} + \begin{bmatrix} \dot{v}(t) \\ -\dot{v}(t) \\ -v(t) \end{bmatrix}$$

of the original size and with the original variables.

Bibliographical remarks

Different methods for index computation and index reduction have been widely studied in the literature. Before the work of Gear [90] and often still today, the method to solve differential-algebraic equations was to use differentiation and transformation to turn the system into a system of ordinary differential equations. For linear constant coefficient systems this can be obtained via the computation of the Kronecker canonical form or rather its variations under unitary transformations, see [69], [70], [216], [226].

Approaches for index reduction without reducing to an ordinary differential equation were introduced in [45], [91], [92], [145] using derivatives of the constraint equations or in [100] using the concept of matrix chains. Index reduction for multibody systems is studied in [3], [4], [24], [27], [55], [79], [94], [205], for circuit simulation in [17], [18], [19], [63], [84], [103], [104], [130], [213], and for chemical engineering in [20], [61], [162].

A breakthrough in index reduction methods came through the work on derivative arrays [46], [48], [50], [51], [58]. This approach also forms the basis for the methods that we have presented and that were derived in [126], [128] for regular systems and extended to general over- and underdetermined systems and control systems in [39], [40], [116], [117], [129], [132], [183]. These methods are also used for symbolic computation [59], [134], [171], [172].

Index reduction by introducing dummy variables in derivative arrays has been introduced in [154] and modified in [130]. It is often based on the computation of a generic or structural index, see for example [76], [161], [188], [189]. Differential-geometric approaches for index reduction and integration methods are discussed for example in [166], [169], [170], [175], [178], [181], [182]. Regularization and index reduction via feedback in control problems has been discussed for example in [32], [33], [34], [62], [129], [132].

Exercises

1. Determine analytically an equivalent strangeness-free system for the differential-algebraic equation

$$\begin{bmatrix} 0 & 0 \\ 1 & \eta t \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -1 & -\eta t \\ 0 & -(1+\eta) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \sin(t) \\ \exp(t) \end{bmatrix}.$$

   What happens if we solve the so obtained problem by means of the implicit Euler method? Compare with Example 5.15.

2. Implement the index reduction procedure of Section 6.1 in the following way. Suppose that the characteristic values μ, a, and d as well as the functions M_μ, N_μ, and g_μ are

supplied. In order to compute suitable matrices $\tilde{Z}_2$, $\tilde{T}_2$, and $\tilde{Z}_1$ for given t_i, use an appropriate QR-decomposition of the corresponding matrices. Test your program with the help of the problem of Example 1.

3. Combine the implemented index reduction of Exercise 2 with the implicit Euler method using a constant stepsize. Test your program with the help of the problem of Example 1.
4. Apply your program of Exercise 3 to the differential-algebraic equation

$$\begin{bmatrix} 0 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 \\ \sin(t) \end{bmatrix}.$$

5. Consider the problem

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \begin{bmatrix} x_1(0) \\ x_2(0) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

Obviously, the given initial value is not consistent. Determine a consistent initial value using (6.18) as well as (6.21), choosing in the latter case the part A_{22} of A_2 as that belonging to x_2.

6. Implement the ordinary Gauß–Newton method for the solution of a given underdetermined system $F_\mu(t_0, x_0, y_0) = 0$ with respect to (x_0, y_0). Use finite differences to approximate the necessary Jacobians. Apply your program to the problem

$$\begin{bmatrix} 1 & 0 \\ \dot{x}_1 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_2 - x_1 - \sin(t) \\ x_2 - \exp(t) \end{bmatrix} = 0.$$

with $\mu = 1$ in order to determine consistent initial values.

7. Extend your program of Exercise 6 by the computation of a suitable matrix $\tilde{Z}_1$ at a computed solution of $F_\mu(t_0, x_0, y_0) = 0$, supposing that a and d are given. Test your program with the help of the problem of Example 1.
8. Extend your program of Exercise 7 by the implicit Euler method according to (6.30), using a constant stepsize. Test your program with the help of the problem of Example 1.
9. Apply your program of Exercise 8 to the differential-algebraic equation of Exercise 6.
10. Show that the differential-algebraic equation

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_1, \quad x_1\dot{x}_4 + x_3 = 1, \quad x_4 = 2$$

satisfies Hypothesis 4.2 with $\mu = 1$, $a = 2$, and $d = 2$. Apply your program of Exercise 8 for various initial guesses $(\tilde{x}_0, \tilde{y}_0)$ such as

$$\tilde{x}_0 = (0, 0, 1, 2), \quad \tilde{y}_0 = (0, 0, 0, 0, 0, 0, 0, 0),$$

or

$$\tilde{x}_0 = (0, 1, 1, 2), \quad \tilde{y}_0 = (1, 0, 0, 0, 0, 0, 0, 0).$$

11. Show that the differential-algebraic equation

$$\dot{x}_1 = 1, \quad \dot{x}_2 = (\exp(x_3 - 1) + 1)/2, \quad x_2 - x_1 = 0$$

satisfies Hypothesis 4.2 with $\mu = 1$, $a = 2$, and $d = 1$. Apply your program of Exercise 8.

12. Consider the control problem

$$\begin{aligned} I_R \ddot{\varphi} &= u, \\ m_G \ddot{z}_G + d_1(\dot{z}_G - \dot{z}_Z) + c_1(z_G - z_Z) &= \lambda, \\ m_Z \ddot{z}_Z + d_1(\dot{z}_Z - \dot{z}_G) + c_1(z_Z - z_G) &= 0, \\ \dot{z}_G &= v_U \varphi, \end{aligned}$$

modeling a multibody system with generalized positions φ, z_G, z_Z and a non-holonomic constraint, i.e., with a constraint that involves first derivatives of the generalized positions. Written as a first order system, the (scalar) unknowns are given by φ, z_G, z_Z, $\dot{\varphi}$, $\dot{z}_G$, $\dot{z}_Z$, λ, u, where λ is the Lagrange multiplier belonging to the constraint and u is the control. Show that for any proportional state feedback the resulting closed loop system satisfies Hypothesis 3.48 with $\mu = 1$, $a = 2$, and $d = 5$.

13. Apply your program of Exercise 3 to the differential-algebraic equation of Exercise 12 choosing the control u according to $u(t) = 0.001$ and the parameters according to $I_R = 0.002$, $m_G = 3$, $m_Z = 10$, $c_1 = 250$, $d_1 = 10$, and $v_U = 2.8$.

14. Derive a strangeness-free differential-algebraic equation that is equivalent to the problem of Exercise 13 and apply your program of Exercise 3.

15. Show that the circuit problem in Example 6.17 possesses the strangeness index $\mu = 1$.

16. Determine the analytic solution for the circuit problem in Example 6.17 for the choice $v(t) = \sin(100t)$.

17. Derive a strangeness-free formulation of the form (5.50) for the circuit problem in Example 6.17 and solve the problem numerically on the interval $[0.0, 0.1]$ by one of the implementations of Chapter 5 choosing $v(t) = \sin(100t)$ and consistent initial values with $e_2(0) = 0$.

18. Apply your program of Exercise 3 to the circuit problem in Example 6.17 in the original formulation with $\mu = 1$, $d = 1$, and $a = 2$, in the minimally extended strangeness-free formulation with $\mu = 0$, $d = 1$, and $a = 3$, and in the strangeness-free formulation of Example 17 with $\mu = 0$, $d = 1$, and $a = 2$. Compare the results.

Chapter 7
Boundary value problems

In this chapter, we study general nonlinear boundary value problems for differential-algebraic equations, i.e., problems of the form

$$F(t, x, \dot{x}) = 0, \tag{7.1a}$$
$$b(x(\underline{t}), x(\bar{t})) = 0 \tag{7.1b}$$

in an interval $\mathbb{I} = [\underline{t}, \bar{t}] \subset \mathbb{R}$. In view of Chapter 4, we assume that $F\colon \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{R}^n$ with $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ open, satisfies Hypothesis 4.2. The boundary conditions are described by a function $b\colon \mathbb{D}_x \times \mathbb{D}_x \to \mathbb{R}^d$. In particular, the number of (scalar) boundary conditions coincides with the number d of differential equations imposed by (7.1a). As in the previous chapters, we require that all involved functions are sufficiently often continuously differentiable. For ease of notation, we write tuples also as columns and vice versa. Again, we only discuss real problems, since we can consider the real and imaginary part of a complex problem separately.

Example 7.1. In [108], the model of a periodically driven electronic amplifier is given. The equations in the unknowns $U_1, \dots, U_5$ have the form

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

with

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

The problem can be shown to satisfy Hypothesis 4.2 with $\mu = 0$, $a = 2$, and $d = 3$. If we ask for the periodic response of the amplifier, a possible set of boundary conditions is given by

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

In the area of ordinary differential equations, there are two major approaches to solve boundary value problems, namely shooting techniques and collocation methods. In this chapter, we generalize both approaches to differential-algebraic equations. But at first, we discuss existence and uniqueness of solutions of boundary value problems for differential-algebraic equations.

7.1 Existence and uniqueness of solutions

In the context of general nonlinear problems, the existence of solutions is usually shown via the application of fix point theorems or Newton–Kantorovich-like convergence theorems for iterative methods. Typically, the assumptions that have to be made are very strong and difficult to verify. A standard ingredient in the investigation of numerical methods for boundary value problems is the assumption that a solution of the given problem does exist. Let F_μ be the derivative array associated with (7.1a). According to Remark 4.15, we therefore require that there exists a sufficiently smooth solution $x^* \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (7.1) in the sense that

$$F(t, x^*(t), \dot{x}^*(t)) = 0 \quad \text{for all } t \in \mathbb{I}, \tag{7.2a}$$
$$F_\mu(t, x^*(t), \mathcal{P}(t)) = 0 \quad \text{for all } t \in \mathbb{I}, \tag{7.2b}$$
$$b(x^*(\underline{t}), x^*(\bar{t})) = 0, \tag{7.2c}$$

where $\mathcal{P} : \mathbb{I} \to \mathbb{R}^{(\mu+1)n}$ is some smooth function arising from the parameterization of the solution set $\mathbb{L}_\mu$ that coincides with $\dot{x}^*$ in the first n components.

Restricting the functions Z_1, Z_2 and T_2 of Hypothesis 4.2 to the path $(t, x^*(t), \mathcal{P}(t))$, which lies in $\mathbb{L}_\mu$ due to (7.2b), we obtain functions

$$Z_1 : \mathbb{I} \to \mathbb{R}^{n,d}, \quad Z_2 : \mathbb{I} \to \mathbb{R}^{(\mu+1)n,a}, \quad T_2 : \mathbb{I} \to \mathbb{R}^{n,d}, \tag{7.3}$$

using the same notation for the restricted functions. Note that due to the constant rank assumptions in Hypothesis 4.2, these functions can be chosen to be smooth on the whole interval $\mathbb{I}$. By definition, they satisfy

$$Z_2(t)^T F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(t, x^*(t), \mathcal{P}(t)) = 0 \quad \text{for all } t \in \mathbb{I}, \tag{7.4a}$$
$$Z_2(t)^T F_{\mu;x}(t, x^*(t), \mathcal{P}(t)) T_2(t) = 0 \quad \text{for all } t \in \mathbb{I}, \tag{7.4b}$$
$$\operatorname{rank} Z_1(t)^T F_{\dot{x}}(t, x^*(t), \dot{x}^*(t)) T_2(t) = d \quad \text{for all } t \in \mathbb{I}. \tag{7.4c}$$

In addition, there exist smooth functions

$$\begin{aligned} Z_2' &: \mathbb{I} \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, & T_1 &: \mathbb{I} \to \mathbb{R}^{(\mu+1)n,a}, \\ T_2' &: \mathbb{I} \to \mathbb{R}^{n,a}, & T_1' &: \mathbb{I} \to \mathbb{R}^{(\mu+1)n,(\mu+1)n-a}, \end{aligned} \tag{7.5}$$

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

$$Z_2'(t)^T F_{\mu;\dot{x},\dots,x^{(\mu+1)}}(t, x^*(t), \mathcal{P}(t)) T_1(t) = 0 \quad \text{for all } t \in \mathbb{I}. \tag{7.6}$$

Following Section 4.1, we know that for every $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ in a neighborhood of $(t_0, x^*(t_0), \mathcal{P}(t_0))$, the differential-algebraic equation (7.1a) is locally solvable and thus defines a function x from a neighborhood of t_0 into $\mathbb{R}^n$. This solution can be extended until the boundary of the domain of F_μ is reached. Since

$$(t_0, x^*(t_0), \mathcal{P}(t_0)) \in \mathbb{L}_\mu, \quad t_0 \in \mathbb{I}$$

defines a solution on the whole interval $\mathbb{I}$, the same holds for every $(t_0, x_0, y_0) \in \mathbb{L}_\mu$ in a sufficiently small neighborhood of $(t_0, x^*(t_0), \mathcal{P}(t_0))$. Thus, the solution x of (7.1a) can locally be seen as a function of (t_0, x_0, y_0). In this way, the value $x(\bar{t})$ becomes a function of $(\underline{t}, \underline{x}, \underline{y}) \in \mathbb{L}_\mu$ in a neighborhood of $(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t}))$. This means that also the boundary condition of a boundary value problem becomes a function of $(\underline{t}, \underline{x}, \underline{y})$.

To take the restriction of $(\underline{t}, \underline{x}, \underline{y})$ to $\mathbb{L}_\mu$ and the non-uniqueness of the parameterization $\mathcal{P}$ into account, we locally define a (nonlinear) projection onto $\mathbb{L}_\mu$ by considering the nonlinear system

$$F_\mu(\underline{t}, \hat{x}, \hat{y}) = 0, \tag{7.7a}$$
$$T_2(\underline{t})^T(\hat{x} - x) = 0, \tag{7.7b}$$
$$T_1(\underline{t})^T(\hat{y} - y) = 0, \tag{7.7c}$$

in the unknowns $(x, y, \hat{x}, \hat{y})$. If we write (7.7) as

$$H(x, y, \hat{x}, \hat{y}) = 0, \tag{7.8}$$

then a solution of this system is given by $(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t}))$. Since the Jacobian of H with respect to $\hat{x}$, $\hat{y}$ satisfies

$$\begin{aligned}
&\operatorname{rank} H_{\hat{x},\hat{y}}(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t})) \\
&\quad = \operatorname{rank} \begin{bmatrix} F_{\mu;x}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) & F_{\mu;y}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix} \\
&\quad = \operatorname{rank} \begin{bmatrix} Z_2'(\underline{t})^T F_{\mu;x}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) & Z_2'(\underline{t})^T F_{\mu;y}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) \\ Z_2(\underline{t})^T F_{\mu;x}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) & 0 \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix},
\end{aligned}$$

and since by construction the matrices

$$\begin{bmatrix} Z_2'(t)^T F_{\mu;y}(t, x^*(t), \mathcal{P}(t)) \\ T_1(t)^T \end{bmatrix}, \quad \begin{bmatrix} Z_2(t)^T F_{\mu;x}(t, x^*(t), \mathcal{P}(t)) \\ T_2(t)^T \end{bmatrix}$$

are nonsingular for all $t \in \mathbb{I}$, it follows that $H_{\hat{x},\hat{y}}(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t}))$ is nonsingular. We can therefore solve locally for $(\hat{x}, \hat{y})$ and obtain a function S according to

$$(\hat{x}, \hat{y}) = S(x, y). \tag{7.9}$$

Since $F_\mu(\underline{t}, S(x, y)) = 0$, we have that $(\underline{t}, S(x, y)) \in \mathbb{L}_\mu$ for every (x, y) in a neighborhood of $(x^*(\underline{t}), \mathcal{P}(\underline{t}))$. Observing that the initial value problem for (7.1a) together with $(\underline{t}, S(x^*(\underline{t}), \mathcal{P}(\underline{t}))) \in \mathbb{L}_\mu$ is solvable on the whole interval $\mathbb{I}$, the initial value problem remains solvable on the whole interval $\mathbb{I}$ with an initial condition given by $(\underline{t}, \underline{x}, \underline{y})$ from a neighborhood $\mathbb{V} = \mathbb{L}_\mu \cap \tilde{\mathbb{V}}$ of $(\underline{t}, x^*(\underline{t}), P(\underline{t}))$, cp. the notation of Section 4.1. Thus, the differential-algebraic equation defines a flow

$$\Phi : \mathbb{V} \to \mathbb{R}^n, \tag{7.10}$$

that maps $(\underline{t}, \underline{x}, \underline{y}) \in \mathbb{V}$ to the final value $x(\bar{t})$ of the solution x of the associated initial value problem, cp. Figure 7.1. Since the value $\underline{t}$ is kept fixed during the

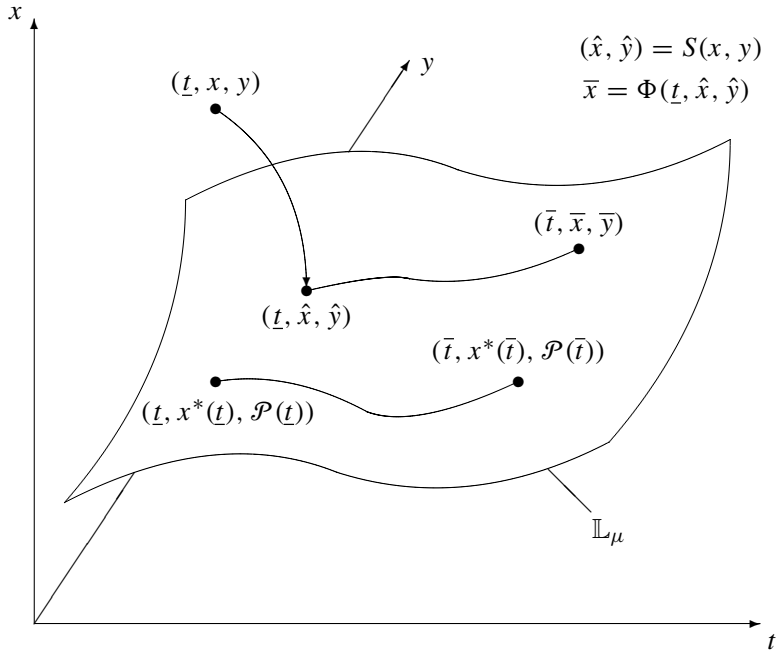


Figure 7.1. Construction of a (local) flow

following discussion, we will, for convenience, no longer state $\underline{t}$ in the argument list of Φ.

For later use, we will need the derivatives of S in (7.9) at $(x^*(\underline{t}), \mathcal{P}(\underline{t}))$. These are given by the solution of the linear system

$$\begin{aligned} &H_{\hat{x},\hat{y}}(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t}))S_{x,y}(x^*(\underline{t}), \mathcal{P}(\underline{t})) \\ &\quad = -H_{x,y}(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t})), \end{aligned}$$

i.e.,

$$\begin{bmatrix} F_{\mu,x}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) & F_{\mu;y}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t})) \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix} S_{x,y}(x^*(\underline{t}), \mathcal{P}(\underline{t})) = \begin{bmatrix} 0 & 0 \\ T_2(\underline{t})^T & 0 \\ 0 & T_1(\underline{t})^T \end{bmatrix}.$$

Let W be a matrix with orthonormal columns that span kernel $F_{\mu;x,y}(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t}))$. Setting

$$\tilde{W} = \begin{bmatrix} T_2(\underline{t}) & 0 \\ 0 & T_1(\underline{t}) \end{bmatrix}, \tag{7.11}$$

we see that $\tilde{W}^T W$ is nonsingular, since $H_{\hat{x},\hat{y}}(x^*(\underline{t}), \mathcal{P}(\underline{t}), x^*(\underline{t}), \mathcal{P}(\underline{t}))$ is nonsingular, and we get

$$S_{x,y}(x^*(\underline{t}), \mathcal{P}(\underline{t})) = W(\tilde{W}^T W)^{-1}\tilde{W}^T. \tag{7.12}$$

We then have the following theorem on the local uniqueness of solutions of boundary value problems for differential-algebraic equations.

Theorem 7.2. *The function x^* in (7.2) is a locally unique solution of the boundary value problem (7.1) in the sense that $(x^*(\underline{t}), \mathcal{P}(\underline{t}))$ is a solution of*

$$F_\mu(\underline{t}, \underline{x}, \underline{y}) = 0, \tag{7.13a}$$

$$T_1(\underline{t})^T(\underline{y} - \mathcal{P}(\underline{t})) = 0, \tag{7.13b}$$

$$b(\underline{x}, \Phi(S(\underline{x}, \underline{y}))) = 0, \tag{7.13c}$$

with nonsingular Jacobian, if and only if

$$\mathcal{E} = \big(C + D\Phi_{x,y}(x^*(\underline{t}), \mathcal{P}(\underline{t}))S_x(x^*(\underline{t}), \mathcal{P}(\underline{t}))\big)\, T_2(\underline{t}) \tag{7.14}$$

is nonsingular, where $C = b_{x_l}(x^*(\underline{t}), x^*(\bar{t}))$ and $D = b_{x_r}(x^*(\underline{t}), x^*(\bar{t}))$, the subscripts x_l and x_r denoting differentiation of b with respect to its first and second argument.

Proof. Obviously, $(\underline{x}, \underline{y}) = (x^*(\underline{t}), \mathcal{P}(\underline{t}))$ is a solution of (7.13). Moreover, the Jacobian J of (7.13) is given by

$$J = \begin{bmatrix} F_{\mu;x} & F_{\mu;y} \\ 0 & T_1^T \\ C + D\Phi_{x,y} S_x & D\Phi_{x,y} S_y \end{bmatrix},$$

where we have omitted the arguments $(\underline{t}, x^*(\underline{t}), \mathcal{P}(\underline{t}))$. We then get that the rank of J is equal to the rank of

$$\begin{bmatrix} Z_2'^T F_{\mu;x} T_2' & Z_2'^T F_{\mu;x} T_2 & Z_2'^T F_{\mu;y} T_1' & 0 \\ Z_2^T F_{\mu;x} T_2' & 0 & 0 & 0 \\ 0 & 0 & 0 & I \\ (C + D\Phi_{x,y} S_x) T_2' & (C + D\Phi_{x,y} S_x) T_2 & D\Phi_{x,y} S_y T_1' & D\Phi_{x,y} S_y T_1 \end{bmatrix}.$$

Since

$$S_x = W(\tilde{W}^T W)^{-1} \begin{bmatrix} T_2^T \\ 0 \end{bmatrix}, \quad S_y = W(\tilde{W}^T W)^{-1} \begin{bmatrix} 0 \\ T_1^T \end{bmatrix}$$

by (7.12), we have $S_x T_2' = 0$ and $S_y T_1' = 0$ from the definition of T_1' and T_2'. Moreover, $Z_2^T F_{\mu;x} T_2'$ and $Z_2'^T F_{\mu;y} T_1'$ are nonsingular by construction. Thus, J is nonsingular if and only if

$$\mathcal{E} = (C + D\Phi_{x,y} S_x) T_2$$

is nonsingular. □

System (7.13) is a straightforward generalization of the so-called (single) shooting method known from the treatment of boundary value problems for ordinary differential equations, see, e.g., [8]. The only difference is that here we must guarantee that initial conditions are consistent in order to fix a (unique) solution of the differential-algebraic equation.

Example 7.3. The classical shooting problem was to fire a cannon at a given target. The task then was to align the cannon in order to hit the target. Ignoring ballistic effects, we can model this problem by

$$\ddot{x} = 0, \quad \ddot{y} = -g,$$

where x, y are the Cartesian coordinates of the cannon ball and g is the gravity constant. At the beginning, the cannon together with the cannon ball is assumed to be located at the origin, whereas the target is located in a distance of L at a height H. After an unknown flight time T, the cannon ball is wanted to be in the same place

as the target. Finally, the initial velocity of the cannon ball shall be known to be some given v. This leads to the boundary conditions

$$x(0) = 0, \quad y(0) = 0, \quad \dot{x}(0)^2 + \dot{y}(0)^2 = v^2,$$
$$x(T) = L, \quad y(T) = H.$$

In the spirit of the shooting method, we solve the differential equations in terms of the initial values

$$x(0) = x_0, \quad \dot{x}(0) = \dot{x}_0, \quad y(0) = y_0, \quad \dot{y}(0) = \dot{y}_0.$$

We thus obtain

$$x(t) = x_0 + \dot{x}_0 t, \quad y(t) = y_0 + \dot{y}_0 t - \tfrac{1}{2} g t^2.$$

Utilizing these representations in the boundary conditions gives

$$x_0 = 0, \quad y_0 = 0, \quad \dot{x}_0^2 + \dot{y}_0^2 = v^2,$$
$$x_0 + \dot{x}_0 T = L, \quad y_0 + \dot{y}_0 T - \tfrac{1}{2} g T^2 = H,$$

which constitutes a nonlinear system of equations for the determination of x_0, $\dot{x}_0$, y_0, $\dot{y}_0$, T. Every regular solution of this system, i.e., every solution where the Jacobian is nonsingular, provides a regular solution of the boundary value problem.

Of course, we would have had to transform the unknown interval $[0, T]$ to a fixed interval $[0, 1]$ by scaling time and to introduce a trivial differential equation $\dot{T} = 0$ in order to bring this example into the form (7.1a). But this would not have essentially altered the shooting approach and the related computations.

Remark 7.4. In the case of linear boundary value problems, i.e., problems (7.1), where F and b are linear, local existence and uniqueness immediately yields global existence and uniqueness of the solution x^*, as it is typical in all (continuous) linear problems.

7.2 Multiple shooting

Although the previous section was only intended to give sufficient conditions for a solution of (7.1) to be locally unique, it also provides a possible numerical approach by simply trying to solve (7.13). In this context, the evaluation of the function Φ, which involves the solution of an initial value problem, is typically assumed to be exact because we can (at least theoretically) keep the discretization errors as small as we want by choosing sufficiently small stepsizes. Using the single shooting method,

one is, however, faced with the difficulty that the arising initial value problems may be unstable in the sense that small changes in the initial value may lead to large changes in the solution. This may have the effect that the computed trajectories become large even though the solution of the boundary value problem is nicely bounded. Even worse, it may happen that the trajectories do not extend until $\bar{t}$ due to errors in the initial guess.

Example 7.5. It is well-known that a linear initial value problem

$$\dot{x} = A(t)x + f(t), \quad x(t_0) = x_0,$$

with $A \in C([t_0, \infty), \mathbb{R}^{n,n})$ and $f \in C([t_0, \infty), \mathbb{R}^n)$ always possesses a unique solution $x \in C^1([t_0, \infty), \mathbb{R}^n)$. If we, however, consider the (scalar) nonlinear initial value problem

$$\dot{x} = x^2, \quad x(0) = x_0 > 0,$$

the situation is different. Still, there exists a unique solution locally in a neighborhood of t_0 due to the theorem of Picard and Lindelöf [220]. This local solution can be determined by separation of the variables. From

$$\int_0^t \frac{\dot{x}(s)}{x(s)^2}\, ds = \int_0^t s\, ds,$$

we obtain that

$$\frac{1}{x(0)} - \frac{1}{x(t)} = t,$$

hence

$$x(t) = \frac{x_0}{1 - x_0 t}.$$

Obviously, the solution cannot be extended up to $t = 1/x_0$.

To overcome these difficulties, the solution interval $\mathbb{I}$ is split beforehand into smaller subintervals according to

$$\underline{t} = t_0 < t_1 < \cdots < t_{N-1} < t_N = \bar{t}, \quad N \in \mathbb{N}. \tag{7.15}$$

Given initial guesses

$$(x_i, y_i) \in \mathbb{R}^{(\mu+2)n}, \quad i = 0, \ldots, N, \tag{7.16}$$

for points $(t_i, \hat{x}_i, \hat{y}_i) \in \mathbb{L}_\mu$, the idea of *multiple shooting* for differential-algebraic equations then is to project (t_i, x_i, y_i) onto $\mathbb{L}_\mu$ and to solve the associated initial value problems on the smaller intervals $[t_i, t_{i+1}]$, requiring additionally that the pieces match to a continuous function on the whole interval and that the boundary

condition is satisfied. In order to develop a method which can actually be implemented, it is not possible to use functions such as Z_1, Z_2, or T_2 in the definition of the procedure. Instead, we must look for computationally available quantities. Note that this is in contrast to the previous section which was merely dedicated to a theoretical investigation.

Given (t_i, x_i, y_i) as initial guess for a point on the set $\mathbb{L}_\mu$, we can solve $F_\mu(t_i, x, y) = 0$ by the Gauß–Newton method to obtain $(t_i, \tilde{x}_i, \tilde{y}_i) \in \mathbb{L}_\mu$. Of course, we must require that the guess (t_i, x_i, y_i) is sufficiently accurate to guarantee convergence. Applying Hypothesis 4.2, we can then compute matrices $\tilde{Z}_{2,i}$ and $\tilde{T}_{2,i}$, so that the columns form orthonormal bases of corange $F_{\mu;y}(t_i, \tilde{x}_i, \tilde{y}_i)$ and kernel $\tilde{Z}_{2,i}^T F_{\mu;x}(t_i, \tilde{x}_i, \tilde{y}_i)$, respectively. Subsequently, we can determine matrices $\tilde{Z}'_{2,i}$ and $\tilde{T}'_{2,i}$ so that the columns complement the columns of $\tilde{Z}_{2,i}$ and $\tilde{T}_{2,i}$ to bases of the full space. Analogous to (7.5) and (7.6), we finally can define $\tilde{T}_{1,i}$ and $\tilde{T}'_{1,i}$.

Similar to (7.7), the system

$$F_\mu(t_i, \hat{x}_i, \hat{y}_i) = 0, \tag{7.17a}$$

$$\tilde{T}_{2,i}^T(\hat{x}_i - x_i) = 0, \tag{7.17b}$$

$$\tilde{T}_{1,i}^T(\hat{y}_i - y_i) = 0 \tag{7.17c}$$

locally defines functions S_i according to

$$(\hat{x}_i, \hat{y}_i) = S_i(x_i, y_i) \tag{7.18}$$

in such a way that $(t_i, S_i(x_i, y_i)) \in \mathbb{L}_\mu$. Defining W_i to have columns that form an orthonormal basis of kernel $F_{\mu;x,y}(t_i, \hat{x}_i, \hat{y}_i)$ with $(t_i, \hat{x}_i, \hat{y}_i) \in \mathbb{L}_\mu$ and setting

$$\tilde{W}_i = \begin{bmatrix} \tilde{T}_{2,i} & 0 \\ 0 & \tilde{T}_{1,i} \end{bmatrix}, \tag{7.19}$$

we obtain

$$S_{i;x,y}(\hat{x}_i, \hat{y}_i) = W_i(\tilde{W}_i^T W_i)^{-1}\tilde{W}_i^T \tag{7.20}$$

similar to (7.12), as long as $\tilde{W}_i^T W_i$ is invertible. In the same way as with Φ in (7.10), we define flows Φ_i that map initial values $(t_i, \hat{x}_i, \hat{y}_i) \in \mathbb{L}_\mu$ to the value $x(t_{i+1})$ of the solution x of the corresponding initial value problem. As for Φ, we here omit the argument t_i of Φ_i for simplicity. Again, we assume that we can evaluate Φ_i exactly, i.e., without discretization errors.

The multiple shooting method is then given by

$$F_\mu(t_i, x_i, y_i) = 0, \qquad i = 0, \dots, N, \tag{7.21a}$$

$$\tilde{T}_{2,i+1}^T(x_{i+1} - \Phi_i(S_i(x_i, y_i))) = 0, \quad i = 0, \dots, N-1, \tag{7.21b}$$

$$b(x_0, x_N) = 0. \tag{7.21c}$$

Comparing with the single shooting method of Section 7.1, the condition (7.13a) is now required in (7.21a) at all mesh points t_i with corresponding unknowns (x_i, y_i). Besides the boundary condition (7.21c), we impose continuity conditions for the differential components in (7.21b). Condition (7.13b), which was responsible for local uniqueness of the solution in (7.13), cannot be used here because it involves knowledge of the actual solution. Thus, in the present form, system (7.21) is underdetermined. To solve this system, we apply the following Gauß–Newton-like method, where in the course of the presentation, we will select a suitable generalized inverse of the Jacobian by additional conditions that will turn out to be the appropriate replacement for (7.13b).

Given approximations (x_i, y_i), the Gauß–Newton-like method is defined by the corrections $(\Delta x_i, \Delta y_i)$ that are added to (x_i, y_i) to get updated approximations. In the (underdetermined) ordinary Gauß–Newton method, these corrections satisfy the linearized equations

$$F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + F_{\mu;y}(t_i, x_i, y_i)\Delta y_i = -F_\mu(t_i, x_i, y_i), \tag{7.22a}$$

$$\begin{aligned}\tilde{T}^T_{2,i+1}(\Delta x_{i+1} - \Phi_{i;x,y}(S_i(x_i, y_i))(S_{i;x}(x_i, y_i)\Delta x_i + S_{i;y}(x_i, y_i)\Delta y_i)) \\ = -\tilde{T}^T_{2,i+1}(x_{i+1} - \Phi_i(S_i(x_i, y_i))),\end{aligned} \tag{7.22b}$$

$$b_{x_l}(x_0, x_N)\Delta x_0 + b_{x_r}(x_0, x_N)\Delta x_N = -b(x_0, x_N). \tag{7.22c}$$

For an efficient numerical method, however, the structure and the properties of the Jacobian should be utilized. In the following, we will perturb the coefficient matrix in such a way that the system decouples into smaller systems of reasonable size. In particular, we will choose perturbations that tend to zero when the (x_i, y_i) converge to a solution of (7.21). This property then implies that the resulting Gauß–Newton-like process will show superlinear convergence, see, e.g., [71].

Definition 7.6. Let z^m, $m \in \mathbb{N}_0$, from some normed vector space form a convergent sequence with limit point z^*. We say that the z^m converge *superlinearly* to z^*, if there exist $\zeta_m > 0$, $m \in \mathbb{N}_0$, with $\zeta_m \to 0$ and $m^* \in \mathbb{N}_0$ such that $\|z^{m+1} - z^*\| \leq \zeta_m \|z^m - z^*\|$ for $m \geq m^*$.

In a solution of (7.21), the matrices $F_{\mu;y}(t_i, x_i, y_i)$ will have rank deficiency a. We therefore perturb $F_{\mu;y}(t_i, x_i, y_i)$ to matrices $\tilde{M}_i$ with rank deficiency a. The only condition that we must require is that these perturbations tend to zero when the matrices $F_{\mu;y}(t_i, x_i, y_i)$ tend to matrices with rank deficiency a. One possibility to achieve this is to replace $F_{\mu;y}(t_i, x_i, y_i)$ by the nearest matrix of rank deficiency a, by neglecting the a smallest singular values of $F_{\mu;y}(t_i, x_i, y_i)$, see, e.g., [99]. The equations (7.22a) are thus replaced by

$$F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + \tilde{M}_i \Delta y_i = -F_\mu(t_i, x_i, y_i). \tag{7.23}$$

Let the columns of $Z_{2,i}$ form an orthonormal basis of corange $\tilde{M}_i$ and let $[\, Z'_{2,i}\ Z_{2,i}\,]$ be orthogonal. Relation (7.23) then splits into

$$Z'^T_{2,i} F_{\mu;x}(t_i, x_i, y_i)\Delta x_i + Z'^T_{2,i}\tilde{M}_i \Delta y_i = -Z'^T_{2,i} F_\mu(t_i, x_i, y_i), \tag{7.24a}$$

$$Z^T_{2,i} F_{\mu;x}(t_i, x_i, y_i)\Delta x_i = -Z^T_{2,i} F_\mu(t_i, x_i, y_i). \tag{7.24b}$$

Requiring in addition that

$$\tilde{T}^T_{1,i}\Delta y_i = 0 \tag{7.25}$$

as substitute for (7.13b) and observing that

$$\begin{bmatrix} Z'^T_{2,i}\tilde{M}_i \\ \tilde{T}^T_{1,i} \end{bmatrix}$$

is nonsingular for sufficiently good initial guesses (x_i, y_i), it follows that we can solve (7.24a) with (7.25) for Δy_i in terms of Δx_i.

Let the columns of the matrix $T_{2,i}$ form an orthonormal basis of the space kernel $Z^T_{2,i} F_{\mu;x}(t_i, x_i, y_i)$. For sufficiently good initial guesses (x_i, y_i) also $\tilde{T}^T_{2,i} T_{2,i}$ is nonsingular. Thus, there exists a matrix $T'_{2,i}$ such that $[\, T'_{2,i}\ \ T_{2,i}\,]$ is nonsingular and

$$\tilde{T}^T_{2,i} T'_{2,i} = 0. \tag{7.26}$$

Defining $\Delta v'_i$ and Δv_i by the relation

$$\Delta x_i = T'_{2,i}\Delta v'_i + T_{2,i}\Delta v_i, \tag{7.27}$$

equation (7.24b) becomes

$$Z^T_{2,i} F_{\mu;x}(t_i, x_i, y_i) T'_{2,i}\Delta v'_i = -Z^T_{2,i} F_\mu(t_i, x_i, y_i). \tag{7.28}$$

Since $Z^T_{2,i} F_{\mu;x}(t_i, x_i, y_i) T'_{2,i}$ is nonsingular by construction, (7.28) can be solved for $\Delta v'_i$.

Turning to (7.22b), we know that at a solution of (7.22) the relations

$$S_{i;x}(x_i, y_i)\Delta x_i = W_i(\tilde{W}_i^T W_i)^{-1}\begin{bmatrix} \tilde{T}^T_{2,i} \\ 0 \end{bmatrix} T_{2,i}\Delta v_i = S_{i;x}(x_i, y_i) T_{2,i}\Delta v_i, \tag{7.29a}$$

$$S_{i;y}(x_i, y_i)\Delta y_i = W_i(\tilde{W}_i^T W_i)^{-1}\begin{bmatrix} 0 \\ \tilde{T}^T_{1,i} \end{bmatrix}\Delta y_i = 0 \tag{7.29b}$$

hold because of (7.25) and (7.26). Thus, we replace (7.22b) by

$$\begin{aligned} &\tilde{T}^T_{2,i+1} T_{2,i+1}\Delta v_{i+1} - \tilde{T}^T_{2,i+1}\Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i) T_{2,i}\Delta v_i \\ &\qquad = -\tilde{T}^T_{2,i+1}(x_{i+1} - \Phi_i(S_i(x_i, y_i))), \end{aligned} \tag{7.30}$$

which is again a perturbation that tends to zero when the iteration converges. Due to (7.30) we only need the derivative $\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x}(x_i, y_i)$ in the direction of the d columns of $T_{2,i}$. In particular, if we use numerical differentiation to approximate this derivative, then we only need to solve d initial value problems.

Finally, we write (7.22c) in the form

$$\begin{aligned} &b_{x_l}(x_0, x_N)T_{2,0}\Delta v_0 + b_{x_r}(x_0, x_N)T_{2,N}\Delta v_N \\ &\quad = -b(x_0, x_N) - b_{x_l}(x_0, x_N)T'_{2,0}\Delta v'_0 - b_{x_r}(x_0, x_N)T'_{2,N}\Delta v'_N. \end{aligned} \tag{7.31}$$

Setting

$$G_i = \tilde{T}^T_{2,i+1}\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x}(x_i, y_i)T_{2,i}, \quad i = 0, \ldots, N-1, \tag{7.32a}$$

$$J_i = \tilde{T}^T_{2,i}T_{2,i}, \qquad i = 1, \ldots, N, \tag{7.32b}$$

$$\tilde{C} = b_{x_l}(x_0, x_N)T_{2,0}, \ \tilde{D} = b_{x_r}(x_0, x_N)T_{2,N}, \tag{7.32c}$$

the remaining linear system that we have to solve for the unknowns Δv_i has the shooting-like coefficient matrix

$$\tilde{\mathcal{E}}_N = \begin{bmatrix} -G_0 & J_1 & & & \\ & -G_1 & J_2 & & \\ & & \ddots & \ddots & \\ & & & -G_{N-1} & J_N \\ \tilde{C} & & & & \tilde{D} \end{bmatrix}. \tag{7.33}$$

This system can be solved by standard methods such as Gauß elimination with pivoting. Since the blocks J_i are invertible for sufficiently good initial guesses (x_i, y_i), it follows that the matrix $\tilde{\mathcal{E}}_N$ is nonsingular if and only if the condensed matrix

$$\mathcal{E}_N = \tilde{C} + \tilde{D}(J_N^{-1}G_{N-1})(J_{N-1}^{-1}G_{N-2})\ldots(J_2^{-1}G_1)(J_1^{-1}G_0) \tag{7.34}$$

is nonsingular. Thus, for the method to work locally, it suffices to show that this is the case at least at the solution and therefore in some neighborhood of it.

At a solution $(x_i, y_i) = (x^*(t_i), y_i^*)$, the matrix $\mathcal{E}_N$ takes the form

$$\begin{aligned} \mathcal{E}_N = CT_{2,0} + DT_{2,N} \prod_{i=N-1}^{i=0} \Big[&(\tilde{T}^T_{2,i+1}T_{2,i+1})^{-1}\tilde{T}^T_{2,i+1} \\ &\cdot \Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x}(x_i, y_i)T_{2,i} \Big]. \end{aligned} \tag{7.35}$$

To take into account that $\Phi_i(S_i(x, y))$ is consistent at t_{i+1} for (x, y) in a neighborhood of (x_i, y_i) as the value of a solution of the differential-algebraic equation on

$[t_i, t_{i+1}]$, we consider the nonlinear system

$$F_\mu(t_i, x, \hat{y}) + Z_{2,i}\alpha = 0, \tag{7.36a}$$

$$\tilde{T}_{1,i}^T(\hat{y} - y_i) = 0. \tag{7.36b}$$

Writing this as

$$H_i(x, \hat{y}, \alpha) = 0, \tag{7.37}$$

we know that $H_i(x_i, y_i, 0) = 0$. Since

$$\operatorname{rank} H_{i;\hat{y},\alpha}(x_i, y_i, 0) = \operatorname{rank} \begin{bmatrix} F_{\mu;y}(t_i, x_i, y_i) & Z_{2,i} \\ \tilde{T}_{1,i}^T & 0 \end{bmatrix} = \operatorname{rank} \begin{bmatrix} Z_{2,i}'^T F_{\mu;y}(t_i, x_i, y_i) & 0 \\ 0 & I \\ \tilde{T}_{1,i}^T & 0 \end{bmatrix},$$

the construction of $Z_{2,i}'$ and $\tilde{T}_{1,i}$ guarantees that the matrix $H_{i;\hat{y},\alpha}(x_i, y_i, 0)$ is nonsingular. Thus, via the implicit function theorem, (7.36) locally defines functions K_i and L_i according to

$$\hat{y} = K_i(x), \quad \alpha = L_i(x). \tag{7.38}$$

For all x with $L_i(x) = 0$, we have $F_\mu(t_i, x, K_i(x)) = 0$ and x is consistent at t_i. Furthermore, differentiating

$$F_\mu(t_i, x, K_i(x)) + Z_{2,i}L_i(x) = 0$$

with respect to x, evaluating at x_i, and multiplying by $Z_{2,i}^T$ yields

$$L_{i;x}(x_i) = -Z_{2,i}^T F_{\mu;x}(t_i, x_i, y_i).$$

Hence, $L_{i;x}$ has full row rank in a neighborhood of x_i and the set of solutions of $L_i(x) = 0$ forms a manifold of dimension $d = n - a$, which is a submanifold of the manifold of consistent values at point t_i. Due to the results of Section 4.1, the consistency of x is locally described by the condition $\hat{F}_2(t, x) = 0$ in (4.23) which fixes a manifold of dimension d. The submanifold of the solutions of $L_i(x) = 0$ must therefore coincide with the manifold of consistent values at t_i.

Thus, given an x that is consistent at t_i, the function K_i yields a $\hat{y}$ such that $(t_i, x, \hat{y}) \in \mathbb{L}_\mu$, while $L_i(x) = 0$. In particular,

$$F_\mu(t_{i+1}, \Phi_i(S_i(x, y)), K_{i+1}(\Phi_i(S_i(x, y)))) = 0 \tag{7.39}$$

holds in a neighborhood of (x_i, y_i). Differentiating this relation with respect to (x, y) and setting $(x, y) = (x_i, y_i)$, we obtain

$$\begin{aligned} &F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1})\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) \\ &\quad + F_{\mu;y}(t_{i+1}, x_{i+1}, y_{i+1})K_{i+1;x}(x_{i+1})\Phi_{i;x,y}(S_i(x_i, y_i))S_{i;x,y}(x_i, y_i) = 0. \end{aligned}$$

Multiplying with $Z_{2,i+1}^T$ from the left finally yields

$$Z_{2,i+1}^T F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1}) \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i) = 0. \tag{7.40}$$

Hence, the columns of $\Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i)$ lie in the kernel of the matrix $Z_{2,i+1}^T F_{\mu;x}(t_{i+1}, x_{i+1}, y_{i+1})$, which in turn is spanned by the columns of $T_{2,i+1}$. Since the expression $T_{2,i+1}(\tilde{T}_{2,i+1}^T T_{2,i+1})^{-1} \tilde{T}_{2,i+1}^T$ is a projector onto this kernel, we have

$$\begin{aligned} &T_{2,i+1}(\tilde{T}_{2,i+1}^T T_{2,i+1})^{-1} \tilde{T}_{2,i+1}^T \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i) \\ &\quad = \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i). \end{aligned} \tag{7.41}$$

Thus, (7.35) reduces to

$$\mathcal{E}_N = C T_{2,0} + D\Big(\prod_{i=N-1}^{i=0} \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x}(x_i, y_i) \Big) T_{2,0}. \tag{7.42}$$

Finally, defining

$$\Psi_i(x) = (x, K_i(x)) \tag{7.43}$$

and using $\tilde{T}_{1,i}^T K_{i;x}(x_i) = 0$ which holds due to (7.36b), we find that

$$\begin{aligned} S_{i;x,y}(x_i, y_i) \Psi_{i;x}(x_i) &= W_i (\tilde{W}_i^T W_i)^{-1} \begin{bmatrix} \tilde{T}_{2,i}^T & 0 \\ 0 & \tilde{T}_{1,i}^T \end{bmatrix} \begin{bmatrix} I \\ K_{i;x}(x_i) \end{bmatrix} \\ &= W_i (\tilde{W}_i^T W_i)^{-1} \begin{bmatrix} \tilde{T}_{2,i}^T \\ 0 \end{bmatrix} = S_{i;x}(x_i, y_i). \end{aligned}$$

Hence, (7.42) becomes

$$\mathcal{E}_N = C T_{2,0} + D\Big(\prod_{i=N-1}^{i=0} \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i) \Psi_{i;x}(x_i) \Big) T_{2,0}. \tag{7.44}$$

Comparing with (7.14), the term

$$\prod_{i=N-1}^{i=0} \Phi_{i;x,y}(S_i(x_i, y_i)) S_{i;x,y}(x_i, y_i) \Psi_{i;x}(x_i)$$

in (7.44) is nothing else than the derivative $\Phi_{x,y} \tilde{S}_x$ of $\Phi \circ \tilde{S}$ decomposed according to

$$\Phi \circ \tilde{S} = (\Phi_{N-1} \circ S_{N-1}) \circ (\Psi_{N-1} \circ \Phi_{N-2} \circ S_{N-2}) \circ \cdots \circ (\Psi_1 \circ \Phi_0 \circ S_0), \tag{7.45}$$

where $\tilde{S}$ differs from S by replacing $T_1(\underline{t})$, $T_2(\underline{t})$ with $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$ in (7.7). This means that for sufficiently good initial guesses, the matrix $\mathcal{E}_N$ in (7.44) is nonsingular when $\mathcal{E}$ of (7.14) is nonsingular, i.e., when there exists a locally unique solution of the boundary value problem in the sense of Theorem 7.2.

Summarizing the obtained results, we have the following convergence theorem.

Theorem 7.7. *Suppose that the boundary value problem* (7.1) *satisfies Hypothesis* 4.2 *and that* (7.1) *has a locally unique solution according to Theorem* 7.2. *Then, for sufficiently good initial guesses, the iterates of the Gauß–Newton-like procedure developed in the course of this section converge superlinearly to a solution of* (7.21).

Proof. Writing the Gauß–Newton-like procedure for (7.21) in the form

$$z^{m+1} = z^m - \mathcal{M}_m^- \mathcal{F}(z^m),$$

where $\mathcal{M}_m$ is the chosen perturbation of $\mathcal{F}_z(z^m)$ and $\mathcal{M}_m^-$ denotes the chosen pseudoinverse due to (7.25), we have

$$\mathcal{M}_m \mathcal{M}_m^- = I,$$

since multiplication with $\mathcal{M}_m^-$ yields a solution of the perturbed linear system. Thus, we get

$$\begin{aligned}
z^{m+1} - z^m &= -\mathcal{M}_m^- \mathcal{F}(z^m) \\
&= -\mathcal{M}_m^- [\mathcal{F}(z^m) - \mathcal{F}(z^{m-1}) - \mathcal{M}_{m-1}(z^m - z^{m-1})] \\
&= -\mathcal{M}_m^- \big[\mathcal{F}(z^{m-1} + s(z^m - z^{m-1}))|_{s=0}^{s=1} - \mathcal{F}_z(z^{m-1})(z^m - z^{m-1}) \\
&\qquad - (\mathcal{M}_{m-1} - \mathcal{F}_z(z^{m-1}))(z^m - z^{m-1})\big] \\
&= -\mathcal{M}_m^- \Big[\int_0^1 (\mathcal{F}_z(z^{m-1} + s(z^m - z^{m-1})) - \mathcal{F}_z(z^{m-1}))(z^m - z^{m-1})\, ds \\
&\qquad - (\mathcal{M}_{m-1} - \mathcal{F}_z(z^{m-1}))(z^m - z^{m-1})\Big].
\end{aligned}$$

Introducing constants β, γ, and δ_m according to

$$\|\mathcal{M}_m^-\| \le \beta, \quad \|\mathcal{F}_z(u) - \mathcal{F}_z(v)\| \le \gamma \|u - v\|, \quad \|\mathcal{M}_m - \mathcal{F}_z(z^m)\| \le \delta_m$$

for some vector norm and its associated matrix norm, recalling that we assume sufficient smoothness for the data, we obtain the estimate

$$\|z^{m+1} - z^m\| \le \tfrac{1}{2}\beta\gamma \|z^m - z^{m-1}\|^2 + \beta\delta_{m-1}\|z^m - z^{m-1}\|. \tag{7.46}$$

For a sufficiently good initial guess z^0, we have that

$$\tfrac{1}{2}\beta\gamma\|z^1 - z^0\| + \beta\delta_0 \le L < 1.$$

Due to the construction of $\mathcal{M}_m$, we may assume that $\delta_m \le \delta_0$ for $z^m \in \overline{S}(z^0, \rho)$, $\rho = \frac{1}{1-L}$, where $\overline{S}(z^0, \rho)$ denotes the closed ball of radius ρ around z^0. It follows then by induction that

$$\|z^{m+1} - z^m\| \le L\|z^m - z^{m-1}\|$$

and that

$$\|z^m - z^0\| \le \tfrac{1}{1-L}\|z^1 - z^0\|.$$

Hence, we stay in $\overline{S}(z^0, \rho)$. Moreover, since

$$\|z^{m+k} - z^m\| \le \tfrac{L^m}{1-L}\|z^1 - z^0\|,$$

the iterates z^m form a Cauchy series and thus converge to a $z^* \in \overline{S}(z^0, \rho)$. Since the iterates satisfy $\mathcal{M}_m(z^{m+1} - z^m) = -\mathcal{F}(z^m)$ where the left hand side tends to zero, we get $\mathcal{F}(z^m) \to 0$ and the continuity of $\mathcal{F}$ yields $\mathcal{F}(z^*) = 0$.

By assumption, we now can utilize that $\delta_m \to 0$ in order to show superlinear convergence. We first observe that (7.46) then says that

$$\|z^{m+1} - z^m\| \le \varepsilon_{m-1}\|z^m - z^{m-1}\|, \quad \varepsilon_{m-1} = \tfrac{1}{2}\beta\gamma\|z^m - z^{m-1}\| + \beta\delta_{m-1}$$

with $\varepsilon_m \to 0$. This gives the estimate

$$\begin{aligned}\|z^{m+1} - z^*\| &\le \frac{1}{1-L}\|z^{m+2} - z^{m+1}\| \le \frac{\varepsilon_m}{1-L}\|z^{m+1} - z^m\| \\ &\le \frac{\varepsilon_m}{1-L}(\|z^{m+1} - z^*\| + \|z^m - z^*\|).\end{aligned}$$

Since $1 - L - \varepsilon_m$ is positive for sufficiently large m, we obtain for such values of m that

$$\|z^{m+1} - z^*\| \le \zeta_m\|z^m - z^*\|, \quad \zeta_m = \frac{\varepsilon_m}{1 - L - \varepsilon_m}$$

with $\zeta_m \to 0$. This finally proves superlinear convergence of the given Gauß–Newton-like method. □

Remark 7.8. The main difficulty in the construction of multiple shooting methods for differential-algebraic equations is to deal with inconsistent iterates. In the method presented here, we have used (locally defined) nonlinear projections on $\mathbb{L}_\mu$ to get consistent initial values. A second possibility would have been to shift the manifold $\mathbb{L}_\mu$ in such a way that the given inconsistent iterate then lies in the shifted manifold. In the case of single shooting, we could define

$$\hat{\mathbb{L}}_\mu = \{(t, x, y) \in \mathbb{R}^{(\mu+2)n+1} \mid F_\mu(t, x - x_0 + \hat{x}, y - y_0 + \hat{y}) = 0\}$$

with $(\hat{x}, \hat{y}) = S(x_0, y_0)$ and solve the arising initial value problem with respect to $\hat{\mathbb{L}}_\mu$. The same idea is used in [197] in the form of so-called *relaxed algebraic*

constraints when these are explicitly available. One can show that using the technique of shifting the manifold would yield a method with the same properties as the method that we have presented here. However, the method of shifting the manifold has the disadvantage that it requires to modify F_μ for the use in the initial value solver.

Remark 7.9. In the case of linear boundary value problems

$$E(t)\dot{x} = A(t)x + f(t), \quad Cx(\underline{t}) + Dx(\bar{t}) = w,$$

we can work with the reduced differential-algebraic equation (3.60). In particular, we can directly parameterize the space of consistent values x_i at a point t_i by

$$x_i = -\hat{A}_2(t_i)^+ \hat{f}_2(t_i) + T_2(t_i)v_i, \quad v_i \in \mathbb{R}^d.$$

Thus, in this case there is no need of a projection to obtain consistent initial values. Moreover, no values y_i are needed in order to integrate a linear differential-algebraic equation. Hence, in the linear case we can omit (7.21a) and rewrite (7.21b) and (7.21c) in terms of the unknowns $v_0, \ldots, v_N$. Since the given problem is linear, (7.21b) and (7.21c) then constitute a linear system for $v_0, \ldots, v_N$ whose coefficient matrix has the form (7.33).

7.3 Collocation

The multiple shooting method replaces the boundary value problem by a sequence of initial value problems. But it is well-known already from the case of ordinary differential equations, see, e.g. [8], that a boundary value problem may be well conditioned while the corresponding initial value problems are not. In such cases, multiple shooting is not an adequate approach.

Example 7.10. Consider the (scalar) boundary value problem

$$\ddot{x} = \lambda^2 x, \quad x(0) = 1, \quad x(1) = 1.$$

with $\lambda > 0$. Introducing $x_1 = x$ and $x_2 = \dot{x}$, we can rewrite the differential equation of second order as a system of differential equations of first order according to

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \lambda^2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad x_1(0) = 1, \quad x_1(1) = 1.$$

All solutions of the differential equation have the form

$$x(t) = c_1 \begin{bmatrix} \exp(\lambda t) \\ \lambda \exp(\lambda t) \end{bmatrix} + c_2 \begin{bmatrix} \exp(-\lambda t) \\ -\lambda \exp(-\lambda t) \end{bmatrix},$$

with $c_1, c_2 \in \mathbb{R}$.

If we provide (nontrivial) initial values $x_1(0) = \delta_1$, $x_2(0) = \delta_1$, we get the conditions

$$c_1 + c_2 = \delta_1, \quad \lambda c_1 - \lambda c_2 = \delta_2,$$

which yield

$$c_1 = \frac{\lambda\delta_1 + \delta_2}{2\lambda}, \quad c_2 = \frac{\lambda\delta_1 - \delta_2}{2\lambda}.$$

The corresponding solution x given by

$$x(t) = \frac{\lambda\delta_1 + \delta_2}{2\lambda} \begin{bmatrix} \exp(\lambda t) \\ \lambda \exp(\lambda t) \end{bmatrix} + \frac{\lambda\delta_1 - \delta_2}{2\lambda} \begin{bmatrix} \exp(-\lambda t) \\ -\lambda \exp(-\lambda t) \end{bmatrix}$$

can be seen as a perturbation of the solution of the given boundary value problem with respect to perturbations of the initial values. Thus, using the maximum norm in $\mathbb{R}^2$ and the related norm in $C([0, 1], \mathbb{R}^2)$, the quotient $\|x\|/\|\delta\|$ behaves like $\lambda \exp(\lambda)$ for $\lambda \to \infty$. The same holds by symmetry if we provide initial values at $t = 1$.

Considering on the other hand the (nontrivial) boundary conditions $x_1(0) = \delta_1$, $x_1(1) = \delta_2$, we get the conditions

$$c_1 + c_2 = \delta_1, \quad \exp(\lambda)c_1 + \exp(-\lambda)c_2 = \delta_2,$$

which yield

$$c_1 = \frac{\delta_2 - \exp(-\lambda)\delta_1}{\exp(\lambda) - \exp(-\lambda)}, \quad c_2 = \frac{\exp(\lambda)\delta_1 - \delta_2}{\exp(\lambda) - \exp(-\lambda)}.$$

The corresponding solution x given by

$$x(t) = \frac{\delta_2 - \exp(-\lambda)\delta_1}{\exp(\lambda) - \exp(-\lambda)} \begin{bmatrix} \exp(\lambda t) \\ \lambda \exp(\lambda t) \end{bmatrix} + \frac{\exp(\lambda)\delta_1 - \delta_2}{\exp(\lambda) - \exp(-\lambda)} \begin{bmatrix} \exp(-\lambda t) \\ -\lambda \exp(-\lambda t) \end{bmatrix}$$

can be seen as a perturbation of the solution of the given boundary value problem with respect to perturbations of the boundary values. In this case, the quotient $\|x\|/\|\delta\|$ behaves like λ for $\lambda \to \infty$.

Since the numbers $\|x\|/\|\delta\|$ describe how errors in the initial condition or in the boundary condition are amplified, they represent corresponding condition numbers. If we choose for example $\lambda = 100$, then we get $\lambda \exp(\lambda) > 10^{45}$ and we cannot expect to get reasonable results when we solve the corresponding initial value problems numerically. The boundary value problem on the other hand possesses a nicely bounded condition number.

An alternative class of methods is given by so-called *collocation methods*. In contrast to multiple shooting methods, for collocation methods an approximate

solution is sought in a finite dimensional subspace of $C(\mathbb{I}, \mathbb{R}^n)$ in such a way that it satisfies the given equation at selected points. The preferred spaces in this context are spaces of piecewise polynomial functions with respect to a prescribed mesh.

In order to formulate and investigate collocation methods for (7.1), we first assume that it has the special form

$$\hat{F}_1(t, x, \dot{x}) = 0, \tag{7.47a}$$

$$\hat{F}_2(t, x) = 0 \tag{7.47b}$$

according to (4.23). The results for the general case then follow from the local transformation to the form (7.47) that we have presented in Section 4.1.

For the analysis, it is convenient to write (7.1) as an operator equation. For the choice of the correct spaces, we must not only observe that (7.47a) and (7.47b) have different smoothness properties but also that the collocation solution is required to be piecewise polynomial and globally continuous. Given a mesh

$$\begin{aligned} &\pi : \ \underline{t} = t_0 < t_1 < \cdots < t_{N-1} < t_N = \bar{t}, \quad N \in \mathbb{N}, \\ &h_i = t_{i+1} - t_i, \quad h = \max_{i=0,\ldots,N-1} h_i, \quad h \le K \min_{i=0,\ldots,N-1} h_i, \end{aligned} \tag{7.48}$$

where $K > 0$ is some fixed constant when we consider $h \to 0$, we define the spaces

$$\mathbb{X} = C^1_\pi(\mathbb{I}, \mathbb{R}^n) \cap C^0(\mathbb{I}, \mathbb{R}^n), \tag{7.49a}$$

$$\mathbb{Y} = C^0_\pi(\mathbb{I}, \mathbb{R}^d) \times C^1_\pi(\mathbb{I}, \mathbb{R}^a) \cap C^0(\mathbb{I}, \mathbb{R}^a) \times \mathbb{R}^d. \tag{7.49b}$$

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

If we equip the spaces in (7.49) with the norms

$$\|x\|_{\mathbb{X}} = \max_{t\in\mathbb{I}} \|x(t)\|_\infty + \max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|\dot{x}(t)\|_\infty \}, \tag{7.50a}$$

$$\begin{aligned} \|(f_1, f_2, w)\|_{\mathbb{Y}} = &\max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|f_1(t)\|_\infty \} + \max_{t\in\mathbb{I}} \|f_2(t)\|_\infty \\ &+ \max_{i=0,\ldots,N-1} \{ \max_{t\in[t_i,t_{i+1}]} \|\dot{f}_2(t)\|_\infty \} + \|w\|_\infty, \end{aligned} \tag{7.50b}$$

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

The boundary value problem (7.1) then takes the form of an operator equation

$$L(x) = 0, \tag{7.51}$$

with

$$L\colon \mathbb{X} \to \mathbb{Y}, \tag{7.52a}$$

$$x \mapsto \begin{bmatrix} \hat{F}_1(t, x(t), \dot{x}(t)) \\ \hat{F}_2(t, x(t)) \\ b(x(\underline{t}), x(\bar{t})) \end{bmatrix}. \tag{7.52b}$$

Since x^* solves (7.1) according to (7.2), we have $L(x^*) = 0$. In the construction of a Newton-like method for the solution of (7.51), we will later need the Fréchet derivative $DL[z]$ of L at $z \in \mathbb{X}$, which is given by

$$DL[z]\colon \mathbb{X} \to \mathbb{Y}, \tag{7.53a}$$

$$x \mapsto \begin{bmatrix} \hat{F}_{1;x}(t, z(t), \dot{z}(t))x(t) + \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))\dot{x}(t) \\ \hat{F}_{2;x}(t, z(t))x(t) \\ b_{x_l}(z(\underline{t}), z(\bar{t}))x(\underline{t}) + b_{x_r}(z(\underline{t}), z(\bar{t}))x(\bar{t}) \end{bmatrix}. \tag{7.53b}$$

In order to select a finite dimensional version of (7.51), we introduce the finite dimensional spaces

$$\mathbb{X}_\pi = \mathbb{P}_{k+1,\pi} \cap C^0(\mathbb{I}, \mathbb{R}^n), \tag{7.54a}$$

$$\mathbb{Y}_\pi = \mathbb{R}^{kNd} \times \mathbb{R}^{(kN+1)a} \times \mathbb{R}^d, \tag{7.54b}$$

where $\mathbb{P}_{k+1,\pi}$ denotes the space of piecewise polynomials of maximal degree k, or, synonymously, of maximal order $k+1$. Observe that $\mathbb{Y}_\pi$ is chosen such that

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

Hence, we must require $(kN+1)n$ scalar conditions in order to fix a unique approximate solution in $\mathbb{X}_\pi$. For collocation methods, these conditions (besides the boundary conditions) are of the form that the given differential-algebraic equation is satisfied at some selected points. In view of (7.47), the choice of the points should reflect the different smoothness properties of (7.47a) and (7.47b). Moreover, we want the approximations at the mesh points to be consistent, i.e., to satisfy the algebraic constraints. We therefore use a Gauß-type scheme for the differential part and a Lobatto-type scheme for the algebraic part. These schemes are given by nodes

$$0 < \varrho_1 < \cdots < \varrho_k < 1, \tag{7.55a}$$

$$0 = \sigma_0 < \cdots < \sigma_k = 1, \quad k \in \mathbb{N}, \tag{7.55b}$$

respectively, and define the collocation points

$$t_{ij} = t_i + h_i \varrho_j, \quad j = 1, \ldots, k, \tag{7.56a}$$

$$s_{ij} = t_i + h_i \sigma_j, \quad j = 0, \ldots, k. \tag{7.56b}$$

Observe that if it is desired that the resulting method is symmetric with respect to time reversion, then we must choose the nodes to lie symmetric in [0, 1].

The *collocation discretization* of (7.51) is then given by the nonlinear (discrete) operator equation

$$L_\pi(x_\pi) = 0, \tag{7.57}$$

with

$$L_\pi : \mathbb{X} \to \mathbb{Y}_\pi, \tag{7.58a}$$

$$x \mapsto \begin{bmatrix} \hat{F}_1(t_{ij}, x(t_{ij}), \dot{x}(t_{ij})) \\ \hat{F}_2(s_{ij}, x(s_{ij})) \\ b(x(\underline{t}), x(\bar{t})) \end{bmatrix}, \tag{7.58b}$$

and we seek a solution $x_\pi \in \mathbb{X}_\pi$. For ease of notation, we have omitted in (7.58b) that the indices i and j must run over the values $i = 0, \dots, N-1$, $j = 1, \dots, k$ in the first component and $i = 0, \dots, N-1$, $j = 1, \dots, k$ together with $i = 0$, $j = 0$ in the second component. Note that in the second component the indices $i = 1, \dots, N-1$, $j = 0$ must be omitted, since the space $\mathbb{X}_\pi$ includes continuity of the solution. We will use this kind of abbreviation in the remainder of this section.

We will also need the Fréchet derivative $DL_\pi[z]$ of the discretized operator L_π at $z \in \mathbb{X}$, which is given by

$$DL_\pi[z] : \mathbb{X} \to \mathbb{Y}_\pi, \tag{7.59a}$$

$$x \mapsto \begin{bmatrix} \hat{F}_{1;x}(t_{ij}, z(t_{ij}), \dot{z}(t_{ij}))x(t_{ij}) + \hat{F}_{1;\dot{x}}(t_{ij}, z(t_{ij}), \dot{z}(t_{ij}))\dot{x}(t_{ij}) \\ \hat{F}_{2;x}(s_{ij}, z(s_{ij}))x(s_{ij}) \\ b_{x_l}(z(\underline{t}), z(\bar{t}))x(\underline{t}) + b_{x_r}(z(\underline{t}), z(\bar{t}))x(\bar{t}) \end{bmatrix}. \tag{7.59b}$$

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

$$R_\pi : \mathbb{Y} \to \mathbb{Y}_\pi, \tag{7.60a}$$

$$\begin{bmatrix} f_1 \\ f_2 \\ w \end{bmatrix} \mapsto \begin{bmatrix} f_1(t_{ij}) \\ f_2(s_{ij}) \\ w \end{bmatrix}. \tag{7.60b}$$

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

Example 7.11. Consider the model of a physical pendulum

$$\begin{aligned} \dot{p}_1 &= v_1, & \dot{v}_1 &= -2p_1\lambda, \\ \dot{p}_2 &= v_2, & \dot{v}_2 &= -2p_2\lambda - g, \\ p_1^2 + p_2^2 &= 1, \end{aligned}$$

where $g = 9.81$ is the gravity constant, cp. Example 1.3. According to Example 4.27, it satisfies Hypothesis 4.2 with characteristic values $\mu = 2$, $a = 3$, and $d = 2$. We are thus allowed to impose two scalar boundary conditions. Following [137], we take

$$v_2(0) = 0, \quad p_1(0.55) = 0.$$

To get the hidden constraints, we differentiate the given constraint twice and eliminate the arising derivatives with the help of the other equations. In this way, we obtain

$$2p_1v_1 + 2p_2v_2 = 0$$

and

$$2v_1^2 - 4p_1^2\lambda + 2v_2^2 - 4p_2^2\lambda - 2gp_2 = 0.$$

Obviously, the latter constraint can always be solved for λ. The two other constraints possess the Jacobian $[\,2p_1\ 2p_2\,]$ with respect to p_1, p_2 in the first case and v_1, v_2 in the second case. The kernel vector $[\,-p_2\ p_1\,]^T$ can then be used to select a suitable differential part. This leads to a possible reduced differential-algebraic equation of the form

$$\begin{aligned}
&p_1\dot{p}_2 - p_2\dot{p}_1 = p_1v_2 - p_2v_1,\\
&p_1\dot{v}_2 - p_2\dot{v}_1 = -gp_1,\\
&p_1^2 + p_2^2 = 1,\\
&2p_1v_1 + 2p_2v_2 = 0,\\
&2v_1^2 - 4p_1^2\lambda + 2v_2^2 - 4p_2^2\lambda - 2gp_2 = 0.
\end{aligned}$$

Renaming the variables according to

$$(x_1, x_2, x_3, x_4, x_5) = (p_1, p_2, v_1, v_2, \lambda),$$

this system, after some trivial simplifications, can be written as

$$\begin{aligned}
&x_1(\dot{x}_2 - x_4) - x_2(\dot{x}_1 - x_3) = 0,\\
&x_1\dot{x}_4 - x_2\dot{x}_3 + gx_1 = 0,\\
&x_1^2 + x_2^2 - 1 = 0,\\
&x_1x_3 + x_2x_4 = 0,\\
&x_3^2 + x_4^2 - gx_2 - 2x_5 = 0.
\end{aligned}$$

Hence, the corresponding boundary value problem has the form $L(x) = 0$ with

$$L(x) = \begin{bmatrix} f_1 \\ f_2 \\ w \end{bmatrix},$$

where

$$f_1(t) = \begin{bmatrix} x_1(t)(\dot{x}_2(t) - x_4(t)) - x_2(t)(\dot{x}_1(t) - x_3(t)) \\ x_1(t)\dot{x}_4(t) - x_2(t)\dot{x}_3(t) + g x_1(t) \end{bmatrix},$$

$$f_2(t) = \begin{bmatrix} x_1(t)^2 + x_2(t)^2 - 1 \\ x_1(t)x_3(t) + x_2(t)x_4(t) \\ x_3(t)^2 + x_4(t)^2 - g x_2(t) - 2x_5(t) \end{bmatrix},$$

$$w = \begin{bmatrix} x_4(0) \\ x_1(0.55) \end{bmatrix}.$$

The Fréchet derivative $DL[z]$ of L at some z can be obtained via its defining relation

$$L(z + x) = L(z) + DL[z]x + R(x), \quad R(x)/\|x\| \to 0 \text{ for } \|x\| \to 0.$$

In the first component we thus get (omitting the argument t)

$$\begin{aligned} &(z_1 + x_1)((\dot{z}_2 + \dot{x}_2) - (z_4 + x_4)) \\ &\quad - (z_2 + x_2)((\dot{z}_1 + \dot{x}_1) - (z_4 + x_3)) - z_1(\dot{z}_2 - z_4) + z_2(\dot{z}_1 - z_3) \\ &\quad = z_1(\dot{x}_2 - x_4) + x_1(\dot{z}_2 - z_4) - z_2(\dot{x}_1 - x_3) - x_2(\dot{z}_1 - z_3) + \mathcal{O}(\|x\|^2). \end{aligned}$$

Treating the other parts in the same way, we get

$$DL[z](x) = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{w} \end{bmatrix},$$

where

$$\tilde{f}_1(t) = \begin{bmatrix} z_1(t)(\dot{x}_2(t) - x_4(t)) + x_1(t)(\dot{z}_2(t) - z_4(t)) - \\ \qquad -z_2(t)(\dot{x}_1(t) - x_3(t)) - x_2(t)(\dot{z}_1(t) - z_3(t)) \\ z_1(t)\dot{x}_4(t) + x_1(t)\dot{z}_4(t) - z_2(t)\dot{x}_3(t) - x_2(t)\dot{z}_3(t) + g x_1(t) \end{bmatrix},$$

$$\tilde{f}_2(t) = \begin{bmatrix} 2z_1(t)x_1(t) + 2z_2(t)x_2(t) \\ z_1(t)x_3(t) + x_1(t)z_3(t) + z_2(t)x_4(t) + x_2(t)z_4(t) \\ 2z_3(t)x_3(t) + 2z_4(t)x_4(t) - g x_2(t) - 2x_5(t) \end{bmatrix},$$

$$\tilde{w} = \begin{bmatrix} x_4(0) \\ x_1(0.55) \end{bmatrix}.$$

A representation of the collocation operator L_π and its Fréchet derivative $DL_\pi[z]$ can then simply be obtained by utilizing $L_\pi = R_\pi L$ and $DL_\pi[z] = R_\pi DL[z]$ with (7.60).

The first step in the analysis of collocation methods is to show that if x^* satisfies some regularity condition that guarantees that x^* is locally unique, then equation

(7.57) is solvable in $\mathbb{X}_\pi$ at least for sufficiently small h. In addition, we derive orders of convergence for $h \to 0$.

To do so, we follow [8, pp. 222–226], where a corresponding result is shown in the case of ordinary differential equations. In the context of differential-algebraic equations, we consider the iterative process

$$x_\pi^{m+1} = x_\pi^m - DL_\pi[x^*]^{-1} L_\pi(x_\pi^m) \tag{7.61}$$

and prove that under suitable assumptions it generates a sequence $\{x_\pi^m\}$ in $\mathbb{X}_\pi$ that converges to a solution of (7.57). Note that the iteration (7.61) is only a tool for the theoretical analysis. It cannot be used as a numerical method, since the value of the Fréchet derivative at the exact solution x^* is not available.

Recall the basic properties of the iterative process (7.61) given in Theorem 5.7. In the present context, however, we are interested in properties of (7.61) for $h \to 0$. Thus, we must consider families of iterations (7.61) with the maximum mesh sizes tending to zero. For these we must show that certain constants are independent of h. Unfortunately, in the standard formulation of Theorem 5.7 and its proof this does not hold for the constants β and γ. The main task of the following considerations is therefore to replace the standard definition of β and γ in Theorem 5.7 by more appropriate quantities and to show that then the crucial estimates in the proof of Theorem 5.7 still hold. Since the modified quantities will play the same role as the original constants β and γ, we will keep the same notation.

To derive these results, we first investigate in detail the linear boundary value problem associated with the Fréchet derivative $DL_\pi[x^*]$. Introducing the Jacobians

$$\begin{aligned} \hat{E}_1(t) &= \hat{F}_{1;\dot{x}}(t, x^*(t), \dot{x}^*(t)), \qquad C = b_{x_l}(x^*(\underline{t}), x^*(\bar{t})), \\ \hat{A}_1(t) &= -\hat{F}_{1;x}(t, x^*(t), \dot{x}^*(t)), \quad D = b_{x_r}(x^*(\underline{t}), x^*(\bar{t})), \\ \hat{A}_2(t) &= -\hat{F}_{2;x}(t, x^*(t)), \end{aligned} \tag{7.62}$$

we have that

$$DL[x^*] \colon \mathbb{X} \to \mathbb{Y}, \tag{7.63a}$$

$$x \mapsto \begin{bmatrix} \hat{E}_1(t)\dot{x}(t) - \hat{A}_1(t)x(t) \\ -\hat{A}_2(t)x(t) \\ Cx(\underline{t}) + Dx(\bar{t}) \end{bmatrix} \tag{7.63b}$$

and $DL_\pi[x^*] = R_\pi DL[x^*]$.

Applying the transformation of Lemma 5.6 to the (linear) boundary value problem

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t), \tag{7.64a}$$

$$Cx(\underline{t}) + Dx(\bar{t}) = \hat{w} \tag{7.64b}$$

with some $g = (\hat{f}, \hat{w}) \in \mathbb{Y}$ yields that

$$\begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} + \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix}, \tag{7.65a}$$

$$\begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} \tilde{x}_1(\underline{t}) \\ \tilde{x}_2(\underline{t}) \end{bmatrix} + \begin{bmatrix} D_1 & D_2 \end{bmatrix} \begin{bmatrix} \tilde{x}_1(\bar{t}) \\ \tilde{x}_2(\bar{t}) \end{bmatrix} = \hat{w}, \tag{7.65b}$$

where

$$x = Q \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix}, \quad P\hat{f} = \begin{bmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{bmatrix}, \quad CQ(\underline{t}) = \begin{bmatrix} C_1 & C_2 \end{bmatrix}, \quad DQ(\bar{t}) = \begin{bmatrix} D_1 & D_2 \end{bmatrix}.$$

The solutions of the differential-algebraic equation (7.65a) therefore have the form

$$\tilde{x}_1(t) = \tilde{x}_1(\underline{t}) + \int_{\underline{t}}^{s} \tilde{f}_1(s)\, ds, \quad \tilde{x}_2(t) = -\tilde{f}_2(t).$$

Inserting this into the boundary condition (7.65b) gives a linear equation for $\tilde{x}_1(\underline{t})$ with coefficient matrix $C_1 + D_1$. We therefore have proved the following result, compare with Theorem 7.2.

Lemma 7.12. *The boundary value problem (7.64) is uniquely solvable if and only if $C_1 + D_1 \in \mathbb{R}^{d,d}$ is nonsingular, with C_1, D_1 as defined in (7.65b).*

Writing (7.64) in operator form as $DL[x^*]x = g$, the corresponding collocation discretization reads $DL_\pi[x^*]x_\pi = R_\pi g$ or

$$\hat{E}_1(t_{ij})\dot{x}_\pi(t_{ij}) - \hat{A}_1(t_{ij})x_\pi(t_{ij}) = \hat{f}_1(t_{ij}), \tag{7.66a}$$

$$-\hat{A}_2(s_{ij})x_\pi(s_{ij}) = \hat{f}_2(s_{ij}), \tag{7.66b}$$

$$Cx_\pi(\underline{t}) + Dx_\pi(\bar{t}) = \hat{w}. \tag{7.66c}$$

In order to determine a solution $x_\pi \in \mathbb{X}_\pi$ of (7.66), we set $x_{\pi,i} = x_\pi|_{[t_i, t_{i+1}]}$, which is a polynomial of maximal degree k for $i = 0, \ldots, N-1$, and require the continuity of x_π explicitly as $x_{\pi,i-1}(t_i) = x_{\pi,i}(t_i)$ for $i = 1, \ldots, N-1$. This, however, would lead to an overdetermined system when combined with (7.66b), since it yields that

$$\hat{A}_2(t_i)(x_{\pi,i-1}(t_i) - x_{\pi,i}(t_i)) = 0, \quad i = 1, \ldots, N-1,$$

because of $s_{i,0} = s_{i-1,k} = t_i$. Observe that

$$\begin{bmatrix} \hat{A}_2(t_i) \\ T_2(t_i)^T \end{bmatrix}$$

is nonsingular for $i = 0, \dots, N$. It is therefore sufficient to require that $T_2(t_i)^T (x_{\pi,i-1}(t_i) - x_{\pi,i}(t_i)) = 0$ for $i = 1, \dots, N-1$. Hence, the collocation method for a linear differential-algebraic equation consists in determining polynomials $x_{\pi,i}$ of maximal degree k that satisfy the linear system

$$\hat{E}_1(t_{ij})\dot{x}_{\pi,i}(t_{ij}) - \hat{A}_1(t_{ij})x_{\pi,i}(t_{ij}) = \hat{f}_1(t_{ij}), \tag{7.67a}$$

$$-\hat{A}_2(s_{ij})x_{\pi,i}(s_{ij}) = \hat{f}_2(s_{ij}), \tag{7.67b}$$

$$T_2(t_i)^T (x_{\pi,i-1}(t_i) - x_{\pi,i}(t_i)) = 0, \tag{7.67c}$$

$$Cx_\pi(\underline{t}) + Dx_\pi(\bar{t}) = \hat{w}. \tag{7.67d}$$

Note that the indices i and j range over all those values such that (7.67a) and (7.67b) are required for all collocation points and (7.67c) is required for all interior mesh points $t_1, \dots, t_{N-1}$.

For the representation of the arising polynomials, we use Lagrange interpolation with respect to both sets of nodes. In particular, we define the Lagrange interpolation polynomials

$$L_l(\xi) = \prod_{\substack{j=0 \\ j\neq l}}^{k} \frac{\xi - \sigma_j}{\sigma_l - \sigma_j}, \quad \tilde{L}_l(\xi) = \prod_{\substack{m=1 \\ m\neq l}}^{k} \frac{\xi - \varrho_m}{\varrho_l - \varrho_m}. \tag{7.68}$$

The polynomials $x_{\pi,i}$ can then be written as

$$x_{\pi,i}(t) = \sum_{l=0}^{k} x_{i,l} L_l\left(\frac{t - t_i}{h_i}\right), \tag{7.69}$$

with $x_{i,l} = x_{\pi,i}(s_{il})$. Defining $v_{jl} = \dot{L}_l(\varrho_j)$ and $u_{jl} = L_l(\varrho_j)$ for $l = 0, \dots, k$, $j = 1, \dots, k$, we get

$$\dot{x}_{\pi,i}(t_{ij}) = \tfrac{1}{h_i} \sum_{l=0}^{k} v_{jl} x_{i,l}, \quad x_{\pi,i}(t_{ij}) = \sum_{l=0}^{k} u_{jl} x_{i,l}.$$

If we set

$$w_{jl} = \int_0^{\sigma_j} \tilde{L}_l(\xi)\, d\xi, \quad j, l = 1, \dots, k, \tag{7.70}$$

then we see that $V = [v_{jl}]_{j,l}$ is nonsingular with $V^{-1} = [w_{jl}]_{j,l}$, cp. Exercise 7. Finally, we introduce $x_N = x_{N,0} = x_{\pi,N-1}(t_N)$.

Inserting the representation of $x_{\pi,i}$ and $\dot{x}_{\pi,i}$ into (7.67), we get a linear system for the values $x_{i,j}$ given by

$$\frac{1}{h_i}\sum_{l=0}^{k} v_{jl}\hat{E}_1(t_{ij})x_{i,l} - \sum_{l=0}^{k} u_{jl}\hat{A}_1(t_{ij})x_{i,l} = \hat{f}_1(t_{ij}), \tag{7.71a}$$

$$-\hat{A}_2(s_{ij})x_{i,j} = \hat{f}_2(s_{ij}), \tag{7.71b}$$

$$x_{i,k} - x_{i+1,0} = 0, \tag{7.71c}$$

$$Cx_{0,0} + Dx_{N,0} = \hat{w}, \tag{7.71d}$$

$$-\hat{A}_2(t_0)x_{0,0} = \hat{f}_2(t_0), \tag{7.71e}$$

with $j = 1, \ldots, k$ and $i = 0, \ldots, N-1$. Here, we have used the fact that the matrix

$$\begin{bmatrix} \hat{A}_2(t_i) \\ T_2(t_i)^T \end{bmatrix}$$

is nonsingular and have combined the equations in (7.67b) for $j = 0$ with those in (7.67c) to obtain (7.71c).

In order to prove that the collocation method is well defined, we have to show that the linear system (7.71) has a unique solution. To do so, we proceed in two steps. First we consider the local systems of the form

$$B_i \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,k} \end{bmatrix} = a_i x_{i,0} + b_i, \quad i = 0, \ldots, N-1, \tag{7.72}$$

which consist of the collocation conditions (7.71a) and (7.71b) for $j = 1, \ldots, k$. Their solvability is examined in Lemma 7.13.

Solving these systems in terms of the other variables and inserting the solution into the other equations leads to relations

$$x_{i,k} = [\,0 \;\cdots\; 0 \;\; I\,]B_i^{-1}a_i x_{i,0} + [\,0 \;\cdots\; 0 \;\; I\,]B_i^{-1}b_i, \tag{7.73}$$

which, together with (7.71c), can be written as

$$x_{i+1} = W_i x_i + g_i. \tag{7.74}$$

Representations for W_i and g_i are given in Lemma 7.14.

In the second step, we consider the global system

$$K_h \begin{bmatrix} x_0 \\ \vdots \\ x_N \end{bmatrix} = g_h, \tag{7.75}$$

representing the continuity conditions (7.71c), the boundary condition (7.71d), and the consistency condition (7.71e), with the solutions (7.74) of the local systems (7.72) already inserted. Its solvability is examined in Lemma 7.15.

Setting $\hat{E}_{1,j} = \hat{E}_1(t_{ij})$, $\hat{A}_{1,j} = \hat{A}_1(t_{ij})$, $\hat{A}_{2,j} = \hat{A}_2(s_{ij})$, $\hat{f}_{1,j} = \hat{f}_1(t_{ij})$ and $\hat{f}_{2,j} = \hat{f}_2(s_{ij})$ for selected fixed i, the local system (7.72) is given by

$$B_i = \left[\begin{array}{c|c|c|c} \frac{v_{11}}{h_i}\hat{E}_{1,1} - u_{11}\hat{A}_{1,1} & \frac{v_{12}}{h_i}\hat{E}_{1,1} - u_{12}\hat{A}_{1,1} & \dots & \frac{v_{1k}}{h_i}\hat{E}_{1,1} - u_{1k}\hat{A}_{1,1} \\ -\hat{A}_{2,1} & 0 & & 0 \\ \hline \frac{v_{21}}{h_i}\hat{E}_{1,2} - u_{21}\hat{A}_{1,2} & \frac{v_{22}}{h_i}\hat{E}_{1,2} - u_{22}\hat{A}_{1,2} & \dots & \frac{v_{2k}}{h_i}\hat{E}_{1,2} - u_{2k}\hat{A}_{1,2} \\ 0 & -\hat{A}_{2,2} & & 0 \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \frac{v_{k1}}{h_i}\hat{E}_{1,k} - u_{k1}\hat{A}_{1,k} & \frac{v_{k2}}{h_i}\hat{E}_{1,k} - u_{k2}\hat{A}_{1,k} & \dots & \frac{v_{kk}}{h_i}\hat{E}_{1,k} - u_{kk}\hat{A}_{1,k} \\ 0 & 0 & & -\hat{A}_{2,k} \end{array}\right]$$

and

$$a_i = \left[\begin{array}{c} -\frac{v_{10}}{h_i}\hat{E}_{1,1} + u_{10}\hat{A}_{1,1} \\ 0 \\ \hline \vdots \\ \hline -\frac{v_{k0}}{h_i}\hat{E}_{1,k} + u_{k0}\hat{A}_{1,k} \\ 0 \end{array}\right], \quad b_i = \left[\begin{array}{c} \hat{f}_{1,1} \\ \hat{f}_{2,1} \\ \hline \vdots \\ \hline \hat{f}_{1,k} \\ \hat{f}_{2,k} \end{array}\right],$$

where $B_i \in \mathbb{R}^{kn,kn}$, $a_i \in \mathbb{R}^{kn,n}$, and $b_i \in \mathbb{R}^{kn}$. In the following lemma, we prove the nonsingularity of B_i for sufficiently small h_i, using multiplications from the left and from the right, respectively, with block matrices

$$T_P = \operatorname{diag}\left(\begin{bmatrix} P_{11}(t_{ij}) & P_{12}(s_{ij}) \\ 0 & P_{22}(s_{ij}) \end{bmatrix}\right)_{j=1,\dots,k}, \quad T_Q = \operatorname{diag}\Big(Q(s_{ij})\Big)_{j=1,\dots,k},$$

where P, Q transform the differential-algebraic equation into the canonical form (5.11). We also need to reorder block rows and columns with the help of $U_k \in \mathbb{R}^{kn,kn}$ given by

$$U_k = \left[\begin{array}{c|c|c|c|c|c|c|c} I_d & & & & 0 & & & \\ 0 & & & & I_a & & & \\ \hline & I_d & & & & 0 & & \\ & 0 & & & & I_a & & \\ \hline & & \ddots & & & & \ddots & \\ \hline & & & I_d & & & & 0 \\ & & & 0 & & & & I_a \end{array}\right]. \tag{7.76}$$

For the following estimates, we use the ℓ_∞-norm and the associated matrix norm. Recall again that we assume all functions to be sufficiently smooth.

Lemma 7.13. *Consider the block matrix*

$$\Delta_i = \begin{bmatrix} \Delta_i^1 & \Delta_i^2 \\ 0 & 0 \end{bmatrix} \tag{7.77}$$

with

$$\Delta_i^s = \Big[h_i \sum_{l=1}^{k} w_{jl} G_{lm}^s \Big]_{j,m=1,\dots,k}, \quad s = 1, 2,$$

and

$$[\, G_{lm}^1 \;\; G_{lm}^2 \,] = \begin{cases} (v_{ll}(\sigma_l - \varrho_l) - 1)(P_{11}\hat{E}_1\dot{Q})(t_{il}) \\ \qquad -(u_{ll} - 1)(P_{11}\hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i), & l = m, \\ v_{lm}(\sigma_m - \varrho_l)(P_{11}\hat{E}_1\dot{Q})(t_{il}) \\ \qquad - u_{lm}(P_{11}\hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i), & l \neq m, \end{cases}$$

for $m = 0, \dots, k$ and l, $j = 1, \dots, k$. Then the representation

$$B_i = T_P^{-1} U_k \begin{bmatrix} \frac{1}{h_i} V \otimes I & 0 \\ 0 & -I \end{bmatrix} (I + \Delta_i) U_k^T T_Q^{-1} \tag{7.78}$$

holds and, for sufficiently small h_i, the matrix B_i is nonsingular with

$$B_i^{-1} = T_Q U_k (I - \Delta_i + \mathcal{O}(h_i^2)) \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P. \tag{7.79}$$

Proof. Taylor expansion yields

$$Q(s_{im}) = Q(t_{il}) + h_i(\sigma_m - \varrho_l)\dot{Q}(t_{il}) + \mathcal{O}(h_i^2) = Q(t_{il}) + \mathcal{O}(h_i)$$

and

$$(P_{12}\hat{A}_2 Q)(s_{il}) = (P_{12}\hat{A}_2 Q)(t_{il}) + \mathcal{O}(h_i) = (P_{11}\hat{E}_1\dot{Q} - P_{11}\hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i).$$

This leads to

$$\begin{aligned} &[\, P_{11}(t_{il}) \;\; P_{12}(s_{il}) \,] \begin{bmatrix} \frac{v_{lm}}{h_i}\hat{E}_{1,l} - u_{lm}\hat{A}_{1,l} \\ 0 \end{bmatrix} Q(s_{im}) \\ &\quad = \frac{v_{lm}}{h_i}(P_{11}\hat{E}_1)(t_{il})Q(s_{im}) - u_{lm}(P_{11}\hat{A}_1)(t_{il})Q(s_{im}) \\ &\quad = \frac{v_{lm}}{h_i}(P_{11}\hat{E}_1 Q)(t_{il}) + v_{lm}(\sigma_m - \varrho_l)(P_{11}\hat{E}_1\dot{Q})(t_{il}) \\ &\qquad - u_{lm}(P_{11}\hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i) \\ &\quad = \frac{v_{lm}}{h_i}[\, I \;\; 0 \,] + [\, G_{lm}^1 \;\; G_{lm}^2 \,] \end{aligned}$$

for $m \neq l$. Analogously, we get

$$\begin{aligned}
&[\, P_{11}(t_{il}) \;\; P_{12}(s_{il}) \,] \begin{bmatrix} \frac{v_{lm}}{h_i} \hat{E}_{1,l} - u_{lm} \hat{A}_{1,l} \\ -\hat{A}_{2,l} \end{bmatrix} Q(s_{il}) \\
&\quad = \frac{v_{ll}}{h_i} (P_{11} \hat{E}_1)(t_{il}) Q(s_{il}) - u_{ll} (P_{11} \hat{A}_1)(t_{il}) Q(s_{il}) - (P_{12} \hat{A}_2 Q)(s_{il}) \\
&\quad = \frac{v_{ll}}{h_i} (P_{11} \hat{E}_1 Q)(t_{il}) + v_{ll} (\sigma_l - \varrho_l)(P_{11} \hat{E}_1 \dot{Q})(t_{il}) \\
&\qquad - u_{ll} (P_{11} \hat{A}_1 Q)(t_{il}) - (P_{11} \hat{E}_1 \dot{Q} - P_{11} \hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i) \\
&\quad = \frac{v_{ll}}{h_i} [\, I \;\; 0 \,] + [\, G^1_{ll} \;\; G^2_{ll} \,].
\end{aligned}$$

Multiplying B_i with T_P from the left and T_Q from the right and reordering the rows and columns using U_k as in (7.76), we obtain that

$$U_k^T T_P B_i T_Q U_k = \begin{bmatrix} \frac{1}{h_i} V \otimes I & 0 \\ 0 & -I \end{bmatrix} + \begin{bmatrix} G^1 & G^2 \\ 0 & 0 \end{bmatrix},$$

with $G^s = [G^s_{lm}]_{l,m}$. Since V is nonsingular with $V^{-1} = [w_{jl}]_{j,l}$, we have

$$\begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P B_i T_Q U_k = I + \Delta_i, \tag{7.80}$$

with Δ_i as defined in (7.77). Multiplication with the inverses yields the desired representation of B_i.

Since the matrices G^s_{lm} stay bounded for all l, m and $s = 1, 2$ when $h_i \to 0$, we have $\|\Delta_i\| = \mathcal{O}(h_i)$. Thus, $I + \Delta_i$ is nonsingular for sufficiently small h_i with the inverse satisfying $(I + \Delta_i)^{-1} = I - \Delta_i + \mathcal{O}(h_i^2)$. By this and (7.80), we see that B_i is nonsingular for sufficiently small h_i and that B_i^{-1} has the stated representation. □

Having shown that the local systems (7.72) are uniquely solvable at least for sufficiently small h_i, we now derive representations for W_i and g_i in (7.74).

Lemma 7.14. *In equation* (7.74)*, the coefficients* W_i, g_i *have the representations*

$$W_i = Q(t_{i+1}) \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1}, \tag{7.81}$$

with $F_{i1} = \mathcal{O}(h_i^2)$, $F_{i2} = \mathcal{O}(h_i)$, *and*

$$g_i = Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22} \hat{f}_2)(t_{i+1}) \end{bmatrix}, \tag{7.82}$$

where $c_i = \mathcal{O}(h_i)$.

Proof. Using the representation of B_i^{-1} given in Lemma 7.13, we compute

$$W_i\, Q(t_i) = [\, 0 \ \cdots \ 0 \ \ I \,]\, B_i^{-1} a_i\, Q(t_i).$$

With $Q(t_i) = Q(t_{il}) + \mathcal{O}(h_i) = Q(t_{il}) - \varrho_l h_i \dot{Q}(t_{il}) + \mathcal{O}(h_i^2)$, we have

$$\begin{aligned}
&[\, P_{11}(t_{il}) \ \ P_{12}(s_{il}) \,] \begin{bmatrix} & -\frac{v_{l0}}{h_i}\hat{E}_{1,l} + u_{l0}\hat{A}_{1,l} \\ 0 & \end{bmatrix} Q(t_i) \\
&\quad = -\frac{v_{l0}}{h_i}(P_{11}\hat{E}_1 Q)(t_{il}) + v_{l0}\varrho_l (P_{11}\hat{E}_1 \dot{Q})(t_{il}) + u_{l0}(P_{11}\hat{A}_1 Q)(t_{il}) + \mathcal{O}(h_i) \\
&\quad = -\frac{v_{l0}}{h_i}[\, I \ \ 0 \,] - [\, G_{l0}^1 \ \ G_{l0}^2 \,].
\end{aligned}$$

Hence, we have

$$U_k^T T_P\, a_i\ Q(t_i) = -\frac{1}{h_i}\begin{bmatrix} v_0 \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} G_0^1 & G_0^2 \\ 0 & 0 \end{bmatrix},$$

with $v_0 = [v_{l0}]_{l=1,\ldots,k}$ and $G_0^s = [G_{l0}^s]_{l=1,\ldots,k}$ for $s = 1, 2$.

Applying the next factor in the representation of B_i^{-1} and using $v_0 = -Ve$ with $e = [\, 1 \ \cdots \ 1 \,]^T$, we get

$$\begin{aligned}
&\begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P\, a_i\ Q(t_i) \\
&\quad = \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} \left(\begin{bmatrix} \frac{1}{h_i} Ve \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} G_0^1 & G_0^2 \\ 0 & 0 \end{bmatrix} \right) \\
&\quad = \begin{bmatrix} e \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} h_i V^{-1} \otimes G_0^1 & h_i V^{-1} \otimes G_0^2 \\ 0 & 0 \end{bmatrix}.
\end{aligned}$$

The next factor that has to be applied due to the representation of B_i^{-1} is given by

$$I - \Delta_i + \mathcal{O}(h_i^2) = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} \Delta_i^1 & \Delta_i^2 \\ 0 & 0 \end{bmatrix} + \mathcal{O}(h_i^2).$$

Setting

$$\begin{aligned}
H_i = &\left(\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} - \begin{bmatrix} \Delta_i^1 & \Delta_i^2 \\ 0 & 0 \end{bmatrix} + \mathcal{O}(h_i^2) \right) \\
&\cdot \left(\begin{bmatrix} e \otimes I & 0 \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} h_i V^{-1} \otimes G_0^1 & h_i V^{-1} \otimes G_0^2 \\ 0 & 0 \end{bmatrix} \right),
\end{aligned}$$

and observing that $\Delta^s_{j0} = h_i \sum_{l=1}^k w_{jl} G^s_{l0} = \mathcal{O}(h_i)$ according to Lemma 7.13, we obtain that

$$H_i = \begin{bmatrix} I & 0 \\ \vdots & \vdots \\ I & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \Delta^1_{1,0} & \Delta^2_{1,0} \\ \vdots & \vdots \\ \Delta^1_{k,0} & \Delta^2_{k,0} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} - \begin{bmatrix} \sum_{m=0}^k \Delta^1_{1,m} & 0 \\ \vdots & \vdots \\ \sum_{m=0}^k \Delta^1_{k,m} & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ \vdots & \vdots \\ \mathcal{O}(h_i^2) & \mathcal{O}(h_i^2) \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Thus,

$$H_i = \begin{bmatrix} * & * \\ \vdots & \vdots \\ * & * \\ I - F_{i,1} & -F_{i,2} \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix},$$

with $F_{i1} = \sum_{m=0}^k \Delta^1_{k,m} + \mathcal{O}(h_i^2)$, $F_{i2} = \Delta^2_{k,0} + \mathcal{O}(h_i^2)$.

Altogether, we obtain

$$W_i Q(t_i) = [\,0\ \cdots\ 0\ I\,] B_i^{-1} a_i Q(t_i) = [\,0\ \cdots\ 0\ I\,] \operatorname{diag}(Q(s_{ij})) U_k H_i$$

and hence, since $s_{ik} = t_{i+1}$, we have

$$W_i = Q(t_{i+1}) \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix} Q(t_i)^{-1}.$$

In order to show that $F_{i1} = \mathcal{O}(h_i^2)$, we use interpolation of the polynomials $p(t) = 1$, $q(t) = t$ at the points $\sigma_0, \ldots, \sigma_k$ to obtain

$$\sum_{m=0}^k L_m(\varrho_l) = 1, \quad \sum_{m=0}^k \dot{L}_m(\varrho_l) = 0, \quad \sum_{m=0}^k \dot{L}_m(\varrho_l)\sigma_m = 1.$$

By inserting the terms defined in Lemma 7.13, we see that

$$\begin{aligned}
\sum_{m=0}^{k} \Delta_{km}^{1} &= \sum_{m=0}^{k} h_i \sum_{l=1}^{k} w_{kl} G_{lm}^{1} \\
&= h_i \sum_{l=1}^{k} w_{kl} \Big(\sum_{\substack{m=0 \\ m \neq l}}^{k} \big[v_{lm}(\sigma_m - \varrho_l)(P_{11}\hat{E}_1 \dot{Q}_1)(t_{il}) - u_{lm}(P_{11}\hat{A}_1 Q_1)(t_{il}) \big] \\
&\quad + (v_{ll}(\sigma_l - \varrho_l) - 1)(P_{11}\hat{E}_1 \dot{Q}_1)(t_{il}) - (u_{ll} - 1)(P_{11}\hat{A}_1 Q_1)(t_{il}) + \mathcal{O}(h_i) \Big) \\
&= h_i \sum_{l=1}^{k} w_{kl} \Big(\Big[\sum_{m=0}^{k} \dot{L}_m(\varrho_l)(\sigma_m - \varrho_l) - 1 \Big] (P_{11}\hat{E}_1 \dot{Q}_1)(t_{il}) \\
&\quad - \Big[\sum_{m=0}^{k} L_m(\varrho_l) - 1 \Big] (P_{11}\hat{A}_1 Q_1)(t_{il}) + \mathcal{O}(h_i) \Big) = \mathcal{O}(h_i^2)
\end{aligned}$$

and therefore

$$F_{i1} = \sum_{m=0}^{k} \Delta_{km}^{1} + \mathcal{O}(h_i^2) = \mathcal{O}(h_i^2).$$

Looking at the definition of Δ_{k0}^{2}, it is obvious that $F_{i2} = \mathcal{O}(h_i)$.

Analogously, we insert the representation for B_i^{-1} given in Lemma 7.13 into $g_i = [\,0 \ \cdots \ 0 \ I\,] B_i^{-1} b_i$. We first observe that

$$\begin{aligned}
U_k^T T_P b_i &= U_k^T \begin{bmatrix} \begin{bmatrix} (P_{11}\hat{f}_1)(t_{ij}) + (P_{12}\hat{f}_2)(t_{ij}) \\ (P_{22}\hat{f}_2)(t_{ij}) \end{bmatrix}_{j=1,\ldots,k} \end{bmatrix} \\
&= \begin{bmatrix} \big[(P_{11}\hat{f}_1)(t_{ij}) + (P_{12}\hat{f}_2)(t_{ij}) \big]_{j=1,\ldots,k} \\ \big[(P_{22}\hat{f}_2)(t_{ij}) \big]_{j=1,\ldots,k} \end{bmatrix}.
\end{aligned}$$

Applying then the remaining factors of the representation of B_i^{-1} gives

$$\begin{aligned}
g_i &= [\,0 \ \cdots \ 0 \ I\,] T_Q U_k (I - \Delta_i + \mathcal{O}(h_i^2)) \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P b_i \\
&= [\,0 \ \cdots \ 0 \ Q(s_{ik})\,] U_k (I - \Delta_i + \mathcal{O}(h_i^2)) \begin{bmatrix} \mathcal{O}(h_i) \\ \big[-(P_{22}\hat{f}_2)(t_{ij}) \big]_{j=1,\ldots,k} \end{bmatrix} \\
&= [\,0 \ \cdots \ 0 \ Q(t_{i+1})\,] U_k \begin{bmatrix} \mathcal{O}(h_i) \\ \big[-(P_{22}\hat{f}_2)(t_{ij}) \big]_{j=1,\ldots,k} \end{bmatrix} \\
&= Q(t_{i+1}) \begin{bmatrix} c_i \\ -(P_{22}\hat{f}_2)(t_{i+1}) \end{bmatrix},
\end{aligned}$$

with c_i as in (7.82). □

The global system (7.75) is given by $K_h \in \mathbb{R}^{(N+1)n,(N+1)n}$ and $g_h \in \mathbb{R}^{(N+1)n}$, where

$$K_h = \begin{bmatrix} C & & & & D \\ -\hat{A}_2(t_0) & & & & \\ W_0 & -I & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & W_{N-1} & -I \end{bmatrix}, \quad g_h = \begin{bmatrix} \hat{w} \\ \hat{f}_2(t_0) \\ -g_0 \\ \vdots \\ \vdots \\ -g_{N-1} \end{bmatrix}. \tag{7.83}$$

To prove that K_h is nonsingular and that $K_h^{-1} g_h$ is bounded for $h \to 0$, we multiply from the left and from the right, respectively, with

$$T_l = \operatorname{diag}(I, P_{22}(t_0), Q(t_1)^{-1}, \ldots, Q(t_N)^{-1}), \quad T_r = \operatorname{diag}(Q(t_0)), \ldots, Q(t_N)),$$

where P, Q are given by Lemma 5.6. We also use $U_N \in \mathbb{R}^{(N+1)n,(N+1)n}$, which is defined analogously to U_k in (7.76), to reorder rows and columns. Finally, we set

$$M_h = \begin{bmatrix} C_1 & & & D_1 \\ I & -I & & \\ & \ddots & \ddots & \\ & & I & -I \end{bmatrix}, \quad N_h = \begin{bmatrix} C_2 & & & D_2 \\ -F_{0,2} & 0 & & \\ & \ddots & \ddots & \\ & & -F_{N-1,2} & 0 \end{bmatrix},$$

$$D_h = \begin{bmatrix} 0 & & & \\ -F_{0,1} & 0 & & \\ & \ddots & \ddots & \\ & & -F_{N-1,1} & 0 \end{bmatrix},$$

with C_1, C_2, D_1, D_2 given in (7.65b) and F_{i1}, F_{i2} given in Lemma 7.14, and define

$$A_h = \begin{bmatrix} M_h & N_h \\ 0 & -I \end{bmatrix}, \quad \Delta_h = \begin{bmatrix} D_h & 0 \\ 0 & 0 \end{bmatrix}.$$

Lemma 7.15. *The matrix K_h of the global system (7.75) given in (7.83) has the representation*

$$K_h = T_l^{-1} U_N (A_h + \Delta_h) U_N^T T_r^{-1}. \tag{7.84}$$

For a uniquely solvable boundary value problem (7.64), the matrix K_h is invertible for sufficiently small h, with

$$K_h^{-1} = T_r^{-1} U_N (I - A_h^{-1} \Delta_h + \mathcal{O}(h^2)) A_h^{-1} U_N^T T_l. \tag{7.85}$$

Proof. Multiplying with T_l from the left and T_r from the right, we get blockwise

$$\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix} \begin{bmatrix} C \\ -\hat{A}_2(t_0) \end{bmatrix} Q(t_0) = \begin{bmatrix} C_1 & C_2 \\ 0 & -I \end{bmatrix},$$
$$\begin{bmatrix} I & 0 \\ 0 & P_{22}(t_0) \end{bmatrix} \begin{bmatrix} D \\ 0 \end{bmatrix} Q(t_N) = \begin{bmatrix} D_1 & D_2 \\ 0 & 0 \end{bmatrix},$$
$$Q(t_{i+1})^{-1} W_i Q(t_i) = \begin{bmatrix} I - F_{i1} & -F_{i2} \\ 0 & 0 \end{bmatrix},$$

if we use the representation of W_i given in Lemma 7.14. Reordering of the rows and columns yields

$$U_N^T T_l K_h T_r U_N = A_h + \Delta_h$$

and thus (7.84).

By Lemma 7.12, the matrix $S = C_1 + D_1$ is regular. Hence, M_h is regular with inverse

$$M_h^{-1} = \begin{bmatrix} S^{-1} & & \\ & \ddots & \\ & & S^{-1} \end{bmatrix} \begin{bmatrix} I & D_1 & \cdots & D_1 \\ \vdots & -C_1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & D_1 \\ I & -C_1 & \cdots & -C_1 \end{bmatrix}.$$

Using Lemma 7.14, it follows that

$$\|M_h^{-1} D_h\|_\infty \le \|S^{-1}\|_\infty \max\left\{\|C_1\|_\infty, \|D_1\|_\infty\right\} \sum_{i=0}^{N-1} \|F_{i1}\|_\infty = \mathcal{O}(h).$$

Since M_h is nonsingular, the same holds for A_h. In particular, we obtain

$$\|A_h^{-1} \Delta_h\|_\infty = \|M_h^{-1} D_h\|_\infty = \mathcal{O}(h).$$

Thus, $A_h + \Delta_h$ is invertible for sufficiently small h and

$$(A_h + \Delta_h)^{-1} = (I - A_h^{-1} \Delta_h + \mathcal{O}(h^2)) A_h^{-1}.$$

This proves the invertibility of K_h and the representation of K_h^{-1}. □

We then combine Lemmas 7.12, 7.14, and 7.15 to investigate the solvability of the linear collocation system (7.66).

Theorem 7.16. *Let the boundary value problem (7.64) be uniquely solvable. Then, for sufficiently small h, there exists a unique solution $x_\pi \in \mathbb{X}_\pi$ of (7.66). Moreover, writing (7.66) in the form $DL_\pi[x^*]x_\pi = R_\pi g$, the estimate*

$$\|x_\pi\|_{\mathbb{X}} \le \beta \|g\|_{\mathbb{Y}} \tag{7.86}$$

holds, where β does not depend on g and h.

Proof. Using (7.82) and observing that $c_i = \mathcal{O}(h_i)$ and $F_{i2} = \mathcal{O}(h_i)$ by Lemma 7.14, the representations of K_h^{-1} and M_h^{-1} yield that

$$A_h^{-1} U_N^T T_l g_h = \begin{bmatrix} M_h^{-1} & M_h^{-1} N_h \\ 0 & -I \end{bmatrix} \begin{bmatrix} \hat{w} \\ -c_0 \\ \vdots \\ -c_{N-1} \\ (P_{22}\hat{f}_2)(t_0) \\ \vdots \\ (P_{22}\hat{f}_2)(t_N) \end{bmatrix} = \begin{bmatrix} M_h^{-1} d_h \\ -(P_{22}\hat{f}_2)(t_0) \\ \vdots \\ -(P_{22}\hat{f}_2)(t_N) \end{bmatrix}$$

with

$$d_h = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_N \end{bmatrix} = \begin{bmatrix} \hat{w} + C_2(P_{22}\hat{f}_2)(t_0) + D_2(P_{22}\hat{f}_2)(t_N) \\ -c_0 - F_{0,2}(P_{22}\hat{f}_2)(t_0) \\ \vdots \\ -c_{N-1} - F_{N-1,2}(P_{22}\hat{f}_2)(t_{N-1}) \end{bmatrix}.$$

In particular, $\|d_i\|_\infty = \mathcal{O}(h)$ for $i = 1, \dots, N$. Hence, with $S = C_1 + D_1$ we obtain that

$$\|M_h^{-1} d_h\|_\infty \le \|S^{-1}\|_\infty \Big(\|d_0\|_\infty + \max\big\{\|C_1\|_\infty, \|D_1\|_\infty\big\} \sum_{i=1}^{N} \|d_i\|_\infty \Big),$$

and therefore

$$\begin{aligned} \|K_h^{-1} g_h\|_\infty &\le \|T_r U_N (I - A_h^{-1}\Delta_h + \mathcal{O}(h^2)) A_h^{-1} U_N^T T_l g_h\|_\infty \\ &\le \|T_r U_N (I - A_h^{-1}\Delta_h + \mathcal{O}(h^2))\|_\infty \\ &\qquad \cdot \max\big\{\|M_h^{-1} d_h\|_\infty, \|(P_{22}\hat{f}_2)(t_0)\|_\infty, \dots, \|(P_{22}\hat{f}_2)(t_N)\|_\infty\big\}. \end{aligned}$$

Recalling the definition of the quantity c_i in the proof of Lemma 7.14, we have that $\|c_i\|_\infty \le \beta h_i \|g\|_{\mathbb{Y}}$, $i = 0, \dots, N-1$, where β does not depend on h_i and i. Then, the quantities d_i, $i = 0, \dots, N$, are also bounded according to $\|d_i\|_\infty \le \beta h_i \|g\|_{\mathbb{Y}}$ with possibly increased constant β. Thus, we get

$$\max_{i=0,\dots,N} \|x_i\|_\infty = \|K_h^{-1} g_h\|_\infty \le \beta \|g\|_{\mathbb{Y}},$$

again with possibly increased constant β. The relations (7.72) and (7.69) then show that a similar estimate holds for all values $x_{i,j}$ and hence uniformly on $\mathbb{I}$. In this way, we get

$$\max_{t \in \mathbb{I}} \|x_\pi(t)\|_\infty \le \beta \|g\|_{\mathbb{Y}}$$

with a (possibly increased) constant β which does not depend on h.

To get the estimate for the derivative $\dot{x}_\pi$ on $[t_i, t_{i+1}]$ involved in (7.86), we observe that

$$\begin{aligned}\hat{E}_1(t_{ij})\dot{x}_\pi(t_{ij}) &= \hat{A}_1(t_{ij})x_\pi(t_{ij}) + \hat{f}_1(t_{ij}), \quad j = 1, \dots, k,\\ 0 &= \hat{A}_2(s_{ij})x_\pi(s_{ij}) + \hat{f}_2(s_{ij}), \quad j = 0, \dots, k.\end{aligned}$$

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

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

compare (7.69). Hence,

$$\begin{aligned}\hat{A}_2(t_{ij})\dot{x}_\pi(t_{ij}) &= \sum_{l=0}^{k} \hat{A}_2(t_{ij})x_\pi(s_{il})\dot{L}_l\left(\frac{t_{ij} - t_i}{h_i}\right)\frac{1}{h_i}\\ &= \sum_{l=0}^{k} (\hat{A}_2(s_{il}) + \mathcal{O}(h))x_\pi(s_{il})\dot{L}_l(\varrho_j)\frac{1}{h_i}\\ &= \sum_{l=0}^{k} \mathcal{O}(h)\cdot x_\pi(s_{il})\dot{L}_j(\varrho_j)\frac{1}{h_i} - \sum_{l=0}^{k} \hat{f}_2(s_{il})\dot{L}_l(\varrho_j)\frac{1}{h_i}\\ &= \sum_{l=0}^{k} \mathcal{O}(h)\cdot x_\pi(s_{il})\dot{L}_l(\varrho_j)\frac{1}{h_i} - \sum_{l=0}^{k} (\hat{f}_2(s_{il}) - \hat{f}_2(t_i))\dot{L}_l(\varrho_j)\frac{1}{h_i},\end{aligned}$$

where the latter identity follows from

$$\sum_{l=0}^{k} \dot{L}_l\left(\frac{t - t_i}{h_i}\right) = 0$$

for all $t \in [t_i, t_{i+1}]$. Since $\hat{f}_2$ is continuously differentiable on $[t_i, t_{i+1}]$, there exist points $\theta_{ij} \in [t_i, t_{i+1}]$ which satisfy

$$\hat{f}_2(s_{il}) - \hat{f}_2(t_i) = h_i \sigma_l \dot{\hat{f}}_2(\theta_{ij}).$$

Therefore, possibly increasing the constant β, we have that

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

Together with

$$\|\hat{E}_1(t_{ij})\dot{x}_\pi(t_{ij})\|_\infty \le \beta \|g\|_{\mathbb{Y}},$$

it then follows that

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

since

$$\begin{bmatrix} \hat{E}_1 \\ \hat{A}_2 \end{bmatrix}$$

is a (smooth) pointwise nonsingular matrix function, compare Lemma 5.6. Using Lagrange interpolation of $\dot{x}_\pi$ at the points t_{ij} with $\tilde{L}_j$ as in (7.68), we have

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

Thus, it follows that

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

and finally

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

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

Remark 7.17. Theorem 7.16 shows that the collocation discretization given by $DL_\pi[x^*]$ is stable in the sense that the operator norm of the linear operator $DL_\pi[x^*]^{-1} R_\pi : \mathbb{Y} \to \mathbb{X}$ is bounded according to

$$\|DL_\pi[x^*]^{-1} R_\pi\|_{\mathbb{X} \leftarrow \mathbb{Y}} \le \beta \tag{7.87}$$

with β independent of h.

The next question then is how good the collocation solution approximates the true solution. Recall again that we assume that all coefficient functions and hence the solution are sufficiently smooth.

Theorem 7.18. *Let x be the unique solution of the boundary value problem (7.64) and let h be sufficiently small. Then the collocation solution $x_\pi \in \mathbb{X}_\pi$ of (7.66) satisfies*

$$\|x - x_\pi\|_{\mathbb{X}} \le \Gamma h^k, \tag{7.88}$$

where the constant Γ is independent of h.

Proof. Interpolation of x analogous to (7.69) yields

$$x(t) = \sum_{l=0}^{k} x(s_{il}) L_l \left(\frac{t - t_i}{h_i} \right) + \psi_i(t), \quad \psi_i(t) = \frac{x^{(k+1)}(\theta_i(t))}{(k+1)!} \prod_{j=0}^{k} (t - s_{ij}) \tag{7.89}$$

with $\theta_i(t) \in [t_i, t_{i+1}]$. Inserting this representation into the differential-algebraic equation and evaluating at the collocation points t_{ij} and s_{ij} gives

$$B_i \begin{bmatrix} x(s_{i1}) \\ \vdots \\ x(s_{ik}) \end{bmatrix} = a_i x(t_i) + b_i - \begin{bmatrix} \tau_{i,1} \\ \vdots \\ \tau_{i,k} \end{bmatrix}, \quad \tau_{i,j} = \begin{bmatrix} (\hat{E}_1 \dot{\psi}_i - \hat{A}_1 \psi_i)(t_{ij}) \\ 0 \end{bmatrix},$$

compare (7.72). Obviously, we have $\psi_i(t_{ij}) = \mathcal{O}(h_i^{k+1})$ and $\dot{\psi}_i(t_{ij}) = \mathcal{O}(h_i^k)$, thus $\tau_{i,j} = \mathcal{O}(h_i^k)$. Since the collocation problem is uniquely solvable for sufficiently small h, it follows that the matrix B_i is nonsingular and that we can solve for $x(s_{ik}) = x(t_{i+1})$ to get

$$x(t_{i+1}) = W_i x(t_i) + g_i - \tau_i, \quad \tau_i = [\, 0 \ \cdots \ I \,] B_i^{-1} \begin{bmatrix} \tau_{i,1} \\ \vdots \\ \tau_{i,k} \end{bmatrix},$$

compare (7.74). This then yields the representation

$$\tau_i = Q(t_{i+1}) \begin{bmatrix} \varphi_i \\ 0 \end{bmatrix}, \quad \varphi_i = \mathcal{O}(h_i^{k+1})$$

of the local error τ_i analogously to that of g_i given in Lemma 7.14. The continuity, boundary and consistency conditions for x lead to the global system

$$K_h \begin{bmatrix} x(t_0) \\ \vdots \\ x(t_N) \end{bmatrix} = g_h + \tau_h, \quad \tau_h = \begin{bmatrix} 0 \\ \tau_0 \\ \vdots \\ \tau_{N-1} \end{bmatrix},$$

compare (7.75). Due to the unique solvability of the collocation problem for sufficiently small h, the matrix K_h is nonsingular and the difference of the global systems for x and x_π, respectively, gives

$$K_h \begin{bmatrix} x(t_0) - x_0 \\ \vdots \\ x(t_N) - x_N \end{bmatrix} = \tau_h. \tag{7.90}$$

Replacing g_h by τ_h in the proof of Lemma 7.15 and observing that $\tau_h = \mathcal{O}(h^{k+1})$, we have $K_h^{-1} \tau_h = \mathcal{O}(h^k)$, i.e.,

$$\max_{i=0,\ldots,N} \|x(t_i) - x_i\|_\infty = \mathcal{O}(h^k).$$

Similarly, we obtain

$$\begin{bmatrix} x(s_{i1}) - x_{i,1} \\ \vdots \\ x(s_{ik}) - x_{i,k} \end{bmatrix} = B_i^{-1} a_i (x(t_i) - x_i) - B_i^{-1} \begin{bmatrix} \tau_{i,1} \\ \vdots \\ \tau_{i,k} \end{bmatrix} = \mathcal{O}(h^k) + \mathcal{O}(h_i^k) \quad (7.91)$$

for the difference in the local systems. Thus, we have shown that

$$\max_{j=0,\ldots,k} \|x(s_{ij}) - x_{i,j}\|_\infty = \mathcal{O}(h^k).$$

Looking at the differences of the interpolation representations for x and x_π, respectively, we then find that

$$\max_{t\in[t_i,t_{i+1}]} \|x(t) - x_\pi(t)\|_\infty \le \Gamma h^k.$$

To get an estimate for the derivative $\dot{x} - \dot{x}_\pi$ on $[t_i, t_{i+1}]$, we observe that

$$\begin{aligned} \hat{E}_1(t_{ij})(\dot{x}(t_{ij}) - \dot{x}_\pi(t_{ij})) &= \hat{A}_1(t_{ij})(x(t_{ij}) - x_\pi(t_{ij})), \quad j = 1, \ldots, k, \\ 0 &= \hat{A}_2(s_{ij})(x(s_{ij}) - x_\pi(s_{ij})), \quad j = 0, \ldots, k. \end{aligned}$$

Applying Lagrange interpolation of $x - x_\pi$ at the points $s_{ij} = t_i + h_i\sigma_j$ gives

$$x(t) - x_\pi(t) = \sum_{l=0}^{k} (x(s_{il}) - x_\pi(s_{il})) L_l\left(\frac{t - t_i}{h_i}\right) + \mathcal{O}(h^{k+1}).$$

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

$$\begin{aligned} &\hat{A}_2(t_{ij})(\dot{x}(t_{ij}) - \dot{x}_\pi(t_{ij})) \\ &\quad = \hat{A}_2(t_{ij}) \sum_{l=0}^{k} (x(s_{il}) - x_\pi(s_{il})) \dot{L}_l\left(\frac{t_{ij} - t_i}{h_i}\right) \frac{1}{h_i} + \mathcal{O}(h^k) \\ &\quad = \sum_{l=0}^{k} (\hat{A}_2(s_{il}) + \mathcal{O}(h))(x(s_{il}) - x_\pi(s_{il})) \dot{L}_l(\varrho_j) \frac{1}{h_i} + \mathcal{O}(h^k) \\ &\quad = \sum_{l=0}^{k} (0 + \mathcal{O}(h) \cdot \mathcal{O}(h^k)) \cdot \mathcal{O}(1) \cdot \mathcal{O}(h^{-1}) + \mathcal{O}(h^k) \\ &\quad = \mathcal{O}(h^k), \end{aligned}$$

where all constants do not depend on h. Together with the relation

$$\hat{E}_1(t_{ij})(\dot{x}(t_{ij}) - \dot{x}_\pi(t_{ij})) = \mathcal{O}(h^k),$$

this implies

$$\|\dot{x}(t_{ij}) - \dot{x}_\pi(t_{ij})\|_\infty \le \Gamma h^k,$$

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

$$\dot{x}(t) - \dot{x}_\pi(t) = \sum_{l=1}^{k} (\dot{x}(t_{il}) - \dot{x}_\pi(t_{il}))\tilde{L}_l\left(\frac{t - t_i}{h_i}\right) + \mathcal{O}(h^k).$$

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

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

possibly increasing again the constant Γ. □

Having performed the detailed analysis for the linear boundary value problems associated with the Fréchet derivative $DL_\pi[x^*]$, we return to the nonlinear case. As already mentioned, we want to apply Theorem 5.7, but we must replace the constants β and γ by quantities which do not depend on the maximal mesh width h. With the interpretation of the constant β from (7.86) according to Remark 7.17, we already have an appropriate substitute for the constant β. In order to obtain a corresponding replacement for γ, we must consider the dependence of the operator $DL[z]$ on z.

Lemma 7.19. *Let L from (7.51) be defined on a convex and compact neighborhood $\mathbb{D} \subseteq \mathbb{X}$ of x^*. Then there exists a constant γ that is independent of h such that*

$$\|L(x) - L(y) - DL[z](x - y)\|_{\mathbb{Y}} \le \tfrac{1}{2}\gamma\|x - y\|_{\mathbb{X}}(\|x - z\|_{\mathbb{X}} + \|y - z\|_{\mathbb{X}}) \quad (7.92)$$

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

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

$$\begin{aligned}
&\|f_1(t)\|_\infty \\
&\quad = \|\hat{F}_1(t, x(t), \dot{x}(t)) - \hat{F}_1(t, y(t), \dot{y}(t)) \\
&\qquad - \hat{F}_{1;x}(t, z(t), \dot{z}(t))(x(t) - y(t)) - \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))(\dot{x}(t) - \dot{y}(t))\|_\infty \\
&\quad = \|\hat{F}_1(t, u(t; s), \dot{u}(t; s))|_{s=0}^{s=1} \\
&\qquad - \hat{F}_{1;x}(t, z(t), \dot{z}(t))(x(t) - y(t)) - \hat{F}_{1;\dot{x}}(t, z(t), \dot{z}(t))(\dot{x}(t) - \dot{y}(t))\|_\infty
\end{aligned}$$

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

with all the constants $\gamma_1, \ldots, \gamma_4$ and γ being independent of t, x, y, z, and h. Analogously, we get for the second component f_2

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

possibly increasing γ. Furthermore,

$$
\begin{aligned}
\|\dot{f}_2(t)\|_\infty &= \Big\| \int_0^1 [(\hat{F}_{2;tx}(t, u(t;s)) + \hat{F}_{2;xx}(t, u(t;s))(\dot{u}(t;s)) \\
&\qquad - \hat{F}_{2;tx}(t, z(t)) - \hat{F}_{2;xx}(t, z(t))(\dot{z}(t)))(x(t) - y(t)) \\
&\qquad + (\hat{F}_{2;x}(t, u(t;s)) - \hat{F}_{2;x}(t, z(t)))(\dot{x}(t) - \dot{y}(t))]ds \Big\|_\infty \\
&\le \int_0^1 [(\gamma_1 \|u(t;s) - z(t)\|_\infty + \gamma_2 \|u(t;s) - z(t)\|_\infty \\
&\qquad + \gamma_3 \|\dot{u}(t;s) - \dot{z}(t)\|_\infty)\|x(t) - y(t)\|_\infty \\
&\qquad + \gamma_4 \|u(t;s) - z(t)\|_\infty \|\dot{x}(t) - \dot{y}(t)\|_\infty]ds \\
&\le \gamma \|x - y\|_{\mathbb{X}} \int_0^1 \|u(\,\cdot\,;s) - z\|_{\mathbb{X}} ds,
\end{aligned}
$$

again possibly increasing γ. Finally, for w we get

$$
\begin{aligned}
\|w\|_\infty &= \|b(x(\underline{t}), x(\overline{t})) - b(y(\underline{t}), y(\overline{t})) - b_{x_l}(z(\underline{t}), z(\overline{t}))(x(\underline{t}) \\
&\qquad - y(\underline{t})) - b_{x_r}(z(\underline{t}), z(\overline{t}))(x(\overline{t}) - y(\overline{t}))\|_\infty \\
&= \|b(u(\underline{t};s), u(\overline{t};s))|_{s=0}^{s=1} \\
&\qquad - b_{x_l}(z(\underline{t}), z(\overline{t}))(x(\underline{t}) - y(\underline{t})) - b_{x_r}(z(\underline{t}), z(\overline{t}))(x(\overline{t}) - y(\overline{t}))\|_\infty
\end{aligned}
$$

$$
\begin{aligned}
&= \Big\| \int_0^1 [(b_{x_l}(u(\underline{t};s),u(\bar{t};s)) - b_{x_l}(z(\underline{t}),z(\bar{t})))(x(\underline{t}) - y(\underline{t})) \\
&\qquad + (b_{x_r}(u(\underline{t};s),u(\bar{t};s)) - b_{x_r}(z(\underline{t}),z(\bar{t})))(x(\bar{t}) - y(\bar{t}))]ds \Big\|_\infty \\
&\le \int_0^1 [(\gamma_1\|u(\underline{t};s) - z(\underline{t})\|_\infty + \gamma_2\|u(\bar{t};s) - z(\bar{t})\|_\infty)\|x(\underline{t}) - y(\underline{t})\|_\infty \\
&\qquad + (\gamma_3\|u(\underline{t};s) - z(\underline{t})\|_\infty + \gamma_4\|u(\bar{t};s) - z(\bar{t})\|_\infty)\|x(\bar{t}) - y(\bar{t})\|_\infty]ds \\
&\le \gamma\|x - y\|_{\mathbb{X}} \int_0^1 \|u(\,\cdot\,;s) - z\|_{\mathbb{X}} ds,
\end{aligned}
$$

and thus we have, again possibly increasing γ,

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

□

With these modifications of the constants β and γ, we can show that the crucial estimates in the proof of Theorem 5.7 still hold. In particular, we have

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

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

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

Thus, to apply Theorem 5.7, it only remains to discuss the assumptions concerning the quantities $\alpha = \|x_\pi^1 - x_\pi^0\|_{\mathbb{X}}$ and $\hat{\tau} = -\|x_\pi^0 - x^*\|_{\mathbb{X}}$. Both involve the choice of the starting value x_π^0 of the iteration (7.61). To get an appropriate x_π^0, we consider the linear boundary value problem

$$
DL[x^*]x = DL[x^*]x^* \tag{7.93}
$$

and its collocation discretization

$$
DL_\pi[x^*]x = DL_\pi[x^*]x^*. \tag{7.94}
$$

Since (7.93) has the unique solution x^* by construction, Theorem 7.88 yields that (7.94) has a unique solution $x_\pi^0 \in \mathbb{X}_\pi$ for sufficiently small h satisfying

$$\|x^* - x_\pi^0\|_{\mathbb{X}} \le \Gamma h^k. \tag{7.95}$$

In particular, we can choose h so small that

$$\|x^* - x_\pi^0\|_{\mathbb{X}} \le \frac{1}{2\beta\gamma}, \tag{7.96}$$

cp. Corollary 5.8. Moreover, because of

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

we have

$$\|x_\pi^1 - x_\pi^0\|_{\mathbb{X}} \le \tilde{\Gamma} h^k \tag{7.97}$$

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

$$\alpha = \|x_\pi^1 - x_\pi^0\|_{\mathbb{X}} \le \frac{1}{9\beta\gamma} \tag{7.98}$$

and that $\overline{S}(x_\pi^0, 4\alpha) \subseteq \mathbb{D}$. It follows then inductively, as in the proof of Theorem 5.7, that (7.61) generates a sequence $\{x_\pi^m\}$ with

$$x_\pi^m \in \overline{S}(x_\pi^0, 4\alpha) \cap \mathbb{X}_\pi. \tag{7.99}$$

Since $\mathbb{X}_\pi \subseteq \mathbb{X}$ is closed, the sequence converges to an $x_\pi^* \in \overline{S}(x_\pi^0, 4\alpha) \cap \mathbb{X}_\pi$ with $L_\pi(x_\pi^*) = 0$. Local uniqueness follows, since (7.96) and (7.98) imply that $\rho_- < \rho_+$. Observing that

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

by Corollary 5.8, we have shown the following result.

Theorem 7.20. *Let $x^* \in \mathbb{X}$ be a regular solution of $L(x) = 0$. Then, for sufficiently small h, there exists a locally unique solution $x_\pi^* \in \mathbb{X}_\pi$ of $L_\pi(x_\pi) = 0$. In particular, the estimate*

$$\|x^* - x_\pi^*\|_{\mathbb{X}} \le \Gamma h^k \tag{7.100}$$

holds, with Γ independent of h.

Hence, the proposed collocation discretization given by (7.57) represents a method which is convergent of order k in the sense of (7.100), whenever we choose the nodes according to (7.55).

It is well-known from the case of ordinary differential equations that special choices of the nodes yield even higher order compared to (7.100) at least at the mesh points t_i. This effect is called *superconvergence*. In the following, we will discuss this issue for the situation that we choose Gauß nodes for $\varrho_1, \dots, \varrho_k$ and Lobatto nodes for $\sigma_0, \dots, \sigma_k$. These are given by transforming the interval $[0, 1]$ to $[-1, 1]$ and defining the modified Gauß nodes $\tilde{\varrho}_j = 2\varrho_j - 1$, $j = 1, \dots, k$, to be the zeros of the Legendre polynomial P_k and the modified Lobatto nodes $\tilde{\sigma}_j = 2\sigma_j - 1$, $j = 1, \dots, k-1$, to be the zeros of $\dot{P}_k$, respectively. Since the Legendre polynomials form an orthogonal system in the space of polynomials, we have

$$\int_0^1 q(s) \prod_{j=1}^{k} (s - \varrho_j) ds = 0 \quad \text{for all } q \in \mathbb{P}_k, \tag{7.101a}$$

$$\int_0^1 q(s) \prod_{j=0}^{k} (s - \sigma_j) ds = 0 \quad \text{for all } q \in \mathbb{P}_{k-1}. \tag{7.101b}$$

Moreover, the two sets of nodes are related by

$$\int_0^{\sigma_j} \prod_{l=1}^{k} (s - \varrho_l) ds = 0, \quad j = 0, \dots, k. \tag{7.102}$$

See, e.g., [108, Ch. IV] for more details on this topic. For the cases $k = 1, 2, 3$, the Gauß and Lobatto nodes are given in Tables 7.1 and 7.2, respectively.

We begin the analysis of the collocation methods belonging to this special choice of the nodes again by studying the convergence order in the case of linear boundary value problems.

Theorem 7.21. *Consider Gauß nodes for $\varrho_1, \dots, \varrho_k$ and Lobatto nodes for $\sigma_0, \dots, \sigma_k$. For sufficiently small h, let x_π be the unique solution of the linear boundary value problem (7.66). Then, for the resulting approximations $x_i = x_\pi(t_i)$ the estimate*

$$\max_{i=0,\dots,N} \|x(t_i) - x_i\|_\infty = \mathcal{O}(h^{2k}) \tag{7.103}$$

holds.

Proof. Since x_i is consistent, the initial value problem $\hat{E}\dot{y} = \hat{A}y + f$, $y(t_i) = x_i$ is uniquely solvable and the solution y has a representation

$$(Q^{-1}y)(t) = \begin{bmatrix} [\, I \;\; 0 \,](Q(t_i)^{-1}x_i + \int_{t_i}^{t} (P\hat{f})(s) ds) \\ -(P_{22}\hat{f}_2)(t) \end{bmatrix},$$

Table 7.1. Gauß nodes

ϱ_j	$j=1$	$j=2$	$j=3$
$k=1$	$\frac{1}{2}$		
$k=2$	$\frac{1}{2}-\frac{\sqrt{3}}{6}$	$\frac{1}{2}+\frac{\sqrt{3}}{6}$	
$k=3$	$\frac{1}{2}-\frac{\sqrt{15}}{10}$	$\frac{1}{2}$	$\frac{1}{2}+\frac{\sqrt{15}}{10}$

Table 7.2. Lobatto nodes

σ_j	$j=0$	$j=1$	$j=2$	$j=3$
$k=1$	0	1		
$k=2$	0	$\frac{1}{2}$	1	
$k=3$	0	$\frac{1}{2}-\frac{\sqrt{5}}{10}$	$\frac{1}{2}+\frac{\sqrt{5}}{10}$	1

when we use the transformation (7.65) to canonical form. The approximation x_π is the solution of the initial value problem

$$\hat{E}\dot{y} = \hat{A}y + (\hat{E}\dot{x}_\pi - \hat{A}x_\pi), \quad y(t_i) = x_i,$$

and can therefore be written in the form

$$(Q^{-1}x_\pi)(t) = \begin{bmatrix} [\, I \;\; 0\,](Q(t_i)^{-1}x_i + \int_{t_i}^{t}(P(\hat{E}\dot{x}_\pi - \hat{A}x_\pi))(s)ds) \\ (P_{22}\hat{A}_2 x_\pi)(t) \end{bmatrix}.$$

Since x_π is consistent at the mesh point t_{i+1}, the difference of these representations at $t = t_{i+1}$ gives

$$y(t_{i+1}) - x_{i+1} = Q(t_{i+1}) \begin{bmatrix} \int_{t_i}^{t_{i+1}} \phi_d(s)ds + \int_{t_i}^{t_{i+1}} \phi_a(s)ds \\ 0 \end{bmatrix}, \tag{7.104}$$

with functions

$$\phi_d = P_{11}(\hat{f}_1 - \hat{E}_1\dot{x}_\pi + \hat{A}_1 x_\pi), \quad \phi_a = P_{12}(\hat{f}_2 + \hat{A}_2 x_\pi).$$

Since x_π satisfies the collocation conditions, the collocation points $t_{i1}, \ldots, t_{ik}$ are zeros of ϕ_d and $s_{i0}, \ldots, s_{ik}$ are zeros of ϕ_a, respectively. Hence, there exist smooth functions ω_d and ω_a with

$$\phi_d(s) = \omega_d(s)\prod_{j=1}^{k}(s - t_{ij}), \quad \phi_a(s) = \omega_a(s)\prod_{j=0}^{k}(s - s_{ij}).$$

Taylor expansion yields $\omega_d = \psi_d + \mathcal{O}(h_i^k)$, $\omega_a = \psi_a + \mathcal{O}(h_i^{k-1})$ with polynomials $\psi_d \in \mathbb{P}_k$ and $\psi_a \in \mathbb{P}_{k-1}$, respectively. Inserting this into (7.104) and using the orthogonality properties (7.101) of the Gauß and Lobatto schemes, we obtain

$$\begin{aligned}\int_{t_i}^{t_{i+1}} \phi_d(s)ds &= \int_{t_i}^{t_{i+1}} [\psi_d(s)\prod_{j=1}^{k}(s-t_{ij}) + \mathcal{O}(h_i^{2k})]ds \\ &= h_i^{k+1}\int_0^1 \psi_d(t_i+h_i\xi)\prod_{j=1}^{k}(\xi-\varrho_j)d\xi + \mathcal{O}(h_i^{2k+1}),\\ \int_{t_i}^{t_{i+1}} \phi_a(s)ds &= \int_{t_i}^{t_{i+1}} [\psi_a(s)\prod_{j=0}^{k}(s-s_{ij}) + \mathcal{O}(h_i^{2k})]ds \\ &= h_i^{k+2}\int_0^1 \psi_a(t_i+h_i\xi)\prod_{j=0}^{k}(\xi-\sigma_j)d\xi + \mathcal{O}(h_i^{2k+1}).\end{aligned}$$

Altogether, we then have that

$$\tilde{\tau}_i = y(t_{i+1}) - x_{i+1} = Q(t_{i+1})\begin{bmatrix}\int_{t_i}^{t_{i+1}}\phi_d(s)ds + \int_{t_i}^{t_{i+1}}\phi_a(s)ds\\ 0\end{bmatrix} = \mathcal{O}(h_i^{2k+1}).$$

Considering a fundamental solution $W(\,\cdot\,, t_i)$, i.e., a solution of

$$\hat{E}\dot{W} = \hat{A}W, \quad W(t_i,t_i) = Q(t_i)\begin{bmatrix}I & 0\\ 0 & 0\end{bmatrix}Q(t_i)^{-1},$$

we see that $x(t) - y(t) = W(t,t_i)(x(t_i) - y(t_i))$. Setting $t = t_{i+1}$, we get in particular that

$$W(t_{i+1},t_i)(x(t_i)-x_i) = x(t_{i+1}) - y(t_{i+1}) = x(t_{i+1}) - x_{i+1} - \tilde{\tau}_i$$

for $i = 0,\ldots,N-1$. This, together with the boundary condition and the consistency condition in t_0, gives the linear system

$$\begin{bmatrix} C & & & D\\ -A_2(t_0) & & & 0\\ W(t_1,t_0) & -I & & \\ & \ddots & \ddots & \\ & & W(t_N,t_{N-1}) & -I\end{bmatrix}\begin{bmatrix}x(t_0)-x_0\\ \vdots\\ x(t_N)-x_N\end{bmatrix} = \begin{bmatrix}0\\0\\-\tilde{\tau}_0\\ \vdots\\ -\tilde{\tau}_{N-1}\end{bmatrix},$$

compare (7.90). Proceeding as in the proof of Theorem 7.18, we get

$$\max_{i=0,\ldots,N}\|x(t_i)-x_i\|_\infty = \mathcal{O}(h^{2k}),$$

since the inhomogeneity is of size $\mathcal{O}(h^{2k+1})$. □

Corollary 7.22. *Under the assumptions of Theorem 7.22, we have that*

$$\max_{j=0,\dots,k} \|x(s_{ij}) - x_{ij}\|_\infty = \mathcal{O}(h_i^{k+2}) + \mathcal{O}(h^{2k}) \tag{7.105}$$

for $i = 0, \dots, N-1$ *if* $k \geq 2$, *and*

$$\max_{t\in\mathbb{I}} \|x(t) - x_\pi(t)\|_\infty = \mathcal{O}(h^{k+1}). \tag{7.106}$$

Proof. Looking at (7.91) in the proof of Theorem 7.18 and using $x(t_i) - x_i = \mathcal{O}(h^{2k})$ due to Theorem 7.22, it is sufficient to show that $B_i^{-1}[\tau_{i,j}]_j = \mathcal{O}(h_i^{k+2})$ in order to prove the first assertion. Recall the definition of ψ_i in (7.89). Transformation to the canonical form of Lemma 5.6 yields

$$\begin{aligned} P_{11}(\hat{E}_1\dot{\psi}_i - \hat{A}_1\psi_i) &= (P_{11}\hat{E}_1 Q)\tfrac{d}{dt}(Q^{-1}\psi_i) - (P_{11}\hat{A}_1 Q - P_{11}\hat{E}_1\dot{Q})(Q^{-1}\psi_i) \\ &= [\, I \;\; 0\,]\tfrac{d}{dt}(Q^{-1}\psi_i) + (P_{12}\hat{A}_2 Q)(Q^{-1}\psi_i) = \dot{\varphi} + \mathcal{O}(h_i^{k+1}), \end{aligned}$$

when we define $\varphi = [\, I \;\; 0\,](Q^{-1}\psi_i)$. By interpolation of $\dot{\varphi}$ with respect to ϱ_l and by Taylor expansion of the interpolation error, we obtain

$$\begin{aligned} \sum_{l=1}^{k} \tilde{L}_l\big(\tfrac{t-t_i}{h_i}\big)\dot{\varphi}(t_{il}) &= \dot{\varphi}(t) - \frac{\dot{\varphi}^{(k)}(\theta(t))}{k!}\prod_{l=1}^{k}(t - t_{il}) \\ &= \dot{\varphi}(t) - K\prod_{l=1}^{k}(t - t_{il}) + \mathcal{O}(h_i^{k+1}) \end{aligned}$$

with $K = \frac{1}{k!}\varphi^{(k+1)}(t_i)$. Inserting the definition of w_{jl} given in (7.70) leads to

$$\begin{aligned} \sum_{l=1}^{k} w_{jl}\dot{\varphi}(t_{il}) &= \int_0^{\sigma_j} \sum_{l=1}^{k} \tilde{L}_l(\xi)\dot{\varphi}(t_{il})d\xi = \tfrac{1}{h_i}\int_{t_i}^{s_{ij}} \sum_{l=1}^{k} \tilde{L}_l\big(\tfrac{t-t_i}{h_i}\big)\dot{\varphi}(t_{il})dt \\ &= \tfrac{1}{h_i}\int_{t_i}^{s_{ij}} \dot{\varphi}(t)dt - \tfrac{K}{h_i}\int_{t_i}^{s_{ij}} \prod_{l=1}^{k}(t - t_{il})dt + \mathcal{O}(h_i^{k+1}) \\ &= \frac{\varphi(s_{ij}) - \varphi(t_i)}{h_i} - Kh_i^k\int_0^{\sigma_j}\prod_{l=1}^{k}(\xi - \varrho_l)d\xi + \mathcal{O}(h_i^{k+1}) \\ &= \mathcal{O}(h_i^{k+1}), \end{aligned}$$

since $s_{ij}, t_i = s_{i0}$ are zeros of ψ_i and thus of φ, and since the second term is zero due to (7.102). Altogether, recalling that $V^{-1} = [w_{jl}]_{j,l}$, we have

$$U_k^T T_P[\tau_{i,j}]_j = \begin{bmatrix} \big[(P_{11}(\hat{E}_1\dot{\psi}_i - \hat{A}_1\psi_i))(t_{ij})\big]_j \\ 0 \end{bmatrix} = \begin{bmatrix} \big[\dot{\varphi}(t_{ij}) + \mathcal{O}(h_i^{k+1})\big]_j \\ 0 \end{bmatrix},$$

which implies that

$$\begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P [\tau_{i,j}]_j = h_i \begin{bmatrix} \left[\sum_{l=1}^k w_{jl} \dot{\varphi}(t_{il})\right]_j \\ 0 \end{bmatrix} + \mathcal{O}(h_i^{k+2}) = \mathcal{O}(h_i^{k+2})$$

and therefore

$$B_i^{-1}[\tau_{i,j}]_j = T_Q U_k (I - \Delta_i + \mathcal{O}(h_i^2)) \begin{bmatrix} h_i V^{-1} \otimes I & 0 \\ 0 & -I \end{bmatrix} U_k^T T_P [\tau_{i,j}]_j = \mathcal{O}(h_i^{k+2}).$$

The convergence order $k+1$ for any $t \in \mathbb{I}$ can now be proved by considering the difference of the interpolation representations for x and x_π. □

In order to transfer the estimates of superconvergence from the linear to the nonlinear case, we observe that

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

Hence,

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

and it follows with (7.100) and (7.94) that

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

where once more the involved constants in the remainders are independent of h. Application of (7.86) yields

$$x_\pi^* = x_\pi^0 + \mathcal{O}(h^{2k}). \tag{7.107}$$

In particular, we have

$$\begin{aligned} x_\pi^*(t) - x^*(t) &= (x_\pi^*(t) - x_\pi^0(t)) + (x_\pi^0(t) - x^*(t)) \\ &= x_\pi^0(t) - x^*(t) + \mathcal{O}(h^{2k}) \end{aligned} \tag{7.108}$$

for all $t \in \mathbb{I}$ and the estimates (7.103), (7.105), and (7.106) applied to (7.94) carry over to the nonlinear case. Thus, we have proved the following superconvergence result.

Theorem 7.23. *Let the assumptions of Theorem 7.20 hold and let $\varrho_1, \dots, \varrho_k$ and $\sigma_0, \dots, \sigma_k$ be Gauß and Lobatto nodes, respectively. Then*

$$\max_{i=0,\dots,N} \|x^*(t_i) - x_\pi^*(t_i)\|_\infty = \mathcal{O}(h^{2k}), \tag{7.109a}$$

$$\max_{j=0,\dots,k} \|x^*(s_{ij}) - x_\pi^*(s_{ij})\|_\infty = \mathcal{O}(h^{k+2}) \quad \text{for } k \geq 2, \tag{7.109b}$$

$$\max_{t \in \mathbb{I}} \|x^*(t) - x_\pi^*(t)\|_\infty = \mathcal{O}(h^{k+1}), \tag{7.109c}$$

with constants independent of h.

Up to now, we have considered collocation only for the special case of boundary value problems for nonlinear differential-algebraic equations of the form (7.47). To transfer the problem $L_\pi(x_\pi) = 0$, written in more detail as

$$\begin{aligned} &\hat{F}_1(t_{ij}, x_\pi(t_{ij}), \dot{x}_\pi(t_{ij})) = 0, \\ &\hat{F}_2(s_{ij}, x_\pi(s_{ij})) = 0, \\ &b(x_\pi(\underline{t}), x_\pi(\bar{t})) = 0, \end{aligned} \tag{7.110}$$

to a nonlinear problem (7.1), where F satisfies Hypothesis 4.2, we proceed as in Section 6.2. The function $\hat{F}_1$ represents a suitable differential part of F, which we must select depending on the collocation point t_{ij}. For this, we choose suitable matrices $\tilde{Z}_{1,ij} \in \mathbb{R}^{n,d}$. The function $\hat{F}_2$ represents the algebraic constraints of the differential-algebraic equation. Following Section 6.2, consistency of $x_\pi(s_{ij})$ at s_{ij} is given when there exists a y_{ij} such that $(s_{ij}, x_\pi(s_{ij}), y_{ij}) \in \mathbb{L}_\mu$. In summary, we therefore replace (7.110) by

$$\begin{aligned} &\tilde{Z}_{1,ij}^T F(t_{ij}, x_\pi(t_{ij}), \dot{x}_\pi(t_{ij})) = 0, \\ &F_\mu(s_{ij}, x_\pi(s_{ij}), y_{ij}) = 0, \\ &b(x_\pi(\underline{t}), x_\pi(\bar{t})) = 0, \end{aligned} \tag{7.111}$$

which constitutes an underdetermined nonlinear system in the unknowns

$$(x_\pi, y_{ij}) \in \mathbb{P}_{k+1,\pi} \cap C^0(\mathbb{I}, \mathbb{R}^n) \times \mathbb{R}^{(Nk+1)(\mu+1)n}$$

with $i = 0, \ldots, N-1$, $j = 1, \ldots, k$ and $i = 0$, $j = 0$.

The iteration process of choice for the numerical solution of (7.111) is a Gauß–Newton-like method of the form

$$z^{m+1} = z^m - \mathcal{M}_m^+ \mathcal{F}(z^m), \quad z^m = (x_\pi^m, y_{ij}^m), \tag{7.112}$$

when we write (7.111) as $\mathcal{F}(z) = 0$. In contrast to the ordinary Gauß–Newton method, we replace the Jacobian $\mathcal{F}_z(z^m)$ by a perturbed matrix $\mathcal{M}_m$ in order to get a more efficient procedure. In particular, we determine $\mathcal{M}_m$ from $\mathcal{F}_z(z_m)$ in such a way that we replace the block entries $F_{\mu;\dot{x},\ldots,x^{(\mu+1)}}(s_{ij}, x_\pi^m(s_{ij}), y_{ij}^m)$ by matrices of rank deficiency a, e.g., by ignoring the a smallest singular values. This decouples the determination of $\Delta y_{ij}^m = y_{ij}^{m+1} - y_{ij}^m$ for each i, j from the corrections for x_π^m and leaves a linear system, which has the same form as the linear system arising in the case of a linear boundary value problem. Thus, we can use the presented technique of solving first local systems of the form (7.72) and then a global system of the form (7.75). Having computed the corrections for x_π^m, it then remains the solution of the decoupled underdetermined linear systems for the Δy_{ij}^m. Taking the Moore–Penrose pseudoinverse to select a solution realizes the Moore–Penrose pseudoinverse of the overall system due to the decoupling. Since the applied perturbations tend to zero when z^m converges to a solution, we get superlinear convergence as in Theorem 7.7.

Theorem 7.24. *Suppose that the boundary value problem* (7.1) *satisfies Hypothesis* 4.2 *and that* (7.1) *has a locally unique solution according to Theorem* 7.2. *Then, for sufficiently good initial guesses and sufficiently small h, the iterates of the Gauß–Newton-like procedure developed in the course of this section converge superlinearly to a solution of* (7.111).

Remark 7.25. In contrast to multiple shooting where it was assumed that we can solve initial value problems exactly and where we therefore did not need to consider discretization errors, the error of the collocation solution depends on the quality of the selected mesh. As in the solution of initial value problems, it is therefore necessary for real-life applications to adapt the mesh during the numerical solution of the problem. For strategies in the context of boundary value problems, see, e.g., [8].

Bibliographical remarks

Boundary value problems for ordinary differential equations are well covered in the literature, see for example [8]. For differential-algebraic equations, collocation methods are studied for example in [10], [12], [13], [14], [15], [21], [111], [118], see also [11]. Shooting methods for certain classes of differential-algebraic equations were presented in [81], [137], [145], [197].

The results of this section are based on [136], [207], [208] in the linear case and [130], [131] in the nonlinear case. The convergence analysis for Gauß–Newton-like methods can be found in [71], [160].

Exercises

1. Show that the Radau IIA methods with $s = 1$ and $s = 2$, see Table 5.2, applied to (7.47) can be written in the form

$$\hat{F}_1\left(t_i + \varrho_j h, u_{j0} x_i + \sum_{l=1}^{k} u_{jl} x_{i,j}, \frac{1}{h}\left(v_{j0} x_i + \sum_{l=1}^{k} v_{jl} x_{i,j}\right)\right) = 0,$$
$$\hat{F}_2(t_i + \sigma_j h, x_{i,j}) = 0,$$

   with $j = 1, \ldots, k$ and $x_{i+1} = x_{i,k}$.

2. Since initial value problems are special boundary value problems, the collocation method given by $L_\pi(x_\pi) = 0$ defines a one-step method for the solution of (7.47). Using (7.69),

this one-step method has the form

$$\hat{F}_1\Big(t_i + \varrho_j h, u_{j0} x_i + \sum_{l=1}^{k} u_{jl} x_{i,j}, \frac{1}{h}\Big(v_{j0} x_i + \sum_{l=1}^{k} v_{jl} x_{i,j}\Big)\Big) = 0,$$
$$\hat{F}_2(t_i + \sigma_j h, x_{i,j}) = 0,$$

with $j = 1, \ldots, k$ and $x_{i+1} = x_{i,k}$. Determine the coefficients when we choose Gauß and Lobatto nodes for $k = 1$ and $k = 2$ according to Table 7.1 and Table 7.2.

3. Implement the methods of Exercise 1 and Exercise 2 for the solution of initial value problems with a differential-algebraic equation of the form (7.47) using a constant stepsize. Use Newton's method and finite differences to approximate the Jacobians for the solution of the arising nonlinear systems. Apply the program to the regular strangeness-free problem

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} \exp(t) + \cos(t) \\ \exp(t) \end{bmatrix}, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and verify the claimed orders numerically.

4. Implement the single shooting method as special case of the multiple shooting method for $N = 1$. Use the simplification that the differential-algebraic equation is already given in the form (7.47) and that suitable matrices $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$ and a suitable function S_0 are available. Furthermore, set $x_1 = \Phi_0(S_0(x_0, y_0))$ such that only the unknowns (x_0, y_0) must be considered. Take the program of Exercise 3 to solve the arising initial value problems and use Newton's method and finite differences to approximate the Jacobians for the solution of the arising nonlinear systems. Write the problem of Exercise 3 as a boundary value problem and determine suitable quantities $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$, and S_0. Test your implementation with the help of the so obtained problem.
5. Determine suitable quantities $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$, and S_0 for the problem of Example 7.1 and apply the program developed in Exercise 4.
6. Determine suitable quantities $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$, and S_0 for the problem of Example 7.11 and apply the program developed in Exercise 4, using the initial guess $(p_{1,0}, p_{2,0}, v_{1,0}, v_{2,0}) = (1, 0, 0, 0)$.
7. Show that $V = [v_{jl}]_{j,l=1,\ldots,k}$ is nonsingular with $V^{-1} = [w_{jl}]_{j,l=1,\ldots,k}$ as in (7.70) by verifying that $V^{-1}V = I_k$.
8. Verify (7.101) and (7.102) for $k = 1, 2, 3$ using Tables 7.1 and 7.2.
9. Implement the collocation method of Section 7.3 using equidistant mesh points $t_i = t_0 + ih$, $i = 0, \ldots, N$, with $h = (\bar{t} - \underline{t})/N$, and Gauß and Lobatto nodes for $k = 1$. Assuming that the differential-algebraic equation is already given in the form (7.47), the collocation system can be written as

$$\begin{aligned} &\hat{F}_1\Big(t_i + \frac{1}{2}h, \frac{1}{2}(x_i + x_{i+1}), \frac{1}{h}(x_{i+1} - x_i)\Big) = 0, && i = 0, \ldots, N-1, \\ &\hat{F}_2(t_i, x_i) = 0, && i = 0, \ldots, N, \\ &b(x_0, x_N) = 0. \end{aligned}$$

Take the program of Exercise 3 to generate suitable initial guesses for x_i, $i = 0, \ldots, N$, and use Newton's method and finite differences to approximate the Jacobians for the solution of the arising nonlinear systems. Write the problem of Exercise 3 as a boundary value problem and test your implementation with the help of the so obtained problem.

10. Apply the program developed in Exercise 9 to the problem of Example 7.1.
11. Apply the program developed in Exercise 9 to the problem of Example 7.11.
12. Apply the programs developed in Exercise 4 and in Exercise 9 to the linear problem of Example 7.10. How large can you choose the parameter λ to still get a numerical solution?
13. The simplest model for a predator-prey interaction is given by the so-called Lotka–Volterra system
$$\dot{x}_1 = \alpha x_1 - \beta x_1 x_2, \quad \dot{x}_2 = -\gamma x_2 + \delta x_1 x_2,$$
where x_1, x_2 are measures for the population size of prey and predator, respectively, and $\alpha, \beta, \gamma, \delta > 0$. Show that this system can be transformed by
$$\tilde{x}_1(\tau) = \xi_1 x_1(\lambda\tau), \quad \tilde{x}_2(\tau) = \xi_2 x_2(\lambda\tau)$$
with appropriate constants $\lambda, \xi_1, \xi_2 > 0$ into
$$\dot{\tilde{x}}_1 = \tilde{x}_1(1 - \tilde{x}_2), \quad \dot{\tilde{x}}_2 = -c\tilde{x}_2(1 - \tilde{x}_1)$$
with some constant $c > 0$.
14. Consider the Lotka–Volterra system
$$\dot{x}_1 = x_1(1 - x_2), \quad \dot{x}_2 = -cx_2(1 - x_1)$$
with a given constant $c > 0$.
    (a) Show that the function $H : \mathbb{D} \to \mathbb{R}$ with $\mathbb{D} = \{(x_1, x_2) \mid x_1, x_2 > 0\}$ and
    $$H(x_1, x_2) = c(x_1 - \log x_1) + (x_2 - \log x_2)$$
    stays constant along every solution of the transformed system which starts in $\mathbb{D}$.
    (b) Show that H, restricted to the curve $x_2 = x_1^q$, $q \in \overline{\mathbb{R}}$, with the convention $x_1 = 1$ for $q = \pm\infty$, has a minimum $z_{\min}$ at $(x_1, x_2) = (1, 1)$ and assumes every value $z > z_{\min}$ exactly twice. Prove on this basis that all solutions of the transformed system with $x_1(t_0), x_2(t_0) > 0$ stay in $\mathbb{D}$ and are periodic.
15. In order to determine a periodic solution of the Lotka–Volterra system of Exercise 14, we consider the boundary value problem
$$\begin{aligned} &(1 - x_1)\dot{x}_2 - c(1 - x_2)\dot{x}_1 + cx_2(1 - x_1)^2 + cx_1(1 - x_2)^2 = 0, \\ &c(x_1 - \log x_1) + (x_2 - \log x_2) - z = 0, \\ &x_1(0) = x_1(T), \quad x_1(0) = 1, \end{aligned}$$
where z is the desired level of the function H and T is the unknown period of the solution. The first boundary condition requires the solution to be periodic, whereas the second boundary condition is used to fix the phase of the solution. The latter is necessary, since shifting a solution of an autonomous problem in time yields again a solution.

(a) Show that the given differential-algebraic equation satisfies Hypothesis 4.2, provided that we exclude the point $(x_1, x_2) = (1, 1)$.

(b) Transform the given boundary value problem to the interval $[0, 1]$ by scaling time and using $x_3 = T$ as further unknown.

16. Determine suitable quantities $\tilde{T}_{1,0}$, $\tilde{T}_{2,0}$, and S_0 for the problem of Exercise 15 and apply the program developed in Exercise 4, using the initial guess $(x_{1,0}, x_{2,0}, x_{3,0}) = (1.0, 0.6, 6.0)$.

17. Apply the program developed in Exercise 9 to the problem of Exercise 15.

Chapter 8
Software for the numerical solution of differential-algebraic equations

Due to the great importance of differential-algebraic equations in applications, many of the numerical techniques for differential-algebraic equations have been implemented in software packages. In this section, we will give a survey over the available codes and briefly describe their particular features.

Most of the available software packages for differential-algebraic equations are FORTRAN or C subroutine libraries and can be found via the internet from software repositories or homepages. We have listed all relevant web addresses in Table 8.1.

The most important software repositories for differential-algebraic equations are the NETLIB server ① and the test set for initial value problems ②. Some solvers are also commercially available from the NAG ③ or IMSL library ④ or in mathematical software environments like MATLAB ⑤, SCILAB ⑥ or OCTAVE ⑦. Special codes for multibody systems are available in commercial packages like Simpack ⑲, ADAMS ⑳, or DYMOLA ㉑. Also some symbolic computation packages such as MAPLE ⑧, MATHEMATICA ⑨ contain routines for the solution of differential-algebraic equations.

We will present now a brief overview over some of the existing non-commercial codes.

For problems of the form

$$E\dot{x} = f(t, x), \quad x(t_0) = x_0, \tag{8.1}$$

where E is a constant square matrix, several alternative codes are available.

A generalized Adams Method is the code GAMD ② by Iavernaro and Mazzia [114], [115]. Widely used are the 2-stage Radau IIA method, implemented in RADAU5 ⑩, a diagonally-implicit Runge–Kutta method of order 4, implemented in SDIRK4 ⑩, and a Rosenbrock method of order 4(3), implemented in RODAS ⑩, all due to Hairer and Wanner, see [108]. A BDF code for the same problem is MEBDFDAE ① of Cash [60]. Some of these codes also contain variants that apply to the more general problem

$$E(t, x)\dot{x} = f(t, x), \quad x(t_0) = x_0, \tag{8.2}$$

for which also an extrapolation code based on the linear-implicit Euler method called LIMEX ⑪ of Deuflhard, Hairer, and Zugck [73] is available. For (8.2) the code MEBDFV ⑫, an extension of MEBDFDAE, was developed by Abdulla and

①	www.netlib.org	NETLIB
②	pitagora.dm.uniba.it/~testset/	IVP Testset
③	www.nag.co.uk/	NAG
④	http://absoft.com/	IMSL
⑤	www.mathworks.com/	MATLAB
⑥	scilabsoft.inria.fr/	SCILAB
⑦	www.octave.org/	OCTAVE
⑧	www.maplesoft.com/	MAPLE
⑨	www.wolfram.com/	MATHEMATICA
⑩	www.unige.ch/math/folks/hairer/	E. Hairer
⑪	www.zib.de/Software/	Zuse Inst. Berlin
⑫	www.ma.ic.ac.uk/~jcash/	J. R. Cash
⑬	www.cwi.nl/cwi/projects/PSIDE	PSIDE
⑭	www.math.tu-berlin.de/numerik/mt/NumMat/	TU Berlin
⑮	synmath.synoptio.de/	SynOptio
⑯	www.llnl.gov/CASC/sundials/	SUNDIALS
⑰	www1.iwr.uni-heidelberg.de/	IWR Heidelberg
⑱	www-m2.ma.tum.de/~simeon/numsoft.html	B. Simeon
⑲	www.simpack.de/	SIMPACK
⑳	www.mscsoftware.com/	ADAMS
㉑	www.dynasim.com/	DYMOLA

Table 8.1. Web pages

Cash. A further extension of MEBDFV is MEBDFI ① by the same authors for the general implicit problem

$$F(t, x, \dot{x}) = 0, \quad x(t_0) = x_0. \tag{8.3}$$

An implementation of the 2-stage Radau IIA method for (8.3) is the code PSIDE ⑬ of Lioen, Swart and van der Veen [141], [142] which is specifically designed for the use on parallel computers.

For this general problem, the BDF code DASSL ① of Petzold [164] and several of its variations and extensions (including root finding and iterative solution methods for large scale problems) are all included in the code DASKR ①.

For different classes of quasilinear problems, a package DAESOLVE ① has been implemented by Rheinboldt [181], [193].

All the above listed codes have limitations to the index that they can handle, which is typically a differentiation index of at most three.

For general linear problems with variable coefficients of arbitrary strangeness index, i.e., for systems of the form

$$E(t)\dot{x} = A(t)x + f(t), \quad x(t_0) = x_0, \tag{8.4}$$

two versions of the code GELDA by the authors [133] are available. Version 1.0 ① is for the square case and Version 2.0 ⑭ for general nonsquare problems. Both versions use adaptations of the codes RADAU5 and DASSL as integrators.

A general nonlinear code GENDA ⑭ for square problems of the form (8.3) of arbitrary index has been developed by the authors [135]. A MATLAB ⑤ interface solvedae ⑮ for both codes GELDA and GENDA including symbolic generation of the derivative arrays is commercially available.

A very recent development is the SUite of Nonlinear and DIfferential/ALgebraic equation Solvers SUNDIALS ⑯ which builds on many of the existing methods for ordinary differential equations and differential-algebraic equations and includes serial and parallel versions and a MATLAB interface.

There are also many special algorithms that have been implemented for structured problems. These include the codes MEXAX ⑪ of Lubich, Nowak, Pöhle and Engstler [144], MBSSIM ⑰ of von Schwerin and Winckler [218], ODASSL of Führer [85] which is available from the author, see also [79], GEOMS ⑭ of Steinbrecher [205], and MBSpack ⑱ of Simeon [201] for multibody systems.

Not much software has been developed for boundary value problems associated with differential-algebraic equations. Essentially, the only code is COLDAE ① of Ascher and Spiteri [12] for semi-explicit systems of index $\nu = 2$ of the form

$$\begin{aligned} \dot{x} &= f(t, x, y), \\ 0 &= g(t, x, y). \end{aligned}$$

Remark 8.1. Several of the available codes are in a state of flux. This concerns, in particular, better implementations on modern computer architectures, the use of preconditioned iterative methods for the solution of the linear systems that arise at every iteration step as well as the use of symbolic or automatic differentiation packages to determine the necessary Jacobians and derivative arrays.

As is standard in the design of modern software packages, the implementation of most of these codes is based on existing basic linear algebra subroutines BLAS, see [75], [139], which are usually provided with the computer architecture, and use high quality linear algebra packages such as LAPACK ① or SCALAPACK ①, see [1] and [26], respectively, for the solution of problems like the singular value decomposition, linear system solution or the solution of least squares problems. A possible solver for nonlinear problems is NLSCON ⑪, see [158].

Furthermore, in the solution of the linear and nonlinear systems that arise in the time stepping procedures, most of the methods need appropriate scaling routines and the design of such routines is essential in getting the codes to work.

Remark 8.2. Almost all the available codes contain order and step size control mechanisms. For BDF codes, order and stepsize control is described in detail in [29] and for Runge–Kutta methods such techniques are discussed in [106], [108].

Bibliographical remarks

There is a multitude of numerical methods for differential-algebraic equations that we have not described here. Several of these have also been implemented in software that we have mentioned above. For more details on some of these methods, see, e.g., [11], [29], [106], [108], [181] and the described web pages.

Exercises

1. Download and implement the codes `DASSL`, `RADAU5`, and `GELDA`. Check your implementation with the help of the problem

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} \exp(t) + \cos(t) \\ \exp(t) \end{bmatrix}, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

2. Compare the behavior of the codes `DASSL`, `RADAU5`, and `GELDA` for the linear problem with constant coefficients

$$\begin{bmatrix} 0 & 2 & 1 & & \\ 0 & 1 & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & 1 \\ & & \ddots & \ddots & 2 \\ & & & 0 & 1 \end{bmatrix} \dot{x} = \begin{bmatrix} 2 & 1 & & & \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & 1 \\ & & & 1 & 2 \end{bmatrix} x + \begin{bmatrix} -3t^2 + 6t \\ -4t^2 + 8t \\ \vdots \\ -4t^2 + 8t \\ -4t^2 + 6t \\ -3t^2 + 2t \end{bmatrix},$$

   $x(0) = 0$, in the interval $[0, 1]$ with respect to accuracy and computing time.

3. Apply the codes `DASSL`, `RADAU5`, and `GELDA` to the problem of Example 3.1

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{x} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} x, \quad x(0) = 0$$

   and analyze how the codes handle the nonuniqueness of the solution.

4. Apply the codes DASSL, RADAU5, and GELDA to the problem of Example 3.2 with E, A, and f given by

$$E(t) = \begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix}, \quad A(t) = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix}, \quad f(t) = \begin{bmatrix} \sin(t) \\ \exp(t) \end{bmatrix}$$

and analyze how the codes handle the problem that the matrix pair $(E(t), A(t))$ is singular for all t.

5. Apply the codes DASSL, RADAU5, and GELDA to the problem (5.33) as well as to the transformed system (5.34) and to the equivalent strangeness-free system for various choices of η and compare the numerical results.

6. Apply the codes DASSL, RADAU5, and GELDA to the regular strangeness-free problem

$$\begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix} \dot{x} = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix} x + \begin{bmatrix} \exp(t) + \cos(t) \\ \exp(t) \end{bmatrix}, \quad x(0) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

and compare the numerical results.

7. Download and implement the code GENDA.
Check your implementation with the help of the problem of Exercise 1

8. Apply the codes DASSL, RADAU5, and GENDA to the problem

$$\begin{bmatrix} 1 & 0 \\ \dot{x}_1 & 1 \end{bmatrix} \begin{bmatrix} \dot{x}_2 - x_1 - \sin(t) \\ x_2 - \exp(t) \end{bmatrix} = 0.$$

with $\mu = 1$ and compare the numerical results.

9. Apply the codes DASSL, RADAU5, and GENDA to the problem

$$\dot{x}_1 = x_2, \quad \dot{x}_2 = x_1, \quad x_1 \dot{x}_4 + x_3 = 1, \quad x_4 = 2$$

with $\mu = 1$, $a = 2$, and $d = 2$. Use the initial values $x_0 = (0, 0, 1, 2)$ and $x_0 = (0, 1, 1, 2)$.

10. Apply the codes DASSL, RADAU5, and GENDA to the problem

$$\dot{x}_1 = 1, \quad \dot{x}_2 = (\exp(x_3 - 1) + 1)/2, \quad x_2 - x_1 = 0$$

with $\mu = 1$, $a = 2$, and $d = 1$.

11. Apply the codes DASSL, RADAU5, and GENDA to the various formulations of Example 6.17.

Final remarks

With the current state of the theory of differential-algebraic equations and their numerical treatment, this topic can now be covered in standard graduate courses. This textbook is designed for such courses presenting a systematic analysis of initial and boundary value problems for differential-algebraic equations, together with numerical methods and software for the solution of these problems. In particular, it covers linear and nonlinear problems, over- and underdetermined problems as well as control problems and problems with structure.

Many important research topics concerning differential-algebraic equations, however, are still in their infancy. These include for example the stability analysis of differential-algebraic equations, the study of optimal control problems and, in particular, the coupling of differential-algebraic equations with other types of equations, such as ordinary or partial differential equations, via complex networks.

We hope that this textbook will provide the basis for future research in these topics.

Bibliography

[1] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammerling, A. McKenney, S. Ostrouchov, and D. Sorenson. *LAPACK Users' Guide*. SIAM Publications, Philadelphia, PA, 2nd edition, 1995. 354

[2] J. D. Aplevich. *Implicit Linear Systems*. Springer-Verlag, Berlin, 1991. 52

[3] M. Arnold. Index and stability of differential-algebraic systems. In U. Helmke, R. Mennicken, and J. Saurer, editors, *Systems and Networks: Mathematical Theory and Applications II*, volume 79 of Mathematical Research, pages 41–44. Akademie Verlag, Berlin, 1994. 295

[4] M. Arnold, V. Mehrmann, and A. Steinbrecher. Index reduction of linear equations of motions in industrial multibody system simulation. Technical Report 146, DFG Research Center MATHEON, TU Berlin, Berlin, Germany, 2004. 295

[5] M. Arnold and B. Simeon. Pantograph and catenary dynamics: a benchmark problem and its numerical solution. *Appl. Numer. Math.*, 34:345–362, 2000. 10

[6] V. I. Arnold. *Ordinary Differential Equations*. Springer-Verlag, Berlin, 1992. 195

[7] U. M. Ascher, H. Chin, and S. Reich. Stabilization of DAEs and invariant manifolds. *Numer. Math.*, 67:131–149, 1994. 274

[8] U. M. Ascher, R. Mattheij, and R. Russell. *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*. SIAM Publications, Philadelphia, PA, 2nd edition, 1995. 303, 314, 321, 348

[9] U. M. Ascher and L. R. Petzold. Projected implicit Runge-Kutta methods for differential algebraic equations. *SIAM J. Numer. Anal.*, 28:1097–1120, 1991. 274

[10] U. M. Ascher and L. R. Petzold. Projected collocation for higher-order higher-index differential-algebraic equations. *SIAM J. Numer. Anal.*, 43:1635–1657, 1992. 348

[11] U. M. Ascher and L. R. Petzold. *Computer Methods for Ordinary Differential and Differential-Algebraic Equations*. SIAM Publications, Philadelphia, PA, 1998. 4, 10, 270, 348, 355

[12] U. M. Ascher and R. J. Spiteri. Collocation software for boundary value differential-algebraic equations. *SIAM J. Sci. Comput.*, 15:938–952, 1994. 348, 354

[13] U. M. Ascher and R. Weiss. Collocation for singular perturbation problems I: First order systems with constant coefficients. *SIAM J. Numer. Anal.*, 20:537–557, 1983. 348

[14] U. M. Ascher and R. Weiss. Collocation for singular perturbation problems III: Nonlinear problems without turning points. *SIAM J. Sci. Statist. Comput.*, 5:811–829, 1984. 348

[15] U. M. Ascher and R. Weiss. Collocation for singular perturbation problems II: Linear first order systems without turning points. *Math. Comp.*, 43:157–187, 1984. 348

[16] M. Athans and P. L. Falb. *Optimal Control*. McGraw-Hill, New York, NY, 1966. 50, 120

[17] S. Bächle. Index reduction for differential-algebraic equations in circuit simulation. Technical Report 141, DFG Research Center MATHEON, TU Berlin, Berlin, Germany, 2004. 293, 295

[18] S. Bächle and F. Ebert. Element-based topological index reduction for differential-algebraic equations in circuit simulation. Technical Report 246, DFG Research Center MATHEON, TU Berlin, Berlin, Germany, 2004. Submitted for patenting. 8, 293, 294, 295

[19] S. Bächle and F. Ebert. Graph theoretical algorithms for index reduction in circuit simulation. Technical Report 245, DFG Research Center MATHEON, TU Berlin, Berlin, Germany, 2004. 293, 294, 295

[20] R. Bachmann, L. Brüll, Th. Mrziglod, and U. Pallaske. On methods for reducing the index of differential algebraic equations. *Comp. Chem. Eng.*, 14:1271–1273, 1990. 295

[21] Y. Bai. A perturbed collocation method for boundary-value problems in differential algebraic equations. *Appl. Math. Comput.*, 45:269–291, 1991. 348

[22] K. Balla and R. März. Linear differential algebraic equations of index 1 and their adjoint equations. *Res. in Math.*, 37:13–35, 2000. 147

[23] A. Barrlund. Constrained least squares methods for linear time varying DAE systems. *Numer. Math.*, 60:145–161, 1991. 273

[24] J. Baumgarte. Stabilization of constraints and integrals of motion in dynamical systems. *Comp. Meth. Appl. Mech. Eng.*, 1:1–16, 1972. 273, 295

[25] A. Ben-Israel and T. N. E. Greville. *Generalized Inverses: Theory and Applications*. Springer-Verlag, New York, NY, 2nd edition, 2003. 114, 115

[26] L. S. Blackford, J. Choi, A. Cleary, E. D'Azevedo, J. Demmel, I. Dhillon, J. Dongarra, S. Hammarling, G. Henry, A. Petitet, K. Stanley, D. Walker, and R. C. Whaley. *ScaLAPACK Users' Guide*. SIAM Publications, Philadelphia, PA, 1997. 354

[27] W. Blajer. Index of differential-algebraic equations governing the dynamics of constrained systems. *Appl. Math. Modelling*, 16:70–77, 1992. 295

[28] R. K. Brayton, F. G. Gustavson, and G. D. Hachtel. A new efficient algorithm for solving differential-algebraic systems using implicit backward differentiation formulas. *Proc. IEEE*, 60:98–108, 1972. 270

[29] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*. SIAM Publications, Philadelphia, PA, 2nd edition, 1996. 4, 6, 52, 147, 210, 217, 269, 270, 355

[30] R. L. Brown and C. W. Gear. Documentation for DFASUB — a program for the solution of simultaneous implicit differential and nonlinear equations. Technical Report UIUCDCS-R-73-575, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL, 1973. 270

[31] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Numerical computation of an analytic singular value decomposition of a matrix valued function. *Numer. Math.*, 60:1–40, 1991. 276

[32] A. Bunse-Gerstner, R. Byers, V. Mehrmann, and N. K. Nichols. Feedback design for regularizing descriptor systems. *Lin. Alg. Appl.*, 299:119–151, 1999. 295

[33] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols. Regularization of descriptor systems by derivative and proportional state feedback. *SIAM J. Matr. Anal. Appl.*, 13:46–67, 1992. 53, 295

[34] A. Bunse-Gerstner, V. Mehrmann, and N. K. Nichols. Regularization of descriptor systems by output feedback. *IEEE Trans. Automat. Control*, 39:1742–1748, 1994. 53, 295

[35] K. Burrage. *Parallel and Sequential Methods for Ordinary Differential Equations*. Oxford University Press, Oxford, 1995. 270

[36] K. Burrage and L. R. Petzold. On order reduction for Runge-Kutta methods applied to differential/algebraic systems and to stiff systems of ODEs. *SIAM J. Numer. Anal.*, 27:447–456, 1988. 270

[37] J. C. Butcher. Implicit Runge-Kutta processes. *Math. Comp.*, 18:50–64, 1964. 225

[38] J. C. Butcher. *The Numerical Analysis of Ordinary Differential Equations: Runge-Kutta and General Linear Methods*. Wiley, Chichester, 1987. 226, 270

[39] R. Byers, T. Geerts, and V. Mehrmann. Descriptor systems without controllability at infinity. *SIAM J. Cont.*, 35:462–479, 1997. 53, 295

[40] R. Byers, P. Kunkel, and V. Mehrmann. Regularization of linear descriptor systems with variable coefficients. *SIAM J. Cont.*, 35:117–133, 1997. 141, 295

[41] S. L. Campbell. Linear systems of differential equations with singular coefficients. *SIAM J. Math. Anal.*, 8:1057–1066, 1977. 52

[42] S. L. Campbell. *Singular Systems of Differential Equations I*. Pitman, San Francisco, CA, 1980. 4, 22, 32, 52

[43] S. L. Campbell. *Singular Systems of Differential Equations II*. Pitman, San Francisco, CA, 1982. 4, 22, 52

[44] S. L. Campbell. One canonical form for higher index linear time varying singular systems. *Circ. Syst. Signal Process.*, 2:311–326, 1983. 5, 147, 210

[45] S. L. Campbell. Regularizations of linear time varying singular systems. *Automatica*, 20:365–370, 1984. 295

[46] S. L. Campbell. The numerical solution of higher index linear time-varying singular systems of differential equations. *SIAM J. Sci. Statist. Comput.*, 6:334–338, 1985. 210, 295

[47] S. L. Campbell. Comment on controlling generalized state-space (descriptor) systems. *Internat. J. Control*, 46:2229–2230, 1987. 81

[48] S. L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM J. Math. Anal.*, 18:1101–1115, 1987. 96, 147, 210, 273, 295

[49] S. L. Campbell. Least squares completions for nonlinear differential algebraic equations. *Numer. Math.*, 65:77–94, 1993. 273

[50] S. L. Campbell. Numerical methods for unstructured higher index DAEs. *Annals of Numer. Math.*, 1:265–278, 1994. 295

[51] S. L. Campbell. High index differential algebraic equations. *J. Mech. of Struct. Mach.*, 23:199–222, 1995. 295

[52] S. L. Campbell. Linearization of DAE's along trajectories. *Z. Angew. Math. Phys.*, 46:70–84, 1995. 210

[53] S. L. Campbell and C. W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995. 7, 147, 154, 182, 210

[54] S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM J. Sci. Comput.*, 16:257–270, 1995. 182, 210

[55] S. L. Campbell and B. Leimkuhler. Differentiation of constraints in differential algebraic equations. *J. Mech. of Struct. Mach.*, 19:19–40, 1991. 295

[56] S. L. Campbell and C. D. Meyer. *Generalized Inverses of Linear Transformations*. Pitman, San Francisco, CA, 1979. 24, 47, 80, 114

[57] S. L. Campbell, C. D. Meyer, and N. J. Rose. Applications of the Drazin inverse to linear systems of differential equations with singular constant coefficients. *SIAM J. Appl. Math.*, 31:411–425, 1976. 52

[58] S. L. Campbell and E. Moore. Constraint preserving integrators for general nonlinear higher index DAEs. *Numer. Math.*, 69:383–399, 1995. 273, 295

[59] S. L. Campbell, E. Moore, and Y. Zhong. Utilization of automatic differentiation in control algorithms. *IEEE Trans. Automat. Control*, 39:1047–1052, 1994. 295

[60] J. R. Cash and S. Cosidine. A MEBDF code for stiff initial value problems. *ACM Trans. Math. Software*, 18:142–158, 1992. 352

[61] Y. Chung and W. Westerberg. A proposed numerical algorithm for solving nonlinear index problems. *Ind. Eng. Chem. Res.*, 29:1234–1239, 1990. 295

[62] D. Chu, V. Mehrmann, and N. K. Nichols. Minimum norm regularization of descriptor systems by output feedback. *Lin. Alg. Appl.*, 296:39–77, 1999. 295

[63] K. D. Clark. A structural form for higher index semistate equations I: Theory and applications to circuit and control. *Lin. Alg. Appl.*, 98:169–197, 1988. 295

[64] J. D. Cobb. On the solutions of linear differential equations with singular coefficients. *J. Diff. Equations*, 46:310–323, 1982. 53

[65] E. A. Coddington and N. Levinson. *Theory of Ordinary Differential Equations*. Tata McGraw-Hill, New Dehli, 6th edition, 1982. 17, 66, 70

[66] L. Conlon. *Differentiable Manifolds — A First Course*. Birkhäuser, Boston, MA, 1993. 195

[67] L. Dai. *Singular Control Systems*. Springer-Verlag, Berlin, 1989. 48, 52

[68] K. Deimling. *Nonlinear Functional Analysis*. Springer-Verlag, Berlin, 1985. 164, 195

[69] J. W. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil $\lambda A - B$, Part I. *ACM Trans. Math. Software*, 19:160–174, 1993. 52, 295

[70] J. W. Demmel and B. Kågström. The generalized Schur decomposition of an arbitrary pencil $\lambda A - B$, Part II. *ACM Trans. Math. Software*, 19:185–201, 1993. 52, 295

[71] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*. Springer-Verlag, Berlin, 2004. 280, 284, 307, 348

[72] P. Deuflhard and F. Bornemann. *Scientific Computing with Ordinary Differential Equations*. Springer-Verlag, New York, NY, 2002. 4, 270

[73] P. Deuflhard, E. Hairer, and J. Zugck. One step and extrapolation methods for differential-algebraic systems. *Numer. Math.*, 51:501–516, 1987. 270, 352

[74] C. de Boor. An empty exercise. *ACM Signum Newsletter*, 25:2–6, 1990. 20

[75] J. J. Dongarra, J. Du Croz, S. Hammerling, and R. J. Hanson. Algorithm 656: An extended set of FORTRAN basic linear algebra subprograms. *ACM Trans. Math. Software*, 14:18–32, 1988. 354

[76] I. Duff and C. W. Gear. Computing the structural index. *SIAM J. Alg. Discr. Meth.*, 7:594–603, 1986. 295

[77] A. Edelman, E. Elmroth, and B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils. Part I: Versal deformations. *SIAM J. Matr. Anal. Appl.*, 18:653–692, 1997. 52

[78] A. Edelman, E. Elmroth, and B. Kågström. A geometric approach to perturbation theory of matrices and matrix pencils. Part II: A stratification-enhanced staircase algorithm. *SIAM J. Matr. Anal. Appl.*, 20(3):667–699, 1999. 52

[79] E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Systems*. Teubner Verlag, Stuttgart, 1998. 4, 9, 210, 217, 270, 273, 276, 286, 287, 295, 354

[80] R. Engelking. *General Topology*. Polish Scientific Publishers, Warszawa, 1977. 129

[81] R. England, R. Lamour, and J. Lopez-Estrada. Multiple shooting using a dichotomically stable integrator for solving DAEs. *Appl. Numer. Math.*, 42:117–131, 2003. 348

[82] D. Estévez-Schwarz. *Consistent initialization for index-2 differential-algebraic equations and its application circuit simulation*. Dissertation, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2000. 278

[83] D. Estévez-Schwarz, U. Feldmann, R. März, S. Sturtzel, and C. Tischendorf. Finding beneficial DAE structures in circuit simulation. Technical Report 00-7, Institut für Mathematik, Humboldt Universität zu Berlin, Berlin, Germany, 2000. 8, 290

[84] D. Estévez-Schwarz and C. Tischendorf. Structural analysis for electrical circuits and consequences for MNA. *Internat. J. Circ. Theor. Appl.*, 28:131–162, 2000. 8, 290, 295

[85] C. Führer. *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen*. Dissertation, Mathematisches Institut, TU München, München, Germany, 1988. 273, 354

[86] C. Führer and B. J. Leimkuhler. Numerical solution of differential-algebraic equations for constrained mechanical motion. *Numer. Math.*, 59:55–69, 1991. 273

[87] F. R. Gantmacher. *The Theory of Matrices I*. Chelsea Publishing Company, New York, NY, 1959. 16, 18

[88] F. R. Gantmacher. *The Theory of Matrices II*. Chelsea Publishing Company, New York, NY, 1959. 6, 14, 52

[89] C. W. Gear. *Numerical Initial Value Problems in Ordinary Differential Equations*. Prentice-Hall, Engelwood Cliffs, NJ, 1971. 270

[90] C. W. Gear. The simultaneous numerical solution of differential-algebraic equations. *IEEE Trans. Circ. Theor.*, CT-18:89–95, 1971. 4, 295

[91] C. W. Gear. Maintaining solution invariants in the numerical solution of ODEs. *SIAM J. Sci. Statist. Comput.*, 7:734–743, 1986. 295

[92] C. W. Gear. Differential-algebraic equation index transformations. *SIAM J. Sci. Statist. Comput.*, 9:39–47, 1988. 7, 147, 210, 295

[93] C. W. Gear. Differential-algebraic equations, indices, and integral equations. *SIAM J. Numer. Anal.*, 27:1527–1534, 1990. 210

[94] C. W. Gear, B. Leimkuhler, and G. K. Gupta. Automatic integration of Euler-Lagrange equations with constraints. *J. Comput. Appl. Math.*, 12/13:77–90, 1985. 273, 287, 295

[95] C. W. Gear and L. R. Petzold. Differential/algebraic systems and matrix pencils. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, pages 75–89. Springer-Verlag, Berlin, 1983. 6, 147

[96] T. Geerts. Solvability conditions, consistency, and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case. *Lin. Alg. Appl.*, 181:111–130, 1993. 33, 46, 53, 147

[97] T. Geerts and V. Mehrmann. Linear differential equations with constant coefficients: A distributional approach. Technical Report SFB 343/90–073, Fakultät für Mathematik, Universität Bielefeld, Bielefeld, Germany, 1990. 53, 147

[98] I. Gohberg, P. Lancaster, and L. Rodman. *Matrices and Indefinite Scalar Products*. Birkhäuser, Basel, 1983. 52

[99] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, MD, 3rd edition, 1996. 14, 62, 275, 281, 283, 307

[100] E. Griepentrog and R. März. *Differential-Algebraic Equations and their Numerical Treatment*. Teubner Verlag, Leipzig, 1986. 4, 7, 52, 119, 147, 210, 270, 273, 295

[101] C. Groetsch. *Generalized Inverses of Linear Operators. Representation and Approximation*. Marcel-Dekker, New York, NY, 1977. 114

[102] M. Günther. A joint DAE/PDE model for interconnected electrical networks. *Math. Comp. Mod. of Dyn. Syst.*, 6:114–128, 2000. 10

[103] M. Günther and U. Feldmann. CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations. *Surv. Math. Ind.*, 8:97–129, 1999. 8, 210, 286, 290, 295

[104] M. Günther and U. Feldmann. CAD-based electric-circuit modeling in industry II. Impact of circuit configurations and parameters. *Surv. Math. Ind.*, 8:131–157, 1999. 8, 210, 286, 290, 295

[105] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Springer-Verlag, Berlin, 1989. 4, 7, 52, 147, 217, 238, 270

[106] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Springer-Verlag, Berlin, 2nd edition, 1993. 17, 166, 218, 225, 226, 228, 242, 254, 270, 355

[107] E. Hairer and G. Wanner. On the instability of the BDF formulas. *SIAM J. Numer. Anal.*, 20:1206–1209, 1983. 259

[108] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer-Verlag, Berlin, 2nd edition, 1996. 4, 11, 52, 131, 147, 180, 217, 218, 226, 228, 231, 254, 270, 273, 274, 276, 298, 342, 352, 355

[109] P. Hamann. Modellierung und Simulation von realen Planetengetrieben. Diplomarbeit, Institut für Mathematik, TU Berlin, Berlin, Germany, 2003. 276

[110] P. Hamann and V. Mehrmann. Numerical solution of hybrid differential-algebraic equations. Technical Report 263, DFG Research Center MATHEON, TU Berlin, Berlin, Germany, 2005. 276

[111] M. Hanke. On a least-squares collocation method for linear differential-algebraic equations. *Numer. Math.*, 54:79–90, 1988. 348

[112] M. L. J. Hautus and L. M. Silverman. System structure and singular control. *Lin. Alg. Appl.*, 50:369–402, 1983. 33

[113] M. R. Hesteness. *Calculus of Variations and Optimal Control Theory*. John Wiley and Sons, New York, NY, 1966. 120, 121

[114] F. Iavernaro and F. Mazzia. Block-boundary value methods for the solution of ordinary differential equation. *SIAM J. Sci. Comput.*, 21:323–339, 1998. 270, 352

[115] F. Iavernaro and F. Mazzia. Solving ordinary differential equations by generalized Adams methods: properties and implementation techniques. *Appl. Numer. Math.*, 28:107–126, 1998. 270, 352

[116] A. Ilchmann and V. Mehrmann. A behavioural approach to linear time-varying descriptor system. Part 1. General theory. *SIAM J. Cont.*, 44:1725–1747, 2005. 10, 295

[117] A. Ilchmann and V. Mehrmann. A behavioural approach to linear time-varying descriptor system. Part 2. Descriptor systems. *SIAM J. Cont.*, 44:1748–1765, 2005. 10, 295

[118] L. O. Jay. Collocation methods for differential-algebraic equations of index three. *Numer. Math.*, 65:407–421, 1993. 348

[119] T. Kato. *Perturbation Theory for Linear Operators*. Springer-Verlag, Berlin, 2nd edition, 1976. 102

[120] H. W. Knobloch and H. Kwakernaak. *Lineare Kontrolltheorie*. Springer-Verlag, Berlin, 1985. 123

[121] L. Kronecker. Algebraische Reduction der Schaaren bilinearer Formen. *Sitzungsber. Akad. der Wiss., Berlin*, pages 763–776, 1890. 4, 13, 52

[122] P. Kunkel and V. Mehrmann. Smooth factorizations of matrix valued functions and their derivatives. *Numer. Math.*, 60:115–132, 1991. 147, 276

[123] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comput. Appl. Math.*, 56:225–259, 1994. 5, 7, 147

[124] P. Kunkel and V. Mehrmann. A new look at pencils of matrix valued functions. *Lin. Alg. Appl.*, 212/213:215–248, 1994. 5, 147

[125] P. Kunkel and V. Mehrmann. Analysis und Numerik linearer differentiell-algebraischer Gleichungen. In J. Herzberger, editor, *Wissenschaftliches Rechnen. Eine Einführung in das Scientific Computing*, pages 233–278. Akademie Verlag, Berlin, 1995. 132, 270

[126] P. Kunkel and V. Mehrmann. A new class of discretization methods for the solution of linear differential algebraic equations with variable coefficients. *SIAM J. Numer. Anal.*, 33:1941–1961, 1996. 147, 270, 295

[127] P. Kunkel and V. Mehrmann. The linear quadratic control problem for linear descriptor systems with variable coefficients. *Math. Control, Signals, Sys.*, 10:247–264, 1997. 147, 270

[128] P. Kunkel and V. Mehrmann. Regular solutions of nonlinear differential-algebraic equations and their numerical determination. *Numer. Math.*, 79:581–600, 1998. 5, 7, 210, 270, 295

[129] P. Kunkel and V. Mehrmann. Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems. *Math. Control, Signals, Sys.*, 14:233–256, 2001. 5, 7, 210, 295

[130] P. Kunkel and V. Mehrmann. Index reduction for differential-algebraic equations by minimal extension. *Z. Angew. Math. Mech.*, 84:579–597, 2004. 273, 287, 292, 295, 348

[131] P. Kunkel and V. Mehrmann. Characterization of classes of singular linear differential-algebraic equations. *Electr. J. Lin. Alg.*, 13:359–386, 2005. 5, 348

[132] P. Kunkel, V. Mehrmann, and W. Rath. Analysis and numerical solution of control problems in descriptor form. *Math. Control, Signals, Sys.*, 14:29–61, 2001. 5, 7, 10, 53, 147, 270, 295

[133] P. Kunkel, V. Mehrmann, W. Rath, and J. Weickert. A new software package for linear differential–algebraic equations. *SIAM J. Sci. Comput.*, 18:115–138, 1997. 354

[134] P. Kunkel, V. Mehrmann, and S. Seidel. A MATLAB package for the numerical solution of general nonlinear differential-algebraic equations. Technical Report 16/2005, Institut für Mathematik, TU Berlin, Berlin, Germany, 2005. 295

[135] P. Kunkel, V. Mehrmann, and I. Seufer. GENDA: A software package for the numerical solution of general nonlinear differential-algebraic equations. Technical Report 730, Institut für Mathematik, TU Berlin, Berlin, Germany, 2002. 354

[136] P. Kunkel and R. Stöver. Symmetric collocation methods for linear differential-algebraic boundary value problems. *Numer. Math.*, 91:475–501, 2002. 348

[137] R. Lamour. A shooting method for fully implicit index-2 DAEs. *SIAM J. Sci. Comput.*, 18:94–114, 1997. 319, 348

[138] S. Lang. *Analysis I*. Addison-Wesley, Reading, MA, 3rd edition, 1973. 101

[139] C. L. Lawson, R. J. Hanson, D. Kincaid, and F. T. Krogh. Basic linear algebra subprograms for FORTRAN usage. *ACM Trans. Math. Software*, 5:308–323, 1979. 354

[140] W. Liniger. Multistep and one-leg methods for implicit mixed differential algebraic systems. *IEEE Trans. Circ. and Syst.*, CAS-26:755–762, 1979. 270

[141] W. M. Lioen, J. J. B. de Swart, and W. A. van der Veen. PSIDE users' guide. Report MAS-R9834, CWI, Amsterdam, 1998. 353

[142] W. M. Lioen, J. J. B. de Swart, and W. A. van der Veen. Specification of PSIDE. Report MAS-R9833, CWI, Amsterdam, 1998. 353

[143] C. Lubich. Extrapolation integrators for constrained multibody systems. *Impact Comput. Sci. Eng.*, 3:212–234, 1991. 274

[144] C. Lubich, U. Nowak, U. Pöhle, and C. Engstler. MEXX – numerical software for the integration of constrained mechanical multibody systems. Preprint SC 92-12, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Berlin, Germany, 1992. 354

[145] R. März. On difference and shooting methods for boundary value problems in differential-algebraic equations. Preprint 24, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 1982. 295, 348

[146] R. März. A matrix chain for analyzing differential-algebraic equations. Preprint 162, Sektion Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 1987. 147

[147] R. März. Numerical methods for differential-algebraic equations I: Characterizing DAEs. Seminarberichte 91-32/I, Fachbereich Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 1991. 7

[148] R. März. The index of linear differential algebraic equations with properly stated leading terms. *Res. in Math.*, 42:308–338, 2002. 6, 7, 147

[149] R. März. Solvability of linear differential algebraic equations with properly stated leading terms. *Res. in Math.*, 45:88–105, 2004. 6, 7, 147

[150] R. März. Characterizing differential algebraic equations without the use of derivative arrays. *Computers Math. Appl.*, 50:1141–1156, 2005. 7, 147

[151] R. März and R. Riaza. On linear differential-algebaic equations with properly stated leading term. I: Regular points. Preprint 04-22, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2004. 210

[152] R. März and R. Riaza. On linear differential-algebaic equations with properly stated leading term. II: Critical points. Preprint 04-23, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2004. 210

[153] R. M. M. Mattheij and P. M. E. J. Wijckmans. Sensitivity of solutions of linear DAE to perturbations of the system matrices. *Numer. Algorithms*, 19:159–171, 1998. 281

[154] S. Mattsson and G. Söderlind. Index reduction in differential-algebraic equations using dummy derivatives. *SIAM J. Sci. Statist. Comput.*, 14:677–692, 1993. 273, 287, 295

[155] V. Mehrmann. *The Autonomous Linear Quadratic Control Problem*. Springer-Verlag, Berlin, 1991. 52

[156] V. Mehrmann and C. Shi. Analysis of higher order linear differential-algebraic systems. Preprint 17/2004, Institut für Mathematik, TU Berlin, Berlin, Germany, 2004. 7

[157] K. Nomizu. Characteristic roots and vectors of a differentiable family of symmetric matrices. *Lin. Multilin. Algebra*, 2:159–162, 1973. 62, 147

[158] U. Nowak and L. Weimann. A Family of Newton Codes for Systems of Highly Nonlinear Equations — Algorithm, Implementation, Application. Report TR 90-11, Konrad-Zuse-Zentrum für Informationstechnik Berlin, Berlin, Germany, 1990. 354

[159] R. E. O'Malley. *Introduction to Singular Perturbations*. Academic Press, New York, NY, 1974. 11

[160] J. Ortega and W. C. Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. SIAM Publications, Philadelphia, PA, 2nd edition, 2000. 280, 284, 348

[161] C. C. Pantelides. The consistent initialization of differential-algebraic systems. *SIAM J. Sci. Statist. Comput.*, 9:213–231, 1988. 7, 9, 273, 295

[162] C. C. Pantelides, D. Gritsis, K. R. Morison, and R. W. H. Sargen. The mathematical modeling of transient systems using differential-algebraic equations. *Computers Chem. Engng.*, 12:449–454, 1988. 273, 295

[163] L. R. Petzold. Differential/algebraic equations are not ODEs. *SIAM J. Sci. Statist. Comput.*, 3:367–384, 1982. 270

[164] L. R. Petzold. A description of DASSL: A differential/algebraic system solver. In R. S. Stepleman et al., editors, *IMACS Trans. Scient. Comp. Vol. 1*, pages 65–68. North-Holland, Amsterdam, 1983. 270, 353

[165] L. R. Petzold. Order results for implicit Runge-Kutta methods applied to differential/algebraic systems. *SIAM J. Numer. Anal.*, 23:837–852, 1986. 270

[166] L. R. Petzold and F. A. Potra. ODAE methods for the numerical solution of Euler-Lagrange equations. *Appl. Numer. Math.*, 43:243–259, 1992. 295

[167] J. W. Polderman and J. C. Willems. *Introduction to Mathematical Systems Theory: A Behavioural Approach*. Springer-Verlag, New York, NY, 1998. 10, 49, 52, 138

[168] L. S. Pontryagin, V. Boltyanskii, R. Gamkrelidze, and E. Mishenko. *The Mathematical Theory of Optimal Processes*. Interscience, New York, NY, 1962. 120

[169] F. A. Potra and W. C. Rheinboldt. On the numerical solution of Euler-Lagrange equations. *J. Mech. of Struct. Mach.*, 19:1–18, 1991. 274, 295

[170] F. A. Potra and J. Yen. Implicit numerical integration for Euler-Lagrange equations via tangent space parameterization. *J. Mech. of Struct. Mach.*, 19:77–98, 1991. 295

[171] M. P. Quéré and G. Villard. An algorithm for the reduction of linear DAE. In A. H. M. Levelt, editor, *Proceedings of the 1995 International Symposium on Symbolic and Algebraic Computation (ISSAC'95)*, pages 223–231. ACM Press, New York, NY, 1995. 295

[172] M. P. Quéré-Stucklik. *Algorithmique des Faisceaux Lineaires de matrices. Application á la Theorie des Systemes Lineaires et á la Resolution d'Equations Algebro-Differentielles*. Thèse de doctorat, Departement d'Informatique, l'Université Paris VI, Paris, France, 1997. 295

[173] P. J. Rabier and W. C. Rheinbold. On a computational method for the second fundamental tensor and its application to bifurcation problems. *Numer. Math.*, 57:681–694, 1990. 210

[174] P. J. Rabier and W. C. Rheinboldt. A general existence and uniqueness theorem for implicit differential algebraic equations. *Diff. Int. Eqns.*, 4:563–582, 1991. 195, 210

[175] P. J. Rabier and W. C. Rheinbold. A geometric treatment of implicit differential-algebraic equations. *J. Diff. Equations*, 109:110–146, 1994. 195, 210, 295

[176] P. J. Rabier and W. C. Rheinbold. On impasse points of quasilinear differential algebraic equations. *J. Math. Anal. Appl.*, 181:429–454, 1994. 210

[177] P. J. Rabier and W. C. Rheinbold. On the computation of impasse points of quasilinear differential algebraic equations. *Math. Comp.*, 62:133–154, 1994. 210

[178] P. J. Rabier and W. C. Rheinbold. On the numerical solution of the Euler-Lagrange equations. *SIAM J. Numer. Anal.*, 32:318–329, 1995. 295

[179] P. J. Rabier and W. C. Rheinboldt. Classical and generalized solutions of time-dependent linear differential-algebraic equations. *Lin. Alg. Appl.*, 245:259–293, 1996. 53, 62, 112, 132, 147

[180] P. J. Rabier and W. C. Rheinboldt. Time-dependent linear DAEs with discontinuous inputs. *Lin. Alg. Appl.*, 247:1–29, 1996. 41, 44, 53, 132, 147

[181] P. J. Rabier and W. C. Rheinboldt. *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*. SIAM Publications, Philadelphia, PA, 2000. 4, 210, 270, 273, 287, 295, 353, 355

[182] P. J. Rabier and W. C. Rheinboldt. *Theoretical and Numerical Analysis of Differential-Algebraic Equations*, volume VIII of *Handbook of Numerical Analysis*. Elsevier Publications, Amsterdam, 2002. 4, 5, 53, 295

[183] W. Rath. Derivative and proportional state feedback for linear descriptor systems with variable coefficients. *Lin. Alg. Appl.*, 260:273–310, 1997. 147, 295

[184] W. Rath. *Feedback Design and Regularization for Linear Descriptor Systems with Variable Coefficients*. Dissertation, TU Chemnitz, Chemnitz, Germany, 1997. 147

[185] S. Reich. *Differential-Algebraic Equations and Vector Fields on Manifolds*. Dissertation, TU Dresden, Dresden, Germany, 1988. 210

[186] S. Reich. On a geometric interpretation of differential-algebraic equations. *Circ. Syst. Signal Process.*, 9:367–382, 1990. 210

[187] S. Reich. On an existence and uniqueness theory for nonlinear differential-algebraic equations. *Circ. Syst. Signal Process.*, 10:343–359, 1991. 210

[188] K. J. Reinschke. Graph-theoretic approach to symbolic analysis of linear descriptor systems. *Lin. Alg. Appl.*, 197:217–244, 1994. 295

[189] G. Reißig, W. S. Martinson, and P. I. Barton. Differential-algebraic equations of index 1 may have an arbitrarily high structural index. *SIAM J. Sci. Comput.*, 21:1987–1990, 2000. 7, 273, 295

[190] W. C. Rheinboldt. Differential-algebraic systems as differential equations on manifolds. *Math. Comp.*, 43:473–482, 1984. 5, 7, 210, 289

[191] W. C. Rheinboldt. On the computation of multi-dimensional solution manifolds of parameterized equations. *Numer. Math.*, 53:165–181, 1988. 147

[192] W. C. Rheinboldt. On the existence and uniqueness of solutions of nonlinear semi-implicit differential algebraic equations. *Nonlinear Anal.*, 16:647–661, 1991. 210

[193] W. C. Rheinboldt. MANPACK: A set of algorithms for computations on implicitly defined manifolds. *Appl. Math. Comput.*, 27:15–28, 1996. 353

[194] M. Roche. Runge-Kutta methods for differential algebraic equations. *SIAM J. Numer. Anal.*, 26:963–975, 1989. 270

[195] T. Rübner-Petersen. An efficient algorithm using backward time-scaled differences for solving differential-algebraic equations. Technical report, Institute of Circuit Theory and Telecommunication, Technical University of Denmark, Lyngby, Denmark, 1973. 270

[196] W. Schiehlen. *Multibody Systems Handbook*. Springer-Verlag, Heidelberg, 1990. 9, 287

[197] V. Schulz, H. G. Bock, and M. C. Steinbach. Exploiting invariants in the numerical solution of multipoint boundary value problems for DAE. *SIAM J. Sci. Comput.*, 19:440–446, 1998. 313, 348

[198] I. Schumilina. *Charakterisierung der Algebro-Differentialgleichungen mit Traktabilitätsindex 3*. Dissertation, Institut für Mathematik, Humboldt-Universität zu Berlin, Berlin, Germany, 2004. 147

[199] L. Schwartz. *Theorie des Distributions*. Hermann, Paris, 1978. 33, 34, 37, 38

[200] I. Seufer. *Generalized Inverses of Differential-Algebraic Operators and their Numerical Discretization*. Dissertation, Institut für Mathematik, TU Berlin, Berlin, Germany, 2005. 147

[201] B. Simeon. MBSPACK — numerical integration software for constrained mechanical motion. *Surv. Math. Ind.*, 5:169–202, 1995. 9, 273, 286, 354

[202] B. Simeon. *Numerische Simulation gekoppelter Systeme von partiellen und differential-algebraischen Gleichungen in der Mehrkörperdynamik*. Habilitationsschrift, Fakultät für Mathematik, Universität Karlsruhe, Karlsruhe, Germany, 1999. 10

[203] R. F. Sincovec, A. M. Erisman, E. L. Yip, and M. A. Epton. Analysis of descriptor systems using numerical algorithms. *IEEE Trans. Automat. Control*, AC-26:139–147, 1981. 270

[204] G. Söderlind. DASP3 — A program for the numerical integration of partitioned stiff ODE's and differential-algebraic systems. Technical Report TRITA-NA-8008, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, Sweden, 1980. 270

[205] A. Steinbrecher. *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*. Dissertation, Institut für Mathematik, TU Berlin, Berlin, Germany, 2006. 9, 295, 354

[206] G. W. Stewart. An updating algorithm for subspace tracking. Technical Report UMIACS-TR-90-86, CS-TR 2494, Department of Computer Science, University of Maryland, College Park, MD, 1991. 275

[207] R. Stöver. *Numerische Lösung von linearen differential-algebraischen Randwertproblemen*. Dissertation, Fachbereich Mathematik, Universität Bremen, Bremen, Germany, 1999. 348

[208] R. Stöver. Collocation methods for solving linear differential-algebraic boundary value problems. *Numer. Math.*, 88:771–795, 2001. 274, 348

[209] G. Strang and G. Fix. *An Analysis of the Finite Element Method*. Prentice-Hall, Englewood Cliffs, NJ, 1973. 10

[210] K. Strehmel and R. Weiner. *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung*. Teubner Verlag, Stuttgart, 1992. 218, 226, 270

[211] J. C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth and Brooks/Cole, Pacific Grove, CA, 1989. 10

[212] R. C. Thompson. Pencils of complex and real symmetric and skew matrices. *Lin. Alg. Appl.*, 147:323–371, 1991. 52

[213] C. Tischendorf. Topological index-calculation of DAEs in circuit simulation. *Z. Angew. Math. Mech.*, 3:1103–1104, 1998. 295

[214] C. Tischendorf. Topological index calculation of differential-algebraic equations in circuit simulation. *Surv. Math. Ind.*, 8:187–199, 1999. 8, 210, 278, 286, 290

[215] J. Tuomela. On singular points of quasilinear differential and differential-algebraic equations. Research Report A369, Institute of Mathematics, Helsinki University of Technology, Espoo, Finland, 1996. 210

[216] P. Van Dooren. The computation of Kronecker's canonical form of a singular pencil. *Lin. Alg. Appl.*, 27:103–141, 1979. 52, 295

[217] G. C. Verghese, B. C. Lévy, and T. Kailath. A general state space for singular systems. *IEEE Trans. Automat. Control*, AC-26:811–831, 1981. 53

[218] R. von Schwerin and M. Winckler. A guide to the integrator library MBSSIM-version 1.00. Technical report, IWR, Universität Heidelberg, Heidelberg, Germany, 1997. 354

[219] W. Walter. *Einführung in die Theorie der Distributionen*. Bibliographisches Institut, Mannheim, 1974. 38, 39

[220] W. Walter. *Gewöhnliche Differentialgleichungen*. Springer-Verlag, Berlin, 4th edition, 1990. 66, 305

[221] W. Wasov. *Asymptotic Expansions for Ordinary Differential Equations*. John Wiley and Sons, New York, NY, 1965. 11

[222] J. Weickert. *Applications of the Theory of Differential–Algebraic Equations to Partial Differential Equations of Fluid Dynamics*. Dissertation, Fakultät für Mathematik, TU Chemnitz, Chemnitz, Germany, 1997. 10

[223] K. Weierstraß. Über ein die homogenen Funktionen zweiten Grades betreffendes Theorem, nebst Anwendung desselben auf die Theorie der kleinen Schwingungen. *Monatsh. Akad. der Wissensch., Berlin*, pages 207–220, 1858. 4, 13, 52

[224] K. Weierstraß. Zur Theorie der bilinearen quadratischen Formen. *Monatsh. Akad. der Wissensch., Berlin*, pages 310–338, 1867. 4, 52

[225] P. Wesseling. *Principles of Computational Fluid Dynamics*. Springer-Verlag, Germany, 2001. 10

[226] J. H. Wilkinson. Kronecker's canonical form and the QZ-algorithm. *Lin. Alg. Appl.*, 28:285–303, 1979. 52, 295

[227] S. J. Wright. Stable parallel algorithms for two-point boundary value problems. *SIAM J. Sci. Comput.*, 13:742–764, 1992. 276

[228] J. Yen. Constrained equations of motion in multibody dynamics as ODEs on manifolds. *SIAM J. Numer. Anal.*, 30:553–568, 1993. 274

Index