Differential-Algebraic Equations Forum

DAE-F

René Lamour
Roswitha März
Caren Tischendorf

Differential-Algebraic Equations: A Projector Based Analysis

Springer

Differential-Algebraic Equations Forum

Differential-Algebraic Equations Forum

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

For further volumes:
www.springer.com/series/11221

René Lamour · Roswitha März · Caren Tischendorf

Differential-Algebraic Equations: A Projector Based Analysis

Springer

René Lamour
Department of Mathematics
Humboldt-University of Berlin
Berlin, Germany

Roswitha März
Department of Mathematics
Humboldt-University of Berlin
Berlin, Germany

Caren Tischendorf
Mathematical Institute
University of Cologne
Cologne, Germany

ISBN 978-3-642-27554-8 ISBN 978-3-642-27555-5 (eBook)
DOI 10.1007/978-3-642-27555-5
Springer Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013930590

Mathematics Subject Classification (2010): 34A09, 34-02, 65L80, 65-02, 15A22, 65L05, 49K15, 34D05, 34G99

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Foreword by the Editors

We are very pleased to write the Foreword of this book by René Lamour, Roswitha März, and Caren Tischendorf. This book appears as the first volume in the recently established series "FORUM DAEs"—a forum which aims to present different directions in the widely expanding field of differential-algebraic equations (DAEs).

Although the theory of DAEs can be traced back earlier, it was not until the 1960s that mathematicians and engineers started to study seriously various aspects of DAEs, such as computational issues, mathematical theory, and applications. DAEs have developed today, half a century later, into a discipline of their own within applied mathematics, with many relationships to mathematical disciplines such as algebra, functional analysis, numerical analysis, stochastics, and control theory, to mention but a few. There is an intrinsic mathematical interest in this field, but this development is also supported by extensive applications of DAEs in chemical, electrical and mechanical engineering, as well as in economics.

Roswitha März' group has been at the forefront of the development of the mathematical theory of DAEs since the early 1980s; her valuable contribution was to introduce—with a Russian functional analytic background—the method now known as the "projector approach" in DAEs. Over more than 30 years, Roswitha März established a well-known group within the DAE community, making many fundamental contributions. The projector approach has proven to be valuable for a huge class of problems related to DAEs, including the (numerical) analysis of models for dynamics of electrical circuits, mechanical multibody systems, optimal control problems, and infinite-dimensional differential-algebraic systems.

Broadly speaking, the results of the group have been collected in the present textbook, which comprises 30 years of development in DAEs from the viewpoint of projectors. It contains a rigorous and stand-alone introduction to the projector approach to DAEs. Beginning with the case of linear constant coefficient DAEs, this approach is then developed stepwise for more general types, such as linear DAEs with variable coefficients and nonlinear problems. A central concept in the theory of DAEs is the "index", which is, roughly speaking, a measure of the difficulty of

(numerical) solution of a given DAE. Various index concepts exist in the theory of DAEs; and the one related to the projector approach is the "tractability index". Analytical and numerical consequences of the tractability index are presented. In addition to the discussion of the analytical and numerical aspects of different classes of DAEs, this book places special emphasis on DAEs which are explicitly motivated by practice: The "functionality" of the tractability index is demonstrated by means of DAEs arising in models for the dynamics of electrical circuits, where the index has an explicit interpretation in terms of the topological structure of the interconnections of the circuit elements. Further applications and extensions of the projector approach to optimization problems with DAE constraints and even coupled systems of DAEs and partial differential equations (the so-called "PDAEs") are presented.

If one distinguishes strictly between a textbook and a monograph, then we consider the present book to be the second available textbook on DAEs. Not only is it complementary to the other textbook in the mathematical treatment of DAEs, this book is more research-oriented than a tutorial introduction; novel and unpublished research results are presented. Nonetheless it contains a self-contained introduction to the projector approach. Also various relations and substantial cross-references to other approaches to DAEs are highlighted.

This book is a textbook on DAEs which gives a rigorous and detailed mathematical treatment of the subject; it also contains aspects of computations and applications. It is addressed to mathematicians and engineers working in this field, and it is accessible to students of mathematics after two years of study, and also certainly to lecturers and researchers. The mathematical treatment is complemented by many examples, illustrations and explanatory comments.

Ilmenau, Germany *Achim Ilchmann*
Hamburg, Germany *Timo Reis*
June 2012

Preface

We assume that differential-algebraic equations (DAEs) and their more abstract versions in infinite-dimensional spaces comprise *great potential for future mathematical modeling*. To an increasingly large extent, in applications, DAEs are automatically generated, often by coupling various subsystems with large dimensions, but *without manifested mathematically useful structures*. Providing tools to uncover and to monitor mathematical DAE structures is one of the current challenges. What is needed are criteria in terms of the original data of the given DAE. The projector based DAE analysis presented in this monograph is intended to address these questions.

We have been working on our theory of DAEs for quite some time. This theory has now achieved a certain maturity. Accordingly, it is time to record these developments in one coherent account. From the very beginning we were in the fortunate position to communicate with colleagues from all over the world, advancing different views on the topic, starting with Linda R. Petzold, Stephen L. Campbell, Werner C. Rheinboldt, Yuri E. Boyarintsev, Ernst Hairer, John C. Butcher and many others not mentioned here up to John D. Pryce, Ned Nedialkov, Andreas Griewank. We thank all of them for stimulating discussions.

For years, all of us have taught courses, held seminars, supervised diploma students and PhD students, and gained fruitful feedback, which has promoted the progress of our theory. We are indebted to all involved students and colleagues, most notably the PhD students.

Our work was inspired by several fascinating projects and long term cooperation, in particular with Roland England, Uwe Feldmann, Claus Führer, Michael Günther, Francesca Mazzia, Volker Mehrmann, Peter C. Müller, Peter Rentrop, Ewa Weinmüller, Renate Winkler.

We very much appreciate the joint work with Katalin Balla, who passed away too early in 2005, and the colleagues Michael Hanke, Immaculada Higueras, Galina Kurina, and Ricardo Riaza. All of them contributed essential ideas to the projector based DAE analysis.

We are indebted to the German Federal Ministry of Education and Research (BMBF) and the German Research Foundation (DFG), in particular the research center MATHEON in Berlin, for supporting our research in a lot of projects.

We would like to express our gratitude to many people for their support in the preparation of this volume. In particular we thank our colleague Jutta Kerger.

Last but not least, our special thanks are due to Achim Ilchmann and Timo Reis, the editors of the DAE Forum. We appreciate very much their competent counsel for improving the presentation of the theory.

We are under obligations to the staff of Springer for their careful assistance.

René Lamour Roswitha März Caren Tischendorf

Contents

Notations

Abbreviations

ADAE	abstract DAE
BDF	backward differentiation formula
DAE	differential-algebraic equation
GLM	general linear method
IERODE	inherent explicit regular ODE
IESODE	inherent explicit singular ODE
IVP	initial value problem
MNA	modified nodal analysis
ODE	ordinary differential equation
PDAE	partial DAE
SCF	standard canonical form
SSCF	strong SCF

Common notation

$\mathbb{N}$	natural numbers
$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$\mathbb{K}$	alternatively $\mathbb{R}$ or $\mathbb{C}$
$\mathbb{K}^n$	n-dimensional vector space
$M \in \mathbb{K}^{m,n}$	matrix with m rows and n columns
$M \in L(\mathbb{K}^n, \mathbb{K}^m)$	linear mapping from $\mathbb{K}^n$ into $\mathbb{K}^m$, also for $M \in \mathbb{K}^{m,n}$
$L(\mathbb{K}^m)$	shorthand for $L(\mathbb{K}^m, \mathbb{K}^m)$
M^{T}	transposed matrix
M^*	transposed matrix with real or complex conjugate entries
M^{-1}	inverse matrix
M^-	reflexive generalized inverse of M
M^+	Moore–Penrose inverse of M
$\ker M$	kernel of M, $\ker M = \{z \mid Mz = 0\}$

$\operatorname{im} M$	image of M, $\operatorname{im} M = \{z \mid z = My,\ y \in \mathbb{R}^n\}$
$\operatorname{ind} M$	index of M, $\operatorname{ind} M = \min\{k : \ker M^k = \ker M^{k+1}\}$
$\operatorname{rank} M$	rank of M
$\det M$	determinant of M
span	linear hull of a set of vectors
dim	dimension of a (sub)space
diag	diagonal matrix
$\{0\}$	set containing the zero element only
$M \cdot \mathcal{N}$	$= \{z \mid z = My,\ y \in \mathcal{N}\}$
$\forall$	for all
$\perp$	orthogonal set, $\mathcal{N}^\perp = \{z \mid \langle n, z\rangle = 0, \forall n \in \mathcal{N}\}$
$\otimes$	Kronecker product
$\oplus$	direct sum
$\ominus$	$\mathcal{X} = \mathcal{N}_i \ominus \mathcal{N}_j \Leftrightarrow \mathcal{N}_i = \mathcal{X} \oplus \mathcal{N}_j$
$\{A, B\}$	ordered pair
$\lvert\cdot\rvert$	vector and matrix norms in $\mathbb{R}^m$
$\lvert\lvert\cdot\rvert\rvert$	function norm
$\langle\cdot,\cdot\rangle$	scalar product in $\mathbb{K}^m$, dual pairing
$(\cdot\lvert\cdot)_H$	scalar product in Hilbert space H
I, I_d	identity matrix (of dimension d)
$\mathcal{I}$	interval of independent variable
$(\ )'$	total time derivative, total derivative in jet variables
$(\ )_x$	(partial) derivative with respect to x
$\mathcal{C}(\mathcal{I}, \mathbb{R}^m)$	set of continuous functions
$\mathcal{C}^k(\mathcal{I}, \mathbb{R}^m)$	set of k-times continuously differentiable functions
$L_2(\mathcal{I}, \mathbb{R}^m)$	Lebesgue space
$H^1(\mathcal{I}, \mathbb{R}^m)$	Sobolev space

Special notation

$\mathcal{M}_0(t)$	obvious constraint
G_i	member of admissible matrix function sequence
r_i	$\operatorname{rank} G_i$, see Definition 1.17
S_j	$S_j = \ker \mathcal{W}_j B$, see Theorem 2.8 and following pages
N_j	$N_j = \ker G_j$, in Chapter 9: N_j subspace of $\ker G_j$
$\widehat{N}_i$	intersection: $N_0 + \cdots + N_{i-1} \cap N_i$, see (1.12)
N_{can}	canonical subspace, see Definition 2.36
$N_{can\,\mu}$	canonical subspace of an index μ DAE
S_{can}	canonical subspace (Definition 2.36)
$M_{can,q}$	set of consistent values, see (2.98)
$\mathcal{I}_{reg}$	set of regular points, see Definition 2.74
$X(\cdot, t_0)$	fundamental solution matrix normalized at t_0
dom_f	definition domain of f
$\mathcal{C}^k$-subspace	smooth subspace (cf. Section A.4)
$\mathcal{C}^\nu_*(\mathcal{G})$	set of reference functions, see Definition 3.17

$\mathcal{C}_D^1$	$\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) := \{x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : Dx \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\}$, see (1.78)
H_D^1	$H_D^1(\mathcal{I},\mathbb{R}^m) := \{x \in \mathcal{L}_2(\mathcal{I},\mathbb{R}^m) :\ Dx \in H^1(\mathcal{I},\mathbb{R}^n)\}$
$\mathcal{C}^{ind\ \mu}$	function space, see (2.104)
$\mathcal{G}$	regularity region

For projectors we usually apply the following notation:

Q	nullspace projector of a matrix G, $\operatorname{im} Q = \ker G$, $GQ = 0$
P	complementary projector, $P = I - Q$, $GP = G$
$\mathcal{W}$	projector along the image of G, $\ker \mathcal{W} = \operatorname{im} G$, $\mathcal{W}G = 0$
$P_i \cdots P_j$	ordered product, $\prod_{k=i}^{j} P_k$
Π_i	$\Pi_i := P_0 P_1 \cdots P_i$
Π_{can} $\Pi_{can\ \mu}$	canonical projector (of an index-μ DAE)
P_{dich}	dichotomic projector, see Definition 2.56

Introduction

Ordinary differential equations (ODEs) define relations concerning function values and derivative values of an unknown vector valued function in one real independent variable often called time and denoted by t. An explicit ODE

$$x'(t) = g(x(t),t)$$

displays the derivative value $x'(t)$ explicitly in terms of t and $x(t)$. An implicit ODE

$$f(x'(t),x(t),t) = 0$$

is said to be regular, if all its line-elements (x^1,x,t) are regular. A triple (x^1,x,t) belonging to the domain of interest is said to be a regular line-element of the ODE, if $f_{x^1}(x^1,x,t)$ is a nonsingular matrix, and otherwise a singular line-element. This means, in the case of a regular ODE, the derivative value $x'(t)$ is again fully determined in terms of t and $x(t)$, but in an implicit manner.

An ODE having a singular line-element is said to be a singular ODE. In turn, singular ODEs comprise quite different classes of equations. For instance, the linear ODE

$$tx'(t) - Mx(t) = 0$$

accommodates both regular line-elements for $t \neq 0$ and singular ones for $t = 0$. In contrast, the linear ODE

$$\begin{bmatrix} 1&0&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 0&0&0&0&0 \end{bmatrix} x'(t) + \begin{bmatrix} -\alpha&-1&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix} x(t) - \begin{bmatrix} 0\\0\\0\\0\\ \gamma(t) \end{bmatrix} = 0 \qquad (0.1)$$

has solely singular line-elements. A closer look at the solution flow of the last two ODEs shows a considerable disparity.

The ODE (0.1) serves as a prototype of a differential-algebraic equation (DAE). The related equation $f(x^1,x,t)=0$ determines the components x_1^1,x_3^1,x_4^1, and x_5^1 of x^1 in terms of x and t. The component x_2^1 is not at all given. In addition, there arises the consistency condition $x_5-\gamma(t)=0$ which restricts the flow.

DAEs constitute—in whatever form they are given—somehow uniformly singular ODEs: In common with all ODEs, they define relations concerning function values and derivative values of an unknown vector valued function in one real independent variable. However, in contrast to explicit ODEs, in DAEs these relations are implicit, and, in contrast to regular implicit ODEs, these relations determine just a part of the derivative values. A DAE is an implicit ODE which has solely singular line-elements.

The solutions of the special DAE (0.1) feature an ambivalent nature. On the one hand they are close to solutions of regular ODEs in the sense that they depend smoothly on consistent initial data. On the other hand, tiny changes of γ may yield monstrous variations of the solutions, and the solution varies discontinuously with respect to those changes. We refer to the figures in Example 1.5 to gain an impression of this ill-posed behavior.

The ambivalent nature of their solutions distinguishes DAE as being extraordinary to a certain extent.

DAEs began to attract significant research interest in applied and numerical mathematics in the early 1980s, no more than about three decades ago. In this relatively short time, DAEs have become a widely acknowledged tool to model processes subject to constraints, in order to simulate and to control these processes in various application fields.

The two traditional physical application areas, network simulation in electronics and the simulation of multibody mechanics, are repeatedly addressed in textbooks and surveys (e.g. [96, 25, 189]). Special monographs [194, 63, 188] and much work in numerical analysis are devoted to these particular problems. These two application areas and related fields in science and engineering can also be seen as the most important impetus to begin with systematic DAE research, since difficulties and failures in respective numerical simulations have provoked the analysis of these equations first.

The equations describing electrical networks have the form

$$A(d(x(t),t))'+b(x(t),t)=0, \tag{0.2}$$

with a singular constant matrix A, whereas constrained multibody dynamics is described by equations showing the particular structure

$$x_1'(t)+b_1(x_1(t),x_2(t),x_3(t),t)=0, \tag{0.3}$$

$$x_2'(t)+b_2(x_1(t),x_2(t),t)=0, \tag{0.4}$$

$$b_3(x_2(t),t)=0. \tag{0.5}$$

Those DAEs usually have large dimension. Multibody systems often comprise hundreds of equations and electric network systems even gather up to several millions of equations.

Many further physical systems are naturally described as DAEs, for instance, chemical process modeling, [209]. We agree with [189, p. 192] that DAEs arise probably more often than (regular) ODEs, and many of the well-known ODEs in application are actually DAEs that have been additionally explicitly reduced to ODE form.

Further DAEs arise in mathematics, in particular, as intermediate reduced models in singular perturbation theory, as extremal conditions in optimization and control, and by means of semidiscretization of partial differential equation systems.

Besides the traditional application fields, conducted by the generally increasing role of numerical simulation in science and technology, currently more and more new applications come along, in which different physical components are coupled via a network.

We believe that DAEs and their more abstract versions in infinite-dimensional spaces comprise *great potential for future mathematical modeling*. To an increasingly large extent, in applications, DAEs are automatically generated, often by coupling various subsystems, with large dimensions, but *without manifested mathematically useful structures*. Different modeling approaches may result in different kinds of DAEs. Automatic generation and coupling of various tools may yield quite opaque DAEs. Altogether, this produces the challenging task to *bring to light and to characterize the inherent mathematical structure of DAEs*, to provide test criteria such as index observers and eventually hints for creating better qualified model modifications. For a reliable practical treatment, which is the eventual aim, for numerical simulation, sensitivity analysis, optimization and control, and last but not least practical upgrading models, one needs pertinent information concerning the mathematical structure. Otherwise their procedures may fail or, so much the worse, generate wrong results. In consequence, providing practical assessment tools to uncover and to monitor mathematical DAE structures is one of the actual challenges. What are needed are criteria in terms of the original data of the given DAE. The projector based DAE analysis presented in this monograph is intended to address these questions.

Though DAEs have been popular among numerical analysts and in various application fields, so far they play only a marginal role in contiguous fields such as nonlinear analysis and dynamical systems. However, an input from those fields would be desirable. It seems, responsible for this shortage is the quite common view of DAEs as in essence nothing other than implicitly written regular ODEs or vector fields on manifolds, making some difficulties merely in numerical integration. The latter somehow biased opinion is still going strong. It is fortified by the fact that almost all approaches to DAEs suppose that the DAE is eventually reducible to an ODE as a basic principle. This opinion is summarized in [189, p. 191] as follows: *It is a fact, not a mere point of view, that a DAE eventually reduces to an ODE on a manifold. The attitude of acknowledging this fact from the outset leads to a reduc-*

tion procedure suitable for the investigation of many problems .... The mechanism of the geometric reduction procedure completely elucidates the "algebraic" and the "differential" aspects of a DAE. The algebraic part consists in the characterization of the manifold over which the DAE becomes an ODE and, of course, the differential part provides the reduced ODE. Also in [130] the explicit reduction of the general DAE

$$\mathfrak{f}(x'(t),x(t),t) = 0, \tag{0.6}$$

with a singular partial Jacobian $\mathfrak{f}_{x'}$, into a special reduced form plays a central role. Both monographs [189, 130] concentrate on related reduction procedures which naturally suppose higher partial derivatives of the function $\mathfrak{f}$, either to provide sequences of smooth (sub)manifolds or to utilize a so-called derivative array system. The differential geometric approach and the reduction procedures represent powerful tools to analyze and to solve DAEs. Having said that, we wonder about the misleading character of this purely geometric view, which underlines the closedness to regular ODEs, but loses sight of the ill-posed feature.

So far, most research concerning general DAEs is addressed to equation (0.6), and hence we call this equation a *DAE in standard form*. Usually, a solution is then supposed to be at least continuously differentiable.
In contrast, in the present monograph we investigate equations of the form

$$f((d(x(t),t))',x(t),t) = 0, \tag{0.7}$$

which show the derivative term involved by means of an extra function d. We see the network equation (0.2) as the antetype of this form. Also the system (0.3)–(0.5) has this form

$$\begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} \right)' + \begin{bmatrix} b_1(x_1(t),x_2(t),x_3(t),t) \\ b_2(x_1(t),x_2(t),t) \\ b_3(x_2(t),t) \end{bmatrix} = 0 \tag{0.8}$$

a priori. It appears that in applications actually DAEs in the form (0.7) arise, which precisely indicates the involved derivatives. The DAE form (0.7) is comfortable; it involves the derivative by the extra nonlinear function d, whereby $x(t) \in \mathbb{R}^m$ and $d(x(t),t) \in \mathbb{R}^n$ may have different sizes, as is the case in (0.8). A particular instance of DAEs (0.7) is given by the so-called *conservative form* DAEs [52]. Once again, the idea for version (0.7) originates from circuit simulation problems, in which this form is well approved (e.g. [75, 168]).

However, though equation (0.7) represents a more precise model, one often transforms it to standard form (0.6), which allows to apply results and tools from differential geometry, numerical ODE methods, and ODE software.

Turning from the model (0.7) to a standard form DAE one veils the explicit precise information concerning the derivative part. With this background, we are confronted with the question of what a DAE solution should be. Following the classical sense of differential equations, we ask for continuous functions being *as smooth as necessary*, which satisfy the DAE pointwise on the interval of interest. This is

a common understanding. However, there are different opinions on the meaning of the appropriate smoothness. Having regular ODEs in mind one considers continuously differentiable functions $x(\cdot)$ to be the right candidates for solutions. Up to now, most DAE researchers adopt this understanding of the solution which is supported by the standard DAE formulation. Furthermore, intending to apply formal integrability concepts, differential geometry and derivative array approaches one is led to yet another higher smoothness requirement. In contrast, the multibody system (0.8) suggests, as solutions, continuous functions $x(\cdot)$ having just continuously differentiable components $x_1(\cdot)$ and $x_2(\cdot)$.

An extra matrix figuring out the derivative term was already used much earlier (e.g. [153, 152, 154]); however, this approach did not win much recognition at that time. Instead, the following interpretation of standard form DAEs (e.g. [96]) has been accepted to a larger extent: Assuming the nullspace of the partial Jacobian $\mathfrak{f}_{x'}(x',x,t)$ associated with the standard form DAE (0.6) to be a $\mathcal{C}^1$-subspace, and to be independent of the variables x' and x, one interprets the standard form DAE (0.6) as a short description of the equation

$$\mathfrak{f}((P(t)x(t))' - P'(t)x(t), x(t), t) = 0, \tag{0.9}$$

whereby $P(\cdot)$ denotes any continuously differentiable projector valued function such that the nullspaces $\ker P(\cdot)$ and $\ker \mathfrak{f}_{x'}(x',x,\cdot)$ coincide. This approach is aligned with continuous solutions $x(\cdot)$ having just continuously differentiable products $(Px)(\cdot)$. Most applications yield even constant nullspaces $\ker \mathfrak{f}_{x'}$, and hence constant projector functions P as well. In particular, this is the case for the network equations (0.2) and the multibody systems (0.8).

In general, for a DAE given in the form (0.7), a solution $x(\cdot)$ should be a continuous function such that the superposition $u(\cdot) := d(x(\cdot),\cdot)$ is continuously differentiable. For the particular system (0.8) this means that the components $x_1(\cdot)$ and $x_2(\cdot)$ are continuously differentiable, whereas one accepts a continuous $x_3(\cdot)$.

The question in which way the data functions f and d should be related to each other leads to the notions of *DAEs with properly stated leading term or properly involved derivative*, but also to *DAEs with quasi-proper leading term*. During the last 15 years, the idea of using an extra function housing the derivative part within a DAE has been emphatically pursued. This discussion amounts to the content of this monograph. Formulating DAEs with properly stated leading term yields, in particular, symmetries of linear DAEs and their adjoints, and further favorable consequences concerning optimization problems with DAE constraints. Not surprisingly, numerical discretization methods may perform better than for standard form DAEs. And last, but not least, this approach allows for appropriate generalizations to apply to abstract differential- algebraic systems in Hilbert spaces enclosing PDAEs. We think that, right from the design or modeling stage, it makes sense to look for properly involved derivatives.

This monograph comprises an elaborate analysis of DAEs (0.7), which is accompanied by the consideration of essential numerical aspects. We regard DAEs from an analytical point of view, rather than from a geometric one. Our main ob-

jective consists in the structural and qualitative characterization of DAEs as they are given a priori, without supposing any knowledge concerning solutions and constraints. Afterwards, having the required knowledge of the DAE structure, also solvability assertions follow. Only then do we access the constraints. In contrast, other approaches concede full priority of providing constraints and solutions, as well as transformations into a special form, which amounts to solving the DAE.

We believe in the great potential of our concept in view of the further analysis of classical DAEs and their extensions to abstract DAEs in function spaces. We do not at all apply derivative arrays and prolongated systems, which are commonly used in DAE theory. Instead, certain admissible matrix function sequences and smartly chosen admissible projector functions formed only from the first partial derivatives of the given data function play their role as basic tools. Thereby, continuity properties of projector functions depending on several variables play their role, which is not given if one works instead with basises. All in all, this allows an analysis on a low smoothness level. We pursue a fundamentally alternative approach and present the first rigorous structural analysis of general DAEs in their originally given form without the use of derivative arrays, without supposing any knowledge concerning constraints and solutions.

The concept of a projector based analysis of general DAEs was sketched first in [160, 171, 48], but it has taken its time to mature. Now we come up with a unique general theory capturing constant coefficient linear problems, variable coefficient linear problems and fully nonlinear problems in a hierarchic way. We address a further generalization to abstract DAEs. It seems, after having climbed the (at times seemingly pathless) mountain of projectors, we are given transparency and beautiful convenience. By now the projector based analysis is approved to be a prospective way to investigate DAEs and also to yield reasonable open questions for future research.

The central idea of the present monograph consists in a rigorous definition of regularity of a DAE, accompanied with certain characteristic values including the tractability index, which is related to an open subset of the definition domain of the data function f, a so-called *regularity region*. Regularity is shown to be stable with respect to perturbations. Close relations of regularity regions and linearizations are proved. In general, one has to expect that the definition domain of f decomposes into several regularity regions whose borders consist of critical points. Solutions do not necessarily stay in one of these regions; solutions may cross the borders and undergo bifurcation, etc.

The larger part of the presented material is new and as yet unpublished. Parts were earlier published in journals, and just the regular linear DAE framework (also critical points in this context) is available in the book [194].

The following basic types of DAEs can reasonably be discerned:

✓ fully implicit nonlinear DAE with nonlinear derivative term

$$f((d(x(t),t))',x(t),t) = 0, \tag{0.10}$$

✓ fully implicit nonlinear DAE with linear derivative term

$$f((D(t)x(t))',x(t),t) = 0, \tag{0.11}$$

✓ quasi-linear DAE with nonlinear derivative term (involved linearly)

$$A(x(t),t)(d(x(t),t))' + b(x(t),t) = 0, \tag{0.12}$$

✓ quasi-linear DAE with linear derivative term

$$A(x(t),t)(D(t)x(t))' + b(x(t),t) = 0, \tag{0.13}$$

✓ linear DAE with variable coefficients

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \tag{0.14}$$

✓ linear DAE with constant coefficients

$$A(Dx(t))' + Bx(t) = q(t), \tag{0.15}$$

✓ semi-implicit DAE with explicitly given derivative-free equation

$$f_1((d(x(t),t))',x(t),t) = 0, \tag{0.16}$$
$$f_2(x(t),t) = 0, \tag{0.17}$$

✓ semi-implicit DAE with explicitly partitioned variable and explicitly given derivative-free equation

$$f_1(x_1'(t),x_1(t),x_2(t),t) = 0, \tag{0.18}$$
$$f_2(x_1(t),x_2(t),t) = 0, \tag{0.19}$$

✓ semi-explicit DAE with explicitly partitioned variable and explicitly given derivative-free equation

$$x_1'(t) + b_1(x_1(t),x_2(t),t) = 0, \tag{0.20}$$
$$b_2(x_1(t),x_2(t),t) = 0. \tag{0.21}$$

So-called Hessenberg form DAEs of size r, which are described in Section 3.5, form further subclasses of semi-explicit DAEs. For instance, the DAE (0.8) has Hessenberg form of size 3. Note that much work developed to treat higher index DAEs is actually limited to Hessenberg form DAEs of size 2 or 3.

The presentation is divided into Part I to Part IV followed by Appendices A, B, and C.

Part I describes the core of the projector based DAE analysis: the construction of admissible matrix function sequences associated by admissible projector functions and the notion of regularity regions.

Chapter 1 deals with constant coefficient DAEs and matrix pencils only. We reconsider algebraic features and introduce into the projector framework. This can be skipped by readers familiar with the basic linear algebra including projectors.

The more extensive Chapter 2 provides the reader with admissible matrix function sequences and the resulting constructive projector based decouplings. With this background, a comprehensive linear theory is developed, including qualitative flow characterizations of regular DAEs, the rigorous description of admissible excitations, and also relations to several canonical forms and the strangeness index.

Chapter 3 contains the main constructions and assertions concerning general regular nonlinear DAEs, in particular the regularity regions and the practically important theorem concerning linearizations. It is recommended to take a look to Chapter 2 before reading Chapter 3.

We emphasize the hierarchical organization of Part I. The admissible matrix function sequences built for the nonlinear DAE (0.10) generalize those for the linear DAE (0.14) with variable coefficients, which, in turn, represent a generalization of the matrix sequences made for constant coefficient DAEs (0.15).

Part IV continues the hierarchy in view of different further aspects. Chapter 9 about quasi-regular DAEs (0.10) incorporates a generalization which relaxes the constant-rank conditions supporting admissible matrix function sequences. Chapter 10 on nonregular DAEs (0.11) allows a different number of equations and of unknown components. Finally, in Chapter 12, we describe abstract DAEs in infinite-dimensional spaces and include PDAEs.

Part IV contains the additional Chapter 11 conveying results on minimization with DAE constraints obtained by means of the projector based technique.

Part II is a self-contained index-1 script. It comprises in its three chapters the analysis of regular index-1 DAEs (0.11) and their numerical integration, addressing also stability topics such as contractivity and stability in Lyapunov's sense. Part II constitutes in essence an up-to-date improved and completed version of the early book [96]. While the latter is devoted to standard form DAEs via the interpretation (0.9), now the more general equations (0.11) are addressed.

Part III adheres to Part I giving an elaborate account of computational methods concerning the practical construction of projectors and that of admissible projector functions in Chapter 7. A second chapter discusses several aspects of the numerical treatment of regular higher index DAEs such as consistent initialization and numerical integration.

Appendix B contains technically involved costly proofs. Appendices A and C collect and provide basic material concerning linear algebra and analysis, for instance the frequently used C^1-subspaces.

Plenty of reproducible small academic examples are integrated into the explanations for easier reading, illustrating and confirming the features under consideration. To this end, we emphasize that those examples are always too simple. They bring to light special features, but they do not really reflect the complexity of DAEs.

The material of this monograph is much too comprehensive to be taught in a standard graduate course. However different combinations of selected chapters should be well suited for those courses. In particular, we recommend the following:

- Projector based DAE analysis (Part I, possibly without Chapter 1).
- Analysis of index-1 DAEs and their numerical treatment (Part II, possibly plus Chapter 8).
- Matrix pencils, theoretical and practical decouplings (Chapters 1 and 7).
- General linear DAEs (Chapter 2, material on the linear DAEs of Chapters 10 and 9).

Advanced courses communicating Chapter 12 or Chapter 11 could be given to students well grounded in DAE basics (Parts I and II) and partial differential equations, respectively optimization.

Part I
Projector based approach

Part I describes the core of the projector based DAE analysis, the construction of admissible matrix function sequences and the notions of regular points and regularity regions of general DAEs

$$f((d(x(t),t))',x(t),t) = 0$$

in a hierarchical manner starting with constant coefficient linear DAEs, then turning to linear DAEs with variable coefficients, and, finally, considering fully implicit DAEs.

Chapter 1 deals with constant coefficient DAEs and matrix pencils. We reconsider algebraic features and introduce them into the projector framework. This shows how the structure of the Weierstraß–Kronecker form of a regular matrix pencil can be depicted by means of admissible projectors.

The extensive Chapter 2 on linear DAEs with variable coefficients characterizes regular DAEs by means of admissible matrix function sequences and associated projectors and provides constructive projector based decouplings of regular linear DAEs.

Then, with this background, a comprehensive linear theory of regular DAEs is developed, including qualitative flow properties and a rigorous description of admissible excitations. Moreover, relations to several canonical forms and other index notions are addressed.

Chapter 3 contains the main constructions and assertions concerning general regular nonlinear DAEs, in particular the regularity regions and the practically important theorem concerning linearizations. Also local solvability assertions and perturbation results are proved.

We emphasize the hierarchical organization of the approach. The admissible matrix function sequences built for the nonlinear DAE (0.10) generalize those for the linear DAE (0.14) with variable coefficients, which, in turn, represent a generalization of the matrix sequences made for constant coefficient DAEs (0.15). Part IV continues the hierarchy with respect to different views.

Chapter 1
Linear constant coefficient DAEs

Linear DAEs with constant coefficients have been well understood by way of the theory of matrix pencils for quite a long time, and this is the reason why they are only briefly addressed in monographs. We consider them in detail here, not because we believe that the related linear algebra has to be invented anew, but as we intend to give a sort of guide for the subsequent extensive discussion of linear DAEs with time-varying coefficients and of nonlinear DAEs.

This chapter is organized as follows. Section 1.1 records well-known facts on regular matrix pairs and describes the structure of the related DAEs. The other sections serve as an introduction to the projector based analysis. Section 1.2 first provides the basic material of this analysis: the admissible matrix sequences and the accompanying admissible projectors and characteristic values in Subsection 1.2.1, the decoupling of regular DAEs by arbitrary admissible projectors in Subsection 1.2.2, and the complete decoupling in Subsection 1.2.3. The two subsequent Subsections 1.2.5 and 1.2.6 are to clarify the relations to the Weierstraß–Kronecker form. Section 1.3 provides the main result concerning the high consistency of the projector based approach and the DAE structure by the Weierstraß–Kronecker form, while Section 1.4 collects practically useful details on the topic. Section 1.5 develops proper formulations of the leading term of the DAE by means of two well-matched matrices. The chapter ends with notes and references.

1.1 Regular DAEs and the Weierstraß–Kronecker form

In this section we deal with the equation

$$Ex'(t) + Fx(t) = q(t),\ t \in \mathcal{I}, \tag{1.1}$$

formed by the ordered pair $\{E, F\}$ of real valued $m \times m$ matrices E, F. For given functions $q : \mathcal{I} \to \mathbb{R}^m$ being at least continuous on the interval $\mathcal{I} \subseteq \mathbb{R}$, we are looking for continuous solutions $x : \mathcal{I} \to \mathbb{R}^m$ having a continuously differentiable com-

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_1,

ponent Ex. We use the notation $Ex'(t)$ for $(Ex)'(t)$. Special interest is directed to homogeneous equations

$$Ex'(t) + Fx(t) = 0, \quad t \in \mathbb{R}. \tag{1.2}$$

For $E = I$, the special case of explicit ODEs is covered. Now, in the more general setting, the ansatz $x_*(t) = e^{\lambda_* t} z_*$ well-known for explicit ODEs, yields

$$Ex'_*(t) + Fx_*(t) = e^{\lambda_* t}(\lambda_* E + F)z_*.$$

Hence, x_* is a nontrivial particular solution of the DAE (1.2) if λ_* is a zero of the polynomial $p(\lambda) := \det(\lambda E + F)$, and $z_* \neq 0$ satisfies the relation $(\lambda_* E + F)z_* = 0$. Then λ_* and z_* are called generalized eigenvalue and eigenvector, respectively. This shows the meaning of the polynomial $p(\lambda)$ and the related family of matrices $\lambda E + F$ named the *matrix pencil* formed by $\{E, F\}$.

Example 1.1 (A solvable DAE). The DAE

$$\begin{aligned} x_1' - x_1 &= 0, \\ x_2' + x_3 &= 0, \\ x_2 &= 0, \end{aligned}$$

is given by the matrices

$$E = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \text{ and } F = \begin{bmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

yielding

$$p(\lambda) = \det(\lambda E + F) = \det \begin{bmatrix} \lambda - 1 & 0 & 0 \\ 0 & \lambda & 1 \\ 0 & 1 & 0 \end{bmatrix} = 1 - \lambda.$$

The value $\lambda_* = 1$ is a generalized eigenvalue and the vector $z_* = (1\,0\,0)^{\mathrm{T}}$ is a generalized eigenvector. Obviously, $x_*(t) = e^{\lambda_* t} z_* = (e^t\,0\,0)^{\mathrm{T}}$ is a nontrivial solution of the differential-algebraic equation. □

If E is nonsingular, the homogeneous equation (1.2) represents an implicit regular ODE and its fundamental solution system forms an m-dimensional subspace in $\mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$. What happens if E is singular? Is there a class of equations, such that equation (1.2) has a finite-dimensional solution space? The answer is closely related to the notion of regularity.

Definition 1.2. Given any ordered pair $\{E, F\}$ of matrices $E, F \in L(\mathbb{R}^m)$, the matrix pencil $\lambda E + F$ is said to be *regular* if the polynomial $p(\lambda) := \det(\lambda E + F)$ does not vanish identically. Otherwise the matrix pencil is said to be *singular*.
Both the ordered pair $\{E, F\}$ and the DAE (1.1) are said to be *regular* if the accompanying matrix pencil is regular, and otherwise *nonregular*.

A pair $\{E,F\}$ with a nonsingular matrix E is always regular, and its polynomial $p(\lambda)$ is of degree m. In the case of a singular matrix E, the polynomial degree is lower as demonstrated in Example 1.1.

Proposition 1.3. *For any regular pair $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, there exist nonsingular matrices $L,K \in L(\mathbb{R}^m)$ and integers $0 \le l \le m$, $0 \le \mu \le l$, such that*

$$LEK = \begin{bmatrix} I & \\ & N \end{bmatrix} \begin{matrix} \}m-l \\ \}l \end{matrix}, \quad LFK = \begin{bmatrix} W & \\ & I \end{bmatrix} \begin{matrix} \}m-l \\ \}l \end{matrix}. \tag{1.3}$$

Thereby, N is absent if $l = 0$, and otherwise N is nilpotent of order μ, i.e., $N^\mu = 0$, $N^{\mu-1} \neq 0$. The integers l and μ as well as the eigenstructure of the blocks N and W are uniquely determined by the pair $\{E,F\}$.

Proof. If E is nonsingular, we simply put $l = 0$, $L = E^{-1}$, $K = I$ and the assertion is true.
Assume E to be singular. Since $\{E,F\}$ is a regular pair, there is a number $c \in \mathbb{R}$ such that $cE+F$ is nonsingular. Form $\tilde{E} := (cE+F)^{-1}E$, $\tilde{F} := (cE+F)^{-1}F = I - c\tilde{E}$, $\mu = \operatorname{ind}\tilde{E}$, $r = \operatorname{rank}\tilde{E}^\mu$, $S = [s_1 \dots s_m]$ with $s_1,\dots,s_r$ and $s_{r+1},\dots,s_m$ being bases of $\operatorname{im}\tilde{E}^\mu$ and $\ker\tilde{E}^\mu$, respectively. Lemma A.11 provides the special structure of the product $S^{-1}\tilde{E}S$, namely,

$$S^{-1}\tilde{E}S = \begin{bmatrix} \tilde{M} & 0 \\ 0 & \tilde{N} \end{bmatrix},$$

with a nonsingular $r \times r$ block $\tilde{M}$ and a nilpotent $(m-r)\times(m-r)$ block $\tilde{N}$. $\tilde{N}$ has nilpotency index μ. Compute

$$S^{-1}\tilde{F}S = I - cS^{-1}\tilde{E}S = \begin{bmatrix} I - c\tilde{M} & 0 \\ 0 & I - c\tilde{N} \end{bmatrix}.$$

The block $I - c\tilde{N}$ is nonsingular due to the nilpotency of $\tilde{N}$. Denote

$$L := \begin{bmatrix} \tilde{M}^{-1} & 0 \\ 0 & (I-c\tilde{N})^{-1} \end{bmatrix} S^{-1}(cE+F)^{-1},$$
$$K := S, \qquad N := (I-c\tilde{N})^{-1}\tilde{N}, \qquad W := \tilde{M}^{-1} - cI,$$

so that we arrive at the representation

$$LEK = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad LFK = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}.$$

Since $\tilde{N}$ and $(I-c\tilde{N})^{-1}$ commute, one has

$$N^l = ((I-c\tilde{N})^{-1}\tilde{N})^l = ((I-c\tilde{N})^{-1})^l\tilde{N}^l,$$

and N inherits the nilpotency of $\tilde{N}$. Thus, $N^\mu = 0$ and $N^{\mu-1} \neq 0$. Put $l := m-r$. It remains to verify that the integers l and μ as well as the eigenstructure of N and

W are independent of the transformations L and K. Assume that there is a further collection $\tilde{l}, \tilde{\mu}, \tilde{L}, \tilde{K}$, $\tilde{r} = m - \tilde{l}$ such that

$$\tilde{L}E\tilde{K} = \begin{bmatrix} I_{\tilde{r}} & 0 \\ 0 & \tilde{N} \end{bmatrix}, \quad \tilde{L}F\tilde{K} = \begin{bmatrix} \tilde{W} & 0 \\ 0 & I_{\tilde{l}} \end{bmatrix}.$$

Considering the degree of the polynomial

$$\begin{aligned} p(\lambda) = \det(\lambda E + F) &= \det(L^{-1}) \det(\lambda I_r + W) \det(K^{-1}) \\ &= \det(\tilde{L}^{-1}) \det(\lambda I_{\tilde{r}} + \tilde{W}) \det(\tilde{K}^{-1}) \end{aligned}$$

we realize that the values r and $\tilde{r}$ must coincide, hence $l = \tilde{l}$. Introducing $U := \tilde{L}L^{-1}$ and $V := \tilde{K}^{-1}K$ one has

$$U \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix} = \tilde{L}EK = \begin{bmatrix} I & 0 \\ 0 & \tilde{N} \end{bmatrix} V, \quad U \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} = \tilde{L}FK = \begin{bmatrix} \tilde{W} & 0 \\ 0 & I \end{bmatrix} V,$$

and, in detail,

$$\begin{bmatrix} U_{11} & U_{12}N \\ U_{21} & U_{22}N \end{bmatrix} = \begin{bmatrix} V_{11} & V_{12} \\ \tilde{N}V_{21} & \tilde{N}V_{22} \end{bmatrix}, \quad \begin{bmatrix} U_{11}W & U_{12} \\ U_{21}W & U_{22} \end{bmatrix} = \begin{bmatrix} \tilde{W}V_{11} & \tilde{W}V_{12} \\ V_{21} & V_{22} \end{bmatrix}.$$

Comparing the entries of these matrices we find the relations $U_{12}N = V_{12}$ and $U_{12} = \tilde{W}V_{12}$, which lead to $U_{12} = \tilde{W}U_{12}N = \cdots = \tilde{W}^{\mu}U_{12}N^{\mu} = 0$. Analogously we derive $U_{21} = 0$. Then, the blocks $U_{11} = V_{11}$, $U_{22} = V_{22}$ must be nonsingular. It follows that

$$V_{11}W = \tilde{W}V_{11}, \quad V_{22}N = \tilde{N}V_{22}$$

holds true, that is, the matrices N and $\tilde{N}$ as well as W and $\tilde{W}$ are similar, and in particular, $\mu = \tilde{\mu}$ is valid. □

The real valued matrix N has the eigenvalue zero only, and can be transformed into its Jordan form by means of a real valued similarity transformation. Therefore, in Proposition 1.3, the transformation matrices L and K can be chosen such that N is in Jordan form.

Proposition 1.3 also holds true for complex valued matrices. This is a well-known result of Weierstraß and Kronecker, cf. [82]. The special pair given by (1.3) is said to be *Weierstraß–Kronecker form* of the original pair $\{E, F\}$.

Definition 1.4. The *Kronecker index* of a regular matrix pair $\{E, F\}$, $E, F \in L(\mathbb{R}^m)$, and the *Kronecker index of a regular DAE* (1.1) are defined to be the nilpotency order μ in the Weierstraß–Kronecker form (1.3). We write $\operatorname{ind}\{E, F\} = \mu$.

The Weierstraß–Kronecker form of a regular pair $\{E, F\}$ provides a broad insight into the structure of the associated DAE (1.1). Scaling of (1.1) by L and transforming $x = K \begin{bmatrix} y \\ z \end{bmatrix}$ leads to the equivalent decoupled system

$$y'(t) + Wy(t) = p(t), \quad t \in \mathcal{I}, \tag{1.4}$$
$$Nz'(t) + z(t) = r(t), \quad t \in \mathcal{I}, \tag{1.5}$$

with $Lq =: \begin{bmatrix} p \\ r \end{bmatrix}$. The first equation (1.4) represents a standard explicit ODE. The second one appears for $l > 0$, and it has the only solution

$$z(t) = \sum_{j=0}^{\mu-1} (-1)^j N^j r^{(j)}(t), \tag{1.6}$$

provided that r is smooth enough. The latter one becomes clear after recursive use of (1.5) since

$$z = r - Nz' = r - N(r - Nz')' = r - Nr' + N^2 z'' = r - Nr' + N^2(r - Nz')'' = \cdots$$

Expression (1.6) shows the dependence of the solution x on the derivatives of the source or perturbation term q. The higher the index μ, the more differentiations are involved. Only in the index-1 case do we have $N = 0$, hence $z(t) = r(t)$, and no derivatives are involved. Since numerical differentiations in these circumstances may cause considerable trouble, it is very important to know the index μ as well as details of the structure responsible for a higher index when modeling and simulating with DAEs in practice. The typical solution behavior of ill-posed problems can be observed in higher index DAEs: small perturbations of the right-hand side yield large changes in the solution. We demonstrate this by the next example.

Example 1.5 (Ill-posed behavior in case of a higher index DAE). The regular DAE

$$\underbrace{\begin{bmatrix} 1&0&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \\ 0&0&0&0&0 \end{bmatrix}}_{E} x'(t) + \underbrace{\begin{bmatrix} -\alpha&-1&0&0&0 \\ 0&1&0&0&0 \\ 0&0&1&0&0 \\ 0&0&0&1&0 \\ 0&0&0&0&1 \end{bmatrix}}_{F} x(t) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ \gamma(t) \end{bmatrix},$$

completed by the initial condition $\begin{bmatrix} 1&0&0&0&0 \end{bmatrix} x(0) = 0$, is uniquely solvable for each sufficiently smooth function γ. The identically zero solution corresponds to the vanishing input function $\gamma(t) = 0$. The solution corresponding to the small excitation $\gamma(t) = \varepsilon \frac{1}{n} \sin nt$, $n \in \mathbb{N}$, ε small, is

$$x_1(t) = \varepsilon \int_0^t n^2 e^{\alpha(t-s)} \cos ns \, \mathrm{d}s, \quad x_2(t) = \varepsilon n^2 \cos nt,$$
$$x_3(t) = -\varepsilon n \sin nt, \quad x_4(t) = -\varepsilon \cos nt, \quad x_5(t) = \varepsilon \frac{1}{n} \sin nt.$$

While the excitation tends to zero for $n \to \infty$, the first three solution components grow unboundedly. The solution value at $t = 0$,

$$x_1(0) = 0,\ x_2(0) = \varepsilon n^2,\ x_3(0) = 0,\ x_4(0) = -\varepsilon,\ x_5(0) = 0,$$

moves away from the origin with increasing n, and the origin is no longer a consistent value at $t = 0$ for the perturbed system, as it is the case for the unperturbed one. Figures 1.1 and 1.2 show γ and the response x_2 for $\varepsilon = 0.1$, $n = 1$ and $n = 100$. □

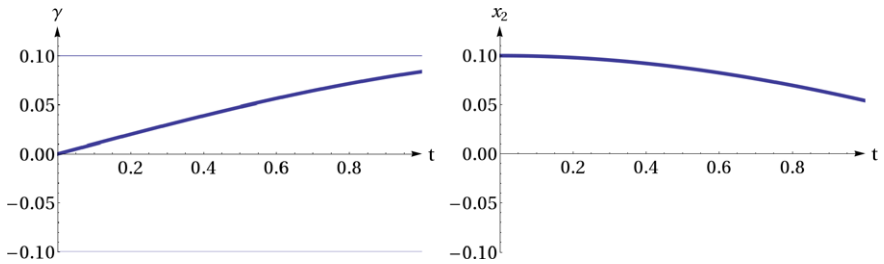


Fig. 1.1 γ and x_2 for $n = 1$

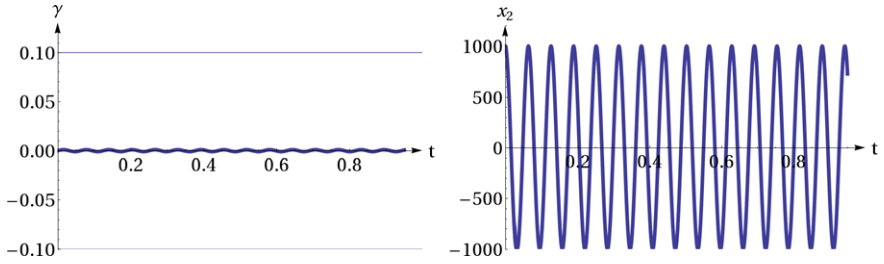


Fig. 1.2 γ and x_2 for $n = 100$

This last little constant coefficient example is relatively harmless. Time-dependent subspaces and nonlinear relations in more general DAEs may considerably amplify the bad behavior. For this reason one should be careful in view of numerical simulations. It may well happen that an integration code seemingly works, however it generates wrong results.

The general solution of a regular homogeneous DAE (1.2) is of the form

$$x(t) = K \begin{bmatrix} e^{-tW} \\ 0 \end{bmatrix} y_0,\ \ y_0 \in \mathbb{R}^{m-l}$$

which shows that the solution space has finite dimension $m - l$ and the solution depends smoothly on the initial value $y_0 \in \mathbb{R}^{m-l}$. Altogether, already for constant coefficient linear DAEs, the solutions feature an ambivalent behavior: they depend smoothly on certain initial values while they are ill-posed with respect to excitations.

The next theorem substantiates the above regularity notion.

Theorem 1.6. *The homogeneous DAE (1.2) has a finite-dimensional solution space if and only if the pair $\{E,F\}$ is regular.*

Proof. As we have seen before, if the pair $\{E,F\}$ is regular, then the solutions of (1.2) form an $(m-l)$-dimensional space. Conversely, let $\{E,F\}$ be a singular pair, i.e., $\det(\lambda E+F)\equiv 0$. For any set of $m+1$ different real values $\lambda_1,\dots,\lambda_{m+1}$ we find nontrivial vectors $\eta_1,\dots,\eta_{m+1}\in\mathbb{R}^m$ such that $(\lambda_i E+F)\eta_i=0$, $i=1,\dots,m+1$, and a nontrivial linear combination $\sum_{i=1}^{m+1}\alpha_i\eta_i=0$.

The function $x(t)=\sum_{i=1}^{m+1}\alpha_i e^{\lambda_i t}\eta_i$ does not vanish identically, and it satisfies the DAE (1.2) as well as the initial condition $x(0)=0$. For disjoint $(m+1)$-element sets $\{\eta_1,\dots,\eta_{m+1}\}$, one always has different solutions, and, consequently, the solution space of a homogeneous initial value problem (IVP) of (1.2) is not finite. □

Example 1.7 (Solutions of a nonregular DAE (cf. [97])). The pair $\{E,F\}$,

$$E=\begin{bmatrix}1&1&0&0\\0&0&0&1\\0&0&0&0\\0&0&1&0\end{bmatrix},\quad F=\begin{bmatrix}0&1&0&0\\0&0&0&0\\0&0&1&0\\0&0&0&0\end{bmatrix},\quad m=4,$$

is singular. In detail, the homogeneous DAE (1.2) reads

$$\begin{aligned}(x_1+x_2)' \quad &+\ x_2 &&= 0,\\ x_4' \quad &&&= 0,\\ &\quad x_3 &&= 0,\\ x_3' \quad &&&= 0.\end{aligned}$$

What does the solution space look like? Obviously, the component x_3 vanishes identically and x_4 is an arbitrary constant function. The remaining equation $(x_1+x_2)'+x_2=0$ is satisfied by any arbitrary continuous x_2, and the resulting expression for x_1 is

$$x_1(t)=c-x_2(t)-\int_0^t x_2(s)ds,$$

c being a further arbitrary constant. Clearly, this solution space does not have finite dimension, which confirms the assertion of Theorem 1.6. Indeed, the regularity assumption is violated since

$$p(\lambda)=\det(\lambda E+F)=\det\begin{bmatrix}\lambda&\lambda+1&0&0\\0&0&0&\lambda\\0&0&1&0\\0&0&\lambda&0\end{bmatrix}=0.$$

Notice that, in the case of nontrivial perturbations q, for the associated perturbed DAE (1.1) the consistency condition $q_3'=q_4$ must be valid for solvability. In practice, such unbalanced models should be avoided. However, in large dimensions m, this might not be a trivial task. □

We take a closer look at the subsystem (1.5) within the Weierstraß–Kronecker form, which is specified by the nilpotent matrix N. We may choose the transformation matrices L and K in such a way that N has Jordan form, say

$$N = \operatorname{diag}[J_1, \ldots, J_s], \tag{1.7}$$

with s nilpotent Jordan blocks

$$J_i = \begin{bmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix} \in L(\mathbb{R}^{k_i}),\ i = 1, \ldots, s,$$

where $k_1 + \cdots + k_s = l$, $\mu = \max\{k_i : i = 1, \ldots, s\}$. The Kronecker index μ equals the order of the maximal Jordan block in N.

The Jordan form (1.7) of N indicates the further decoupling of the subsystem (1.5) in accordance with the Jordan structure into s lower-dimensional equations

$$J_i \zeta_i'(t) + \zeta_i(t) = r_i(t),\ \ i = 1, \ldots, s.$$

We observe that $\zeta_{i,2}, \ldots, \zeta_{i,k_i}$ are components involved with derivatives whereas the derivative of the first component $\zeta_{i,1}$ is not involved. Notice that the value of $\zeta_{i,1}(t)$ depends on the $(k_i - 1)$-th derivative of $r_{i,k_i}(t)$ for all $i = 1, \ldots, s$ since

$$\zeta_{i,1}(t) = r_{i,1}(t) - \zeta_{i,2}'(t) = r_{i,1}(t) - r_{i,2}'(t) + \zeta_{i,3}'(t) = \cdots = \sum_{j=1}^{k_i} (-1)^{j-1} r_{i,j}^{(j-1)}(t).$$

1.2 Projector based decoupling of regular DAEs

1.2.1 Admissible matrix sequences and admissible projectors

Our aim is now a suitable rearrangement of terms within the equation

$$Ex'(t) + Fx(t) = q(t), \tag{1.8}$$

which allows for a similar insight into the structure of the DAE to that given by the Weierstraß–Kronecker form. However, we do not use transformations, but we work in terms of the original equation setting and apply a projector based decoupling concept. The construction is simple. We consider the DAE (1.8) with the coefficients $E, F \in L(\mathbb{R}^m)$.

Put $G_0 := E$, $B_0 := F$, $N_0 := \ker G_0$ and introduce $Q_0 \in L(\mathbb{R}^m)$ as a projector onto N_0. Let $P_0 := I - Q_0$ be the complementary one. Using the basic projector

properties $Q_0^2 = Q_0$, $Q_0P_0 = P_0Q_0 = 0$, $P_0 + Q_0 = I$, $G_0Q_0 = 0$ and $G_0 = G_0P_0$ (see Appendix A), we rewrite the DAE (1.8) consecutively as

$$\begin{aligned}
& G_0x' + B_0x = q \\
\iff \quad & G_0P_0x' + B_0(Q_0 + P_0)x = q \\
\iff \quad & \underbrace{(G_0 + B_0Q_0)}_{=:G_1}(P_0x' + Q_0x) + \underbrace{B_0P_0}_{=:B_1}x = q \\
\iff \quad & G_1(P_0x' + Q_0x) + B_1x = q.
\end{aligned}$$

Next, let Q_1 be a projector onto $N_1 := \ker G_1$, and let $P_1 := I - Q_1$ the complementary one. We rearrange the last equation to

$$\begin{aligned}
& G_1P_1(P_0x' + Q_0x) + B_1(Q_1 + P_1)x = q \\
\iff \quad & \underbrace{(G_1 + B_1Q_1)}_{G_2}\big(P_1(P_0x' + Q_0x) + Q_1x\big) + \underbrace{B_1P_1}_{B_2}x = q
\end{aligned} \tag{1.9}$$

and so on. The goal is a matrix with maximal possible rank in front of the term containing the derivative x'.

We form, for $i \geq 0$,

$$G_{i+1} := G_i + B_iQ_i, \quad N_{i+1} := \ker G_{i+1}, \quad B_{i+1} := B_iP_i \tag{1.10}$$

and introduce $Q_{i+1} \in L(\mathbb{R}^m)$ as a projector onto N_{i+1} with $P_{i+1} := I - Q_{i+1}$. Denote $r_i := \operatorname{rank} G_i$ and introduce the product of projectors $\Pi_i := P_0 \cdots P_i$. These ranks and products of projectors will play a special role later on. From $B_{i+1} = B_iP_i = B_0\Pi_i$ we derive the inclusion $\ker \Pi_i \subseteq \ker B_{i+1}$ as an inherent property of our construction. Since $G_i = G_{i+1}P_i$, the further inclusions

$$\operatorname{im} G_0 \subseteq \operatorname{im} G_1 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq \operatorname{im} G_{i+1},$$

follow, and hence

$$r_0 \leq r_1 \leq \cdots \leq r_i \leq r_{i+1}.$$

An additional inherent property of the sequence (1.10) is given by

$$N_{i-1} \cap N_i \ \subseteq \ N_i \cap N_{i+1}, \ \ i \geq 1. \tag{1.11}$$

Namely, if $G_{i-1}z = 0$ and $G_iz = 0$ are valid for a vector $z \in \mathbb{R}^m$, which corresponds to $P_{i-1}z = 0$ and $P_iz = 0$, i.e., $z = Q_iz$, then we can conclude that

$$G_{i+1}z = G_iz + B_iQ_iz = B_iz = B_{i-1}P_{i-1}z = 0.$$

From (1.11) we learn that a nontrivial intersection $N_{i_*-1} \cap N_{i_*}$ never allows an injective matrix G_i, $i > i_*$. As we will realize later (see Proposition 1.34), such a nontrivial intersection immediately indicates a singular matrix pencil $\lambda E + F$.

Again, we are aiming at a matrix G_κ the rank of which is as high as possible. However, how can we know whether the maximal rank has been reached? Appropriate criteria would be helpful. As we will see later on, for regular DAEs, the sequence terminates with a nonsingular matrix.

Example 1.8 (Sequence for a regular DAE). For the DAE

$$\begin{aligned} x_1' \quad &+x_1+x_2+x_3 = q_1,\\ x_3' \quad &+x_2 \quad\;\; = q_2,\\ x_1 \quad &+x_3 = q_3, \end{aligned}$$

the first matrices of our sequence are

$$G_0 = E = \begin{bmatrix} 1&0&0\\ 0&0&1\\ 0&0&0 \end{bmatrix}, \quad B_0 = F = \begin{bmatrix} 1&1&1\\ 0&1&0\\ 1&0&1 \end{bmatrix}.$$

As a nullspace projector onto $\ker G_0$ we choose

$$Q_0 = \begin{bmatrix} 0&0&0\\ 0&1&0\\ 0&0&0 \end{bmatrix} \text{ and obtain } G_1 = G_0 + B_0Q_0 = \begin{bmatrix} 1&1&0\\ 0&1&1\\ 0&0&0 \end{bmatrix}, \quad B_1 = B_0P_0 = \begin{bmatrix} 1&0&1\\ 0&0&0\\ 1&0&1 \end{bmatrix}.$$

Since G_1 is singular, we turn to the next level. We choose as a projector onto $\ker G_1$

$$Q_1 = \begin{bmatrix} 1&0&0\\ -1&0&0\\ 1&0&0 \end{bmatrix} \text{ and arrive at } G_2 = G_1 + B_1Q_1 = \begin{bmatrix} 3&1&0\\ 0&1&1\\ 2&0&0 \end{bmatrix}.$$

The matrix G_2 is nonsingular, hence the maximal rank is reached and we stop constructing the sequence. Looking at the polynomial $p(\lambda) = \det(\lambda E + F) = 2\lambda$ we know this DAE to be regular. Later on we shall see that a nonsingular matrix G_2 is typical for regularity with Kronecker index 2. Observe further that the nullspaces N_0 and N_1 intersect trivially, and that the projector Q_1 is chosen such that $\Pi_0Q_1Q_0 = 0$ is valid, or equivalently, $N_0 \subseteq \ker \Pi_0Q_1$. □

Example 1.9 (Sequence for a nonregular DAE). We consider the nonregular matrix pair from Example 1.7, that is

$$G_0 = E = \begin{bmatrix} 1&1&0&0\\ 0&0&0&1\\ 0&0&0&0\\ 0&0&1&0 \end{bmatrix}, \; B_0 = F = \begin{bmatrix} 0&1&0&0\\ 0&0&0&0\\ 0&0&1&0\\ 0&0&0&0 \end{bmatrix}.$$

Choosing

$$Q_0 = \begin{bmatrix} 1 & 0\,0\,0 \\ -1 & 0\,0\,0 \\ 0 & 0\,0\,0 \\ 0 & 0\,0\,0 \end{bmatrix} \quad \text{yields} \quad G_1 = \begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,0\,1 \\ 0\,0\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1\,1\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,0 \end{bmatrix}.$$

The matrix G_1 is singular. We turn to the next level. We pick

$$Q_1 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix} \quad \text{which implies} \quad G_2 = G_0.$$

We continue constructing

$$Q_{2j} = Q_0,\ \ G_{2j+1} = G_1,\ \ Q_{2j+1} = Q_1,\ \ G_{2j+2} = G_0,\ \ j \geq 1.$$

Here we have $r_i = 3$ for all $i \geq 0$. The maximal rank is already met by G_0, but there is no criterion which indicates this in time. Furthermore, $N_i \cap N_{i+1} = \{0\}$ holds true for all $i \geq 0$, such that there is no step indicating a singular pencil via property (1.11). Observe that the product $\Pi_0 Q_1 Q_0 = P_0 Q_1 Q_0$ does not vanish as it does in the previous example. □

The rather irritating experience with Example 1.9 leads us to the idea to refine the choice of the projectors by incorporating more information from the previous steps. So far, just the image spaces of the projectors Q_i are prescribed. We refine the construction by prescribing certain appropriate parts of their nullspaces, too. More precisely, we put parts of the previous nullspaces into the current one.

When constructing the sequence (1.10), we now proceed as follows. At any level we decompose

$$N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i, \tag{1.12}$$

where X_i is any complement to $\widehat{N}_i$ in $N_0 + \cdots + N_{i-1}$. We choose Q_i in such a way that the condition

$$X_i \subseteq \ker Q_i \tag{1.13}$$

is met. This is always possible since the subspaces $\widehat{N}_i$ and X_i intersect trivially (see Appendix, Lemma A.7). This restricts to some extent the choice of the projectors. However, a great variety of possible projectors is left. The choice (1.13) implies the projector products Π_i to be projectors again, cf. Proposition 1.13(2). Our structural analysis will significantly benefit from this property. We refer to Chapter 7 for a discussion of practical calculations.

If the intersection $\widehat{N}_i = (N_0 + \cdots + N_{i-1}) \cap N_i$ is trivial, then we have

$$X_i = N_0 + \cdots + N_{i-1} \subseteq \ker Q_i.$$

This is the case in Example 1.8 which shows a regular DAE.

Definition 1.10. For a given matrix pair $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, and an integer $\kappa \in \mathbb{N}$, we call the matrix sequence $G_0,\dots,G_\kappa$ an *admissible matrix sequence*, if it is built by the rule

Set $G_0 := E$, $B_0 := F$, $N_0 := \ker G_0$, and choose a projector $Q_0 \in L(\mathbb{R}^m)$ onto N_0.
For $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1}Q_{i-1},$$
$$B_i := B_{i-1}P_{i-1}$$
$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$
fix a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,
choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,
set $P_i := I - Q_i$, $\Pi_i := \Pi_{i-1}P_i$

The projectors $Q_0,\dots,Q_\kappa$ in an admissible matrix sequence are said to be *admissible*. The matrix sequence $G_0,\dots,G_\kappa$ is said to be *regular admissible*, if additionally,

$$\widehat{N}_i = \{0\}, \quad \forall\, i = 1,\dots,\kappa.$$

Then, also the projectors $Q_0,\dots,Q_\kappa$ are called *regular admissible*.

Admissible projectors are always cross-linked to the matrix function sequence. Changing a projector at a certain level the whole subsequent sequence changes accordingly. Later on we learn that nontrivial intersections $\widehat{N}_i$ indicate a singular matrix pencil.

The projectors in Example 1.8 are admissible but the projectors in Example 1.9 are not. We revisit Example 1.9 and provide admissible projectors.

Example 1.11 (Admissible projectors). Consider once again the singular pair from Examples 1.7 and 1.9. We start the sequence with the same matrices G_0, B_0, Q_0, G_1 as described in Example 1.9 but now we use an admissible projector Q_1. The nullspaces of G_0 and G_1 are given by

$$N_0 = \operatorname{span}\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix} \text{ and } N_1 = \operatorname{span}\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

This allows us to choose

$$Q_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

which satisfies the condition $X_1 \subseteq \ker Q_1$, where $X_1 = N_0$ and $\widehat{N}_1 = N_0 \cap N_1 = \{0\}$. It yields

$$G_2 = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now we find $N_2 = \operatorname{span}\begin{bmatrix}-2 & 1 & 0 & 0\end{bmatrix}^{\mathrm{T}}$ and with

$$N_0 + N_1 = N_0 \oplus N_1 = \operatorname{span}\left(\begin{bmatrix} 1 \\ -1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \right) = \operatorname{span}\left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \right),$$

we have $N_2 \subseteq N_0 + N_1$, $N_0 + N_1 + N_2 = N_0 + N_1$ as well as $\widehat{N}_2 = (N_0 + N_1) \cap N_2 = N_2$. A possible complement X_2 to $\widehat{N}_2$ in $N_0 + N_1$ and an appropriate projector Q_2 are

$$X_2 = \operatorname{span} \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & -2 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

This leads to $G_3 = G_2$, and the nontrivial intersection $N_2 \cap N_3$ indicates (cf. (1.11)) that also all further matrices G_i are singular. Proposition 1.34 below says that this indicates at the same time a singular matrix pencil. In the next steps, for $i \geq 3$, it follows that $N_i = N_2$ and $G_i = G_2$.

For admissible projectors Q_i, not only is their image $\operatorname{im} Q_i = N_i$ fixed, but also a part of $\ker Q_i$. However, there remains a great variety of possible projectors, since, except for the regular case, the subspaces X_i are not uniquely determined and further represent just a part of $\ker Q_i$. Of course, we could restrict the variety of projectors by prescribing special subspaces. For instance, we may exploit orthogonality as much as possible, which is favorable with respect to computational aspects.

Definition 1.12. The admissible projectors $Q_0, \ldots, Q_\kappa$ are called *widely orthogonal* if $Q_0 = Q_0^*$, and

$$X_i = \widehat{N}_i^{\perp} \cap (N_0 + \cdots + N_{i-1}), \tag{1.14}$$

as well as

$$\ker Q_i = [N_0 + \cdots + N_i]^{\perp} \oplus X_i, \quad i = 1, \ldots, \kappa, \tag{1.15}$$

hold true.

The widely orthogonal projectors are completely fixed and they have their advantages. However, in Subsection 2.2.3 we will see that it makes sense to work with sufficiently flexible admissible projectors.

The next assertions collect useful properties of admissible matrix sequences $G_0,\dots,G_\kappa$ and the associated admissible projectors $Q_0,\dots,Q_\kappa$ for a given pair $\{E,F\}$. In particular, the special role of the products $\Pi_i = P_0\cdots P_i$ is revealed. We emphasize this by using mainly the short notation Π_i.

Proposition 1.13. *Let $Q_0,\dots,Q_\kappa$ be admissible projectors for the pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$. Then the following assertions hold true for $i=1,\dots,\kappa$:*

(1) $\ker\Pi_i = N_0+\cdots+N_i$.
(2) *The products $\Pi_i = P_0\cdots P_i$ and $\Pi_{i-1}Q_i = P_0\cdots P_{i-1}Q_i$, are again projectors.*
(3) $N_0+\cdots+N_{i-1}\subseteq\ker\Pi_{i-1}Q_i$.
(4) $B_i = B_i\Pi_{i-1}$.
(5) $\widehat{N}_i\subseteq N_i\cap\ker B_i = N_i\cap N_{i+1}\subseteq\widehat{N}_{i+1}$.
(6) *If $Q_0,\dots,Q_\kappa$ are widely orthogonal, then* $\operatorname{im}\Pi_i = [N_0+\cdots+N_i]^\perp$, $\Pi_i=\Pi_i^*$ *and* $\Pi_{i-1}Q_i = (\Pi_{i-1}Q_i)^*$.
(7) *If $Q_0,\dots,Q_\kappa$ are regular admissible, then* $\ker\Pi_{i-1}Q_i = \ker Q_i$ *and* $Q_iQ_j = 0$ *for* $j=0,\dots,i-1$.

Proof. (1) ($\Rightarrow$) To show $\ker\Pi_i\subseteq N_0+\cdots+N_i$ for $i=1,\dots,\kappa$, we consider an element $z\in\ker\Pi_i$. Then,

$$0=\Pi_i z = P_0\cdots P_i z = \prod_{k=0}^{i}(I-Q_k)z.$$

Expanding the right-hand expression, we obtain

$$z=\sum_{k=0}^{i}Q_kH_kz\ \in N_0+\cdots+N_i$$

with suitable matrices H_k.
($\Leftarrow$) The other direction will be proven by induction. Starting the induction with $i=0$, we observe that $\ker\Pi_0=\ker P_0=N_0$. We suppose that $\ker\Pi_{i-1}=N_0+\cdots+N_{i-1}$ is valid. Because of

$$N_0+\cdots+N_i = X_i+\widehat{N}_i+N_i$$

each $z\in N_0+\cdots+N_i$ can be written as $z=x_i+\bar{z}_i+z_i$ with

$$x_i\in X_i\subseteq N_0+\cdots+N_{i-1}=\ker\Pi_{i-1},\quad \bar{z}_i\in\widehat{N}_i\subseteq N_i,\quad z_i\in N_i.$$

Since Q_i is admissible, we have $X_i\subseteq\ker Q_i$ and $N_i=\operatorname{im}Q_i$. Consequently,

$$\Pi_i z=\Pi_{i-1}(I-Q_i)z=\Pi_{i-1}(I-Q_i)x_i=\Pi_{i-1}x_i=0$$

which implies $N_0+\cdots+N_i\subseteq\ker\Pi_i$ to be true.
(2) From (1) we know that $\operatorname{im}Q_j=N_j\subseteq\ker\Pi_i$ for $j\le i$. It follows that

$$\Pi_i P_j = \Pi_i (I - Q_j) = \Pi_i .$$

Consequently, $\Pi_i^2 = \Pi_i$ and $\Pi_i \Pi_{i-1} = \Pi_i$ imply

$$\begin{aligned}(\Pi_{i-1} Q_i)^2 = \Pi_{i-1}(I - P_i)\Pi_{i-1} Q_i = \Pi_{i-1} Q_i - \Pi_i \Pi_{i-1} Q_i \\ = \Pi_{i-1} Q_i - \Pi_i Q_i = \Pi_{i-1} Q_i .\end{aligned}$$

(3) For any $z \in N_0 + \cdots + N_{i-1}$, we know from (1) that $\Pi_{i-1} z = 0$ and $\Pi_i z = 0$. Thus

$$\Pi_{i-1} Q_i z = \Pi_{i-1} z - \Pi_i z = 0.$$

(4) By construction of B_i (see (1.10)), we find $B_i = B_0 \Pi_{i-1}$. Using (2), we get that

$$B_i = B_0 \Pi_{i-1} = B_0 \Pi_{i-1} \Pi_{i-1} = B_i \Pi_{i-1} .$$

(5) First, we show that $\widehat{N}_i \subseteq N_i \cap \ker B_i$. For $z \in \widehat{N}_i = (N_0 + \cdots + N_{i-1}) \cap N_i$ we find $\Pi_{i-1} z = 0$ from (1) and, hence, $B_i z = B_0 \Pi_{i-1} z = 0$ using (4). Next,

$$N_i \cap \ker B_i = N_i \cap N_{i+1}$$

since $G_{i+1} z = (G_i + B_i Q_i) z = B_i z$ for any $z \in N_i = \operatorname{im} Q_i = \ker G_i$. Finally,

$$\widehat{N}_{i+1} = (N_0 + \cdots + N_i) \cap N_{i+1} \text{ implies immediately that } N_i \cap N_{i+1} \subseteq \widehat{N}_{i+1} .$$

(6) We use induction to show that $\operatorname{im} \Pi_i = [N_0 + \cdots + N_i]^\perp$. Starting with $i = 0$, we know that $\operatorname{im} \Pi_0 = N_0^\perp$ since $Q_0 = Q_0^*$.
Since $X_i \subseteq N_0 + \cdots + N_{i-1}$ (see (1.14)) we derive from (1) that $\Pi_{i-1} X_i = 0$. Regarding (1.15), we find

$$\operatorname{im} \Pi_i = \Pi_{i-1} \operatorname{im} P_i = \Pi_{i-1}([N_0 + \cdots + N_i]^\perp + X_i) = \Pi_{i-1}([N_0 + \cdots + N_i]^\perp).$$

Using $[N_0 + \cdots + N_i]^\perp \subseteq [N_0 + \cdots + N_{i-1}]^\perp = \operatorname{im} \Pi_{i-1}$ we conclude

$$\operatorname{im} \Pi_i = \Pi_{i-1}([N_0 + \cdots + N_i])^\perp = [N_0 + \cdots + N_i]^\perp .$$

In consequence, Π_i is the orthoprojector onto $[N_0 + \cdots + N_i]^\perp$ along $N_0 + \cdots + N_i$, i.e., $\Pi_i = \Pi_i^*$. It follows that

$$\Pi_{i-1} Q_i = \Pi_{i-1} - \Pi_i = \Pi_{i-1}^* - \Pi_i^* = (\Pi_{i-1} - \Pi_i)^* = (\Pi_{i-1} Q_i)^* .$$

(7) Let $\widehat{N}_i = 0$ be valid. Then, $X_i = N_0 + \cdots + N_{i-1} = N_0 \oplus \cdots \oplus N_{i-1}$ and, therefore,

$$\ker \Pi_{i-1} \overset{(1)}{=} N_0 \oplus \cdots \oplus N_{i-1} = X_i \subseteq \ker Q_i .$$

This implies $Q_i Q_j = 0$ for $j = 0, \ldots, i-1$. Furthermore, for any $z \in \ker \Pi_{i-1} Q_i$, we have $Q_i z \in \ker \Pi_{i-1} \subseteq \ker Q_i$, which means that $z \in \ker Q_i$.

□

Remark 1.14. If the projectors $Q_0,\dots,Q_\kappa$ are regular admissible, and the $\Pi_0,\dots,\Pi_\kappa$ are symmetric, then $Q_0,\dots,Q_\kappa$ are widely orthogonal. This is a consequence of the properties

$$\operatorname{im}\Pi_i = (\ker\Pi_i)^\perp = (N_0\oplus\cdots\oplus N_i)^\perp, \quad \ker Q_i = \operatorname{im}\Pi_i \oplus X_i \quad \text{for } i=1,\dots,\kappa.$$

In more general cases, if there are nontrivial intersections $\widehat{N}_i$, widely orthogonal projectors are given, if the Π_i are symmetric and, additionally, the conditions $Q_i\Pi_i = 0$, $P_i(I-\Pi_{i-1}) = (P_i(I-\Pi_{i-1}))^*$ are valid (cf. Chapter 7).

Now we are in a position to provide a result which plays a central role in the projector approach of regular DAEs.

Theorem 1.15. *If, for the matrix pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$, an admissible matrix sequence $(G_i)_{i\ge 0}$ contains an integer μ such that G_μ is nonsingular, then the representations*

$$G_\mu^{-1}E = \Pi_{\mu-1} + (I-\Pi_{\mu-1})G_\mu^{-1}E(I-\Pi_{\mu-1}) \tag{1.16}$$

$$G_\mu^{-1}F = Q_0+\cdots+Q_{\mu-1} + (I-\Pi_{\mu-1})G_\mu^{-1}F\Pi_{\mu-1} + \Pi_{\mu-1}G_\mu^{-1}F\Pi_{\mu-1} \tag{1.17}$$

are valid and $\{E,F\}$ is a regular pair.

Proof. Let G_μ be nonsingular. Owing to Proposition 1.13 we express

$$\begin{aligned} F(I-\Pi_{\mu-1}) &= F(Q_0+\Pi_0Q_1+\cdots+\Pi_{\mu-2}Q_{\mu-1})\\ &= B_0Q_0+B_1Q_1+\cdots+B_{\mu-1}Q_{\mu-1}\\ &= G_\mu Q_0+G_\mu Q_1+\cdots+G_\mu Q_{\mu-1}\\ &= G_\mu(Q_0+Q_1+\cdots+Q_{\mu-1}), \end{aligned}$$

therefore

$$\Pi_{\mu-1}G_\mu^{-1}F(I-\Pi_{\mu-1}) = 0. \tag{1.18}$$

Additionally, we have $G_\mu = E+F(I-\Pi_{\mu-1})$, thus $I = G_\mu^{-1}E+G_\mu^{-1}F(I-\Pi_{\mu-1})$ and $\Pi_{\mu-1} = \Pi_{\mu-1}G_\mu^{-1}E = G_\mu^{-1}E\Pi_{\mu-1}$. From these properties it follows that

$$\Pi_{\mu-1}G_\mu^{-1}E(I-\Pi_{\mu-1}) = 0, \tag{1.19}$$

which proves the expressions (1.16), (1.17).

Denote the finite set consisting of all eigenvalues of the matrix $-\Pi_{\mu-1}G_\mu^{-1}F$ by Λ. We show the matrix $\lambda E+F$ to be nonsingular for each arbitrary λ not belonging to Λ, which proves the matrix pencil to be regular. The equation $(\lambda E+F)z=0$ is equivalent to

$$\begin{aligned} &\lambda G_\mu^{-1}Ez+G_\mu^{-1}Fz = 0 \quad \Longleftrightarrow\\ &\lambda G_\mu^{-1}E\Pi_{\mu-1}z+\lambda G_\mu^{-1}E(I-\Pi_{\mu-1})z+G_\mu^{-1}F\Pi_{\mu-1}z+G_\mu^{-1}F(I-\Pi_{\mu-1})z = 0 \end{aligned} \tag{1.20}$$

Multiplying (1.20) by $\Pi_{\mu-1}$ and regarding (1.18)–(1.19), yields

$$\lambda\Pi_{\mu-1}z+\Pi_{\mu-1}G_\mu^{-1}F\Pi_{\mu-1}z=(\lambda I+\Pi_{\mu-1}G_\mu^{-1}F)\Pi_{\mu-1}z=0,$$

which implies $\Pi_{\mu-1}z=0$ for $\lambda\notin\Lambda$. Using $\Pi_{\mu-1}z=0$, equation (1.20) multiplied by $I-\Pi_{\mu-1}$ reduces to

$$\lambda(I-\Pi_{\mu-1})G_\mu^{-1}E(I-\Pi_{\mu-1})z+(I-\Pi_{\mu-1})G_\mu^{-1}F(I-\Pi_{\mu-1})z=0.$$

Replacing $G_\mu^{-1}E=I-G_\mu^{-1}F(I-\Pi_{\mu-1})$ we find

$$\lambda(I-\Pi_{\mu-1})z+(1-\lambda)(I-\Pi_{\mu-1})G_\mu^{-1}F(I-\Pi_{\mu-1})(I-\Pi_{\mu-1})z=0.$$

If $\lambda=1$ then this immediately implies $z=0$. If $\lambda\neq 1$ it holds that

$$\left(\frac{\lambda}{1-\lambda}I+\underbrace{(I-\Pi_{\mu-1})G_\mu^{-1}F(I-\Pi_{\mu-1})}_{Q_0+\cdots+Q_{\mu-1}}\right)\underbrace{(I-\Pi_{\mu-1})z}_{z}=0.$$

Multiplication by $Q_{\mu-1}$ gives $Q_{\mu-1}z=0$. Then multiplication by $Q_{\mu-2}$ yields $Q_{\mu-2}z=0$, and so on. Finally we obtain $Q_0z=0$ and hence $z=(I-\Pi_{\mu-1})z=Q_0z+\cdots+\Pi_{\mu-2}Q_{\mu-1}z=0$. □

Once more we emphasize that the matrix sequence depends on the choice of the admissible projectors. However, the properties that are important later on are independent of the choice of the projectors, as the following theorem shows.

Theorem 1.16. *For any pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$, the subspaces $N_0+\cdots+N_i$, $\widehat{N}_i$ and $\operatorname{im}G_i$ are independent of the special choice of the involved admissible projectors.*

Proof. All claimed properties are direct and obvious consequences of Lemma 1.18 below. □

Theorem 1.16 justifies the next definition.

Definition 1.17. For each arbitrary matrix pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$, the integers $r_i:=\operatorname{rank}G_i$, $i\geq 0$, $u_i:=\dim\widehat{N}_i$ $i\geq 1$, which arise from an admissible matrix sequence $(G_i)_{i\geq 0}$, are called *structural characteristic values*.

Lemma 1.18. *Let $Q_0,\ldots,Q_\kappa$ and $\bar{Q}_0,\ldots,\bar{Q}_\kappa$ be any two admissible projector sequences for the pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$, and N_j, $\bar{N}_j$, G_j, $\bar{G}_j$, etc. the corresponding subspaces and matrices. Then it holds that:*

(1) $\bar{N}_0+\cdots+\bar{N}_j=N_0+\cdots+N_j$, *for* $j=0,\ldots,\kappa$.

(2) $\bar{G}_j=G_jZ_j$, $\bar{B}_j=B_j+G_j\sum_{l=0}^{j-1}Q_l\mathfrak{A}_{jl}$, *for* $j=0,\ldots,\kappa$,
with nonsingular matrices $Z_0,\ldots,Z_{\kappa+1}$ given by $Z_0:=I$, $Z_{j+1}:=Y_{j+1}Z_j$,

$$Y_1 := I + Q_0(\bar{Q}_0 - Q_0) = I + Q_0\bar{Q}_0P_0,$$
$$Y_{j+1} := I + Q_j(\bar{\Pi}_{j-1}\bar{Q}_j - \Pi_{j-1}Q_j) + \sum_{l=0}^{j-1} Q_l\mathfrak{A}_{jl}\bar{Q}_j,$$

where $\mathfrak{A}_{jl} = \bar{\Pi}_{j-1}$ *for* $l = 0,\dots,j-1$.

(3) $\bar{G}_{\kappa+1} = G_{\kappa+1}Z_{\kappa+1}$ *and* $\bar{N}_0 + \cdots + \bar{N}_{\kappa+1} = N_0 + \cdots + N_{\kappa+1}$.

(4) $(\bar{N}_0 + \cdots + \bar{N}_{j-1}) \cap \bar{N}_j = (N_0 + \cdots + N_{j-1}) \cap N_j$ *for* $j = 1,\dots,\kappa+1$.

Remark 1.19. The introduction of $\mathfrak{A}_{il}$ seems to be unnecessary at this point. We use these extra terms to emphasize the great analogy to the case of DAEs with time-dependent coefficients (see Lemma 2.12). The only difference between both cases is given in the much more elaborate representation of $\mathfrak{A}_{il}$ for time-dependent coefficients.

Proof. We prove (1) and (2) together by induction. For $i = 0$ we have

$$\bar{G}_0 = E = G_0, \quad \bar{B}_0 = F = B_0, \quad \bar{N}_0 = \ker\bar{G}_0 = \ker G_0 = N_0, \quad Z_0 = I.$$

To apply induction we suppose the relations

$$\bar{N}_0 + \cdots + \bar{N}_j = N_0 + \cdots + N_j, \tag{1.21}$$

$$\bar{G}_j = G_jZ_j, \ \bar{B}_j = B_j + G_j\sum_{l=0}^{j-1} Q_l\mathfrak{A}_{jl} \tag{1.22}$$

to be valid for $j \le i$ with nonsingular Z_j as described above, and

$$Z_j^{-1} = I + \sum_{l=0}^{j-1} Q_l\mathfrak{C}_{jl}$$

with certain $\mathfrak{C}_{jl}$. Comparing $\bar{G}_{i+1}$ and G_{i+1} we write

$$\bar{G}_{i+1} = \bar{G}_i + \bar{B}_i\bar{Q}_i = G_iZ_i + \bar{B}_i\bar{Q}_iZ_i + \bar{B}_i\bar{Q}_i(I - Z_i) \tag{1.23}$$

and consider the last term in more detail. We have, due to the form of Y_l, induction assumption (1.21) and $\operatorname{im}(Y_j - I) \subseteq N_0 + \cdots + N_{j-1} = \ker\Pi_{j-1}$ given for all $j \ge 0$ (see Proposition 1.13) that

$$N_0 + \cdots + N_{j-1} \subseteq \ker\Pi_{j-1}Q_j, \quad \bar{N}_0 + \cdots + \bar{N}_{j-1} \subseteq \ker\bar{\Pi}_{j-1}\bar{Q}_j, \quad j \le i, \tag{1.24}$$

and therefore,

$$Y_{j+1} - I = (Y_{j+1} - I)\Pi_{j-1}, \quad j = 1,\dots,i. \tag{1.25}$$

This implies

$$\operatorname{im}(Y_j - I) \subseteq \ker(Y_{j+1} - I), \quad j = 1,\dots,i. \tag{1.26}$$

Concerning $Z_j = Y_jZ_{j-1}$ and using (1.26), a simple induction proof shows

$$Z_j - I = \sum_{l=1}^{j} (Y_l - I), \quad j = 1, \ldots, i,$$

to be satisfied. Consequently,

$$\operatorname{im}(I - Z_i) \subseteq N_0 + \cdots + N_{i-1} = \bar{N}_0 + \cdots + \bar{N}_{i-1} \subseteq \ker \bar{Q}_i.$$

Using (1.23), we get

$$\bar{G}_{i+1} = G_i Z_i + \bar{B}_i \bar{Q}_i Z_i,$$

which leads to

$$\bar{G}_{i+1} Z_i^{-1} = G_i + \bar{B}_i \bar{Q}_i = G_i + B_i Q_i + (\bar{B}_i \bar{Q}_i - B_i Q_i).$$

We apply the induction assumption (1.22) to find

$$\bar{G}_{i+1} Z_i^{-1} = G_{i+1} + B_i (\bar{Q}_i - Q_i) + G_i \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i.$$

Induction assumption (1.21) and Proposition 1.13 imply $\ker \bar{\Pi}_{i-1} = \ker \Pi_{i-1}$ and hence

$$B_i = B_0 \Pi_{i-1} = B_0 \Pi_{i-1} \bar{\Pi}_{i-1} = B_i \bar{\Pi}_{i-1}.$$

Finally,

$$\begin{aligned} \bar{G}_{i+1} Z_i^{-1} &= G_{i+1} + B_i (\bar{\Pi}_{i-1} \bar{Q}_i - \Pi_{i-1} Q_i) + G_{i+1} \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i \\ &= G_{i+1} + B_i Q_i (\bar{\Pi}_{i-1} \bar{Q}_i - \Pi_{i-1} Q_i) + G_{i+1} \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i \;=\; G_{i+1} Y_{i+1}, \end{aligned}$$

which means that

$$\bar{G}_{i+1} = G_{i+1} Y_{i+1} Z_i = G_{i+1} Z_{i+1}. \tag{1.27}$$

Next, we will show Z_{i+1} to be nonsingular. Owing to the induction assumption, we know that Z_i is nonsingular. Considering the definition of Z_{i+1} we have to show Y_{i+1} to be nonsingular. Firstly,

$$\Pi_i Y_{i+1} = \Pi_i \tag{1.28}$$

since $\operatorname{im} Q_j \subseteq \ker \Pi_i$ for $j \leq i$. This follows immediately from the definition of Y_{i+1} and Proposition 1.13 (1). Using the induction assumption (1.21), Proposition 1.13 and Lemma A.3, we find

$$\Pi_j \bar{\Pi}_j = \Pi_j, \quad \bar{\Pi}_j \Pi_j = \bar{\Pi}_j \quad \text{and} \quad \Pi_j \Pi_j = \Pi_j \qquad \text{for } j = 0, \ldots, i.$$

This implies that

$$\Pi_{i-1} (Y_{i+1} - I) = \Pi_{i-1} (Y_{i+1} - I) \Pi_i \tag{1.29}$$

because

$$\begin{aligned}\Pi_{i-1}(Y_{i+1}-I) &\overset{\text{Prop. 1.13(1)}}{=} \Pi_{i-1}Q_i(\bar{\Pi}_{i-1}\bar{Q}_i-\Pi_{i-1}Q_i)\\ &= (\Pi_{i-1}-\Pi_i)(\bar{\Pi}_{i-1}\bar{Q}_i-\Pi_{i-1}Q_i)\\ &= \Pi_{i-1}(\bar{Q}_i-Q_i) \quad = \Pi_{i-1}(P_i-\bar{P}_i)\\ &= \Pi_i-\Pi_{i-1}\bar{\Pi}_{i-1}\bar{P}_i \;= \Pi_i-\Pi_{i-1}\bar{\Pi}_i\\ &= \Pi_i-\Pi_{i-1}\bar{\Pi}_i\Pi_i \quad = (I-\Pi_{i-1}\bar{\Pi}_i)\Pi_i.\end{aligned}$$

Equations (1.28) and (1.29) imply

$$\Pi_{i-1}(Y_{i+1}-I) = \Pi_{i-1}(Y_{i+1}-I)\Pi_i = \Pi_{i-1}(Y_{i+1}-I)\Pi_i Y_{i+1}$$

and, consequently,

$$\begin{aligned}I &= Y_{i+1}-(Y_{i+1}-I) \overset{(1.25)}{=} Y_{i+1}-(Y_{i+1}-I)\Pi_{i-1}\\ &= Y_{i+1}-(Y_{i+1}-I)\Pi_{i-1}\{(I-\Pi_{i-1})Y_{i+1}+\Pi_{i-1}\}\\ &= Y_{i+1}-(Y_{i+1}-I)\Pi_{i-1}\{Y_{i+1}-\Pi_{i-1}(Y_{i+1}-I)\}\\ &= Y_{i+1}-(Y_{i+1}-I)\Pi_{i-1}\{Y_{i+1}-\Pi_{i-1}(Y_{i+1}-I)\Pi_i Y_{i+1}\}\\ &= (I-(Y_{i+1}-I)\{I-\Pi_{i-1}(Y_{i+1}-I)\Pi_i\})Y_{i+1}.\end{aligned}$$

This means that Y_{i+1} is nonsingular and

$$Y_{i+1}^{-1} = I-(Y_{i+1}-I)\{I-\Pi_{i-1}(Y_{i+1}-I)\Pi_i\}.$$

Then also $Z_{i+1} = Y_{i+1}Z_i$ is nonsingular, and

$$Z_{i+1}^{-1} = Z_i^{-1}Y_{i+1}^{-1} = (I+\sum_{l=0}^{i-1} Q_l\mathfrak{C}_{il})Y_{i+1}^{-1} = I+\sum_{l=0}^{i} Q_l\mathfrak{C}_{i+1\,l}$$

with certain coefficients $\mathfrak{C}_{i+1\,l}$. From (1.27) we conclude $\bar{N}_{i+1} = Z_{i+1}^{-1}N_{i+1}$, and, due to the special form of Z_{i+1}^{-1},

$$\bar{N}_{i+1} \subseteq N_0+\cdots+N_{i+1},\ \bar{N}_0+\cdots+\bar{N}_{i+1} \subseteq N_0+\cdots+N_{i+1}.$$

Owing to the property $\operatorname{im}(Z_{i+1}-I) \subseteq N_0+\cdots+N_i = \bar{N}_0+\cdots+\bar{N}_i$, it holds that

$$N_{i+1} = Z_{i+1}\bar{N}_{i+1} = (I+(Z_{i+1}-I))\bar{N}_{i+1} \subseteq \bar{N}_0+\cdots+\bar{N}_{i+1}.$$

Thus, $N_0+\cdots+N_{i+1} \subseteq \bar{N}_0+\cdots+\bar{N}_{i+1}$ is valid. For symmetry reasons we have

$$N_0+\cdots+N_{i+1} = \bar{N}_0+\cdots+\bar{N}_{i+1}.$$

Finally, we derive from the induction assumption that

$$\begin{aligned}\bar{B}_{i+1} &= \bar{B}_i\bar{P}_i = \big(B_i + G_i\sum_{l=0}^{i-1} Q_l\mathfrak{A}_{il}\big)\bar{P}_i \\ &= B_iP_i\bar{P}_i + B_iQ_i\bar{P}_i + G_{i+1}\sum_{l=0}^{i-1} Q_l\mathfrak{A}_{il}\bar{P}_i \\ &= B_iP_i + B_iQ_i\bar{\Pi}_i + G_{i+1}\sum_{l=0}^{i-1} Q_l\mathfrak{A}_{il}\bar{P}_i \;=\; B_{i+1} + G_{i+1}\sum_{l=0}^{i} Q_l\mathfrak{A}_{i+1,l}\end{aligned}$$

with $\mathfrak{A}_{i+1,l} = \mathfrak{A}_{il}\bar{P}_i$, $l = 0,\ldots,i-1$, $\mathfrak{A}_{i+1,i} = \bar{\Pi}_i$, and therefore, for $l \le i-1$,

$$\mathfrak{A}_{i+1,l} = \mathfrak{A}_{il}\bar{P}_i = \mathfrak{A}_{i-1,l}\bar{P}_{i-1}\bar{P}_i = \mathfrak{A}_{l+1,l}\bar{P}_{l+1}\cdots\bar{P}_i = \bar{\Pi}_l\bar{P}_{l+1}\cdots\bar{P}_i = \bar{\Pi}_i.$$

We have proved assertions (1) and (2), and (3) is a simple consequence. Next we prove assertion (4). By assertion (1) from Lemma 1.13, we have $N_0 + \cdots + N_i = \ker \Pi_i$ and

$$\begin{aligned}G_{i+1} &= G_0 + B_0Q_0 + \cdots + B_iQ_i = G_0 + B_0Q_0 + B_1P_0Q_1 + \cdots + B_iP_0\cdots P_{i-1}Q_i \\ &= G_0 + B_0(Q_0 + P_0Q_1 + \cdots + P_0\cdots P_{i-1}Q_i) \\ &= G_0 + B_0(I - P_0\cdots P_i) = G_0 + B_0(I - \Pi_i).\end{aligned}$$

This leads to the description

$$\begin{aligned}\widehat{N}_{i+1} = (N_0 + \cdots + N_i)\cap N_{i+1} &= \{z \in \mathbb{R}^m : \Pi_i z = 0,\ G_0z + B_0(I - \Pi_i)z = 0\} \\ &= \{z \in \mathbb{R}^m : z \in N_0 + \cdots + N_i,\ G_0z + B_0z = 0\} \\ &= \{z \in \mathbb{R}^m : z \in \bar{N}_0 + \cdots + \bar{N}_i,\ \bar{G}_0z + \bar{B}_0z = 0\} \\ &= (\bar{N}_0 + \cdots + \bar{N}_i)\cap\bar{N}_{i+1}.\end{aligned}$$

□

1.2.2 Decoupling by admissible projectors

In this subsection we deal with matrix pairs $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, the admissible matrix sequence $(G_i)_{i\ge0}$ of which reaches a nonsingular matrix G_μ. Those matrix pairs as well as the associated DAEs

$$Ex'(t) + Fx(t) = q(t) \tag{1.30}$$

are regular by Theorem 1.15. They have the structural characteristic values

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m.$$

The nonsingular matrix G_μ allows for a projector based decoupling such that the decoupled version of the given DAE looks quite similar to the Weierstraß–Kronecker form.

We stress that, at the same time, our discussion should serve as a model for a corresponding decoupling of time-dependent linear DAEs for which we do not have a Weierstraß–Kronecker form.

When constructing an admissible matrix function sequence $(G_i)_{i\geq 0}$ we have in mind a *rearrangement of terms within the original DAE* (1.30) such that the solution components $\Pi_{\mu-1}x(t)$ and $(I-\Pi_{\mu-1})x(t)$ are separated as far as possible and the nonsingular matrix G_μ occurs in front of the derivative $(\Pi_{\mu-1}x(t))'$. Let the admissible matrix sequence (Definition 1.10) starting from $G_0=E$, $B_0=F$ be realized up to G_μ which is nonsingular. Let $\mu\in\mathbb{N}$ be the smallest such index.

Consider the involved admissible projectors $Q_0,\dots,Q_\mu$. We have $Q_\mu=0$, $P_\mu=I$, $\Pi_\mu=\Pi_{\mu-1}$ for trivial reasons. Due to Proposition 1.13, the intersections $\widehat{N}_i$ are trivial,

$$\widehat{N}_i = N_i\cap(N_0+\cdots+N_{i-1})=\{0\},\ \ i=1,\dots,\mu-1,$$

and therefore

$$N_0+\cdots+N_{i-1}=N_0\oplus\cdots\oplus N_{i-1},\quad X_i=N_0\oplus\cdots\oplus N_{i-1},\quad i=1,\dots,\mu-1. \tag{1.31}$$

From (1.31) we derive the relations

$$Q_iQ_j=0,\quad j=0,\dots,i-1,\quad i=1,\dots,\mu-1, \tag{1.32}$$

which are very helpful in computations. Recall the properties

$$\begin{aligned} G_iP_{i-1} &= G_{i-1},\quad B_i=B_i\Pi_{i-1},\ \ i=1,\dots,\mu,\\ G_iQ_j &= B_jQ_j,\ \ j=0,\dots,i-1,\quad i=0,\dots,\mu-1,\end{aligned}$$

from Section 1.2 which will be used frequently.

Applying $G_0=G_0P_0=G_0\Pi_0$ we rewrite the DAE (1.30) as

$$G_0(\Pi_0x(t))'+B_0x(t)=q(t), \tag{1.33}$$

and then, with $B_0=B_0P_0+B_0Q_0=B_0\Pi_0+G_1Q_0$, as

$$G_1P_1P_0(\Pi_0x(t))'+B_0\Pi_0x(t)+G_1Q_0x(t)=q(t).$$

Now we use the relation

$$\begin{aligned} G_1P_1P_0 &= G_1\Pi_0P_1P_0+G_1(I-\Pi_0)P_1P_0\\ &= G_1\Pi_1-G_1(I-\Pi_0)Q_1\\ &= G_1\Pi_1-G_1(I-\Pi_0)Q_1\Pi_0Q_1\end{aligned}$$

to replace the first term. This yields

$$G_1(\Pi_1x(t))'+B_1x(t)+G_1\{Q_0x(t)-(I-\Pi_0)Q_1(\Pi_0Q_1x(t))'\}=q(t).$$

Proceeding further by induction we suppose

$$G_i(\Pi_i x(t))' + B_i x(t) + G_i \sum_{l=0}^{i-1} \{Q_l x(t) - (I-\Pi_l)Q_{l+1}(\Pi_l Q_{l+1} x(t))'\} = q(t) \quad (1.34)$$

and, in the next step, using the properties $G_{i+1}P_{i+1}P_i = G_i$, $B_i Q_i = G_{i+1}Q_i$, $G_i Q_l = G_{i+1}Q_l$, $l = 0, \ldots, i-1$, and

$$\begin{aligned} P_{i+1}P_i\Pi_i &= \Pi_i P_{i+1}P_i\Pi_i + (I-\Pi_i)P_{i+1}P_i\Pi_i \\ &= \Pi_{i+1} - (I-\Pi_i)Q_{i+1} \\ &= \Pi_{i+1} - (I-\Pi_i)Q_{i+1}\Pi_i Q_{i+1}, \end{aligned}$$

we reach

$$G_{i+1}(\Pi_{i+1}x(t))' + B_{i+1}x(t) + G_{i+1}\sum_{l=0}^{i}\{Q_l x(t) - (I-\Pi_l)Q_{l+1}(\Pi_l Q_{l+1}x(t))'\} = q(t),$$

so that expression (1.34) can be used for all $i = 1, \ldots, \mu$. In particular, we obtain

$$G_\mu(\Pi_\mu x(t))' + B_\mu x(t) + G_\mu \sum_{l=0}^{\mu-1}\{Q_l x(t) - (I-\Pi_l)Q_{l+1}(\Pi_l Q_{l+1}x(t))'\} = q(t). \quad (1.35)$$

Taking into account that $Q_\mu = 0$, $P_\mu = I$, $\Pi_\mu = \Pi_{\mu-1}$, and scaling with G_μ^{-1} we derive the equation

$$(\Pi_{\mu-1}x(t))' + G_\mu^{-1}B_\mu x(t) + \sum_{l=0}^{\mu-1} Q_l x(t) - \sum_{l=0}^{\mu-2}(I-\Pi_l)Q_{l+1}(\Pi_l Q_{l+1}x(t))' = G_\mu^{-1}q(t). \quad (1.36)$$

In turn, equation (1.36) can be decoupled into two parts, the explicit ODE with respect to $\Pi_{\mu-1}x(t)$,

$$(\Pi_{\mu-1}x(t))' + \Pi_{\mu-1}G_\mu^{-1}B_\mu x(t) = \Pi_{\mu-1}G_\mu^{-1}q(t), \quad (1.37)$$

and the remaining equation

$$(I-\Pi_{\mu-1})G_\mu^{-1}B_\mu x(t) + \sum_{l=0}^{\mu-1} Q_l x(t) - \sum_{l=0}^{\mu-2}(I-\Pi_l)Q_{l+1}(\Pi_l Q_{l+1}x(t))' = (I-\Pi_{\mu-1})G_\mu^{-1}q(t). \quad (1.38)$$

Next, we show that equation (1.38) uniquely defines the component $(I-\Pi_{\mu-1})x(t)$ in terms of $\Pi_{\mu-1}x(t)$. We decouple equation (1.38) once again into μ further parts according to the decomposition

$$I-\Pi_{\mu-1} = Q_0P_1\cdots P_{\mu-1} + Q_1P_2\cdots P_{\mu-1} + \cdots + Q_{\mu-2}P_{\mu-1} + Q_{\mu-1}. \tag{1.39}$$

Notice that $Q_iP_{i+1}\cdots P_{\mu-1}$, $i=0,\ldots,\mu-2$ are projectors, too, and

$$\begin{aligned}
Q_iP_{i+1}\cdots P_{\mu-1}Q_i &= Q_i,\\
Q_iP_{i+1}\cdots P_{\mu-1}Q_j &= 0, \quad \text{if } i\neq j,\\
Q_iP_{i+1}\cdots P_{\mu-1}(I-\Pi_l)Q_{l+1} &= Q_i(I-\Pi_l)Q_{l+1}=0, \quad \text{for } l=0,\ldots,i-1,\\
Q_iP_{i+1}\cdots P_{\mu-1}(I-\Pi_i)Q_{i+1} &= Q_iQ_{i+1}.
\end{aligned}$$

Hence, multiplying (1.38) by $Q_iP_{i+1}\cdots P_{\mu-1}$, $i=0,\ldots,\mu-2$, and $Q_{\mu-1}$ yields

$$\begin{aligned}
&Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}B_\mu x(t) + Q_ix(t) - Q_iQ_{i+1}(\Pi_iQ_{i+1}x(t))'\\
&\quad - \sum_{l=i+1}^{\mu-2} Q_iP_{i+1}\cdots P_lQ_{l+1}(\Pi_lQ_{l+1}x(t))' = Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}q(t)
\end{aligned} \tag{1.40}$$

for $i=0,\ldots,\mu-2$ and

$$Q_{\mu-1}G_\mu^{-1}B_\mu x(t) + Q_{\mu-1}x(t) = Q_{\mu-1}G_\mu^{-1}q(t). \tag{1.41}$$

Equation (1.41) uniquely determines the component $Q_{\mu-1}x(t)$ as

$$Q_{\mu-1}x(t) = Q_{\mu-1}G_\mu^{-1}q(t) - Q_{\mu-1}G_\mu^{-1}B_\mu x(t),$$

and the formula contained in (1.40) for $i=\mu-2$ gives

$$\begin{aligned}
&Q_{\mu-2}x(t) =\\
&\quad Q_{\mu-2}P_{\mu-1}G_\mu^{-1}q(t) - Q_{\mu-2}P_{\mu-1}G_\mu^{-1}B_\mu x(t) - Q_{\mu-2}Q_{\mu-1}(\Pi_{\mu-2}Q_{\mu-1}x(t))',
\end{aligned}$$

and so on, i.e., in a consecutive manner we obtain expressions determining the components $Q_ix(t)$ with their dependence on $\Pi_{\mu-1}x(t)$ and $Q_{i+j}x(t)$, $j=1,\ldots,\mu-1-i$.

To compose an expression for the whole solution $x(t)$ there is no need for the components $Q_ix(t)$ themselves, $i=0,\ldots,\mu-1$. But one can do it with $Q_0x(t)$, $\Pi_{i-1}Q_ix(t)$, $i=1,\ldots,\mu-1$, which corresponds to the decomposition

$$I = Q_0 + \Pi_0Q_1 + \cdots + \Pi_{\mu-2}Q_{\mu-1} + \Pi_{\mu-1}. \tag{1.42}$$

For this purpose we rearrange the system (1.40), (1.41) once again by multiplying (1.41) by $\Pi_{\mu-2}$ and (1.40) for $i=1,\ldots,\mu-2$ by Π_{i-1}. Let us remark that, even though we scale with projectors (which are singular matrices) here, nothing of the equations gets lost. This is due to the relations

$$\begin{aligned} Q_i &= Q_i\Pi_{i-1}Q_i = (\Pi_{i-1} + (I-\Pi_{i-1}))Q_i\Pi_{i-1}Q_i \\ &= (I + (I-\Pi_{i-1})Q_i)\Pi_{i-1}Q_i, \\ \Pi_{i-1}Q_i &= (I-(I-\Pi_{i-1})Q_i)Q_i, \end{aligned} \tag{1.43}$$

which allow a one-to-one translation of the components $Q_i x(t)$ and $\Pi_{i-1}Q_i x(t)$ into each other. Choosing notation according to the decomposition (1.42),

$$v_0(t) := Q_0 x(t),\ \ v_i(t) := \Pi_{i-1}Q_i x(t),\ \ i = 1,\dots,\mu-1,\ \ u(t) := \Pi_{i-1}x(t), \tag{1.44}$$

we obtain the representation, respectively decomposition

$$x(t) = v_0(t) + v_1(t) + \cdots + v_{\mu-1}(t) + u(t) \tag{1.45}$$

of the solution as well as the structured system resulting from (1.37), (1.40), and (1.41):

$$\begin{aligned} &\left[\begin{array}{c|ccc} I & & & \\ \hline 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{array}\right] \left[\begin{array}{c} u'(t) \\ \hline 0 \\ v_1'(t) \\ \vdots \\ v_{\mu-1}'(t) \end{array}\right] \\ &\quad + \left[\begin{array}{c|ccc} \mathcal{W} & & & \\ \hline \mathcal{H}_0 & I & & \\ \vdots & & \ddots & \\ \vdots & & & I \\ \mathcal{H}_{\mu-1} & & & \end{array}\right] \left[\begin{array}{c} u(t) \\ \hline v_0(t) \\ \vdots \\ \vdots \\ v_{\mu-1}(t) \end{array}\right] = \left[\begin{array}{c} \mathcal{L}_d \\ \mathcal{L}_0 \\ \vdots \\ \vdots \\ \mathcal{L}_{\mu-1} \end{array}\right] q(t) \end{aligned} \tag{1.46}$$

with the $m \times m$ blocks

$$\begin{aligned} \mathcal{N}_{01} &:= -Q_0Q_1, \\ \mathcal{N}_{0j} &:= Q_0P_1\cdots P_{j-1}Q_j, && j = 2,\dots,\mu-1, \\ \mathcal{N}_{i,i+1} &:= -\Pi_{i-1}Q_iQ_{i+1}, && i = 1,\dots,\mu-2, \\ \mathcal{N}_{ij} &:= -\Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_j, && j = i+2,\dots,\mu-1,\ i = 1,\dots,\mu-2, \end{aligned}$$

$$\begin{aligned} \mathcal{W} &:= \Pi_{\mu-1}G_\mu^{-1}B_\mu, \\ \mathcal{H}_0 &:= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}B_\mu, \\ \mathcal{H}_i &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}B_\mu, && i = 1,\dots,\mu-2, \\ \mathcal{H}_{\mu-1} &:= \Pi_{\mu-2}Q_{\mu-1}G_\mu^{-1}B_\mu, \end{aligned}$$

and

$$\begin{aligned}
\mathcal{L}_d &:= \Pi_{\mu-1} G_\mu^{-1}, \\
\mathcal{L}_0 &:= Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1}, \qquad i = 1, \ldots, \mu-2, \\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1}.
\end{aligned}$$

System (1.46) almost looks like a DAE in Weierstraß–Kronecker form. However, compared to the latter it is a puffed up system of dimension $(\mu+1)m$. The system (1.46) is equivalent to the original DAE (1.30) in the following sense.

Proposition 1.20. *Let the DAE (1.30), with coefficients $E, F \in L(\mathbb{R}^m)$, have the characteristic values*

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m.$$

(1) *If $x(.)$ is a solution of the DAE (1.30), then the components $u(.), v_0(.), \ldots, v_{\mu-1}(.)$ given by (1.44) form a solution of the puffed up system (1.46).*
(2) *Conversely, if the functions $u(.), v_0(.), \ldots, v_{\mu-1}(.)$ are a solution of the system (1.46) and if, additionally, $u(t_0) = \Pi_{\mu-1} u(t_0)$ holds for a $t_0 \in \mathcal{I}$, then the compound function $x(.)$ defined by (1.45) is a solution of the original DAE (1.30).*

Proof. It remains to verify (2). Due to the properties of the coefficients, for each solution of system (1.46) it holds that $v_i(t) = \Pi_{i-1} Q_i v_i(t)$, $i = 1, \ldots, \mu-1$, $v_0(t) = Q_0 v_0(t)$, which means that the components $v_i(t)$, $i = 0, \ldots, \mu-1$, belong to the desired subspaces.

The first equation in (1.46) is the explicit ODE $u'(t) + \mathcal{W} u(t) = \mathcal{L}_d q(t)$. Let $u_q(.)$ denote the solution fixed by the initial condition $u_q(t_0) = 0$. We have $u_q(t) = \Pi_{\mu-1} u_q(t)$ because of $\mathcal{W} = \Pi_{\mu-1}\mathcal{W}$, $\mathcal{L}_d = \Pi_{\mu-1}\mathcal{L}_d$. However, for each arbitrary constant $c \in \operatorname{im}(I - \Pi_{\mu-1})$, the function $\bar{u}(.) := c + u_q(.)$ solves this ODE but does not belong to $\operatorname{im}\Pi_{\mu-1}$ as we want it to.

With the initial condition $u(t_0) = u_0 \in \operatorname{im}\Pi_{\mu-1}$ the solution can be kept in the desired subspace, which means that $u(t) \in \operatorname{im}\Pi_{\mu-1}$ for all $t \in \mathcal{I}$. Now, by carrying out the decoupling procedure in reverse order and putting things together we have finished the proof. □

System (1.46) is *given in terms of the original DAE*. It shows in some detail the inherent structure of that DAE. It also serves as the idea of an analogous decoupling of time-varying linear DAEs (see Section 2.6).

Example 1.21 (Decoupling of an index-2 DAE). We reconsider the regular index-2 DAE

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x' + \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} x = q$$

from Example 1.8, with the projectors

$$\Pi_1 = P_0P_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; P_0Q_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

The DAE itself can be rewritten without any differentiations of equations as

$$(-x_1 + x_3)' = q_2 + q_3 - q_1, \tag{1.47}$$

$$x_1' + x_2 = (q_1 - q_3), \tag{1.48}$$

$$x_1 + \frac{1}{2}(-x_1 + x_3) = \frac{1}{2}q_3. \tag{1.49}$$

Obviously, $\Pi_1 x$ reflects the proper state variable $-x_1 + x_3$, for which an explicit ODE (1.47) is given. P_0Q_1x refers to the variable x_1 that is described by the algebraic equation (1.49) when the solution $-x_1 + x_3$ is already given by (1.47). Finally, Q_0x reflects the variable x_2 which can be determined by (1.48). Note, that the variable x_1 has to be differentiated here. Simple calculations yield $\mathcal{W} = \Pi_{\mu-1}G_2^{-1}B_0\Pi_{\mu-1} = 0$, $\mathcal{H}_0 = Q_0P_1G_2^{-1}B_0\Pi_{\mu-1} = 0$ and

$$\mathcal{H}_1 = Q_1G_2^{-1}B_0\Pi_{\mu-1} = \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

This way the DAE decouples as

$$(\Pi_1 x)' = \Pi_1 G_2^{-1} q, \tag{1.50}$$

$$-Q_0Q_1(\Pi_0Q_1x)' + Q_0x = Q_0P_1G_2^{-1}q, \tag{1.51}$$

$$\Pi_0Q_1 + \mathcal{H}_1\Pi_1 x = \Pi_0Q_1G_2^{-1}q. \tag{1.52}$$

These equations mean in full detail

$$\left(\begin{bmatrix} 0 \\ 0 \\ -x_1 + x_3 \end{bmatrix} \right)' = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 1 & 1 \end{bmatrix} q,$$

$$\begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} x_1 \\ 0 \\ x_1 \end{bmatrix} \right)' + \begin{bmatrix} 0 \\ x_2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{bmatrix} q,$$

$$\begin{bmatrix} x_1 \\ 0 \\ x_1 \end{bmatrix} + \begin{bmatrix} -\frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ -x_1 + x_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} q.$$

Dropping the redundant equations as well as all zero lines one arrives exactly at the compressed form (1.47)–(1.49). □

1.2.3 Complete decoupling

A special smart choice of the admissible projectors cancels the coefficients $\mathcal{H}_i$ in system (1.46) so that the second part no longer depends on the first part.

Theorem 1.22. *Let $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, be a pair with characteristic values*

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m.$$

Then there are admissible projectors $Q_0, \dots Q_{\mu-1}$ such that the coupling coefficients $\mathcal{H}_0, \dots, \mathcal{H}_{\mu-1}$ in (1.46) vanish, that is, (1.46) decouples into two independent subsystems.

Proof. For any given sequence of admissible projectors $Q_0, \dots, Q_{\mu-1}$ the coupling coefficients can be expressed as $\mathcal{H}_0 = Q_{0*}\Pi_{\mu-1}$ and $\mathcal{H}_i = \Pi_{i-1}Q_{i*}\Pi_{\mu-1}$ for $i = 1, \dots, \mu-1$, where we denote

$$\begin{aligned} Q_{0*} &:= Q_0P_1 \cdots P_{\mu-1}G_\mu^{-1}B_0, \\ Q_{i*} &:= Q_iP_{i+1} \cdots P_{\mu-1}G_\mu^{-1}B_0\Pi_{i-1}, \qquad i = 1, \dots, \mu-2, \\ Q_{\mu-1*} &:= Q_{\mu-1}G_\mu^{-1}B_0\Pi_{\mu-2}. \end{aligned}$$

We realize that $Q_{i*}Q_i = Q_i$, $i = 0, \dots, \mu-1$, since

$$\begin{aligned} Q_{\mu-1*}Q_{\mu-1} = Q_{\mu-1}G_\mu^{-1}B_0\Pi_{\mu-2}Q_{\mu-1} &= Q_{\mu-1}G_\mu^{-1}B_{\mu-1}Q_{\mu-1} \\ &= Q_{\mu-1}G_\mu^{-1}G_\mu Q_{\mu-1} = Q_{\mu-1}, \end{aligned}$$

and so on for $i = \mu-2, \dots, 0$. This implies $(Q_{i*})^2 = Q_{i*}$, i.e., Q_{i*} is a projector onto N_i, $i = 0, \dots, \mu-1$. By construction one has $N_0 + \cdots + N_{i-1} \subseteq \ker Q_{i*}$ for $i = 1, \dots, \mu-1$. The new projectors $\bar{Q}_0 := Q_0, \dots, \bar{Q}_{\mu-2} := Q_{\mu-2}$, $\bar{Q}_{\mu-1} := Q_{\mu-1*}$ are also admissible, but now, the respective coefficient $\bar{\mathcal{H}}_{\mu-1}$ disappears in (1.46). Namely, the old and new sequences are related by

$$\bar{G}_i = G_i, \;\; i = 0, \dots, \mu-1, \;\; \bar{G}_\mu = G_\mu + B_{\mu-1}Q_{\mu-1*} = G_\mu Z_\mu$$

with nonsingular $Z_\mu := I + Q_{\mu^-1}Q_{\mu-1*}P_{\mu-1}$. This yields

$$\begin{aligned} \bar{Q}_{\mu-1*} := \bar{Q}_{\mu-1}\bar{G}_{\mu-1}B_0\Pi_{\mu-2} &= Q_{\mu-1*}Z_\mu^{-1}G_\mu^{-1}B_0\Pi_{\mu-2} \\ &= Q_{\mu-1}G_\mu^{-1}B_0\Pi_{\mu-2} = Q_{\mu-1*} = \bar{Q}_{\mu-1} \end{aligned}$$

because of

$$Q_{\mu-1*}Z_\mu^{-1} = Q_{\mu-1*}(I - Q_{\mu-1}Q_{\mu-1*}P_{\mu-1}) = Q_{\mu-1},$$

and hence

$$\bar{\mathcal{H}}_{\mu-1} := \bar{\Pi}_{\mu-2}\bar{Q}_{\mu-1*}\bar{\Pi}_{\mu-1} = \Pi_{\mu-2}\bar{Q}_{\mu-1}\bar{\Pi}_{\mu-1} = 0.$$

We show by induction that the coupling coefficients disappear stepwise with an appropriate choice of admissible projectors. Assume $Q_0, \dots, Q_{\mu-1}$ to be such that

$$\mathcal{H}_{k+1} = 0, \ldots, \mathcal{H}_{\mu-1} = 0, \tag{1.53}$$

or, equivalently,

$$Q_{k+1*}\Pi_{\mu-1} = 0, \ldots, Q_{\mu-1*}\Pi_{\mu-1} = 0,$$

for a certain k, $0 \le k \le \mu - 2$. We build a new sequence by letting $\bar{Q}_i := Q_i$ for $i = 0, \ldots, k-1$ (if $k \ge 1$) and $\bar{Q}_k := Q_{k*}$. Thus, $Q_k \bar{P}_k = -\bar{Q}_k P_k$ and the projectors $\bar{Q}_0, \ldots, \bar{Q}_k$ are admissible. The resulting two sequences are related by

$$\bar{G}_i = G_i Z_i, \qquad i = 0, \ldots, k+1,$$

with factors

$$Z_0 = I, \quad \ldots, \quad Z_k = I, \quad Z_{k+1} = I + Q_k Q_{k*} P_k, \quad Z_{k+1}^{-1} = I - Q_k Q_{k*} P_k.$$

We form $\bar{Q}_{k+1} := Z_{k+1}^{-1} Q_{k+1} Z_{k+1} = Z_{k+1}^{-1} Q_{k+1}$. Then, $\bar{Q}_0, \ldots, \bar{Q}_{k+1}$ are also admissible. Applying Lemma 1.18 we proceed with

$$\bar{G}_j = G_j Z_j, \quad \bar{Q}_j := Z_j^{-1} Q_j Z_j, \quad j = k+2, \ldots, \mu - 1,$$

and arrive at a new sequence of admissible projectors $\bar{Q}_0, \ldots, \bar{Q}_{\mu-1}$. The invertibility of Z_j is ensured by Lemma 1.18. Putting $Y_{k+1} := Z_{k+1}$ and, exploiting Lemma 1.18,

$$Y_j := Z_j Z_{j-1}^{-1} = I + Q_{j-1}(\bar{\Pi}_{j-2}\bar{Q}_{j-1} - \Pi_{j-2}Q_{j-1}) + \sum_{l=0}^{j-2} Q_l \bar{\Pi}_{j-2}\bar{Q}_{j-1}, \quad j \ge k+2.$$

Additionally, we learn from Lemma 1.18 that the subspaces $N_0 \oplus \cdots \oplus N_j$ and $\bar{N}_0 \oplus \cdots \oplus \bar{N}_j$ coincide. The expression for Y_j, $j \ge k+2$, simplifies to

$$Y_j = I + \sum_{l=0}^{j-2} Q_l \bar{\Pi}_{j-2}\bar{Q}_{j-1} = I + \sum_{l=k}^{j-2} Q_l \bar{\Pi}_{j-2} Q_{j-1}$$

for our special new projectors because the following relations are valid:

$$\begin{aligned} &Q_j Z_j = 0, \quad \bar{Q}_j = Z_j^{-1} Q_j, \quad \bar{\Pi}_{j-2}\bar{Q}_{j-1} = \bar{\Pi}_{j-2} Z_{j-1}^{-1} Q_{j-1} = \bar{\Pi}_{j-2} Q_{j-1}, \\ &Q_{j-1}(\bar{\Pi}_{j-2}\bar{Q}_{j-1} - \Pi_{j-2}Q_{j-1}) = Q_{j-1}(\bar{\Pi}_{j-2}Q_{j-1} - \Pi_{j-2}Q_{j-1}) = 0. \end{aligned}$$

We have to verify that the new coupling coefficients $\bar{\mathcal{H}}_k$ and $\bar{\mathcal{H}}_j$, $j \ge k+1$, disappear. We compute $\bar{Q}_k Z_{k+1}^{-1} = \bar{Q}_k - \bar{Q}_k P_k = \bar{Q}_k Q_k = Q_k$ and

$$Z_{j-1} Z_j^{-1} = Y_j^{-1} = I - \sum_{l=k}^{j-2} Q_l \bar{\Pi}_{j-2} Q_{j-1}, \quad j \ge k+2. \tag{1.54}$$

For $j \ge k+1$ this yields

$$\bar{Q}_{j*}\bar{\Pi}_{\mu-1} = \bar{Q}_j \bar{P}_{j+1} \cdots \bar{P}_{\mu-1} \bar{G}_\mu^{-1} B \bar{\Pi}_{\mu-1} = Z_j^{-1} Q_j Y_{j+1}^{-1} P_{j+1} \cdots Y_{\mu-1}^{-1} P_{\mu-1} Y_\mu^{-1} B \bar{\Pi}_{\mu-1}$$

and, by inserting (1.54) into the last expression,

$$\bar{Q}_{j*}\bar{\Pi}_{\mu-1} = Z_j^{-1}Q_j(I-\sum_{l=k}^{j-1}Q_l\bar{\Pi}_{j-1}Q_j)P_{j+1}\cdots P_{\mu-1}(I-\sum_{l=k}^{\mu-2}Q_l\bar{\Pi}_{\mu-2}Q_{\mu-1})G_\mu^{-1}B\bar{\Pi}_{\mu-1}.$$

Rearranging the terms one finds

$$\begin{aligned}\bar{Q}_{j*}\bar{\Pi}_{\mu-1} = (Z_j^{-1}Q_jP_{j+1}\cdots P_{\mu-1} + \mathcal{C}_{j,j+1}Q_{j+1}P_{j+2}\cdots P_{\mu-1} \\ +\cdots+\mathcal{C}_{j,\mu-2}Q_{\mu-2}P_{\mu-1}+\mathcal{C}_{j,\mu-1}Q_{\mu-1})G_\mu^{-1}B\bar{\Pi}_{\mu-1}.\end{aligned} \tag{1.55}$$

The detailed expression of the coefficients $\mathcal{C}_{j,i}$ does not matter at all. With analogous arguments we derive

$$\begin{aligned}\bar{Q}_{k*}\bar{\Pi}_{\mu-1} = (Q_{k*}P_{k+1}\cdots P_{\mu-1} + \mathcal{C}_{k,j+1}Q_{k+1}P_{k+2}\cdots P_{\mu-1} \\ +\cdots+\mathcal{C}_{k,\mu-2}Q_{\mu-2}P_{\mu-1}+\mathcal{C}_{k,\mu-1}Q_{\mu-1})G_\mu^{-1}B\bar{\Pi}_{\mu-1}.\end{aligned} \tag{1.56}$$

Next we compute

$$\begin{aligned}\bar{\Pi}_{\mu-1} &= \Pi_{k-1}\bar{P}_k\bar{P}_{k+1}\cdots\bar{P}_{\mu-1} = \Pi_{k-1}\bar{P}_kP_{k+1}\cdots P_{\mu-1}\\ &= \Pi_{k-1}(P_k+Q_k)\bar{P}_kP_{k+1}\cdots P_{\mu-1} = \Pi_{\mu-1}-Q_k\bar{Q}_k\Pi_{\mu-1},\end{aligned}$$

and therefore

$$G_\mu^{-1}B\bar{\Pi}_{\mu-1} = G_\mu^{-1}B(\Pi_{\mu-1}-\Pi_{k-1}Q_k\bar{Q}_k\Pi_{\mu-1}) = G_\mu^{-1}B\Pi_{\mu-1}-Q_k\bar{Q}_k\Pi_{\mu-1}.$$

Regarding assumption (1.53) and the properties of admissible projectors we have

$$Q_{\mu-1}G_\mu^{-1}B\bar{\Pi}_{\mu-1} = Q_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1}-Q_{\mu-1}\bar{Q}_k\Pi_{\mu-1} = Q_{\mu-1*}\Pi_{\mu-1} = 0,$$

and, for $i = k+1,\ldots,\mu-2$,

$$Q_iP_{i+1}\cdots P_{\mu-1}B\bar{\Pi}_{\mu-1} = Q_iP_{i+1}\cdots P_{\mu-1}B\Pi_{\mu-1}-Q_i\bar{Q}_k\Pi_{\mu-1} = Q_{i*}\Pi_{\mu-1} = 0.$$

Furthermore, taking into account the special choice of $\bar{Q}_k$,

$$\begin{aligned}Q_kP_{k+1}\cdots P_{\mu-1}B\bar{\Pi}_{\mu-1} &= Q_kP_{k+1}\cdots P_{\mu-1}B\Pi_{\mu-1}-Q_k\bar{Q}_k\Pi_{\mu-1}\\ &= (Q_{k*}-\bar{Q}_k)\Pi_{\mu-1} = 0.\end{aligned}$$

This makes it evident that all single summands on the right-hand sides of the formulas (1.55) and (1.56) disappear, and thus $\bar{Q}_{j*}\bar{\Pi}_{\mu-1} = 0$ for $j = k,\ldots,\mu-1$, that is, the new decoupling coefficients vanish. In consequence, starting with any admissible projectors we apply the above procedure first for $k = \mu-1$, then for $k = \mu-2$ up to $k = 0$. At each level an additional coupling coefficient is canceled, and we finish with a complete decoupling of the two parts in (1.46). □

Definition 1.23. Let the DAE (1.30), with coefficients $E, F \in L(\mathbb{R}^m)$, have the structural characteristic values

$$r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m,$$

and let the system (1.46) be generated by an admissible matrix sequence $G_0, \ldots, G_\mu$. If in (1.46) all coefficients $\mathcal{H}_i$, $i = 0, \ldots, \mu-1$, vanish, then the underlying admissible projectors $Q_0, \ldots, Q_{\mu-1}$ are called *completely decoupling projectors* for the DAE (1.30).

The completely decoupled system (1.46) offers as much insight as the Weierstraß-Kronecker form does.

Example 1.24 (Complete decoupling of an index-2 DAE). We reconsider once more the regular index-2 DAE

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} x' + \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} x = q$$

from Examples 1.8 and 1.21. The previously used projectors do not yield a complete decoupling. We now use a different projector Q_1 such that

$$Q_1 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{2} & 0 & -\frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \quad G_2 = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix},$$

and further

$$\Pi_1 = P_0 P_1 = \begin{bmatrix} \frac{1}{2} & 0 & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; P_0 Q_1 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \end{bmatrix}.$$

The DAE itself can be rewritten without any differentiations of equations as

$$\begin{aligned} (x_1 - x_3)' &= q_1 - q_2 - q_3, \\ (x_1 + x_3)' + 2x_2 &= q1 + q2 - q_3, \\ x_1 + x_3 &= q_3. \end{aligned}$$

Obviously, $\Pi_1 x$ again reflects the proper state variable $-x_1 + x_3$, for which an explicit ODE is given. $P_0 Q_1 x$ refers to the variable $x_1 + x_3$ that is described by the algebraic equation. Finally, $Q_0 x$ reflects the variable x_2. Simple calculations yield $\mathcal{W} = \Pi_{\mu-1} G_2^{-1} B_0 \Pi_{\mu-1} = 0$, $\mathcal{H}_0 = Q_0 P_1 G_2^{-1} B_0 \Pi_{\mu-1} = 0$ and $\mathcal{H}_1 = Q_1 G_2^{-1} B_0 \Pi_{\mu-1} = 0$. In this way the DAE decouples completely as

$$
\begin{aligned}
(\Pi_1 x)' &= \Pi_1 G_2^{-1} q,\\
-Q_0 Q_1 (\Pi_0 Q_1 x)' + Q_0 x &= Q_0 P_1 G_2^{-1} q,\\
\Pi_0 Q_1 &= \Pi_0 Q_1 G_2^{-1} q.
\end{aligned}
$$

These equations mean in full detail

$$
\begin{aligned}
\left(\begin{bmatrix} \frac{1}{2}(x_1 - x_3) \\ 0 \\ -\frac{1}{2}(x_1 - x_3) \end{bmatrix} \right)' &= \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & \frac{1}{2} \end{bmatrix} q,\\
\begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} \frac{1}{2}(x_1 + x_3) \\ 0 \\ \frac{1}{2}(x_1 + x_3) \end{bmatrix} \right)' + \begin{bmatrix} 0 \\ x_2 \\ 0 \end{bmatrix} &= \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} q,\\
\begin{bmatrix} \frac{1}{2}(x_1 + x_3) \\ 0 \\ \frac{1}{2}(x_1 + x_3) \end{bmatrix} &= \begin{bmatrix} 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{2} \end{bmatrix} q.
\end{aligned}
$$

Dropping the redundant equations as well as all zero lines one arrives exactly at the compressed form described above. □

Example 1.25 (Decoupling of the DAE in Example 1.5). The following matrix sequence is admissible for the pair $\{E, F\}$ from Example 1.5 which is regular with index 4:

$$
G_0 = E = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_0 = F = \begin{bmatrix} -\alpha & -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},
$$

$$
G_1 = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi_0 Q_1 = \begin{bmatrix} 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},
$$

$$
G_2 = \begin{bmatrix} 1 & -1 & \alpha & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & 0 & 0 & 1+\alpha & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi_1 Q_2 = \begin{bmatrix} 0 & 0 & 0 & \alpha & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},
$$

$$G_3 = \begin{bmatrix} 1 & -1 & \alpha & -\alpha^2 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & -1-\alpha-\alpha^2 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \Pi_2 Q_3 = \begin{bmatrix} 0 & 0 & 0 & 0 & -\alpha^2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G_4 = \begin{bmatrix} 1 & -1 & \alpha & -\alpha^2 & \alpha^3 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad \Pi_3 = \begin{bmatrix} 1 & 0 & 1 & -\alpha & -\alpha^2 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and the characteristic values are $r_0 = r_1 = r_2 = r_3 = 4, r_4 = 5$ and $\mu = 4$. Additionally, it follows that

$$Q_3 G_4^{-1} B_0 \Pi_3 = 0, \qquad Q_2 P_3 G_4^{-1} B_0 \Pi_3 = 0,$$
$$Q_1 P_2 P_3 G_4^{-1} B_0 \Pi_3 = 0, \qquad Q_0 P_1 P_2 P_3 G_4^{-1} B_0 \Pi_3 = 0,$$

and

$$\Pi_3 G_4^{-1} B_0 \Pi_3 = -\alpha \Pi_3. \tag{1.57}$$

The projectors Q_0, Q_1, Q_2, Q_3 provide a complete decoupling of the given DAE $Ex'(t) + Fx(t) = q(t)$. The projectors Q_0, $\Pi_0 Q_1$, $\Pi_1 Q_2$ and $\Pi_2 Q_3$ represent the variables x_2, x_3, x_4 and x_5, respectively. The projector Π_3 and the coefficient (1.57) determine the inherent regular ODE, namely (the zero rows are dropped)

$$(x_1 + x_3 - \alpha x_4 + \alpha^2 x_5)' - \alpha(x_1 + x_3 - \alpha x_4 + \alpha^2 x_5) = q_1 + q_2 - \alpha q_3 + \alpha^2 q_4 - \alpha^3 q_5.$$

It is noteworthy that no derivatives of the excitation q encroach in this ODE. □

Notice that for DAEs with $\mu = 1$, the completely decoupling projector Q_0 is uniquely determined. It is the projector onto N_0 along $S_0 = \{z \in \mathbb{R}^m : B_0 z \in \operatorname{im} G_0\}$ (cf. Appendix A). However, for higher index $\mu > 1$, there are many complete decouplings, as the next example shows.

Example 1.26 (Diversity of completely decoupling projectors). Let

$$E = G_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad F = B_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

and choose projectors with a free parameter α:

$$Q_0 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad P_0 = \begin{bmatrix} 1 & -\alpha & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad G_1 = \begin{bmatrix} 1 & 1+\alpha & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_1 = P_0,$$

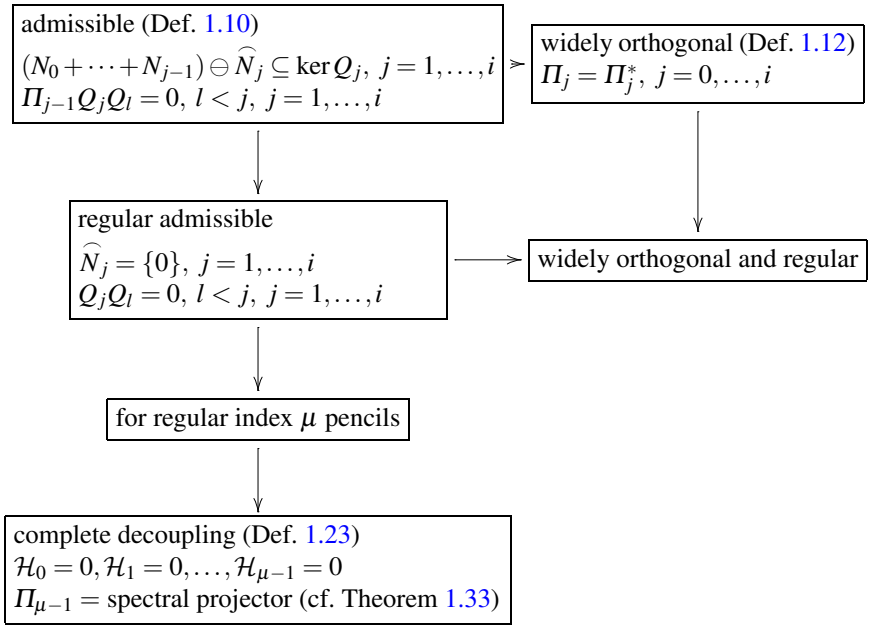
$$Q_1 = \begin{bmatrix} 0 & -(1+\alpha) & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad G_2 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$G_2^{-1} = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad Q_0 P_1 G_2^{-1} B_0 = Q_0,$$

i.e., Q_0 and Q_1 are completely decoupling projectors for each arbitrary value α. However, in contrast, the projector Π_1 is independent of α. □

1.2.4 Hierarchy of projector sequences for constant matrix pencils

The matrices $Q_0, \ldots, Q_i$ are projectors, where Q_j projects onto $N_j = \ker G_j$, $j = 0, \ldots, i$, with $P_0 := I - Q_0$, $\Pi_0 := P_0$ and $P_j := I - Q_j$, $\Pi_j := \Pi_{j-1} P_j$, $\widehat{N}_j := (N_0 + \cdots + N_{j-1}) \cap N_j$, $j = 1, \ldots, i$.



1.2.5 Compression to a generalized Weierstraß–Kronecker form

The DAE (1.30) as well as its decoupled version (1.35) comprise m equations. The advanced decoupled system (1.46) is formally composed of $m(\mu+1)$ equations; however, it can be compressed back on an m-dimensional DAE without losing information. The next lemma records essential properties to be used in the compression procedure.

Lemma 1.27. *The entries $\mathcal{N}_{ij}$ of the decoupled system (1.46) have the following properties for $i = 0, \ldots, \mu-2$:*

$$\begin{aligned} \mathcal{N}_{i,i+1} &= \mathcal{N}_{i,i+1}\Pi_i Q_{i+1}, \\ \mathcal{N}_{ij} &= \mathcal{N}_{ij}\Pi_{j-1}Q_j, \quad j = i+2, \ldots, \mu-1, \\ \ker \mathcal{N}_{i,i+1} &= \ker \Pi_i Q_{i+1}, \\ \operatorname{rank} \mathcal{N}_{i,i+1} &= m - r_{i+1}. \end{aligned}$$

Proof. We use the additional subspaces $S_i := \ker \mathcal{W}_i B_i \subseteq \mathbb{R}^m$ and the projectors $\mathcal{W}_i \in L(\mathbb{R}^m)$ with

$$\ker \mathcal{W}_i = \operatorname{im} G_i,\ i = 0, \ldots, \mu-1.$$

Let G_i^- be the generalized reflexive inverse of G_i with $G_i G_i^- G_i = G_i$, $G_i^- G_i G_i^- = G_i^-$, $G_i G_i^- = I - \mathcal{W}_i$ and $G_i^- G_i = P_i$. We factorize G_{i+1} as

$$G_{i+1} = G_i + B_i Q_i = G_i + \mathcal{W}_i B_i Q_i + G_i G_i^- B_i Q_i = \mathcal{G}_{i+1}\mathcal{F}_{i+1},$$

$$\mathcal{G}_{i+1} := G_i + \mathcal{W}_i B_i Q_i, \quad \mathcal{F}_{i+1} = I + P_i G_i^- B_i Q_i.$$

Since $\mathcal{F}_{i+1}$ is invertible (cf. Lemma A.3), it follows that $\mathcal{G}_{i+1}$ has rank r_{i+1} like G_{i+1}.

Furthermore, it holds that $\ker \mathcal{G}_{i+1} = N_i \cap S_i$. Namely, $\mathcal{G}_{i+1} z = 0$ means that $G_i z = 0$ and $\mathcal{W}_i B_i Q_i z = 0$, i.e., $z = Q_i z$ and $\mathcal{W}_i B_i z = 0$, but this is $z \in N_i \cap S_i$. Therefore, $N_i \cap S_i$ must have the dimension $m - r_{i+1}$. Next we derive the relation

$$N_i \cap S_i = \operatorname{im} Q_i Q_{i+1}. \tag{1.58}$$

If $z \in N_i \cap S_i$ then $z = Q_i z$ and $B_i z = G_i w$ implying $(G_i + B_i Q_i)(P_i w + Q_i z) = 0$, and hence $P_i w + Q_i z = Q_{i+1}(P_i w + Q_i z) = Q_{i+1} w$. Therefore, $z = Q_i z = Q_i Q_{i+1} w$. Consequently, $N_i \cap S_i \subseteq \operatorname{im} Q_i Q_{i+1}$. Conversely, assume $z = Q_i Q_{i+1} y$. Taking into consideration that $(G_i + B_i Q_i) Q_{i+1} = 0$, we derive $z = Q_i z$ and $B_i z = B_i Q_i Q_{i+1} y = -G_i Q_{i+1} y$, i.e., $z \in N_i$ and $z \in S_i$. Thus, relation (1.58) is valid.

Owing to (1.58) we have

$$\operatorname{rank} Q_i Q_{i+1} = \dim N_i \cap S_i = m - r_{i+1}. \tag{1.59}$$

If follows immediately that $\operatorname{rank} \mathcal{N}_{i,i+1} = m - r_{i+1}$, and, since $\operatorname{im} P_{i+1} \subseteq \ker \mathcal{N}_{i,i+1}$, $\operatorname{rank} P_{i+1} = r_{i+1}$, that $\operatorname{im} P_{i+1} = \ker \mathcal{N}_{i,i+1}$. □

We turn to the compression of the large system (1.46) on m dimensions. The projector Q_0 has rank $m-r_0$, the projector $\Pi_{i-1}Q_i$ has rank $m-r_i$ for $i=1,\dots,\mu-1$, and $\Pi_{\mu-1}$ has rank $d:=m-\sum_{j=0}^{\mu-1}(m-r_j)$.

We introduce full-row-rank matrices $\Gamma_i\in L(\mathbb{R}^m,\mathbb{R}^{m-r_i})$, $i=0,\dots,\mu-1$, and $\Gamma_d\in L(\mathbb{R}^m,\mathbb{R}^d)$ such that

$$\begin{aligned}
&\operatorname{im}\Gamma_d\Pi_{\mu-1}=\Gamma_d\operatorname{im}\Pi_{\mu-1}=\mathbb{R}^d, && \ker\Gamma_d=\operatorname{im}(I-\Pi_{\mu-1})=N_0+\cdots+N_{\mu-1},\\
&\Gamma_0N_0=\mathbb{R}^{m-r_0}, && \ker\Gamma_0=\ker Q_0,\\
&\Gamma_i\Pi_{i-1}N_i=\mathbb{R}^{m-r_i}, && \ker\Gamma_i=\ker\Pi_{i-1}Q_i,\ i=1,\dots,\mu-1,
\end{aligned}$$

as well as generalized inverses Γ_d^-,Γ_i^-, $i=0,\dots,\mu-1$, such that

$$\begin{aligned}
&\Gamma_d^-\Gamma_d=\Pi_{\mu-1}, && \Gamma_d\Gamma_d^-=I,\\
&\Gamma_i^-\Gamma_i=\Pi_{i-1}Q_i, && \Gamma_i\Gamma_i^-=I, \qquad i=1,\dots,\mu-1,\\
&\Gamma_0^-\Gamma_0=Q_0, && \Gamma_0\Gamma_0^-=I.
\end{aligned}$$

If the projectors $Q_0,\dots,Q_{\mu-1}$ are widely orthogonal (cf. Proposition 1.13(6)), then the above projectors are symmetric and Γ_d^-, Γ_i^- are the Moore–Penrose generalized inverses. Denoting

$$\tilde{\mathcal{H}}_i:=\Gamma_i\mathcal{H}_i\Gamma_d^-, \qquad \tilde{\mathcal{L}}_i:=\Gamma_i\mathcal{L}_i,\ \ i=0,\dots,\mu-1, \tag{1.60}$$

$$\tilde{\mathcal{W}}:=\Gamma_d\mathcal{W}\Gamma_d^-, \qquad \tilde{\mathcal{L}}_d:=\Gamma_d\mathcal{L}_d, \tag{1.61}$$

$$\tilde{\mathcal{N}}_{ij}:=\Gamma_i\mathcal{N}_{ij}\Gamma_j^-,\ \ j=i+1,\dots,\mu-1,\ i=0,\dots,\mu-2, \tag{1.62}$$

and transforming the new variables

$$\tilde{u}=\Gamma_du, \qquad \tilde{v}_i=\Gamma_iv_i, \qquad i=0,\dots,\mu-1, \tag{1.63}$$

$$u=\Gamma_d^-\tilde{u}, \qquad v_i=\Gamma_i^-\tilde{v}_i, \qquad i=0,\dots,\mu-1, \tag{1.64}$$

we compress the large system (1.46) into the m-dimensional one

$$\left[\begin{array}{c|ccc} I & & & \\ \hline 0 & \tilde{\mathcal{N}}_{01} & \cdots & \tilde{\mathcal{N}}_{0,\mu-1}\\ & \ddots & \ddots & \vdots\\ & & \ddots & \tilde{\mathcal{N}}_{\mu-2,\mu-1}\\ & & & 0 \end{array}\right]\left[\begin{array}{c} \tilde{u}'(t)\\ \hline 0\\ \tilde{v}_1'(t)\\ \vdots\\ \tilde{v}_{\mu-1}'(t) \end{array}\right]$$
$$+\left[\begin{array}{c|ccc} \tilde{\mathcal{W}} & & & \\ \hline \tilde{\mathcal{H}}_0 & I & & \\ \vdots & & \ddots & \\ \vdots & & & I\\ \tilde{\mathcal{H}}_{\mu-1} & & & \end{array}\right]\left[\begin{array}{c} \tilde{u}(t)\\ \hline \tilde{v}_0(t)\\ \vdots\\ \vdots\\ \tilde{v}_{\mu-1}(t) \end{array}\right]=\left[\begin{array}{c} \tilde{\mathcal{L}}_d\\ \tilde{\mathcal{L}}_0\\ \vdots\\ \vdots\\ \tilde{\mathcal{L}}_{\mu-1} \end{array}\right]q \tag{1.65}$$

without losing any information. As a consequence of Lemma 1.27, the blocks $\tilde{\mathcal{N}}_{i,i+1}$ have full column rank $m-r_{i+1}$ for $i=0,\dots,\mu-2$.

Proposition 1.28. *Let the pair $\{E,F\}$, $E,F\in L(\mathbb{R}^m)$ have the structural characteristic values*

$$r_0\le\cdots\le r_{\mu-1}<r_\mu=m.$$

(1) *Then there are nonsingular matrices $L,K\in L(\mathbb{R}^m)$ such that*

$$LEK=\left[\begin{array}{c|ccc} I & & & \\ \hline 0 & \tilde{\mathcal{N}}_{01} & \cdots & \tilde{\mathcal{N}}_{0,\mu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \tilde{\mathcal{N}}_{\mu-2,\mu-1} \\ & & & 0 \end{array}\right],\quad LFK=\left[\begin{array}{c|ccc} \tilde{\mathcal{W}} & & & \\ \hline \tilde{\mathcal{H}}_0 & I & & \\ \vdots & & \ddots & \\ \vdots & & & \ddots \\ \tilde{\mathcal{H}}_{\mu-1} & & & I \end{array}\right],$$

with entries described by (1.60)–(1.62). Each block $\tilde{\mathcal{N}}_{i,i+1}$ has full column rank $m-r_{i+1}$, $i=0,\dots,\mu-2$, and hence the nilpotent part in LEK has index μ.

(2) *By means of completely decoupling projectors, L and K can be built so that the coefficients $\tilde{\mathcal{H}}_0,\dots,\tilde{\mathcal{H}}_{\mu-1}$ disappear, and the DAE transforms into Weierstraß–Kronecker form (1.3) with $l=\sum_{i=0}^{\mu-1}(m-r_i)$.*

Proof. Due to the properties

$$\begin{aligned}
&\mathcal{H}_i=\mathcal{H}_i\Pi_{\mu-1}=\mathcal{H}_i\Gamma_d^-\Gamma_d,\ \ i=0,\dots,\mu-1,\\
&\mathcal{W}=\mathcal{W}\Pi_{\mu-1}=\mathcal{W}\Gamma_d^-\Gamma_d,\\
&\mathcal{N}_{ij}=\mathcal{N}_{ij}\Pi_{j-1}Q_j=\mathcal{N}_{ij}\Gamma_j^-\Gamma_j,\ j=1,\dots,\mu-1,\ i=0,\dots,\mu-2,
\end{aligned}$$

we can recover system (1.46) from (1.65) by multiplying on the left by

$$\Gamma^-:=\left[\begin{array}{c|ccc} \Gamma_d^- & & & \\ \hline & \Gamma_0^- & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1}^- \end{array}\right]\in L(\mathbb{R}^m,\mathbb{R}^{(\mu+1)m})$$

using transformation (1.64) and taking into account that $u=\Gamma_d^-\tilde{u}=\Pi_{\mu-1}u$ and $\Pi_{\mu-1}u'=u'$. The matrix Γ^- is a generalized inverse of

$$\Gamma:=\left[\begin{array}{c|ccc} \Gamma_d & & & \\ \hline & \Gamma_0 & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1} \end{array}\right]\in L(\mathbb{R}^{(\mu+1)m},\mathbb{R}^m)$$

having the properties $\Gamma\Gamma^- = I_m$ and

$$\Gamma^-\Gamma = \left[\begin{array}{c|ccc} \Gamma_d^-\Gamma_d & & & \\ \hline & \Gamma_0^-\Gamma_0 & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1}^-\Gamma_{\mu-1} \end{array}\right] = \left[\begin{array}{c|ccc} \Pi_{\mu-1} & & & \\ \hline & Q_0 & & \\ & & \Pi_0 Q_1 & \\ & & & \Pi_{\mu-2}Q_{\mu-1} \end{array}\right].$$

The product $K := \Gamma \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} = \begin{bmatrix} \boxed{\Gamma_d} \\ \boxed{\Gamma_0} \\ \vdots \\ \boxed{\Gamma_{\mu-1}} \end{bmatrix}$ is nonsingular. Our decomposition now means that

$$\begin{aligned} x &= \Pi_{\mu-1}x + Q_0x + \Pi_0Q_1x + \cdots + \Pi_{\mu-2}Q_{\mu-1}x \\ &= [I \cdots I]\Gamma^-\Gamma \begin{bmatrix} I \\ \vdots \\ I \end{bmatrix} x = [I \cdots I] \begin{bmatrix} u \\ v_0 \\ \vdots \\ v_{\mu-1} \end{bmatrix} \end{aligned}$$

and the transformation (1.63) reads

$$\begin{bmatrix} \tilde{u} \\ \tilde{v}_0 \\ \vdots \\ \tilde{v}_{\mu-1} \end{bmatrix} = \Gamma \begin{bmatrix} u \\ v_0 \\ \vdots \\ v_{\mu-1} \end{bmatrix} = \Gamma\Gamma^-\Gamma \begin{bmatrix} I \\ \vdots \\ \vdots \\ I \end{bmatrix} x = \Gamma \begin{bmatrix} I \\ \vdots \\ \vdots \\ I \end{bmatrix} x = Kx = \tilde{x}.$$

Thus, turning from the original DAE (1.30) to the DAE in the form (1.65) means a coordinate transformation $\tilde{x} = Kx$, with a nonsingular matrix K, combined with a scaling by

$$L := [I \cdots I]\,\Gamma^-\Gamma \begin{bmatrix} \Pi_{\mu-1} & & & & \\ & Q_0P_1\cdots P_{\mu-1} & & & \\ & & \ddots & & \\ & & & Q_{\mu-2}P_{\mu-1} & \\ & & & & Q_{\mu-1} \end{bmatrix} \begin{bmatrix} I \\ \vdots \\ \vdots \\ I \end{bmatrix} G_\mu^{-1}.$$

L is a nonsingular matrix. Namely, $LG_\mu z = 0$ means that

$$\begin{aligned} \Pi_{\mu-1}z + Q_0P_1\cdots P_{\mu-1}z + \Pi_0Q_1P_2\cdots P_{\mu-1}z + \cdots \\ + \Pi_{\mu-3}Q_{\mu-2}P_{\mu-1}z + \Pi_{\mu-2}Q_{\mu-1}z = 0, \end{aligned}$$

and multiplying by $\Pi_{\mu-1}$ yields $\Pi_{\mu-1}z = 0$. Multiplying by $Q_{\mu-1}$ implies $Q_{\mu-1}z = 0$, multiplying by $Q_{\mu-2}P_{\mu-1}$ gives $Q_{\mu-2}P_{\mu-2}z = 0$, and so on. Hence

$$(I - \Pi_{\mu-1})z = Q_{\mu-1}z + Q_{\mu-2}P_{\mu-1}z + \cdots + Q_0P_1\cdots P_{\mu-1}z = 0.$$

The original DAE (1.30) and the system (1.65) are equivalent in the usual sense, which proves the first assertion. Regarding the existence of completely decoupling projectors (see Theorem 1.22), the second assertion immediately follows from the first one. □

1.2.6 Admissible projectors for matrix pairs in a generalized Weierstraß–Kronecker form

Here we deal with the regular matrix pair $\{E,F\}$ given by the $m \times m$ structured matrices

$$E = \left[\begin{array}{c|c} I & 0 \\ \hline 0 & N \end{array}\right]\begin{matrix}\}m-l \\ \}l\end{matrix}, \quad F = \left[\begin{array}{c|c} W & 0 \\ \hline H & I \end{array}\right]\begin{matrix}\}m-l \\ \}l\end{matrix}, \tag{1.66}$$

where $W \in L(\mathbb{R}^{m-l})$, $H =: \begin{bmatrix} H_1 \\ \vdots \\ H_\mu \end{bmatrix} \in L(\mathbb{R}^{m-l}, \mathbb{R}^l)$ and N is a nilpotent, upper triangular $l \times l$ matrix, $l > 0$, of the form

$$N = \begin{bmatrix} 0 & N_{1,2} & \cdots & N_{1,\mu} \\ & \ddots & & \vdots \\ & & \ddots & N_{\mu-1,\mu} \\ & & & 0 \end{bmatrix}\begin{matrix}\}l_1 \\ \\ \\ \}l_{\mu-1} \\ \}l_\mu\end{matrix} \tag{1.67}$$

with $l_1 \geq \cdots \geq l_\mu \geq 1$ and $l_1 + \cdots + l_\mu = l$. The blocks $N_{i,i+1}$ with l_i rows and l_{i+1} columns are assumed to have full column rank, which means, $\ker N_{i,i+1} = \{0\}$ for $i = 1,\ldots,\mu-1$. Then N has nilpotency order μ; that is $N^\mu = 0$, $N^{\mu-1} \neq 0$, and l_i equals the number of its Jordan blocks of order $\geq i$, $i = 1,\ldots,\mu$.

This special form of the nilpotent block is closely related to the tractability index concept, in particular with the decouplings provided by admissible projectors (see Proposition 1.28).

The Jordan form of such a nilpotent matrix N consists of $l_1 - l_2$ (nilpotent) Jordan chains of order one, $l_2 - l_3$ chains of order two, and so on up to $l_{\mu-1} - l_\mu$ chains of order $\mu - 1$, and l_μ chains of order μ. Any nilpotent matrix can be put into the structural form (1.67) by means of a similarity transformation. Thus, without loss of generality we may suppose this special form.

The polynomial $p(\lambda) := \det(\lambda E + F) = \det(\lambda I + W)$ has degree $m - l$. This pair $\{E,F\}$ is regular and represents a slight generalization of the classical Weierstraß–

Kronecker form discussed in Section 1.1 (cf. (1.3)), where the entries of the block H are zeros.

In accordance with the structure of E and F in (1.66) we write $z \in \mathbb{R}^m$ as

$$z = \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_\mu \end{bmatrix}, \quad z_0 \in \mathbb{R}^{m-l}, \quad z_i \in \mathbb{R}^{l_i}, \quad i = 1, \ldots, \mu.$$

Now we construct a matrix sequence (1.10) by admissible projectors. Thereby, in the following pages in the present section, the letter N is used in a twofold sense: N_i, with a single subscript, indicates one of the subspaces, and $N_{j,k}$, with double subscript, means an entry of a matrix.

Put $G_0 = E$, $B_0 = F$. Since $N_0 = \ker G_0 = \{z \in \mathbb{R}^m : z_0 = 0,\ z_\mu = 0,\ \ldots, z_2 = 0\}$ we choose

$$Q_0 = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & I & & \\ & & 0 & \\ & & & \ddots \\ & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \\ \end{matrix}, \quad \Pi_0 = P_0 = \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & I & \\ & & & \ddots \\ & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \\ \end{matrix},$$

which leads to

$$G_1 = \left[\begin{array}{c|cccc} I & & & & \\ \hline I & N_{1,2} & \cdots & \cdots & N_{1,\mu} \\ & 0 & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & N_{\mu-1,\mu} \\ & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \\ \\ \end{matrix}, \quad B_1 = \left[\begin{array}{c|ccc} W & & & \\ \hline H_1 & 0 & & \\ \vdots & & I & \\ & & & \ddots \\ H_\mu & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \\ \end{matrix},$$

and

$$N_1 = \{z \in \mathbb{R}^m : z_0 = 0,\ z_\mu = 0,\ \ldots,\ z_3 = 0,\ z_1 + N_{1,2} z_2 = 0\}, \quad N_1 \cap N_0 = 0.$$

Choosing

$$Q_1 = \left[\begin{array}{c|ccccc} 0 & & & & & \\ \hline & 0 & -N_{1,2} & & & \\ & & I & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \}l_2 \\ \\ \\ \\ \end{matrix}, \quad P_1 = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline & I & N_{1,2} & & & \\ & & 0 & & & \\ & & & I & & \\ & & & & \ddots & \\ & & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \}l_2 \\ \\ \\ \\ \end{matrix},$$

$$\Pi_1 = \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & 0 & \\ & & & I \\ & & & & \ddots \\ & & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \}l_2 \\ \\ \\ \\ \end{matrix},$$

we meet the condition $N_0 \subseteq \ker Q_1$, which means that $Q_1 Q_0 = 0$, and find

$$G_2 = \left[\begin{array}{c|cccc} I & & & & \\ \hline I & N_{1,2} & \cdots & \cdots & N_{1,\mu} \\ & I & N_{2,3} & & \vdots \\ & & 0 & \ddots & \vdots \\ & & & \ddots & N_{\mu-1,\mu} \\ & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \}l_2 \\ \\ \\ \end{matrix}, \quad B_2 = \left[\begin{array}{c|ccc} W & & & \\ \hline H_1 & 0 & & \\ & & 0 & \\ \vdots & & & I \\ & & & & \ddots \\ H_\mu & & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \}l_2 \\ \\ \\ \\ \end{matrix},$$

$N_2 = \{z \in \mathbb{R}^m : z_0 = 0,\ z_\mu = 0, \ldots, z_4 = 0,\ z_2 + N_{23} z_3 = 0,\ z_1 + N_{12} z_2 + N_{13} z_3 = 0\}$, $(N_0 + N_1) \cap N_2 = (\ker \Pi_1) \cap N_2 = \{0\}$. Suppose that we are on level i and that we have $Q_0, \ldots, Q_{i-1}$ being admissible,

$$Q_{i-1} = \left[\begin{array}{c|ccccc} 0 & & & & & \\ \hline & 0 & & * & & \\ & & \ddots & \vdots & & \\ & & & 0 & * & \\ & & & & I & \\ & & & & & 0 \\ & & & & & & \ddots \\ & & & & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \}l_i \\ \\ \\ \\ \end{matrix}, \quad \Pi_{i-1} = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline & 0 & & & & \\ & & \ddots & & & \\ & & & 0 & & \\ & & & & I & \\ & & & & & \ddots \\ & & & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \\ \}l_i \\ \\ \\ \end{matrix},$$

$$Q_{i-1}(N_0 + \cdots + N_{i-2}) = Q_{i-1} \operatorname{im}(I - \Pi_{i-2}) = \operatorname{im} Q_{i-1}(I - \Pi_{i-2}) = \{0\},$$

$$G_i = \left[\begin{array}{c|ccccccc} I & & & & & & & \\ \hline I & N_{1,2} & \cdots & & \cdots & \cdots & N_{1,\mu} \\ & \ddots & \ddots & & & & \vdots \\ & & I & N_{i,i+1} & & & \vdots \\ & & & 0 & \ddots & & \vdots \\ & & & & & \ddots & N_{\mu-1,\mu} \\ & & & & & & 0 \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \}l_i \\ \\ \\ \\ \end{matrix}, \; B_i = \left[\begin{array}{c|ccccc} W & & & & & \\ \hline H_1 & 0 & & & & \\ & & \ddots & & & \\ \vdots & & & 0 & & \\ \vdots & & & & I & \\ & & & & & \ddots \\ H_\mu & & & & & & I \end{array}\right] \begin{matrix} \\ \}l_1 \\ \\ \}l_i \\ \\ \\ \\ \end{matrix}. \tag{1.68}$$

It follows that

$$N_i = \{z \in \mathbb{R}^m : z_0 = 0,\, z_\mu = 0, \ldots,\, z_{i+2} = 0,$$
$$z_i + N_{i,i+1} z_{i+1} = 0, \ldots,\, z_1 + N_{12} z_2 + \cdots + N_{1,i+1} z_{i+1} = 0\},$$
$$(N_0 + \cdots + N_{i-1}) \cap N_i = (\ker \Pi_{i-1}) \cap N_i = \{0\}.$$

Choosing

$$Q_i = \left[\begin{array}{c|ccccc} 0 & & & & & \\ \hline & 0 & & * & & \\ & & \ddots & \vdots & & \\ & & & 0 & * & \\ & & & & I & \\ & & & & & 0 \end{array}\right] \begin{matrix} \}l_1 \\ \\ \\ \\ \}l_{i+1} \\ \\ \end{matrix}, \quad P_i = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline & I & & * & & \\ & & \ddots & \vdots & & \\ & & & I & * & \\ & & & & 0 & \\ & & & & & I \end{array}\right] \begin{matrix} \}l_1 \\ \\ \\ \\ \}l_{i+1} \\ \\ \end{matrix},$$

$$\Pi_i = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline & 0 & & & & \\ & & \ddots & & & \\ & & & 0 & & \\ & & & & I & \\ & & & & & \ddots \\ & & & & & & I \end{array}\right] \begin{matrix} \}l_1 \\ \\ \\ \\ \}l_{i+1} \\ \\ \\ \end{matrix},$$

we meet the admissibility condition (1.13), as $Q_i(I - \Pi_{i-1}) = 0$, and arrive at

$$G_{i+1} = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline I & N_{1,2} & \cdots & \cdots & \cdots & N_{1,\mu} \\ & \ddots & \ddots & & & \vdots \\ & & I & N_{i+1,i+2} & & \vdots \\ & & & 0 & \ddots & \vdots \\ & & & & \ddots & N_{\mu-1,\mu} \\ & & & & & 0 \end{array}\right], B_{i+1} = \left[\begin{array}{c|ccccc} W & & & & & \\ \hline H_1 & 0 & & & & \\ & & \ddots & & & \\ \vdots & & & 0 & & \\ \vdots & & & & I & \\ & & & & & \ddots \\ H_\mu & & & & & & I \end{array}\right].$$

This verifies that formulas (1.68) provide the right pair G_i, B_i at level i, $i \geq 1$, for $\{E, F\}$ as in (1.66). Obviously, we obtain precisely a nonsingular matrix G_μ, but $G_{\mu-1}$ is singular. The characteristic values of our pair $\{E, F\}$ are $u_i = 0$, $i \geq 1$, and

$$r_i = m - \dim N_i = m - l_{i+1} < m,\ i = 0, \ldots, \mu - 1,\ r_\mu = m.$$

The next proposition records this result.

Proposition 1.29. *Each admissible matrix sequence $G_0, \ldots, G_\mu$ for the special pair $\{E, F\}$ given by (1.66), (1.67) consists of singular matrices $G_0, \ldots, G_{\mu-1}$ and*

a nonsingular G_μ. The characteristic values are $u_i = 0$ for $i = 1,\dots,\mu$, and $r_i = m - \dim N_i = m - l_{i+1}$ for $i = 0,\dots,\mu-1$, $r_\mu = m$.

For the associated DAE, and in particular for the DAE in Weierstraß–Kronecker form (1.66) with its structured part N (1.67) and $H = 0$, the decoupling into the basic parts is given a priori (cf. (1.4), (1.5)). The so-called "slow" subsystem

$$y'(t) + Wy(t) = p(t)$$

is a standard explicit ODE, hence an integration problem, whereas the so-called "fast" subsystem

$$Nz'(t) + z(t) = r(t) - Hy(t)$$

contains exclusively algebraic relations and differentiation problems.

The admissible projectors exhibit these two basic structures as well as a further subdivision of the differentiation problems: The proper state variable is comprised by $\Pi_{\mu-1}$ while $I - \Pi_{\mu-1}$ collects all other variables, where

$$\Pi_{\mu-1} = \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & \ddots & \\ & & & 0 \end{array}\right], \quad I - \Pi_{\mu-1} = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & I & & \\ & & \ddots & \\ & & & I \end{array}\right].$$

Those variables that are not differentiated at all and those variables that have to be differentiated i times are given by

$$Q_0 = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & I & & \\ & & 0 & \\ & & & \ddots \\ & & & 0 \end{array}\right] \Big\}l_1 \quad \text{and } \Pi_{i-1}Q_i = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & 0 & & \\ & & \ddots & \\ & & & I \\ & & & & \ddots \\ & & & & & 0 \end{array}\right] \Big\}l_{i+1},$$

respectively.

1.3 Transformation invariance

Here we show the structural characteristic values r_i for $i \geq 0$ to be invariant under transformations. Given a matrix pair $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, and nonsingular matrices $L,K \in L(\mathbb{R}^m)$, we consider the transformed pair $\{\bar{E},\bar{F}\}$, formed by

$$\bar{E} = LEK, \quad \bar{F} = LFK. \tag{1.69}$$

The DAEs associated to the pairs $\{E,F\}$ and $\{\bar{E},\bar{F}\}$ are

$$Ex'(t)+Fx(t)=q(t) \quad \text{and} \quad \bar{E}\bar{x}'(t)+\bar{F}\bar{x}'(t)=\bar{q}(t).$$

They are related to each other by the transformation $x=K\bar{x}$ and premultiplication by L where $\bar{q}=Lq$. In this sense, these DAEs are solution-equivalent.

How are the admissible matrix sequences $(G_i)_{i\geq 0}$ and $(\bar{G}_i)_{i\geq 0}$ as well as the admissible projectors $(Q_i)_{i\geq 0}$ and $(\bar{Q}_i)_{i\geq 0}$ related for $\{E,F\}$ and $\{\bar{E},\bar{F}\}$? The answer is simple.

Theorem 1.30. *If two matrix pairs $\{E,F\}$ and $\{\bar{E},\bar{F}\}$ are related via (1.69), with nonsingular $K,L\in L(\mathbb{R}^m)$, then they have common structural characteristic values*

$$r_i=\bar{r}_i,\ i\geq 0, \quad u_i=\bar{u}_i,\ i\geq 1.$$

If $Q_0,\dots,Q_\kappa$ are admissible projectors for $\{E,F\}$, then the matrices $\bar{Q}_0,\dots,\bar{Q}_\kappa$ with $\bar{Q}_i:=K^{-1}Q_iK$ for $i=0,\dots,\kappa$ are admissible projectors for $\{\bar{E},\bar{F}\}$.

Proof. The transformations $\bar{G}_0=LG_0K$, $\bar{B}_0=LB_0K$, $\bar{N}_0=K^{-1}N_0$ are given to begin with, and $\bar{Q}_0:=K^{-1}Q_0K$ is admissible. Compute $\bar{G}_1=\bar{G}_0+\bar{B}_0\bar{Q}_0=LG_1K$, $\bar{r}_1=r_1$, then

$$\bar{N}_1=K^{-1}N_1,\ \bar{N}_0\cap\bar{N}_1=K^{-1}(N_0\cap N_1).$$

Put $\bar{X}_1:=K^{-1}X_1$ such that $\bar{N}_0=(\bar{N}_0\cap\bar{N}_1)\oplus\bar{X}_1$ and notice that $\bar{Q}_1:=K^{-1}Q_1K$ has the property $\ker\bar{Q}_1\supseteq\bar{X}_1$ implying the sequence $\bar{Q}_0$, $\bar{Q}_1$ to be admissible. At level i, we have

$$\bar{G}_i=LG_iK,\ \bar{N}_0+\cdots+\bar{N}_{i-1}=K^{-1}(N_0+\cdots+N_{i-1}),\ \bar{N}_i=K^{-1}N_i,\ \bar{r}_i=r_i,$$

and $\bar{Q}_i:=K^{-1}Q_iK$ satisfies condition $\ker\bar{Q}_i\supseteq\bar{X}_i$ with

$$\bar{X}_i:=K^{-1}X,\ \bar{N}_0+\cdots+\bar{N}_{i-1}=[(\bar{N}_0+\cdots+\bar{N}_{i-1})\cap\bar{N}_i]\oplus\bar{X}_i.$$

□

Now we are in a position to state the important result concerning the consistency of the projector based approach and the structure described via the Weierstraß–Kronecker form.

Theorem 1.31. *For $E,F\in L(\mathbb{R}^m)$ suppose the pair $\{E,F\}$ to be regular with Kronecker index $\mu\geq 0$. Then the admissible matrix sequence $(G_i)_{i\geq 0}$ exhibits singular matrices $G_0,\dots,G_{\mu-1}$, but a nonsingular G_μ, and vice versa.*

Proof. ($\Rightarrow$) This is a consequence of the existence of the Weierstraß–Kronecker form (cf. Proposition 1.3), Theorem 1.30 and Proposition 1.29.
($\Leftarrow$) Let the pair $\{E,F\}$ have the characteristic values $r_0\leq\cdots\leq r_{\mu-1}<r_\mu=m$. By Theorem 1.22 we can choose completely decoupling projectors. Applying the decoupling and compressing procedure for the associated DAE we arrive at an equivalent DAE of the form

$$\begin{bmatrix} I & \\ & \mathcal{N} \end{bmatrix}\tilde{x}'+\begin{bmatrix} \mathcal{W} & \\ & I \end{bmatrix}\tilde{x}=\tilde{q}. \tag{1.70}$$

The matrix $\tilde{\mathcal{N}}$ is nilpotent with index μ, and it has the structure

$$\tilde{\mathcal{N}} = \begin{bmatrix} 0 & \tilde{\mathcal{N}}_{01} & \cdots & \cdots & \tilde{\mathcal{N}}_{0,\mu-1} \\ & 0 & \ddots & & \vdots \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \tilde{\mathcal{N}}_{\mu-2,\mu-1} \\ & & & & 0 \end{bmatrix} \begin{matrix} \}m-r_0 \\ \\ \\ \}m-r_{\mu-2} \\ \}m-r_{\mu-1} \end{matrix}, \tag{1.71}$$

with full-column rank blocks $\tilde{\mathcal{N}}_{i,i+1}, i = 0,\ldots,\mu-2$.

It turns out that $\{E,F\}$ can be transformed into Weierstraß–Kronecker form with Kronecker index μ, and hence $\{E,F\}$ is a regular pair with Kronecker index μ. □

1.4 Characterizing matrix pencils by admissible projectors

Each regular pair of $m \times m$ matrices with Kronecker index $\mu \geq 1$ can be transformed into the Weierstraß–Kronecker form (cf. Section 1.1).

$$\left\{ \begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix}, \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix} \right\}, \quad J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_s \end{bmatrix},$$

where W is $d \times d$, J is $l \times l$, $d + l = m$, J_i is a nilpotent Jordan block of order k_i, $1 \leq k_i \leq \mu$, and $\max_{i=1,\ldots,s} k_i = \mu$.

As in Section 1.1, let l_i denote the number of all Jordan blocks of order $\geq i$. Then, J has $l_\mu \geq 1$ Jordan blocks of order μ, and $l_i - l_{i+1}$ Jordan blocks of order i, $i = 1,\ldots,\mu-1$, $l_1 + \cdots + l_\mu = l$.

In the present section we show how one can get all this structural information as well as the spectrum of $-W$, that is the finite spectrum of the given matrix pencil, by means of the matrix sequence and the admissible projectors without transforming the given pair into Weierstraß–Kronecker form.

Often the given matrix pair might have a large dimension m but a low Kronecker index μ so that just a few steps in the matrix sequence will do.

The proof of Theorem 1.31, and in particular formula (1.71), show that the detailed structure of the matrix pair can be described by means of admissible projectors. This is the content of the next corollary.

Corollary 1.32. *If $\{E,F\}, E,F \in L(\mathbb{R}^m)$, has the structural characteristic values $r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$, then the nilpotent part in its Weierstraß–Kronecker form contains altogether $s = m - r_0$ Jordan blocks, among them $r_i - r_{i-1}$ Jordan chains of order i, $i = 1,\ldots,\mu$. It holds that $l_i = m - r_{i-1}$, $i = 1,\ldots,\mu$, $d = m - \sum_{j=1}^{\mu}(m - r_{j-1})$.*

Besides the above structural characteristics the matrix sequence also provides the finite spectrum of the matrix pencil as a part of the spectrum of the matrix $\mathcal{W} := \Pi_{\mu-1}G_\mu^{-1}B$.

Theorem 1.33. *Let the pair $\{E,F\}$, $E,F \in L(\mathbb{R}^m)$, be regular with Kronecker index μ, and let the matrix*

$$\mathcal{W} := \Pi_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1} = \Pi_{\mu-1}G_\mu^{-1}B$$

be generated by an admissible matrix sequence $G_0,\dots,G_\mu$. Then the following assertions hold:

(1) *Each finite eigenvalue of $\{E,F\}$ belongs to the spectrum of $-\mathcal{W}$. More precisely, $(\lambda E+F)z=0$ with $z\neq 0$ implies $u := \Pi_{\mu-1}z \neq 0$, and $(\lambda I+\mathcal{W})u=0$.*
(2) *If $(\lambda I+\mathcal{W})u=0$, $\Pi_{\mu-1}u\neq 0$, then λ is a finite eigenvalue of the pair $\{E,F\}$.*
(3) *If $(\lambda I+\mathcal{W})u=0$, $(I-\Pi_{\mu-1})u\neq 0$, then $\lambda=0$ must hold. If, additionally, $\Pi_{\mu-1}u\neq 0$, then $\lambda=0$ is a finite eigenvalue of the pair $\{E,F\}$.*
(4) *$(\lambda I+\mathcal{W})u=0$, $u\neq 0$, $\lambda\neq 0$, implies $\Pi_{\mu-1}u=u$.*
(5) *If $Q_0,\dots,Q_{\mu-1}$ are completely decoupling projectors, then $\mathcal{W}$ simplifies to*

$$\mathcal{W} = G_\mu^{-1}B\Pi_{\mu-1} = G_\mu^{-1}B_\mu,$$

and $\Pi_{\mu-1}$ is the spectral projector of the matrix pair $\{E,F\}$.

Proof. Applying the decoupling procedure (see Subsection 1.2.2) we rewrite the equation $(\lambda E+F)z=0$, with

$$z = u+v_0+\cdots+v_{\mu-1}, \quad u := \Pi_{\mu-1}z, \quad v_0 := Q_0 z, \quad \dots, \quad v_{\mu-1} := \Pi_{\mu-2}Q_{\mu-1}z,$$

as the decoupled system

$$\lambda u + \mathcal{W}u = 0, \tag{1.72}$$

$$\lambda \begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} v_0 \\ \vdots \\ \vdots \\ v_{\mu-1} \end{bmatrix} + \begin{bmatrix} v_0 \\ \vdots \\ \vdots \\ v_{\mu-1} \end{bmatrix} = - \begin{bmatrix} \mathcal{H}_0 \\ \vdots \\ \vdots \\ \mathcal{H}_{\mu-1} \end{bmatrix} u. \tag{1.73}$$

Equation (1.73) leads to the representations

$$\begin{aligned} v_{\mu-1} &= -\mathcal{H}_{\mu-1}u, \\ v_{\mu-2} &= -\mathcal{H}_{\mu-2}u + \lambda\mathcal{N}_{\mu-2,\mu-1}\mathcal{H}_{\mu-1}u, \end{aligned}$$

and so on, showing the linear dependence of u, $v_i = \tilde{\mathcal{H}}_i u$, $i=0,\dots,\mu-1$. The property $\mathcal{H}_i = \mathcal{H}_i\Pi_{\mu-1}$ implies $\tilde{\mathcal{H}}_i = \tilde{\mathcal{H}}_i\Pi_{\mu-1}$.

If $z\neq 0$ then $u\neq 0$ must be true, since otherwise $u=0$ would imply $v_i=0$, $i=0,\dots,\mu-1$, and hence $z=0$. Consequently, λ turns out to be an eigenvalue of $-\mathcal{W}$ and $u=\Pi_{\mu-1}z$ is the corresponding eigenvector. This proves assertion (1).

To verify (2)–(4) we consider

$$(\lambda I + \mathcal{W})\tilde{u} = 0, \quad \tilde{u} = \Pi_{\mu-1}\tilde{u} + (I - \Pi_{\mu-1})\tilde{u} \neq 0.$$

Because $\mathcal{W}(I - \Pi_{\mu-1}) = 0$ and $(I - \Pi_{\mu-1})\mathcal{W} = 0$, our equation decomposes into the following two equations:

$$\lambda(I - \Pi_{\mu-1})\tilde{u} = 0, \quad (\lambda I + \mathcal{W})\tilde{u} = 0. \tag{1.74}$$

Next, if $\Pi_{\mu-1}\tilde{u} \neq 0$ then we put $\tilde{v}_i := \tilde{\mathcal{H}}_i\tilde{u} = \tilde{\mathcal{H}}_i\Pi_{\mu-1}\tilde{u}$, $i = 0, \ldots, \mu-1$. Thus, $\tilde{z} := \Pi_{\mu-1}\tilde{u} + \tilde{v}_0 + \cdots + \tilde{v}_{\mu-1}$ is nontrivial, and it satisfies the condition $(\lambda E + F)\tilde{z} = 0$, and so assertion (2) holds true. Furthermore, if $(I - \Pi_{\mu-1})\tilde{u} \neq 0$, then the first part of (1.74) yields $\lambda = 0$. Together with (2) this validates (3). Assertion (4) is a simple consequence of (1.74).

It remains to show assertion (5). Compute

$$\begin{aligned} G_\mu^{-1}B_\mu - \Pi_{\mu-1}G_\mu^{-1}B_\mu &= (I - \Pi_{\mu-1})G_\mu^{-1}B\Pi_{\mu-1} \\ &= (Q_{\mu-1} + Q_{\mu-2}P_{\mu-1} + \cdots + Q_0P_1\cdots P_{\mu-1})G_\mu^{-1}B\Pi_{\mu-1} \\ &= Q_{\mu-1}\Pi_{\mu-1} + Q_{\mu-2}\Pi_{\mu-1} + \cdots + Q_0\Pi_{\mu-1} = 0. \end{aligned}$$

For the proof that $\Pi_{\mu-1}$ is the spectral projector we refer to [164]. □

The matrix $\mathcal{W} = \Pi_{\mu-1}G_\mu^{-1}B = \Pi_{\mu-1}G_\mu^{-1}B\Pi_{\mu-1}$ resulting from the projector based decoupling procedure contains the finite spectrum of the pencil $\{E, F\}$. The spectrum of $-\mathcal{W}$ consists of the d finite eigenvalues of the pencil $\{E, F\}$ plus $m - d = l$ zero eigenvalues corresponding to the subspace $\operatorname{im}(I - \Pi_{\mu-1}) \subseteq \ker \mathcal{W}$. The eigenvectors corresponding to the nonzero eigenvalues of $\mathcal{W}$ necessarily belong to the subspace $\operatorname{im}\Pi_{\mu-1}$.

Now we have available complete information concerning the structure of the Weierstraß–Kronecker form without computing that form itself. All this information is extracted from the matrix sequence (1.10). Notice that several numerical algorithms to compute the matrix sequence and admissible projectors are addressed in Chapter 7. Using the matrix sequence (1.10), the following characteristics of the matrix pair E, F are obtained:

- $d = m - \sum_{j=1}^{\mu}(m - r_{j-1})$, $l = m - d$ are the basic structural sizes and μ is the Kronecker index,
- $r_{i+1} - r_i$ is the number of Jordan blocks with dimension $i+1$ in the nilpotent part,
- $m - r_i$ is the number of Jordan blocks with dimension $\geq i+1$ in the nilpotent part,
- the finite eigenvalues as described in Theorem 1.33.

There is also an easy regularity criterion provided by the matrix sequence (1.10).

Proposition 1.34. *The pair $\{E, F\}$, $E, F \in L(\mathbb{R}^m)$, is singular if and only if there is a nontrivial subspace among the intersections*

$$N_i \cap N_{i-1}, \quad \widehat{N}_i = N_i \cap (N_0 + \cdots + N_{i-1}), \quad i \geq 1. \tag{1.75}$$

Proof. Owing to the basic property (1.11) and Proposition 1.13, each nontrivial subspace among (1.75) indicates a singular pencil. Conversely, let $\{E,F\}$ be singular. Then all matrices G_i must be singular, their nullspaces N_i have dimensions ≥ 1 and the ranks satisfy the inequality

$$r_0 \leq \cdots \leq r_i \leq \cdots\cdots \leq m-1.$$

There is a maximal rank $r_{max} \leq m-1$ and an integer κ such that $r_i = r_{max}$ for all $i \geq \kappa$. If all above intersections (1.75) are trivial, then it follows that

$$N_0 + \cdots + N_i = N_0 \oplus \cdots \oplus N_i, \quad \dim(N_0 \oplus \cdots \oplus N_i) \geq i+1.$$

However, this contradicts the natural property $N_0 + \cdots + N_i \subseteq \mathbb{R}^m$. □

1.5 Properly stated leading term and solution space

Which kind of solutions is natural for the linear DAE

$$Ex'(t) + Fx(t) = q(t), \tag{1.76}$$

with coefficients $E,F \in L(\mathbb{R}^m)$? Let $\mathcal{I} \subseteq \mathbb{R}$ denote the interval of interest, and let q be at least continuous on $\mathcal{I}$. Should we seek continuously differentiable solutions? The trivial Example 1.35 below reveals that continuous solutions having certain continuously differentiable components seem to be more natural. How can we decide which part of the solution should be continuous only?

By means of any factorization of the leading matrix $E = AD$ with factors $A \in L(\mathbb{R}^n, \mathbb{R}^m))$, $D \in L(\mathbb{R}^m, \mathbb{R}^n)$, the DAE (1.76) can be formally written as

$$A(Dx(t))' + Fx(t) = q(t), \tag{1.77}$$

which suggests seek solutions $x(.)$ having a continuously differentiable part $Dx(.)$. Introduce the function space of relevant functions by

$$\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) := \{x \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m) : Dx \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)\}. \tag{1.78}$$

We emphasize that the structural characteristic values as well as the admissible matrix sequences $(G_i)_{i\geq 0}$ in the previous sections are independent of the special factorization of E since the initial guess of any matrix sequence is $G_0 := E = AD$. The trivial factorization $A = E$, $D = I$ corresponds to the standard form DAE (1.76) itself and make sense, if E is nonsingular. Our goal is to find nontrivial factorizations which reflect possible low-smoothness demands for the solutions.

Example 1.35. Consider the simple system

$$\begin{aligned} x_1'(t) + x_2(t) &= q_1(t), \\ x_2'(t) + x_1(t) &= q_2(t), \\ x_3(t) &= q_3(t), \\ x_4(t) &= q_4(t), \end{aligned}$$

that takes the form (1.76) with

$$E = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad F = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Choosing in the resulting equation (1.77) the factors $A = E$ and $D = \operatorname{diag}(1,1,1,0)$, we are led to the understanding that all first three components $x_1(.), x_2(.), x_3(.)$ should be continuously differentiable. However, a look to the detailed DAE shows that there is no reason to expect the third component to be so smooth. Observe that this matrix D has rank three, while E has rank two, and $\ker D$ is a proper subspace of $\ker E$. Choosing instead $A = I$, $D = E$ or $A = E$, $D = E$ we obtain $\ker D = \ker E$, and a further look at the detailed DAE confirms that now the space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ reflects a natural understanding of the solution.

As we can see in this example, aiming for lowest appropriate smoothness demands and the notion of a natural solution, the rank of D has to be as low as possible, i.e., the dimension of the nullspace $\ker D$ has to be maximal, that is, we are led to the condition $\ker D = \ker E$.

In general, if $E = AD$ and $\ker D = \ker E$, then the intersection $\ker A \cap \operatorname{im} D$ is trivial, and the sum of these two subspaces is a direct sum. Namely, from $z \in \ker A \cap \operatorname{im} D$, that is, $Az = 0$, $z = Dw$, we conclude $Ew = ADw = Az = 0$, and hence $w \in \ker E = \ker D$, thus $z = Dw = 0$.

Moreover, from $E = AD$, $\ker D = \ker E$, it follows that D and E join their rank but A may have a greater one. The direct sum $\ker A \oplus \operatorname{im} D$ may become a proper subspace of $\mathbb{R}^n$. In the particular case of Example 1.35 the choice $A = I, D = E$ leads to $\ker A \oplus \operatorname{im} D = \operatorname{im} E$ being a two-dimensional subspace of $\mathbb{R}^4$.

There is much freedom in choosing the factorizations $E = AD$. We can always arrange things in such a way that $\ker D = \ker E$ and $\operatorname{im} A = \operatorname{im} E$, and hence all three matrices E, A and D have the same rank. Then, the decomposition

$$\ker A \oplus \operatorname{im} D = \mathbb{R}^n \tag{1.79}$$

is valid. Particular factorizations satisfying condition (1.79) are given by means of reflexive generalized inverses E^- and the accompanying projectors $EE^- \in L(\mathbb{R}^k)$, $E^-E \in L(\mathbb{R}^m)$ (cf. Appendix A.2). Namely, with $A = EE^-$, $D = E$, $n = k$, we have $AD = EE^-E = E$ and

$$\ker A \oplus \operatorname{im} D = \ker EE^- \oplus \operatorname{im} E = \ker EE^- \oplus \operatorname{im} EE^- = \mathbb{R}^n.$$

Similarly, letting $A = E$, $D = E^- E$, $n = m$, we obtain

$$\ker A \oplus \operatorname{im} D = \ker E \oplus \operatorname{im} E^- E = \ker E^- E \oplus \operatorname{im} E^- E = \mathbb{R}^n.$$

We refer to Chapter 7 for computational aspects to factorize $E = AD$. We mention just the full rank factorization, that is, both A and D are full rank matrices, by means of a singular value decomposition

$$E = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix} \begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}^* = \underbrace{\begin{bmatrix} U_{11} \\ U_{21} \end{bmatrix}}_{A} \Sigma \underbrace{\begin{bmatrix} V_{11}^* & V_{21}^* \end{bmatrix}}_{D}, \tag{1.80}$$

$\operatorname{rank} \Sigma = \operatorname{rank} E =: r$, $n = r$. The matrix A has full column rank n and D has full row rank n. Later on we shall understand the property $\ker A = 0$, $\operatorname{im} D = \mathbb{R}^n$ to be preferable, in particular for numerical integration methods.

Definition 1.36. The matrices $A \in L(\mathbb{R}^n, \mathbb{R}^k)$ and $D \in L(\mathbb{R}^m, \mathbb{R}^n)$ are said to be *well matched* if the subspaces $\ker A$ and $\operatorname{im} D$ are transversal so that decomposition (1.79) is valid. Equation (1.77) is a *DAE with properly stated leading term* if A and D are well matched.

Given a DAE (1.77) with properly stated leading term, we look for solutions belonging to the function space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. We might be interested in a different formulation with $AD = \bar{A}\bar{D}$. Also, starting with a standard formulation we have to decide on the factorization. So we are confronted with the question of whether the solution space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ depends on the factorization. The following lemma shows the independence.

Lemma 1.37. *If the matrices $D \in L(\mathbb{R}^m, \mathbb{R}^n)$ and $\bar{D} \in L(\mathbb{R}^m, \mathbb{R}^{\bar{n}})$ have a common nullspace $N := \ker D = \ker \bar{D}$, then the corresponding function spaces coincide, that is*

$$\mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m) = \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m).$$

Proof. The orthoprojector $P \in L(\mathbb{R}^m)$ onto $N^\perp$ satisfies $P = D^+ D = \bar{D}^+ \bar{D}$ (cf. Appendix A.2). Therefore, for any $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, we find $\bar{D}x = \bar{D}\bar{D}^+ \bar{D}x = \bar{D}D^+ Dx \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{\bar{n}})$, and hence $x \in \mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m)$. □

1.6 Notes and references

(1) As we have seen in this chapter, the Weierstraß–Kronecker form of a regular matrix pencil is very helpful for understanding the structure of a linear constant coefficient DAE, and, obviously, DAEs and matrix pencils are closely related.

Ever since Weierstraß and Kronecker ([212, 126]) discovered the canonical forms of matrix pencils, and Gantmacher ([81]) pointed out their connection with differential equations, matrix pencils have attracted much interest over and over again for

many years. There are untold publications on this topic; we only mention a few of them and refer to the sources therein.

(2) A large part of the developments concerning matrix pencils and the accompanying differential equations can be found in the rich literature on control and system theory, where the resulting differential equations are called *singular systems* and *descriptor systems* rather than DAEs (e.g. [37, 56, 147, 150]).

On the other hand, there are important contributions coming from the area of generalized eigenvalue problems and generalized matrix inverses in linear algebra (e.g. [38, 21]). In particular, the *Drazin inverse* and spectral projections were applied to obtain expressions for the solution (cf. also [96]). However, it seems that this was a blind alley in the search for a possible treatment of more general DAEs.

(3) About half a century ago, Gantmacher ([81]) and Dolezal [60] first considered models describing linear time-invariant mechanical systems and electrical circuits by linear constant coefficient DAEs. Today, multibody systems and circuit simulation represent the most traditional DAE application fields (e.g. [63, 78, 101]). In between, in about 1980, due to unexpected phenomena in numerical computations (e.g. [202, 180]), DAEs (descriptor systems) became an actual and challenging topic in applied mathematics

(4) It should be mentioned that there are different perceptions concerning the *Weierstraß–Kronecker form* of a regular matrix pencil. For instance, the version applied here is said to be *quasi-Weierstraß form* in [18] and *Kronecker normal form* in [191].

(5) Unfortunately, the transformation to Weierstraß–Kronecker form as well as the Drazin inverse approaches do not allow for modifications appropriate to the treatment of time-varying and nonlinear DAEs. A development with great potential for suitable generalizations is given by the derivative array approach due to Campbell ([41]). Following this proposal, we consider, in addition to the given DAE

$$Ex'(t) + Fx(t) = q(t), \tag{1.81}$$

the extended system

$$\underbrace{\begin{bmatrix} E & 0 & \dots & 0 \\ F & E & 0 & \dots \\ 0 & F & E & \dots \\ \dots & \dots & \dots & \dots \\ \dots & \dots & F & E \end{bmatrix}}_{\mathcal{E}_\mu} \begin{bmatrix} x'(t) \\ x''(t) \\ \dots \\ x^{\mu+1}(t) \end{bmatrix} = - \begin{bmatrix} F \\ 0 \\ \dots \\ 0 \end{bmatrix} x(t) + \begin{bmatrix} q(t) \\ q'(t) \\ \dots \\ q^{(\mu)(t)} \end{bmatrix}, \tag{1.82}$$

which results from (1.81) by differentiating this equation μ times and collecting all these equations. If the $(\mu+1) \times m$ matrix $\mathcal{E}_\mu$ is 1-full, or in other words, if there exists a nonsingular matrix $\mathcal{R}$ such that

$$\mathcal{R}\mathcal{E}_\mu = \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{K} \end{bmatrix},$$

then an explicit ODE, the *completion ODE*, can be extracted from the derivative array system (1.82), say

$$x'(t) = \mathcal{C}x(t) + \sum_{j=0}^{\mu} \mathcal{D}_j q^{(j)}(t). \tag{1.83}$$

The solutions of the DAE (1.81) are embedded into the solutions of the explicit ODE (1.83). If $\{E,F\}$ forms a regular matrix pair with Kronecker index μ, then $\mathcal{E}_\mu$ is 1-full (cf. [40]). Conversely, if μ is the smallest index such that $\mathcal{E}_\mu$ is 1-full, then $\{E,F\}$ is regular with Kronecker index μ. In this context, applying our sequence of matrices built using admissible projectors, we find that the 1-fullness of $\mathcal{E}_\mu$ implies that G_μ is nonsingular, and then using completely decoupling projectors, we obtain a special representation of the scaling matrix $\mathcal{R}$. We demonstrate this just for $\mu = 1,\ 2$.

Case $\mu = 1$: Let $\mathcal{E}_1$ be 1-full, and consider z with $G_1 z = 0$, i.e., $Ez + FQ_0 z = 0$, and so

$$\begin{bmatrix} E & 0 \\ F & E \end{bmatrix} \begin{bmatrix} Q_0 z \\ z \end{bmatrix} = 0,$$

but then, due to the 1-fullness, it follows that $Q_0 z = 0$. This, in turn, gives $Ez = 0$ and then $z = 0$. Therefore, G_1 is nonsingular. Taking the completely decoupling projector Q_0 such that $Q_0 = Q_0 G_1^{-1} F$ holds true, we obtain

$$\underbrace{\begin{bmatrix} P_0 & Q_0 \\ -P_0 G_1^{-1} F & P_0 \end{bmatrix} \begin{bmatrix} G_1^{-1} & 0 \\ 0 & G_1^{-1} \end{bmatrix}}_{\mathcal{R}} \begin{bmatrix} E & 0 \\ F & E \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & P_0 \end{bmatrix}. \tag{1.84}$$

Case $\mu = 2$: Let $\mathcal{E}_2$ be 1-full, and consider z with $G_2 z = 0$, i.e. $Ez + FQ_0 z + FP_0 Q_1 z = 0$. Because $(E + FQ_0)Q_1 = G_1 Q_1 = 0$ we find that $E(Q_0 + P_0 Q_1)z = EQ_1 z = -FQ_0 Q_1 z$, and therefore

$$\begin{bmatrix} E & 0 & 0 \\ F & E & 0 \\ 0 & F & E \end{bmatrix} \begin{bmatrix} Q_0 Q_1 z \\ (Q_0 + P_0 Q_1)z \\ z \end{bmatrix} = 0.$$

Now, the 1-fullness of $\mathcal{E}_2$ implies $Q_0 Q_1 z = 0$, but this yields $EP_0 Q_1 = 0$, so that $P_0 Q_1 z = 0$, and therefore $Q_1 z = 0$ and $FQ_0 z + Ez = 0$. Finally, we conclude that $z = Q_1 z = 0$, which means that G_2 is nonsingular. With completely decoupling projectors Q_0, Q_1 we compute

$$\begin{bmatrix} P_0 P_1 & Q_0 P_1 + P_0 Q_1 & Q_0 Q_1 \\ Q_0 P_1 + P_0 Q_1 & Q_0 Q_1 & P_0 P_1 \\ -P_0 P_1 G_2^{-1} F & P_0 P_1 & P_0 Q_1 \end{bmatrix} \begin{bmatrix} G_2^{-1} & 0 & 0 \\ 0 & G_2^{-1} & 0 \\ 0 & 0 & G_2^{-1} \end{bmatrix} =: \mathcal{R},$$

$$\mathcal{R}\begin{bmatrix} E & 0 & 0 \\ F & E & 0 \\ 0 & F & E \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & P_0P_1G_2^{-1}F & P_0P_1 \\ 0 & P_0 & 0 \end{bmatrix}.$$

The resulting completion ODE (cf. (1.83)) is

$$\begin{aligned} x'(t) + P_0P_1G_2^{-1}Fx(t) = \\ P_0P_1G_2^{-1}q(t) + (Q_0P_1 + P_0Q_1)G_2^{-1}q'(t) + Q_0Q_1G_2^{-1}q''(t), \end{aligned} \tag{1.85}$$

and it decomposes into the three parts

$$\begin{aligned} P_0P_1x'(t) + P_0P_1G_2^{-1}FP_0P_1x(t) &= P_0P_1G_2^{-1}q(t), \\ P_0Q_1x'(t) &= P_0Q_1G_2^{-1}q'(t), \\ Q_0x'(t) &= P_0Q_1G_2^{-1}q'(t) + Q_0Q_1G_2^{-1}q''(t), \end{aligned}$$

while the decoupling procedure described in Section 1.5 yields

$$\begin{aligned} (P_0P_1x)'(t) + P_0Q_1G_2^{-1}FP_0P_1x(t) &= P_0P_1G_2^{-1}q(t), \\ P_0Q_1x(t) &= P_0Q_1G_2^{-1}q(t), \\ Q_0x(t) &= P_0Q_1G_2^{-1}q(t) + Q_0Q_1(P_0Q_1G_2^{-1}q)'(t). \end{aligned}$$

A comparison shows consistency but also differences. In order to recover the DAE solutions from the solutions of the explicit ODE (1.85) one obviously needs consistent initial values. Naturally, more smoothness has to be given when using the derivative array and the completion ODE. Applying derivative array approaches to time-varying linear or nonlinear DAEs one has to ensure the existence of all the higher derivatives occurring when differentiating the original DAE again and again, and in practice one has to provide these derivatives.

(6) The matrix sequence (1.10) for a DAE was first introduced in [156], and part of the material is included in [97]. Completely decoupling projectors, formerly called *canonical projectors*, are provided in [164], and they are applied in Lyapunov type stability criteria, e.g., in [162, 165].

In these earlier papers, the sum spaces $N_0 + \cdots + N_j$ do not yet play their important role as they do in the present material. The central role of these sum spaces is pointed out in [170] where linear time-varying DAEs are analyzed. In the same paper, admissible projectors are introduced for regular DAEs only, which means that trivial intersections $\widehat{N}_i$ are supposed. The present notion of admissible projectors generalizes the previous definition and accepts nontrivial intersections $\widehat{N}_i$. This allows us to discuss also nonregular DAEs, in particular so-called rectangular systems, where the number of equations and the number of unknowns are different.

(7) The projector based decoupling procedure is not at all restricted to square matrices. Although our interest mainly concerns regular DAEs, to be able to con-

sider several aspects of nonregular DAEs one can construct the admissible matrix sequences $(G_i)_{i\geq 0}$ and accompanying projectors $(Q_i)_{i\geq 0}$ in the same way also for ordered pairs $\{E,F\}$ of rectangular matrices $E,F \in L(\mathbb{R}^m,\mathbb{R}^k)$. We address these problems in Chapter 10 on nonregular DAEs.

(8) The Kronecker index is, from our point of view, the most adequate characteristic of a matrix pencil and the associated DAE. In contrast, the widely used *structural index* does not necessarily provide the Kronecker index. This structural index may be arbitrarily higher and also far less than the Kronecker index (see [209, 18]).

(9) Complete decouplings are used to calculate the spectral projector for descriptor systems in an efficient manner (see [213]).

Chapter 2
Linear DAEs with variable coefficients

In this chapter we provide a comprehensive analysis of linear DAEs

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

with properly stated leading term, by taking up the ideas of the projector based decoupling described for constant coefficient DAEs in Chapter 1. To handle the time-varying case, we proceed pointwise on the given interval and generate admissible sequences of matrix functions $G_i(\cdot) = G_{i-1}(\cdot) + B_{i-1}(\cdot)Q_{i-1}(\cdot)$ associated with admissible projector functions $Q_i(\cdot)$, instead of the former admissible matrix sequences and projectors. Thereby we incorporate into the matrix function $B_i(\cdot)$ an additional term that comprises the variations in time. This term is the crucial one of the generalization. Without this term we would be back to the so-called *local matrix pencils* which are known to be essentially inappropriate to characterize time-varying DAEs (e.g., [25, 96]). Aside from the higher technical content in the proofs, the projector based decoupling applies in precisely the same way as for constant coefficient DAEs, and fortunately, most results take the same or only slightly modified form.

In contrast to Chapter 1 which is restricted to square DAE systems, that means, the number of unknowns equals the number of equations, the present chapter is basically valid for systems of k equations and m unknowns. Following the arguments e.g., in [130], so-called rectangular systems may play their role in optimization and control. However, we emphasize that our main interest is directed to regular DAEs, with $m = k$ by definition. Nonregular DAEs, possibly with $m \neq k$, are discussed in more detail in Chapter 10.

We introduce in Section 2.1 the DAEs with properly stated leading term and describe in Section 2.2 our main tools, the admissible matrix function sequences associated to admissible projector functions and characteristic values. Widely orthogonal projector functions in Subsection 2.2.3 form a practically important particular case. The analysis of invariants in Section 2.3 serves as further justification of the concept.

The main objective of this chapter is the comprehensive characterization of *regular DAEs*, in particular, in their decoupling into an *inherent regular explicit ODE* and

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_2,

a subsystem which comprises the *inherent differentiations*. We consider the constructive existence proof of *fine* and *complete* decouplings (Theorem 2.42) to be the most important special result which describes the DAE structure as the basis of our further investigations. This leads to the intrinsic DAE theory in Section 2.6 offering solvability results, flow properties, and the *T-canonical form*. The latter appears to be an appropriate generalization of the Weierstraß–Kronecker form. Several specifications for *regular standard form DAEs* are recorded in Subsection 2.7. Section 2.9 reflects aspects of the critical point discussion and emphasizes the concept of *regularity intervals*.
In Section 2.10 we explain by means of canonical forms and reduction steps how the *strangeness* and the tractability index concepts are related to each other.

2.1 Properly stated leading terms

We consider the equation

$$A(Dx)' + Bx = q, \tag{2.1}$$

with continuous coefficients

$$A \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^k)), \quad D \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n)), \quad B \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^k)),$$

and the excitation $q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^k)$, where $\mathcal{I} \in \mathbb{R}$ is an interval. A solution of this equation is a function belonging to the function space

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

which satisfies the DAE in the classical sense, that is, pointwise on the given interval.

The two coefficient functions A and D are to figure out precisely all those components of the unknown function, the first derivatives of which are actually involved in equation (2.1). For this, A and D are supposed to be well matched in the sense of the following definition, which roughly speaking means that there is no gap and no overlap of the factors within the product AD and the border between A and D is smooth.

Definition 2.1. The leading term in equation (2.1) is said to be *properly stated* on the interval $\mathcal{I}$, if the transversality condition

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \; t \in \mathcal{I}, \tag{2.2}$$

is valid and the projector valued function $R : \mathcal{I} \to L(\mathbb{R}^n)$ defined by

$$\operatorname{im} R(t) = \operatorname{im} D(t), \; \ker R(t) = \ker A(t), \quad t \in \mathcal{I},$$

is continuously differentiable.

The projector function $R \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n))$ is named the *border projector* of the leading term of the DAE.

To shorten the phrase *properly stated leading term*, sometimes we speak of *proper leading terms*.
We explicitly point out that, in a proper leading term, both involved matrix functions A and D have necessarily constant rank. This is a consequence of the smoothness of the border projector R (see Lemma A.14).

Applying the notion of $\mathcal{C}^1$-subspaces (Definition A.19, Appendix A), a proper leading term is given, exactly if $\operatorname{im} D$ and $\ker A$ are transversal $\mathcal{C}^1$-subspaces. Equivalently (see Lemma A.14), one has a proper leading term, if condition (2.2) is satisfied and there are basis functions $\vartheta_i \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$, $i = 1, \ldots, n$, such that

$$\operatorname{im} D(t) = \operatorname{span}\{\vartheta_1(t), \ldots, \vartheta_r(t)\}, \quad \ker A(t) = \operatorname{span}\{\vartheta_{r+1}(t), \ldots, \vartheta_n(t)\}, \quad t \in \mathcal{I}.$$

Having those basis functions available, the border projector R can simply be represented as

$$R := [\vartheta_1 \ldots \vartheta_n] \begin{bmatrix} \underbrace{I}_{r} & \\ & 0 \end{bmatrix} [\vartheta_1 \ldots \vartheta_n]^{-1}. \tag{2.3}$$

If A and D form a properly stated leading term, then the relations

$$\operatorname{im} AD = \operatorname{im} A, \ \ \ker AD = \ker D, \ \ \operatorname{rank} A = \operatorname{rank} AD = \operatorname{rank} D =: r$$

are valid (cf. Lemma A.4), and A, AD and D have *common constant rank* r on $\mathcal{I}$.

Besides the coefficients A, D and the projector R we use a pointwise generalized inverse $D^- \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^m))$ of D satisfying the relations

$$DD^-D = D, \ \ D^-DD^- = D^-, \ \ DD^- = R. \tag{2.4}$$

Such a generalized inverse exists owing to the constant-rank property of D. Namely, the orthogonal projector P_D onto $\ker D^\perp$ along $\ker D$ is continuous (Lemma A.15). If we added the fourth condition $D^-D = P_D$ to (2.4), then the resulting D^- would be uniquely determined and continuous (Proposition A.17), and this ensures the existence of a continuous generalized inverses satisfying (2.4).

By fixing only the three conditions (2.4), we have in mind some more flexibility. Here $D^-D =: P_0$ is always a continuous projector function such that $\ker P_0 = \ker D = \ker AD$. On the other hand, prescribing P_0 we fix, at the same time, D^-.

Example 2.2 (Different choices of P_0 and D^-). Write the semi-explicit DAE

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

with $m_1 + m_2 = m$ equations in the form (2.1) with properly stated leading term as

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; D = \begin{bmatrix} I & 0 \end{bmatrix}, \; B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

such that $\ker A = \{0\}$, $\operatorname{im} D = \mathbb{R}^{m_1}$ and $R = I$. Any continuous projector function P_0 along $\ker D$ and the corresponding generalized inverse D^- have the form

$$P_0 = \begin{bmatrix} I & 0 \\ \mathfrak{A} & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} I \\ \mathfrak{A} \end{bmatrix},$$

with an arbitrary continuous block $\mathfrak{A}$. The choice $\mathfrak{A} = 0$ yields the symmetric projector P_0. □

2.2 Admissible matrix function sequences

2.2.1 Basics

Now we are ready to compose the basic sequence of matrix functions and subspaces to work with. Put

$$G_0 := AD, \;\; B_0 := B, \;\; N_0 := \ker G_0 \tag{2.5}$$

and choose projector functions $P_0, Q_0, \Pi_0 \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$ such that

$$\Pi_0 = P_0 = I - Q_0, \;\; \operatorname{im} Q_0 = N_0.$$

For $i \geq 0$, as long as the expressions exist, we form

$$G_{i+1} = G_i + B_i Q_i, \tag{2.6}$$
$$N_{i+1} = \ker G_{i+1}, \tag{2.7}$$

choose projector functions P_{i+1}, Q_{i+1} such that $P_{i+1} = I - Q_{i+1}$, $\operatorname{im} Q_{i+1} = N_{i+1}$, and put

$$\begin{aligned} \Pi_{i+1} &:= \Pi_i P_{i+1}, \\ B_{i+1} &:= B_i P_i - G_{i+1} D^- (D\Pi_{i+1} D^-)' D\Pi_i. \end{aligned} \tag{2.8}$$

We emphasize that B_{i+1} contains the derivative of $D\Pi_{i+1}D^-$, that is, this term comprises the variation in time. This term disappears in the constant coefficient case, and then we are back at the formulas (1.10) in Chapter 1. The specific form of the new term is motivated in Section 2.4.1 below, where we consider similar decoupling rearrangements for the DAE (2.1) as in Chapter 1 for the constant coefficient case.
We are most interested in continuous matrix functions G_{i+1}, B_{i+1}; in particular we have to take that $D\Pi_{i+1}D^-$ is smooth enough.
Important characteristic values of the given DAE emerge from the rank functions

$$r_j := \operatorname{rank} G_j, \;\; j \geq 0.$$

Example 2.3 (Matrix functions for Hessenberg size-1 and size-2 DAEs). Write the semi-explicit DAE

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

with $m_1 + m_2 = m$ equations in the form (2.1) as

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; D = \begin{bmatrix} I & 0 \end{bmatrix}, \; B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, \; D^- = \begin{bmatrix} I \\ 0 \end{bmatrix}.$$

Then we have a proper leading term and

$$G_0 = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \; G_1 = \begin{bmatrix} I & B_{12} \\ 0 & B_{22} \end{bmatrix}.$$

Case 1:
Let B_{22} be nonsingular on the given interval. Then G_1 is also nonsingular. It follows that $Q_1 = 0$, thus $G_2 = G_1$ and so on. The sequence becomes stationary. All rank functions r_i are constant, in particular $r_0 = m_1$, $r_1 = m$.
Case 2:
Let $B_{22} = 0$, but the product $B_{21}B_{12}$ remains nonsingular. We denote by Ω a projector function onto $\operatorname{im} B_{12}$, and by B_{12}^- a reflexive generalized inverse such that $B_{12}B_{12}^- = \Omega$, $B_{12}^- B_{12} = I$. The matrix function G_1 now has rank $r_1 = m_1$, and a nontrivial nullspace. We choose the next projector functions Q_1 and the resulting $D\Pi_1 D^-$ as

$$Q_1 = \begin{bmatrix} \Omega & 0 \\ -B_{12}^- & 0 \end{bmatrix}, \quad D\Pi_1 D^- = I - \Omega.$$

This makes it clear that, for a continuously differentiable $D\Pi_1 D^-$, we have to assume the range of B_{12} to be a C^1-subspace (cf. A.4). Then we form the matrix functions

$$B_1 = \begin{bmatrix} B_{11} & 0 \\ B_{21} & 0 \end{bmatrix} - \begin{bmatrix} -\Omega' & 0 \\ 0 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} I + (B_{11} + \Omega')\Omega & B_{12} \\ B_{21}\Omega & 0 \end{bmatrix},$$

and consider the nullspace of G_2.
$G_2 z = 0$ means

$$z_1 + (B_{11} + \Omega')\Omega z_1 + B_{12}z_2 = 0, \quad B_{21}\Omega z_1 = 0.$$

The second equation means $B_{21}B_{12}B_{12}^- z_1 = 0$, thus $B_{12}^- z_1 = 0$, and hence $\Omega z_1 = 0$. Now the first equation simplifies to $z_1 + B_{12}z_2 = 0$. Multiplication by B_{12}^- gives $z_2 = 0$, and then $z_1 = 0$. Therefore, the matrix function G_2 is nonsingular, and again

the sequence becomes stationary.
Up to now we have not completely fixed the projector function Ω onto $\operatorname{im} B_{12}$. In particular, we can take the orthoprojector function such that $\Omega = \Omega^*$ and $\ker \Omega = \ker B_{12}^* = \operatorname{im} B_{12}^{\perp}$, which corresponds to $B_{12}^- = B_{12}^+ = (B_{12}^* B_{12})^{-1} B_{12}^*$ and

$$\ker Q_1 = \{z \in \mathbb{R}^{m_1+m_2} : B_{12}^* z_1 = 0\}.$$

□

Example 2.4 (Matrix functions for a transformed regular index-3 matrix pencil). The constant coefficient DAE

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{\bar{E}} \bar{x}'(t) + \bar{x}(t) = q(t), \quad t \in \mathbb{R},$$

has Weierstraß–Kronecker canonical form, and its matrix pencil $\{\bar{E}, I\}$ is regular with Kronecker index 3. By means of the simple factorization

$$\bar{E} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} =: \bar{A}\bar{D}$$

we rewrite the leading term properly as

$$\bar{A}(\bar{D}\bar{x}(t))' + \bar{x}(t) = q(t), \quad t \in \mathbb{R}.$$

Then we transform $\bar{x}(t) = K(t)x(t)$ by means of the smooth matrix function K,

$$K(t) := \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -t & 1 \end{bmatrix}, \quad t \in \mathbb{R},$$

being everywhere nonsingular. This yields the new DAE

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{\tilde{A}} \Big(\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -t & 1 \end{bmatrix}}_{\tilde{D}(t)} x(t) \Big)' + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -t & 1 \end{bmatrix}}_{\tilde{B}(t)} x(t) = q(t), \quad t \in \mathbb{R}. \tag{2.9}$$

Next we reformulate the DAE once again by deriving

$$(\tilde{D}(t)x(t))' = (\tilde{D}(t) \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t))' = \tilde{D}(t) \left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t) \right)' + \tilde{D}'(t) \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t),$$

which leads to the further equivalent DAE

$$\underbrace{\begin{bmatrix}0&1&0\\0&-t&1\\0&0&0\end{bmatrix}}_{A(t)}\left(\underbrace{\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}}_{D}x(t)\right)'+\underbrace{\begin{bmatrix}1&0&0\\0&0&0\\0&-t&1\end{bmatrix}}_{B(t)}x(t)=q(t),\quad t\in\mathbb{R}. \tag{2.10}$$

Observe that the local matrix pencil $\{A(t)D,B(t)\}$ is singular for all $t\in\mathbb{R}$.
We construct a matrix function sequence for the DAE (2.10). The DAE is expected to be regular with index 3, as its equivalent constant coefficient counterpart. A closer look to the solutions strengthens this expectation. We have

$$A(t)=\begin{bmatrix}0&1&0\\0&-t&1\\0&0&0\end{bmatrix},\ D(t)=\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix},\ B(t)=\begin{bmatrix}1&0&0\\0&0&0\\0&-t&1\end{bmatrix},\ G_0(t)=\begin{bmatrix}0&1&0\\0&-t&1\\0&0&0\end{bmatrix},$$

and $R(t)=D(t)$. Set $D(t)^-=D(t)$ and $\Pi_0(t)=P_0(t)=D(t)$. Next we compute $G_1(t)=G_0(t)+B(t)Q_0(t)$ as well as a projector $Q_1(t)$ onto $\ker G_1(t)=N_1(t)$:

$$G_1(t)=\begin{bmatrix}1&1&0\\0&-t&1\\0&0&0\end{bmatrix},\quad Q_1(t)=\begin{bmatrix}0&-1&0\\0&1&0\\0&t&0\end{bmatrix}.$$

This leads to

$$\Pi_1(t)=\begin{bmatrix}0&0&0\\0&0&0\\0&-t&1\end{bmatrix},\ B_1(t)=\begin{bmatrix}0&0&0\\0&1&0\\0&-t&1\end{bmatrix},\ G_2(t)=\begin{bmatrix}1&1&0\\0&1-t&1\\0&0&0\end{bmatrix}.$$

A suitable projector function Q_2 and the resulting B_2 and G_3 are:

$$Q_2(t)=\begin{bmatrix}0&-t&1\\0&t&-1\\0&-t(1-t)&1-t\end{bmatrix},\ \Pi_2(t)=0,\ B_2(t)=\begin{bmatrix}0&0&0\\0&0&0\\0&-t&1\end{bmatrix},\ G_3(t)=\begin{bmatrix}1&1&0\\0&1-t&1\\0&-t&1\end{bmatrix}.$$

The matrix functions G_i, $i=0,1,2$, are singular with constant ranks, and G_3 is the first matrix function that is nonsingular. Later on, this turns out to be typical for regular index-3 DAEs (cf. Definition 10.2), and meets our expectation in comparison with the constant coefficient case (cf. Theorem 1.31). At this place it should be mentioned that here the term $B_0P_0Q_1$ vanishes identically, which corresponds to the singular local matrix pencil. This fact makes the term $G_1D^-(D\Pi_1D)'D\Pi_0Q_1$ crucial for G_2 to incorporate a nontrivial increment with respect to G_1.
Observe that the nullspaces and projectors fulfill the relations

$$N_0(t)\cap N_1(t)=\{0\},\quad (N_0(t)+N_1(t))\cap N_2(t)=\{0\},$$
$$Q_1(t)Q_0(t)=0,\quad Q_2(t)Q_0(t)=0,\quad Q_2(t)Q_1(t)=0.$$

The matrix functions G_i as well as the projector functions Q_i are continuous and it holds that $\operatorname{im}G_0=\operatorname{im}G_1=\operatorname{im}G_2\subset\operatorname{im}G_3$. □

Any matrix function sequence (2.5)–(2.8) generates subspaces

$$\operatorname{im} G_0 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq \operatorname{im} G_{i+1}$$

of nondecreasing dimensions.
To show several useful properties we introduce the additional projector functions $\mathcal{W}_j : \mathcal{I} \to L(\mathbb{R}^k)$ and generalized inverses $G_j^- : \mathcal{I} \to L(\mathbb{R}^k, \mathbb{R}^m)$ of G_j such that

$$\ker \mathcal{W}_j = \operatorname{im} G_j, \tag{2.11}$$

$$G_j G_j^- G_j = G_j,\ \ G_j^- G_j G_j^- = G_j^-,\ \ G_j^- G_j = P_j,\ \ G_j G_j^- = I - \mathcal{W}_j. \tag{2.12}$$

Proposition 2.5. *Let the DAE (2.1) have a properly stated leading term. Then, for each matrix function sequence (2.5)–(2.8) the following relations are satisfied:*

(1) $\ker \Pi_i \subseteq \ker B_{i+1}$,
(2) $\mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_i = \cdots = \mathcal{W}_{i+1} B_0 = \mathcal{W}_{i+1} B$,
$\mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_0 = \mathcal{W}_{i+1} B_0 \Pi_i$,
(3) $G_{i+1} = (G_i + \mathcal{W}_i B Q_i) F_{i+1}$ *with* $F_{i+1} = I + G_i^- B_i Q_i$ *and*
$\operatorname{im} G_{i+1} = \operatorname{im} G_i \oplus \operatorname{im} \mathcal{W}_i B Q_i$,
(4) $N_i \cap \ker B_i = N_i \cap N_{i+1} \subseteq N_{i+1} \cap \ker B_{i+1}$,
(5) $N_{i-1} \cap N_i \subseteq N_i \cap N_{i+1}$,
(6) $\operatorname{im} G_i + \operatorname{im} B_i \subseteq \operatorname{im} [AD, B] = \operatorname{im} [G_0, B_0]$.

Proof. (1) From (2.8) we successively derive an expression for B_{i+1}:

$$\begin{aligned} B_{i+1} &= \Big(B_{i-1} P_{i-1} - G_i D^- (D \Pi_i D^-)' D \Pi_{i-1} \Big) P_i - G_{i+1} D^- (D \Pi_{i+1} D^-)' D \Pi_i \\ &= B_{i-1} P_{i-1} P_i - \sum_{j=i}^{i+1} G_j D^- (D \Pi_j D^-)' D \Pi_i, \end{aligned}$$

hence

$$B_{i+1} = B_0 \Pi_i - \sum_{j=1}^{i+1} G_j D^- (D \Pi_j D^-)' D \Pi_i, \tag{2.13}$$

but this immediately verifies assertion (1).
(2) Because of $\operatorname{im} G_j \subseteq \operatorname{im} G_{i+1}$ for $j \le i+1$, we have $\mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_0 \Pi_i$ due to (2.13). Taking into account also the inclusion $\operatorname{im} B_j Q_j = \operatorname{im} G_{j+1} Q_j \subseteq \operatorname{im} G_{j+1} \subseteq \operatorname{im} G_{i+1}$, for $j \le i$, we obtain from (2.8) that $\mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_i P_i = \mathcal{W}_{i+1} B_i - \mathcal{W}_{i+1} B_i Q_i = \mathcal{W}_{i+1} B_i = \mathcal{W}_{i+1} B_{i-1} P_{i-1} = \mathcal{W}_{i+1} B_{i-1} = \cdots = \mathcal{W}_{i+1} B_0$, which proves assertion (2).
(3) We rearrange G_{i+1} as

$$G_{i+1} = G_i + G_i G_i^- B_i Q_i + (I - G_i G_i^-) B_i Q_i = G_i ((I + G_i^- B_i Q_i) + \mathcal{W}_i B_i Q_i.$$

Because of $Q_i G_i^- = Q_i P_i G_i^- = 0$ the matrix function $F_{i+1} := I + G_i^- B_i Q_i$ remains nonsingular (see Lemma A.3) and the factorization

$$G_{i+1} = (G_i + \mathcal{W}_i B_i Q_i)F_{i+1} = (G_i + \mathcal{W}_i B Q_i)F_{i+1}$$

holds true. This yields assertion (3).
(4) $z \in N_i \cap \ker B_i$, i.e., $G_i z = 0$, $B_i z = 0$, leads to $z = Q_i z$ and $G_{i+1} z = B_i Q_i z = B_i z = 0$, thus $z \in N_i \cap N_{i+1}$. Conversely, $z \in N_i \cap N_{i+1}$ yields $z = Q_i z$, $B_i z = B_i Q_i z = G_{i+1} z = 0$, i.e., $z \in N_i \cap \ker B_i$ and we are done with assertion (4).
(5) From $z \in N_{i-1} \cap N_i$ it follows that $z = Q_{i-1} z$ and $B_i z = B_i Q_{i-1} z = B_i P_{i-1} Q_{i-1} z = 0$ because of $B_i = B_i P_{i-1}$ (cf. (2.13)), hence $z \in N_i \cap \ker B_i = N_i \cap N_{i+1}$.
(6) follows from $\operatorname{im} G_0 + \operatorname{im} B_0 = \operatorname{im}[G_0, B_0]$ by induction. Namely, $\operatorname{im} G_i + \operatorname{im} B_i \subseteq \operatorname{im}[G_0, B_0]$ implies $\operatorname{im} B_i Q_i \subseteq \operatorname{im}[G_0, B_0]$, hence $\operatorname{im} G_{i+1} \subseteq \operatorname{im}[G_i, B_0 Q_i] \subseteq \operatorname{im}[G_0, B_0]$, and further $\operatorname{im} B_{i+1} \subseteq \operatorname{im}[G_{i+1}, B_i] \subseteq \operatorname{im}[G_0, B_0]$. □

2.2.2 Admissible projector functions and characteristic values

In Chapter 1 on constant coefficient DAEs, useful decoupling properties are obtained by restricting the variety of possible projectors Q_i and somehow choosing smart ones, so-called *admissible* ones. Here we take up this idea again, and we incorporate conditions concerning ranks and dimensions to ensure the continuity of the matrix functions associated to the DAE. Possible rank changes will be treated as critical points discussed later on in Section 2.9. The following definition generalizes Definition 1.10.

Definition 2.6. Given the DAE (2.1) with properly stated leading term, Q_0 denotes a continuous projector function onto $\ker D$ and $P_0 = I - Q_0$. The generalized inverse D^- is given by $DD^-D = D$, $D^-DD^- = D^-$, $DD^- = R$, $D^-D = P_0$.
For a given level $\kappa \in \mathbb{N}$, we call the sequence $G_0, \ldots, G_\kappa$ an *admissible matrix function sequence* associated to the DAE on the interval $\mathcal{I}$, if it is built by the rule

Set $G_0 := AD$, $B_0 := B$, $N_0 := \ker G_0$.
For $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1} Q_{i-1},$$
$$B_i := B_{i-1} P_{i-1} - G_i D^- (D \Pi_i D^-)' D \Pi_{i-1}$$
$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$
fix a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,
choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,
set $P_i := I - Q_i$, $\Pi_i := \Pi_{i-1} P_i$

and, additionally,

(a) G_i has constant rank r_i on $\mathcal{I}$, $i = 0, \ldots, \kappa$,

(b) the intersection $\widehat{N}_i$ has constant dimension $u_i := \dim \widehat{N}_i$ on $\mathcal{I}$,

(c) the product function Π_i is continuous on $\mathcal{I}$ and $D\Pi_i D^-$ is there continuously differentiable, $i = 0, \ldots, \kappa$.

The projector functions $Q_0, \ldots, Q_\kappa$ in an admissible matrix function sequence are said to be *admissible* themselves.
An admissible matrix function sequence $G_0, \ldots, G_\kappa$ is said to be *regular admissible*, if

$$\widehat{N}_i = \{0\}, \quad \forall\, i = 1, \ldots, \kappa.$$

Then, also the projector functions $Q_0, \ldots, Q_\kappa$ are called *regular admissible*.

Examples 2.3 and 2.4 already show regular admissible matrix function sequences.
The matrix functions $G_0, \ldots, G_\kappa$ in an admissible sequence are a priori continuous on the given interval.
If $G_0, \ldots, G_\kappa$ are admissible, besides the nullspaces $N_0, \ldots, N_\kappa$ and the intersection spaces $\widehat{N}_1, \ldots, \widehat{N}_\kappa$ also the sum spaces $N_0 + \cdots + N_i$, $i = 1, \ldots, \kappa$, and the complements $X_1, \ldots, X_\kappa$ have constant dimension. Namely, the construction yields

$$N_0 + \cdots + N_{i-1} = X_i \oplus \widehat{N}_i, \quad N_0 + \cdots + N_i = X_i \oplus N_i, \quad i = 1, \ldots \kappa,$$

and hence

$$\begin{aligned} \dim N_0 &= m - r_0, \\ \dim(N_0 + \cdots + N_{i-1}) &= \dim X_i + u_i, \\ \dim(N_0 + \cdots + N_i) &= \dim X_i + m - r_i, \quad i = 1, \ldots, \kappa. \end{aligned}$$

It follows that

$$\begin{aligned} \dim(N_0 + \cdots + N_i) &= \underbrace{\dim(N_0 + \cdots + N_{i-1}) - u_i}_{\dim X_i} + \underbrace{m - r_i}_{\dim N_i} \\ &= \sum_{j=0}^{i-1} (m - r_j - u_{j+1}) + m - r_i = \sum_{j=0}^{i} (m - r_j) - \sum_{j=0}^{i-1} u_{j+1}. \end{aligned}$$

We are most interested in the case of trivial intersections $\widehat{N}_i$, yielding $X_i = N_0 + \cdots + N_{i-1}$, and $u_i = 0$. In particular, all so-called regular DAEs in Section 2.6 belong to this latter class. Due to the trivial intersection $\widehat{N}_i = \{0\}$, the subspace $N_0 + \cdots + N_i$ has dimension $\dim(N_0 + \cdots + N_{i-1}) + \dim N_i$, that is, its increase is maximal at each level.

The next proposition collects benefits from admissible projector functions. Comparing with Proposition 1.13 we recognize a far-reaching conformity. The most important benefit seems to be the fact that Π_i being a product of projector functions is again a projector function, and it projects along the sum space $N_0 + \cdots + N_i$ which now appears to be a $\mathcal{C}$-subspace.

We stress once more that admissible projector functions are always cross-linked with their generating admissible matrix function sequence. Nevertheless, for brevity, we simply speak of admissible projector functions or admissible projectors, dropping this natural background.

Proposition 2.7. *Given a DAE (2.1) with properly stated leading term, and an integer* $\kappa \in \mathbb{N}$.
If $Q_0, \ldots, Q_\kappa$ *are admissible projector functions, then the following eight relations become true for* $i = 1, \ldots, \kappa$.

(1) $\ker \Pi_i = N_0 + \cdots + N_i$,
(2) *the products* $\Pi_i = P_0 \cdots P_i$ *and* $\Pi_{i-1} Q_i = P_0 \cdots P_{i-1} Q_i$, *as well as* $D\Pi_i D^-$ *and* $D\Pi_{i-1} Q_i D^-$, *are projector valued functions, too,*
(3) $N_0 + \cdots + N_{i-1} \subseteq \ker \Pi_{i-1} Q_i$,
(4) $B_i = B_i \Pi_{i-1}$,
(5) $\widehat{N}_i \subseteq N_i \cap N_{i+1}$, *and hence* $\widehat{N}_i \subseteq \widehat{N}_{i+1}$,
(6) $G_{i+1} Q_j = B_j Q_j$, $0 \le j \le i$,
(7) $D(N_0 + \cdots + N_i) = \operatorname{im} DP_0 \cdots P_{i-1} Q_i \oplus \operatorname{im} D\Pi_{i-2} Q_{i-1} \oplus \cdots \oplus \operatorname{im} DP_0 Q_1$,
(8) *the products* $Q_i(I - \Pi_{i-1})$ *and* $P_i(I - \Pi_{i-1})$ *are projector functions onto* $\widehat{N}_i$ *and* X_i, *respectively.*

Additionally, the matrix functions $G_1, \ldots, G_\kappa$, *and* $G_{\kappa+1}$ *are continuous.*
If $Q_0, \ldots, Q_\kappa$ *are regular admissible then it holds for* $i = 1, \ldots, \kappa$ *that*

$$\ker \Pi_{i-1} Q_i = \ker Q_i, \quad \text{and} \quad Q_i Q_j = 0, \quad j = 0, \ldots, i-1.$$

Proof. (1) See the proof of Proposition 1.13 (1).
(2) Due to assertion (1) it holds that $\ker \Pi_i = N_0 + \cdots + N_i$, which means $\Pi_i Q_j = 0$, $j = 0, \ldots, i$. With $0 = \Pi_i Q_j = \Pi_i (I - P_j)$, we obtain $\Pi_i = \Pi_i P_j$, $j = 0, \ldots, i$, which yields $\Pi_i \Pi_i = \Pi_i$. Derive further

$$\begin{aligned}
(\Pi_{i-1} Q_i)^2 &= (\Pi_{i-1} - \Pi_i)(\Pi_{i-1} - \Pi_i) \\
&= \Pi_{i-1} - \underbrace{\Pi_{i-1} \Pi_i}_{= \Pi_{i-1} P_i} - \underbrace{\Pi_i \Pi_{i-1}}_{= \Pi_i} + \Pi_i = \Pi_{i-1} Q_i, \\
(D\Pi_i D^-)^2 &= D\Pi_i \underbrace{D^- D}_{= P_0} \Pi_i D^- = D\Pi_i D^-, \\
(D\Pi_{i-1} Q_i D^-)^2 &= D\Pi_{i-1} Q_i \underbrace{D^- D}_{= P_0} \Pi_{i-1} Q_i D^- = D(\Pi_{i-1} Q_i)^2 D^- = D\Pi_{i-1} Q_i D^-.
\end{aligned}$$

(3) See the proof of Proposition 1.13 (3).
(4) The detailed structure of B_i given in (2.13) and the projector property of Π_{i-1} (cf. (1)) proves the statement.
(5) $z \in N_i \cap (N_0 + \cdots + N_{i-1})$ means that $z = Q_i z$, $\Pi_{i-1} z = 0$, hence

$$G_{i+1} z = G_i z + B_i Q_i z = B_i z = B_i \Pi_{i-1} z = 0.$$

(6) For $0 \le j \le i$, it follows with (4) from

$$\begin{aligned} G_{i+1} &= G_i + B_i Q_i = G_0 + B_0 Q_0 + B_1 Q_1 + \cdots + B_i Q_i \\ &= G_0 + B_0 Q_0 + B_1 P_0 Q_1 + \cdots + B_i P_0 \cdots P_{i-1} Q_i \end{aligned}$$

that

$$G_{i+1} Q_j = (G_0 + B_0 Q_0 + \cdots + B_j P_0 \cdots P_{j-1} Q_j) Q_j = (G_j + B_j Q_j) Q_j = B_j Q_j.$$

(7) From $\ker \Pi_i = N_0 + \cdots + N_i$ it follows that

$$\begin{aligned} D(N_0 + \cdots + N_i) &= D \operatorname{im} (I - \Pi_i) = D \operatorname{im} (Q_0 + P_0 Q_1 + \cdots + \Pi_{i-1} Q_i) \\ &= D\{\operatorname{im} Q_0 \oplus \operatorname{im} P_0 Q_1 \oplus \cdots \oplus \operatorname{im} \Pi_{i-1} Q_i\} \\ &= \operatorname{im} D P_0 Q_1 \oplus \cdots \oplus \operatorname{im} D \Pi_{i-1} Q_i. \end{aligned}$$

This proves assertion (7).
(8) We have (cf. (3))

$$Q_i (I - \Pi_{i-1}) Q_i (I - \Pi_{i-1}) = (Q_i - Q_i \underbrace{\Pi_{i-1} Q_i)(I - \Pi_{i-1})}_{=0} = Q_i (I - \Pi_{i-1}).$$

Further, $z = Q_i (I - \Pi_{i-1}) z$ implies $z \in N_i$, $\Pi_{i-1} z = \Pi_{i-1} Q_i (I - \Pi_{i-1}) z = 0$, and hence $z \in \widehat{N}_i$.
Conversely, from $z \in \widehat{N}_i$ it follows that $z = Q_i z$ and $z = (I - \Pi_{i-1}) z$, thus $z = Q_i (I - \Pi_{i-1}) z$. Similarly, we compute

$$P_i (I - \Pi_{i-1}) P_i (I - \Pi_{i-1}) = P_i (I - \Pi_{i-1}) - P_i (I - \Pi_{i-1}) Q_i (I - \Pi_{i-1}) = P_i (I - \Pi_{i-1}).$$

From $z = P_i (I - \Pi_{i-1}) z$ it follows that $Q_i z = 0$, $\Pi_{i-1} z = \Pi_i (I - \Pi_{i-1}) z = 0$, therefore $z \in X_i$.
Conversely, $z \in X_i$ yields $z = P_i z$, $z = (I - \Pi_{i-1}) z$, and hence $z = P_i (I - \Pi_{i-1}) z$. This verifies (8).
Next we verify the continuity of the matrix functions G_i. Applying the representation (2.13) of the matrix function B_i we express

$$G_{i+1} = G_i + B_0 \Pi_{i-1} Q_i - \sum_{j=1}^{i} G_j D^- (D \Pi_j D^-)' D \Pi_{i-1} Q_i,$$

which shows that, supposing that previous matrix functions $G_0, \ldots, G_i$ are continuous, the continuity of $\Pi_{i-1} Q_i = \Pi_{i-1} - \Pi_i$ implies G_{i+1} is also continuous.
Finally, let $Q_0, \ldots, Q_\kappa$ be regular admissible. $\Pi_{i-1} Q_i z = 0$ implies $Q_i z = (I - \Pi_{i-1}) Q_i z \in N_0 + \cdots + N_{i-1}$, hence $Q_i z \in \widehat{N}_i$, therefore $Q_i z = 0$. It remains to apply (3). □

As in the constant coefficient case, there is a great variety of admissible projector functions, and the matrix functions G_i clearly depend on the special choice of the projector functions Q_j, including the way complements X_j in the decomposition of $N_0 + \cdots + N_{j-1}$ are chosen. Fortunately, there are invariants, in particular, invariant subspaces and subspace dimensions, as shown by the next assertion.

Theorem 2.8. *Let the DAE (2.1) have a properly stated leading term. Then, for a given $\kappa \in \mathbb{N}$, if admissible projector functions up to level κ do at all exist, then the subspaces*

$$\operatorname{im} G_j, \quad N_0 + \cdots + N_j, \quad S_j := \ker \mathcal{W}_j B, \quad j = 0, \ldots, \kappa + 1,$$

as well as the numbers

$$r_j := \operatorname{rank} G_j,\ j = 0, \ldots, \kappa, \quad u_j := \dim \widehat{N}_j,\ j = 1, \ldots, \kappa,$$

and the functions $r_{\kappa+1} : \mathcal{I} \to \mathbb{N} \cup \{0\}$, $u_{\kappa+1} : \mathcal{I} \to \mathbb{N} \cup \{0\}$ are independent of the special choice of admissible projector functions $Q_0, \ldots, Q_\kappa$.

Proof. These assertions are immediate consequences of Lemma 2.12 below at the end of the present section. □

Definition 2.9. If the DAE (2.1) with properly stated leading term has an admissible matrix functions sequence up to level κ, then the integers

$$r_j = \operatorname{rank} G_j,\ j = 0, \ldots, \kappa, \quad u_j = \dim \widehat{N}_j,\ j = 1, \ldots, \kappa,$$

are called *characteristic values* of the DAE.

The characteristic values prove to be invariant under regular transformations and refactorizations (cf. Section 2.3, Theorems 2.18 and 2.21), which justifies this notation. For constant regular matrix pairs, these characteristic values describe the infinite eigenstructure (Corollary 1.32).

The associated subspace $S_0 = \ker \mathcal{W}_0 B$ has its special meaning. At given $t \in \mathcal{I}$, the subspace

$$S_0(t) = \ker \mathcal{W}_0(t) B(t) = \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im} G_0(t) = \operatorname{im} A(t)\}$$

contains all solution values $x(t)$ of the solutions of the homogeneous equation $A(Dx)' + Bx = 0$. As we will see later, for so-called regular index-1 DAEs, the subspace $S_0(t)$ consists at all of those solution values, that means, for each element of $S_0(t)$ there exists a solution passing through it. For regular DAEs with a higher index, the sets of corresponding solution values form proper subspaces of $S_0(t)$.
In general, the associated subspaces satisfy the relations

$$S_{i+1} = S_i + N_i = S_i + N_0 + \cdots + N_i = S_0 + N_0 + \cdots + N_i,\ \ i = 0, \ldots, \kappa.$$

Namely, because of $\operatorname{im} G_i \subseteq \operatorname{im} G_{i+1}$, it holds that $\mathcal{W}_{i+1} = \mathcal{W}_{i+1}\mathcal{W}_i$, hence $S_{i+1} = \ker \mathcal{W}_{i+1}B = \ker \mathcal{W}_{i+1}\mathcal{W}_i B \supseteq \ker \mathcal{W}_i B = S_i$, and Proposition 2.5 (2) yields $S_{i+1} = \ker \mathcal{W}_{i+1}B_{i+1} \supseteq \ker B_{i+1} \supseteq N_0 + \cdots + N_i$.

Summarizing, the following three sequences of subspaces are associated with each admissible matrix function sequence:

$$\operatorname{im} G_0 \subseteq \operatorname{im} G_1 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq \cdots \subseteq \operatorname{im} [AD\ B] \subseteq \mathbb{R}^k, \tag{2.14}$$

$$N_0 \subseteq N_0 + N_1 \subseteq \cdots \subseteq N_0 + \cdots + N_i \subseteq \cdots \subseteq \mathbb{R}^m, \tag{2.15}$$

and

$$S_0 \subseteq S_1 \subseteq \cdots \subseteq S_i \subseteq \cdots \subseteq \mathbb{R}^m. \tag{2.16}$$

All of these subspaces are independent of the special choice of the admissible projector functions. In all three cases, the dimension does not decrease if the index increases. We are looking for criteria indicating that a certain G_μ already has the maximal possible rank. For instance, if we meet an injective matrix G_μ as in Examples 2.3 and 2.4, then the sequence becomes stationary with $Q_\mu = 0$, $G_{\mu+1} = G_\mu$, and so on. Therefore, the smallest index μ such that the matrix function G_μ is injective, indicates at the same time that $\operatorname{im} G_\mu$ is maximal, but $\operatorname{im} G_{\mu-1}$ is a proper subspace, if $\mu \geq 1$. The general case is more subtle. It may happen that no injective G_μ exists. Eventually one reaches

$$\operatorname{im} G_\mu = \operatorname{im} [AD\ B]; \tag{2.17}$$

however, this is not necessarily the case, as the next example shows.

Example 2.10 (Admissible matrix sequence for a nonregular DAE). Set $m = k = 3$, $n = 2$, and consider the constant coefficient DAE

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x \right)' + \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} x = q, \tag{2.18}$$

which is nonregular due to the singular matrix pencil. Here we have $\operatorname{im} [AD\ B] = \mathbb{R}^3$. Compute successively

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{W}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$$G_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

We read off $N_0 = \{z \in \mathbb{R}^3 : z_1 = z_2 = 0\}$, $N_1 = \{z \in \mathbb{R}^3 : z_2 = 0, z_1 + z_3 = 0\}$ and $N_2 = \{z \in \mathbb{R}^3 : z_2 = 0, 2z_1 + z_3 = 0\}$. The intersection $N_0 \cap N_1$ is trivial, and the condition $Q_1 Q_0 = 0$ is fulfilled. We have further

$$N_0 + N_1 = \{z \in \mathbb{R}^3 : z_2 = 0\}, \quad (N_0 + N_1) \cap N_2 = \widehat{N_2} = N_2 \subseteq N_0 + N_1,$$
$$\text{thus } N_0 + N_1 = N_0 + N_1 + N_2 \text{ and } N_0 + N_1 = N_2 \oplus N_0.$$

We can put $X_2 = N_0$, and compute

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 0 & 0 \end{bmatrix}, \text{ with } X_2 \subseteq \ker Q_2, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The projectors Q_0, Q_1, Q_2 are admissible. It holds that $B_2 Q_2 = 0$, $G_3 = G_2$, $N_3 = N_2$, and $\Pi_2 = \Pi_1$, and further

$$S_0 = \{z \in \mathbb{R}^3 : z_2 = 0\}, \quad S_0 = S_1 = S_2 = S_3.$$

We continue the matrix function sequence by $Q_3 := Q_2$, $B_3 = B_2$, $B_3 Q_3 = 0$, $G_4 = G_3$, and so on. It follows that no G_i is injective, and

$$\operatorname{im} G_0 = \cdots = \operatorname{im} G_i = \cdots = \mathbb{R}^2 \times \{0\} \subset \operatorname{im} [AD\ B] = \mathbb{R}^3,$$
$$S_0 = \cdots = S_i = \cdots = \mathbb{R} \times \{0\} \times \mathbb{R},$$
$$N_0 \subset N_0 + N_1 = N_0 + N_1 + N_2 = \cdots = \mathbb{R} \times \{0\} \times \mathbb{R},$$

and the maximal range is already $\operatorname{im} G_0$. A closer look at the DAE (2.18) gives

$$\begin{aligned} x_1' + x_1 + x_3 &= q_1, \\ x_2' + x_2 &= q_2, \\ x_2 &= q_3. \end{aligned}$$

This model is somewhat dubious. It is in parts over- and underdetermined, and much room for interpretations is left (cf. Chapter 10). □

Our next example is much nicer and more important with respect to applications. It is a so-called *Hessenberg form size-3 DAE* and might be considered as the linear prototype of the system describing constrained mechanical motion (see Example 3.41 and Section 3.5).

Example 2.11 (Admissible sequence for the Hessenberg size 3 DAE). Consider the system

$$\begin{bmatrix} x_1' \\ x_2' \\ 0 \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & 0 \\ 0 & B_{32} & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \tag{2.19}$$

with $m = m_1 + m_2 + m_3$ equations, $m_1 \geq m_2 \geq m_3 \geq 1$, $k = m$ components, and a nonsingular product $B_{32} B_{21} B_{13}$. Put $n = m_1 + m_2$,

$$A = \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} I & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & 0 \\ 0 & B_{32} & 0 \end{bmatrix},$$

and write this DAE in the form (2.1).
Owing to the nonsingularity of the $m_3 \times m_3$ matrix function product $B_{32}B_{21}B_{13}$, the matrix functions B_{13} and $B_{21}B_{13}$ have full column rank m_3 each, and B_{32} has full row rank m_3. This yields $\operatorname{im}[AD\ B] = \mathbb{R}^m$. Further, since B_{13} and $B_{21}B_{13}$ have constant rank, there are continuous reflexive generalized inverses B_{13}^- and $(B_{21}B_{13})^-$ such that (see Proposition A.17)

$$\begin{aligned} B_{13}^- B_{13} = I, \quad \Omega_1 &:= B_{13}B_{13}^- && \text{is a projector onto} \quad \operatorname{im} B_{13}, \\ (B_{21}B_{13})^- B_{21}B_{13} = I, \quad \Omega_2 &:= B_{21}B_{13}(B_{21}B_{13})^- && \text{is a projector onto} \quad \operatorname{im} B_{21}B_{13}. \end{aligned}$$

Let the coefficient function B be smooth enough so that the derivatives used below do exist. In particular, Ω_1 and Ω_2 are assumed to be continuously differentiable. We start constructing the matrix function sequence by

$$G_0 = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \quad B_0 = \begin{bmatrix} B_{11} & B_{12} & B_{13} \\ B_{21} & B_{22} & 0 \\ 0 & B_{32} & 0 \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & 0 & B_{13} \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

It follows that

$$\begin{aligned} &N_0 = \{z \in \mathbb{R}^m : z_1 = 0,\ z_2 = 0\}, \quad N_1 = \{z \in \mathbb{R}^m : z_1 + B_{13}z_3 = 0,\ z_2 = 0\}, \\ &\widehat{N}_1 = N_0 \cap N_1 = \{0\}, \quad X_1 = N_0, \\ &N_0 + N_1 = N_0 \oplus N_1 = \{z \in \mathbb{R}^m : z_2 = 0,\ z_1 \in \operatorname{im} B_{13}\}. \end{aligned}$$

The matrix functions G_0 and G_1 have constant rank, $r_0 = r_1 = n$. Compute the projector functions

$$Q_1 = \begin{bmatrix} \Omega_1 & 0 & 0 \\ 0 & 0 & 0 \\ -B_{13}^- & 0 & 0 \end{bmatrix}, \quad D\Pi_1 D^- = \begin{bmatrix} I - \Omega_1 & 0 \\ 0 & I \end{bmatrix},$$

such that $\operatorname{im} Q_1 = N_1$ and $Q_1Q_0 = 0$, that is $\ker Q_1 \supseteq X_1$. Q_1 is continuous, and $D\Pi_1 D^-$ is continuously differentiable. In consequence, Q_0, Q_1 are admissible. Next we form

$$B_1 = \begin{bmatrix} B_{11} + \Omega_1' & B_{12} & 0 \\ B_{21} & B_{22} & 0 \\ 0 & B_{32} & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} I + (B_{11} + \Omega_1')\Omega_1 & 0 & B_{13} \\ B_{21}\Omega_1 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

For $z \in \mathbb{R}^{m_1+m_2+m_3}$ with $z_1 \in \ker \Omega_1$ it holds that $\operatorname{im} G_2 = \begin{bmatrix} z_1 + B_{13} z_3 \\ z_2 \\ 0 \end{bmatrix}$, since $\operatorname{im} B_{13} = \operatorname{im} \Omega_1$. This proves the inclusion

$$\operatorname{im} G_2 \subseteq \mathbb{R}^n \times \{0\} = \{G_2 z : z \in \mathbb{R}^{m_1+m_2+m_3}, z_1 \in \ker \Omega_1\} \subseteq \operatorname{im} G_2,$$

and we obtain $\operatorname{im} G_2 = \mathbb{R}^n \times \{0\}$, and $r_2 = \operatorname{rank} G_2 = m_1 + m_2 = n$. Then we investigate the nullspace of G_2. If $z \in \mathbb{R}^m$ satisfies $G_2 z = 0$, then

$$z_1 + (B_{11} + \Omega_1')\Omega_1 z_1 + B_{13} z_3 = 0, \tag{2.20}$$

$$B_{21}\Omega_1 z_1 + z_2 = 0. \tag{2.21}$$

In turn, equation (2.20) decomposes into

$$\begin{aligned} (I - \Omega_1) z_1 + (I - \Omega_1)(B_{11} + \Omega_1')\Omega_1 z_1 &= 0, \\ B_{13}^-(I + B_{13}^-(B_{11} + \Omega_1'))\Omega_1 z_1 + z_3 &= 0. \end{aligned}$$

Similarly, considering that $\operatorname{im} B_{21} B_{13} = \operatorname{im} B_{21} B_{13} B_{13}^-$ is valid, we derive from (2.21) the relations

$$z_2 = \Omega_2 z_2, \quad B_{13}^- z_1 = -(B_{21} B_{13})^- z_2.$$

Altogether this yields

$$N_2 = \{z \in \mathbb{R}^m : z_2 = \Omega_2 z_2, z_1 = \mathcal{E}_1 \Omega_2 z_2, z_3 = \mathcal{E}_3 \Omega_2 z_2\}, \quad \widehat{N}_2 = \{0\}, \quad X_2 = N_0 + N_1,$$

with

$$\begin{aligned} \mathcal{E}_1 &:= -(I - (I - \Omega_1)(B_{11} + \Omega_1')\Omega_1) B_{13} (B_{21} B_{13})^- \\ &= -(I - (I - \Omega_1)(B_{11} + \Omega_1')) B_{13} (B_{21} B_{13})^-, \\ \mathcal{E}_3 &:= -B_{13}^-(I + (B_{11} + \Omega_1')) B_{13} (B_{21} B_{13})^-. \end{aligned}$$

Notice that $\mathcal{E}_1 = \mathcal{E}_1 \Omega_2$, $\mathcal{E}_3 = \mathcal{E}_3 \Omega_2$. The projector functions

$$Q_2 = \begin{bmatrix} 0 & \mathcal{E}_1 & 0 \\ 0 & \Omega_2 & 0 \\ 0 & \mathcal{E}_3 & 0 \end{bmatrix}, \quad D\Pi_2 D^- = \begin{bmatrix} I - \Omega_1 & -(I - \Omega_1)\mathcal{E}_1 \\ 0 & I - \Omega_2 \end{bmatrix},$$

fulfill the required admissibility conditions, in particular, $Q_2 Q_0 = 0$, $Q_2 Q_1 = 0$, and hence Q_0, Q_1, Q_2 are admissible. The resulting B_2, G_3 have the form:

$$B_2 = \begin{bmatrix} \mathcal{B}_{11} & \mathcal{B}_{12} & 0 \\ \mathcal{B}_{21} & \mathcal{B}_{22} & 0 \\ 0 & B_{32} & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} I + (B_{11} + \Omega_1')\Omega_1 & \mathcal{B}_{11}\mathcal{E}_1 + \mathcal{B}_{12}\Omega_2 & B_{13} \\ B_{21}\Omega_1 & I + \mathcal{B}_{21}\mathcal{E}_1 + \mathcal{B}_{22}\Omega_2 & 0 \\ 0 & B_{32}\Omega_2 & 0 \end{bmatrix}.$$

The detailed form of the entries $\mathcal{B}_{ij}$ does not matter in this context. We show G_3 to be nonsingular. Namely, $G_3 z = 0$ implies $B_{32}\Omega_2 z_2 = 0$, thus $\Omega_2 z_2 = 0$, and further

$B_{21}\Omega_1 z_1 + z_2 = 0$. The latter equation yields $(I-\Omega_2)z_2 = 0$ and $B_{21}\Omega_1 z_1 = 0$, and this gives $\Omega_1 z_1 = 0$, $z_2 = 0$. Now, the first line of the system $G_3 z = 0$ simplifies to $z_1 + B_{13} z_3 = 0$. In turn, $(I-\Omega_1)z_1 = 0$ follows, and hence $z_1 = 0$, $z_3 = 0$. The matrix function G_3 is nonsingular in fact, and we stop the construction.
In summary, our basic subspaces behaves as

$$\begin{aligned} \operatorname{im} G_0 = \operatorname{im} G_1 = \operatorname{im} G_2 \subset \operatorname{im} G_3 = \operatorname{im}[AD\ B] = \mathbb{R}^m, \\ N_0 \subset N_0 + N_1 \subset N_0 + N_1 + N_2 = N_0 + N_1 + N_2 + N_3 \subset \mathbb{R}^m. \end{aligned}$$

The additionally associated projector functions $\mathcal{W}_i$ onto $\operatorname{im} G_i$ and the subspaces $S_i = \ker \mathcal{W}_i B$ are here:

$$\mathcal{W}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \quad \mathcal{W}_0 = \mathcal{W}_1 = \mathcal{W}_2, \quad \mathcal{W}_3 = 0,$$

and

$$S_0 = \{z \in \mathbb{R}^m : B_{32} z_2 = 0\}, \quad S_0 = S_1 = S_2 \subset S_3 = \mathbb{R}^m.$$

The last relation is typical for the large class of DAEs named Hessenberg form DAEs (cf. Section 3.5). While $\operatorname{im} G_3$ and S_3 reach the maximal dimension m, the dimension of the resulting maximal subspace $N_0 + N_1 + N_2$ is less than m.
Notice that the relation $\mathcal{W}_0 B Q_0 = 0$ indicates that $\operatorname{im} G_0 = \operatorname{im} G_1$ holds true, and we can recognize this fact before explicitly computing G_1 (cf. Proposition 2.5(3)). Similarly, $\mathcal{W}_1 B Q_1 = 0$ indicates that $\operatorname{im} G_1 = \operatorname{im} G_2$. Furthermore, we know that $r_3 = r_2 + \operatorname{rank}(\mathcal{W}_2 B Q_2) = n + m_3 = m$ before we compute G_3. □

Now we come to an important auxiliary result which stands behind Theorem 2.8, and which generalizes Lemma 1.18.

Lemma 2.12. *Given the DAE (2.1) with properly stated leading term, if there are two admissible projector function sequences $Q_0, \dots, Q_\kappa$ and $\bar{Q}_0, \dots, \bar{Q}_\kappa$, both admissible on $\mathcal{I}$, then the associated matrix functions and subspaces are related by the following properties:*

(1) $\ker \bar{\Pi}_j = \bar{N}_0 + \cdots + \bar{N}_j = N_0 + \cdots + N_j = \ker \Pi_j,\ j = 0, \dots, \kappa,$

(2) $\bar{G}_j = G_j Z_j,$

$$\bar{B}_j = B_j - G_j Z_j \bar{D}^- (D\bar{\Pi}_j \bar{D}^-)' D\Pi_j + G_j \sum_{l=0}^{j-1} Q_l \mathfrak{A}_{jl},\ j = 1, \dots, \kappa,$$

with nonsingular matrix functions $Z_0, \dots, Z_{\kappa+1}$ given by
$Z_0 := I$, $Z_{i+1} := Y_{i+1} Z_i$, $i = 0, \dots, \kappa$,

$$\begin{aligned} Y_1 &:= I + Q_0(\bar{Q}_0 - Q_0) = I + Q_0 \bar{Q}_0 P_0, \\ Y_{i+1} &:= I + Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1} Q_i) + \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i,\ i = 1, \dots, \kappa, \end{aligned}$$

and certain continuous coefficients $\mathfrak{A}_{il}$ that satisfy condition $\mathfrak{A}_{il} = \mathfrak{A}_{il}\bar{\Pi}_{i-1}$,

(3) $Z_i(\bar{N}_i \cap (\bar{N}_0 + \cdots + \bar{N}_{i-1})) = N_i \cap (N_0 + \cdots + N_{i-1}), \quad i = 1, \ldots, \kappa,$
(4) $\bar{G}_{\kappa+1} = G_{\kappa+1} Z_{\kappa+1}, \quad \bar{N}_0 + \cdots + \bar{N}_{\kappa+1} = N_0 + \cdots + N_{\kappa+1},$
$Z_{\kappa+1}(\bar{N}_{\kappa+1} \cap (\bar{N}_0 + \cdots + \bar{N}_\kappa)) = N_{\kappa+1} \cap (N_0 + \cdots + N_\kappa).$

Proof. We have $G_0 = AD = \bar{G}_0$, $B_0 = B = \bar{B}_0$, $\ker P_0 = N_0 = \bar{N}_0 = \ker \bar{P}_0$, and hence $P_0 = P_0 \bar{P}_0$, $\bar{P}_0 = \bar{P}_0 P_0$.
The generalized inverses D^- and $\bar{D}^-$ of D satisfy the properties $DD^- = D\bar{D}^- = R$, $D^- D = P_0$, $\bar{D}^- D = \bar{P}_0$, and therefore $\bar{D}^- = \bar{D}^- D\bar{D}^- = \bar{D}^- D D^- = \bar{P}_0 D^-$, $D^- = P_0 \bar{D}^-$.
Compare $G_1 = G_0 + B_0 Q_0$ and

$$\begin{aligned} \bar{G}_1 &= \bar{G}_0 + \bar{B}_0 \bar{Q}_0 = G_0 + B_0 \bar{Q}_0 = G_0 + B_0 Q_0 \bar{Q}_0 \\ &= (G_0 + B_0 Q_0)(P_0 + \bar{Q}_0) = G_1 Z_1, \end{aligned}$$

where $Z_1 := Y_1 := P_0 + \bar{Q}_0 = I + Q_0 \bar{Q}_0 P_0 = I + Q_0(\bar{Q}_0 - Q_0)$. Z_1 is invertible; it has the inverse $Z_1^{-1} = I - Q_0 \bar{Q}_0 P_0$.
The nullspaces N_1 and $\bar{N}_1$ are, due to $\bar{G}_1 = G_1 Z_1$, related by $\bar{N}_1 = Z_1^{-1} N_1 \subseteq N_0 + N_1$. This implies $\bar{N}_0 + \bar{N}_1 = N_0 + (Z_1^{-1} N_1) \subseteq N_0 + N_1$. From $N_1 = Z_1 \bar{N}_1 \subseteq N_0 + \bar{N}_1 = \bar{N}_0 + \bar{N}_1$, we obtain $\bar{N}_0 + \bar{N}_1 = N_0 + N_1$.

Since the projectors $P_0 P_1$ and $\bar{P}_0 \bar{P}_1$ have the common nullspace $N_0 + N_1 = \bar{N}_0 + \bar{N}_1$, we may now derive

$$\begin{aligned} D\bar{P}_0\bar{P}_1\bar{D}^- &= D\bar{P}_0\bar{P}_1 \overbrace{P_0 P_1}^{=P_0P_1P_0} \bar{P}_0 D^- = D\bar{P}_0\bar{P}_1 P_0 P_1 D^- = D\bar{P}_0\bar{P}_1\bar{D}^- D P_0 P_1 D^-, \\ DP_0P_1D^- &= DP_0P_1D^- D\bar{P}_0\bar{P}_1\bar{D}^-. \end{aligned}$$

Taking into account the relation $0 = \bar{G}_1 \bar{Q}_1 = G_1 \bar{Q}_1 + G_1(Z_1 - I)\bar{Q}_1$, thus $G_1 \bar{Q}_1 = -G_1(Z_1 - I)\bar{Q}_1$ we obtain (cf. Appendix B for details)

$$\bar{B}_1 = B_1 - G_1 Z_1 \bar{D}^- (D\bar{P}_0\bar{P}_1 D^-)' D.$$

This gives the basis for proving our assertion by induction. The proof is carried out in detail in Appendix B. A technically easier version for the time-invariant case is given in Chapter 1, Lemma 1.18. □

2.2.3 Widely orthogonal projector functions

For each DAE with properly stated leading term, we can always start the matrix function sequence by choosing Q_0 to be the orthogonal projector onto $N_0 = \ker D$, that means, $Q_0 = Q_0^*$, $P_0 = P_0^*$. On the next level, applying the decomposition $\mathbb{R}^m = (N_0 \cap N_1)^\perp \oplus (N_0 \cap N_1)$ we determine X_1 in the decomposition $N_0 = X_1 \oplus (N_0 \cap N_1)$ by $X_1 = N_0 \cap (N_0 \cap N_1)^\perp$. This leads to $N_0 + N_1 = (X_1 \oplus (N_0 \cap N_1)) + N_1 = X_1 \oplus N_1$ and $\mathbb{R}^m = (N_0 + N_1)^\perp \oplus (N_0 + N_1) = (N_0 + N_1)^\perp \oplus X_1 \oplus N_1$. In this way Q_1 is uniquely determined as $\operatorname{im} Q_1 = N_1$, $\ker Q_1 = (N_0 + N_1)^\perp \oplus X_1$.

On the next levels, if $Q_0,\dots,Q_{i-1}$ are admissible, we first apply the decomposition $\mathbb{R}^m = (\widehat{N}_i)^\perp \oplus \widehat{N}_i$, and choose

$$X_i = (N_0+\cdots+N_{i-1}) \cap (\widehat{N}_i)^\perp. \tag{2.22}$$

The resulting decompositions $N_0+\cdots+N_i = X_i \oplus N_i$, and
$\mathbb{R}^m = (N_0+\cdots+N_i)^\perp \oplus (N_0+\cdots+N_i) = (N_0+\cdots+N_i)^\perp \oplus X_i \oplus N_i$ allow for the choice

$$\operatorname{im} Q_i = N_i,\ \ \ker Q_i = (N_0+\cdots+N_i)^\perp \oplus X_i. \tag{2.23}$$

Definition 2.13. Admissible projector functions $Q_0,\dots,Q_\kappa$ are called *widely orthogonal* if $Q_0 = Q_0^*$ and both (2.22) and (2.23) are fulfilled for $i = 1,\dots,\kappa$.

Example 2.14 (Widely orthogonal projectors). The admissible projector functions Q_0, Q_1 built for the Hessenberg size-2 DAE in Example 2.3 with $\Omega = \Omega^*$ are widely orthogonal. In particular, it holds that

$$\ker Q_1 = \{z \in \mathbb{R}^{m_1+m_2} : B_{12}^* z_1 = 0\} = (N_0 \oplus N_1)^\perp \oplus N_0.$$

□

Widely orthogonal projector functions are uniquely fixed by construction. They provide special symmetry properties. In fact, applying widely orthogonal projector functions, the decompositions

$$x(t) = \Pi_i(t)x(t) + \Pi_{i-1}(t)Q_i(t)x(t) + \cdots + \Pi_0(t)Q_1(t)x(t) + Q_0(t)x(t)$$

are orthogonal ones for all t owing to the following proposition.

Proposition 2.15. *If $Q_0,\dots,Q_\kappa$ are widely orthogonal, then Π_i, $i = 0,\dots,\kappa$, and $\Pi_{i-1}Q_i, i = 1,\dots,\kappa$, are symmetric.*

Proof. Let $Q_0,\dots,Q_\kappa$ be widely orthogonal. In particular, it holds that $\Pi_0 = \Pi_0^*$, $\ker \Pi_0 = N_0$, $\operatorname{im} \Pi_0 = N_0^\perp$.
Compute $\operatorname{im} \Pi_1 = \operatorname{im} P_0P_1 = P_0 \operatorname{im} P_1 = P_0((N_0+N_1)^\perp \oplus X_1) = P_0(N_0+N_1)^\perp$
$= P_0(N_0^\perp \cap N_1^\perp) = N_0^\perp \cap N_1^\perp = (N_0+N_1)^\perp$.
To use induction, assume that $\operatorname{im} \Pi_j = (N_0+\cdots+N_j)^\perp$, $j \le i-1$.
Due to Proposition 2.7 (1) we know that $\ker \Pi_i = N_0+\cdots+N_i$ is true; further $\Pi_{i-1}X_i = 0$. From (2.23) it follows that

$$\begin{aligned}\operatorname{im} \Pi_i &= \Pi_{i-1} \operatorname{im} P_i = \Pi_{i-1}((N_0+\cdots+N_i)^\perp \oplus X_i)\\ &= \Pi_{i-1}(N_0+\cdots+N_i)^\perp = \Pi_{i-1}((N_0+\cdots+N_{i-1})^\perp \cap N_i^\perp)\\ &= (N_0+\cdots+N_{i-1})^\perp \cap N_i^\perp = (N_0+\cdots+N_i)^\perp.\end{aligned}$$

Since Π_i is a projector, and $\ker \Pi_i = N_0+\cdots+N_i$, $\operatorname{im} \Pi_i = (N_0+\cdots+N_i)^\perp$, Π_i must be the orthoprojector.
Finally, derive $(\Pi_{i-1}Q_i)^* = (\Pi_{i-1} - \Pi_{i-1}P_i)^* = \Pi_{i-1} - \Pi_{i-1}P_i = \Pi_{i-1}Q_i$. □

Proposition 2.16. *If, for the DAE (2.1) with properly stated leading term, there exist any admissible projector functions $Q_0,\dots,Q_\kappa$, and if $DD^* \in \mathcal{C}^1(\mathcal{I},L(\mathbb{R}^n))$, then also widely orthogonal projector functions can be chosen (they do exist).*

Proof. Let $Q_0,\dots,Q_\kappa$ be admissible. Then, in particular the subspaces $N_0+\cdots+N_i$, $i=0,\dots,\kappa$ are continuous. The subspaces $\operatorname{im} D\Pi_0Q_1,\dots,\operatorname{im} D\Pi_{\kappa-1}Q_\kappa$ belong to the class $\mathcal{C}^1$, since the projectors $D\Pi_0Q_1D^-,\dots,D\Pi_{\kappa-1}Q_\kappa D^-$ do so. Taking Proposition 2.7 into account we know the subspaces $D(N_0+\cdots+N_i)$, $i=1,\dots,\kappa$, to be continuously differentiable.
Now we construct widely orthogonal projectors. Choose $\bar{Q}_0=\bar{Q}_0^*$, and form $\bar{G}_1 = G_0+B_0\bar{Q}_0$. Due to Lemma 2.12 (d) it holds that $\bar{G}_1 = G_1Z_1$, $\bar{N}_0+\bar{N}_1 = N_0+N_1$, $Z_1(\bar{N}_0\cap\bar{N}_1)=N_0\cap N_1$. Since Z_1 is nonsingular, $\bar{G}_1$ has constant rank r_1, and the intersection $\overline{N\mathcal{U}}_1=\bar{N}_1\cap\bar{N}_0$ has constant dimension u_1. Put $\bar{X}_1=\bar{N}_0\cap(\bar{N}_0\cap\bar{N}_1)^\perp$ and fix the projector $\bar{Q}_1$ by means of $\operatorname{im}\bar{Q}_1=\bar{N}_1$, $\ker\bar{Q}_1=\bar{X}_1\oplus(\bar{N}_0+\bar{N}_1)^\perp$. $\bar{Q}_1$ is continuous, but for the sequence $\bar{Q}_0,\bar{Q}_1$ to be admissible, $D\bar{\Pi}_1\bar{D}^-$ has to belong to the class $\mathcal{C}^1$. This projector has the nullspace $\ker D\bar{\Pi}_1\bar{D}^- = D(\bar{N}_0+\bar{N}_1)\oplus\ker R = D(N_0+N_1)\oplus\ker R$, which is already known to belong to $\mathcal{C}^1$. If $D\bar{\Pi}_1\bar{D}^-$ has a range that is a $\mathcal{C}^1$ subspace, then $D\bar{\Pi}_1\bar{D}^-$ itself is continuously differentiable. Derive $\operatorname{im} D\bar{\Pi}_1\bar{D}^- = \operatorname{im} D\bar{\Pi}_1 = D(\bar{N}_0+\bar{N}_1)^\perp = D(N_0+N_1)^\perp = DD^*(D(N_0+N_1))^\perp$. Since $D(N_0+N_1)$ belongs to the class $\mathcal{C}^1$, so does $(D(N_0+N_1))^\perp$. It turns out that $D\bar{\Pi}_1\bar{D}^-$ is in fact continuously differentiable, and hence, $\bar{Q}_0,\bar{Q}_1$ are admissible.
To use induction, assume that $\bar{Q}_0,\dots,\bar{Q}_{i-1}$ are admissible and widely orthogonal. Lemma 2.12 (d) yields $\bar{G}_i = G_iZ_i$, $\bar{N}_0+\cdots+\bar{N}_{i-1} = N_0+\cdots+N_{i-1}$, $\bar{N}_0+\cdots+\bar{N}_i = N_0+\cdots+N_i$, $Z_i(\bar{N}_i\cap(\bar{N}_0+\cdots+\bar{N}_{i-1}))=N_i\cap(N_0+\cdots+N_{i-1})$.
Since Z_i is nonsingular, it follows that $\bar{G}_i$ has constant rank r_i and the intersection $\overline{N\mathcal{U}}_i=\bar{N}_i\cap(\bar{N}_0+\cdots+\bar{N}_{i-1})$ has constant dimension u_i. The involved subspaces are continuous. Put

$$\bar{X}_i = (\bar{N}_0+\cdots+\bar{N}_{i-1})\cap((\bar{N}_0+\cdots+\bar{N}_{i-1})\cap\bar{N}_i)^\perp$$

and choose $\bar{Q}_i$ to be the projector onto $\bar{N}_i$ along $(\bar{N}_0+\cdots+\bar{N}_i)^\perp\oplus\bar{X}_i$.
$\bar{Q}_0,\dots,\bar{Q}_{i-1},\bar{Q}_i$ would be admissible if $D\bar{\Pi}_i\bar{D}^-$ was continuously differentiable. We know $\ker D\bar{\Pi}_i\bar{D}^- = D(N_0+\cdots+N_i)\oplus\ker R$ to be already continuously differentiable. On the other hand, we have $\operatorname{im} D\bar{\Pi}_i\bar{D}^- = D\operatorname{im}\bar{\Pi}_i = D(N_0+\cdots+N_i)^\perp = DD^*(D(N_0+\cdots+N_i))^\perp$, hence $\operatorname{im} D\bar{\Pi}_i\bar{D}^-$ belongs to the class $\mathcal{C}^1$. □

The widely orthogonal projectors have the advantage that they are uniquely determined. This proves its value in theoretical investigations, for instance in verifying Theorem 3.33 on necessary and sufficient regularity conditions for nonlinear DAEs, as well as for investigating critical points. Moreover, in practical calculations, in general, there might be difficulties in ensuring the continuity of the projector functions Π_i. Fortunately, owing to their uniqueness the widely orthogonal projector functions are continuous a priori.
By Proposition 2.16, at least for all DAEs with properly stated leading term, and with a continuously differentiable coefficient D, we may access widely orthogonal

projector functions. However, if D is just continuous, and if DD^* fails to be continuously differentiable as required, then it may happen in fact that admissible projector functions exist but the special widely orthogonal projector functions do not exist for lack of smoothness. The following example shows this situation. At this point we emphasize that most DAEs are given with a smooth D, and our example is rather academic.

Example 2.17 (Lack of smoothness for widely orthogonal projectors). Given the DAE

$$\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\left(\begin{bmatrix}1&\alpha&0\\0&1&0\end{bmatrix}x\right)'+\begin{bmatrix}0&0&0\\0&0&-1\\0&1&0\end{bmatrix}x=q,$$

with a continuous scalar function α, the DAE has the coefficients

$$A=\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix},\ D=\begin{bmatrix}1&\alpha&0\\0&1&0\end{bmatrix},\ B=\begin{bmatrix}0&0&0\\0&0&-1\\0&1&0\end{bmatrix},\ R=\begin{bmatrix}1&0\\0&1\end{bmatrix}.$$

First we construct an admissible matrix function sequence. Set and derive

$$D^-=\begin{bmatrix}1&-\alpha\\0&1\\0&0\end{bmatrix},\ G_0=\begin{bmatrix}1&\alpha&0\\0&1&0\\0&0&0\end{bmatrix},\ Q_0=\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix},\ G_1=\begin{bmatrix}1&\alpha&0\\0&1&-1\\0&0&0\end{bmatrix},\qquad(2.24)$$

and further

$$Q_1=\begin{bmatrix}0&-\alpha&0\\0&1&0\\0&1&0\end{bmatrix},\ Q_1Q_0=0,\ D\Pi_1D^-=\begin{bmatrix}1&0\\0&0\end{bmatrix},\ G_2=\begin{bmatrix}1&\alpha&0\\0&1&-1\\0&1&0\end{bmatrix}.$$

The projector functions Q_0, Q_1 are admissible, and G_2 is nonsingular, such that $Q_2=0$. This sequence is admissible for each arbitrary continuous α; however it fails to be widely orthogonal. Namely, the product Π_0Q_1 is not symmetric.
Next we construct widely orthogonal projector functions. We start with the same matrix functions Q_0, D^- and G_1 (see (2.24)). Compute further

$$N_0\oplus N_1=\operatorname{span}\left\{\begin{bmatrix}0\\0\\1\end{bmatrix},\ \begin{bmatrix}-\alpha\\1\\1\end{bmatrix}\right\},\ (N_0\oplus N_1)^\perp=\operatorname{span}\begin{bmatrix}1\\\alpha\\0\end{bmatrix}.$$

The required projector function onto N_1 along $N_0\oplus(N_0\oplus N_1)^\perp$ is

$$Q_1=\frac{1}{1+\alpha^2}\begin{bmatrix}\alpha^2&-\alpha&0\\-\alpha&1&0\\-\alpha&1&0\end{bmatrix},\ \text{and it follows that}\quad D\Pi_1D^-=\begin{bmatrix}1&0\\\frac{\alpha}{1+\alpha^2}&0\end{bmatrix}.$$

We recognize that, in the given setting, $D\Pi_1 D^-$ is just continuous. If we additionally assume that $\alpha \in \mathcal{C}^1(\mathcal{I},\mathbb{R})$, then Q_0, Q_1 appear to be admissible. Notice that in this case $DD^* = \begin{bmatrix} 1+\alpha^2 & \alpha \\ \alpha & 1 \end{bmatrix}$ is continuously differentiable, which confirms Proposition 2.16 once more.

Let us stress that this special DAE is solvable for arbitrary continuous α. From this point of view there is no need to assume α to be $\mathcal{C}^1$. Namely, the detailed equations are

$$\begin{aligned} (x_1 + \alpha x_2)' &= q_1, \\ x_2' - x_3 &= q_2, \\ x_2 &= q_3, \end{aligned}$$

with the solutions

$$\begin{aligned} x_1(t) + \alpha(t)x_2(t) &= x_1(0) + \alpha(0)x_2(0) + \int_0^t q_3(s)ds, \\ x_2(t) &= q_3(t), \\ x_3(t) &= q_3'(t) - q_2(t). \end{aligned}$$

It turns out that widely orthogonal projectors need some specific slightly higher smoothness which is not justified by solvability. □

2.3 Invariants under transformations and refactorizations

Given the DAE (2.1) with continuous coefficients and properly stated leading term, we premultiply this equation by a nonsingular matrix function $L \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^k))$ and transform the unknown $x = K\bar{x}$ by means of a nonsingular matrix function $K \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$ such that the DAE

$$\bar{A}(\bar{D}\bar{x})' + \bar{B}\bar{x} = \bar{q} \tag{2.25}$$

results, where $\bar{q} := Lq$, and

$$\bar{A} := LA, \quad \bar{D} := DK, \quad \bar{B} := LBK. \tag{2.26}$$

These transformed coefficients are continuous as are the original ones. Moreover, $\bar{A}$ and $\bar{D}$ inherit from A and D the constant ranks, and the leading term of (2.25) is properly stated (cf. Definition 2.1) with the same border projector $\bar{R} = R$ as $\ker \bar{A} = \ker A$, $\operatorname{im} \bar{D} = \operatorname{im} D$.

Suppose that the original DAE (2.1) has admissible projectors $Q_0, \ldots, Q_\kappa$. We form a corresponding matrix function sequence for the transformed DAE (2.25) starting with

$$\bar{G}_0 = \bar{A}\bar{D} = LADK = LG_0K, \quad \bar{B}_0 = \bar{B} = LB_0K,$$
$$\bar{Q}_0 := K^{-1}Q_0K, \quad \bar{D}^- = K^{-1}D^-, \quad \bar{P}_0 = K^{-1}P_0K,$$

such that $\bar{D}\bar{D}^- = DD^- = R$, $\bar{D}^-\bar{D} = \bar{P}_0$, and

$$\bar{G}_1 = \bar{G}_0 + \bar{B}_0\bar{Q}_0 = L(G_0 + B_0Q_0)K = LG_1K.$$

This yields $\bar{N}_0 = K^{-1}N_0$, $\bar{N}_1 = K^{-1}N_1$, $\bar{N}_0 \cap \bar{N}_1 = K^{-1}(N_0 \cap N_1)$. Choose $\bar{Q}_1 := K^{-1}Q_1K$ which corresponds to $\bar{X}_1 := K^{-1}X_1$. Proceeding in this way at each level, $i = 1, \ldots, \kappa$, with

$$\bar{Q}_i := K^{-1}Q_iK$$

it follows that $\bar{\Pi}_i = K^{-1}\Pi_iK$, $\bar{D}\bar{\Pi}_i\bar{D}^- = D\Pi_iD^-$, $\bar{X}_i = K^{-1}X_i$, $\overline{N\mathcal{U}}_i = K^{-1}\widehat{N}_i$, and

$$\bar{G}_{i+1} = LG_{i+1}K, \quad \bar{B}_{i+1} = LB_{i+1}K.$$

This shows that $\bar{Q}_0, \ldots, \bar{Q}_\kappa$ are admissible for (2.25), and the following assertion becomes evident.

Theorem 2.18. *If the DAE (2.1) has an admissible matrix function sequence up to level $\kappa \in \mathbb{N}$, with characteristic values r_i, u_i, $i = 1, \ldots, \kappa$, then the transformed equation (2.25) also has an admissible matrix function sequence up to level κ, with the same characteristic values, i.e., $\bar{r}_i = r_i$, $\bar{u}_i = u_i$, $i = 1, \ldots, \kappa$.*

By Theorem 2.18 the characteristic values are invariant under transformations of the unknown function as well as under premultiplications of the DAE. This feature seems to be rather trivial. The invariance with respect to refactorizations of the leading term, which we verify next, is more subtle.

First we explain what *refactorization* means. For the given DAE (2.1) with properly stated leading term, we consider the product AD to represent a *factorization of the leading term* and we ask whether we can turn to a different factorization $AD = \bar{A}\bar{D}$ such that $\ker \bar{A}$ and $\operatorname{im} \bar{D}$ are again transversal $\mathcal{C}^1$-subspaces. For instance, in Example 2.4, equation (2.10) results from equation (2.9) by taking a different factorization.

In general, we describe the change to a different factorization as follows:
Let $H \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^s, \mathbb{R}^n))$ be given together with a generalized inverse $H^- \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^s))$ such that

$$H^-HH^- = H^-, \quad HH^-H = H, \quad RHH^-R = R. \tag{2.27}$$

H has constant rank greater than or equal to the rank of the border projector R. In particular, one can use any nonsingular $H \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n))$. However, we do not restrict ourselves to square nonsingular matrix functions H.
Due to $AR = ARHH^-R$ we may write

$$\begin{aligned} A(Dx)' = ARHH^-R(Dx)' &= ARH(H^-RDx)' - ARH(H^-R)'Dx \\ &= AH(H^-Dx)' - AH(H^-R)'Dx. \end{aligned}$$

This leads to the new DAE

$$\bar{A}(\bar{D}x)' + \bar{B}x = q \tag{2.28}$$

with the continuous coefficients

$$\bar{A} := AH, \quad \bar{D} := H^- D, \quad \bar{B} := B - ARH(H^- R)' D. \tag{2.29}$$

Because of $\bar{A}\bar{D} = AD$ we call this procedure that changes (2.1) to (2.28) a *refactorization of the leading term*. It holds that

$$\ker \bar{A} = \ker AH = \ker RH, \quad \operatorname{im} \bar{D} = \operatorname{im} H^- D = \operatorname{im} H^- R;$$

further $(H^- RH)^2 = H^- RHH^- RH = H^- RH$. It becomes clear that $H^- RH \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^s))$ is actually the border projector corresponding to the new DAE (2.28), and (2.28) has a properly stated leading term.
We emphasize that the old border space $\mathbb{R}^n$ and the new one $\mathbb{R}^s$ may actually have different dimensions, and this is accompanied by different sizes of the involved matrix functions. Here, the only restriction is $n, s \geq r := \operatorname{rank} D$.

Example 2.19 (A simple refactorization changing the border space dimension). The semi-explicit DAE

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

comprising m_1 and m_2 equations can be written with proper leading term in different ways, for instance as

$$\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix} x \right)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} x = q \tag{2.30}$$

as well as

$$\begin{bmatrix} I \\ 0 \end{bmatrix} \left(\begin{bmatrix} I & 0 \end{bmatrix} x \right)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} x = q. \tag{2.31}$$

The border projector R of the DAE (2.30) as well as H and H^-,

$$R = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad H = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad H^- = \begin{bmatrix} I & 0 \end{bmatrix},$$

satisfy condition (2.27). The DAE (2.31) results from the DAE (2.30) by refactorization of the leading term by means of H. The border projector of the DAE (2.31) is simply $\bar{R} = H^- RH = I$. The dimension of the border space is reduced from $m_1 + m_2$ in (2.30) to m_1 in (2.31). □

Example 2.20 (Nontrivial refactorization). The following two DAEs are given in Example 2.4,

$$\underbrace{\begin{bmatrix}0&1&0\\0&0&1\\0&0&0\end{bmatrix}}_{\tilde{A}}\left(\underbrace{\begin{bmatrix}0&0&0\\0&1&0\\0&-t&1\end{bmatrix}}_{\tilde{D}(t)}x(t)\right)'+\underbrace{\begin{bmatrix}1&0&0\\0&1&0\\0&-t&1\end{bmatrix}}_{\tilde{B}(t)}x(t)=q(t),\quad t\in\mathbb{R}$$

and

$$\underbrace{\begin{bmatrix}0&1&0\\0&-t&1\\0&0&0\end{bmatrix}}_{A(t)}\left(\underbrace{\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}}_{D}x(t)\right)'+\underbrace{\begin{bmatrix}1&0&0\\0&0&0\\0&-t&1\end{bmatrix}}_{B(t)}x(t)=q(t),\quad t\in\mathbb{R}.$$

The border projector of the last DAE is simply $R=D$. The nonsingular matrix function

$$H(t)=\begin{bmatrix}1&0&0\\0&1&0\\0&t&1\end{bmatrix},\quad H(t)^{-}=H(t)^{-1}=\begin{bmatrix}1&0&0\\0&1&0\\0&-t&1\end{bmatrix}$$

fulfills condition (2.27). Comparing the coefficients, one proves that the first DAE results from the refactorization of the second DAE with H. Conversely, one obtains the second DAE by refactorization of the first one with H^{-1}.
Observe that the matrix pencil $\{\tilde{A}\tilde{D}(t),\tilde{B}(t)\}$ is regular with Kronecker index 3, while $\{A(t)D,B(t)\}$ is a singular pencil. This confirms once more the well-known fact that local matrix pencils are inapplicable to characterize time-varying DAEs.
□

Theorem 2.21. *Let the DAE (2.1) have a properly stated leading term and an admissible matrix function sequence up to level $\kappa\in\mathbb{N}$ and characteristic values $r_0,\dots,r_\kappa$, $u_1,\dots,u_\kappa$.*
Let the matrix functions $H\in\mathcal{C}^1(\mathcal{I},L(\mathbb{R}^s,\mathbb{R}^n))$ and $H^-\in\mathcal{C}^1(\mathcal{I},L(\mathbb{R}^n,\mathbb{R}^s))$ satisfy condition (2.27).

(a) *Then the refactorized DAE (2.28) also has a properly stated leading term and an admissible matrix function sequence up to level κ. Its characteristic values coincide with that of (2.1).*
(b) *The subspaces* $\operatorname{im}G_i$, $N_0+\cdots+N_i$, $i=0,\dots,\kappa$, *are invariant.*

Proof. Put $F_1:=I$.
We use induction to show that the following relations are valid:

$$\bar{G}_i=G_iF_i\cdots F_1, \tag{2.32}$$

$$\bar{Q}_i:=(F_i\cdots F_1)^{-1}Q_iF_i\cdots F_1,\quad \bar{\Pi}_{i-1}\bar{Q}_i=\Pi_{i-1}Q_i,\quad \bar{\Pi}_i=\Pi_i, \tag{2.33}$$

$$\bar{B}_i=B_i-G_iD^-H(H^-R)'D\Pi_i+G_i\sum_{j=0}^{i-1}Q_jZ_{ij}\Pi_{i-1}, \tag{2.34}$$

with nonsingular

$$F_i := I + P_{i-1} \sum_{j=0}^{i-2} Q_j Z_{i-1\,j} \Pi_{i-2} Q_{i-1}, \quad i = 1, \ldots, \kappa.$$

The coefficients $Z_{\ell j}$ are continuous matrix functions whose special form does not matter at all.
Since $\bar{G}_0 = \bar{A}\bar{D} = AD = G_0$ we may choose $\bar{D}^- = D^- H$, $\bar{Q}_0 = Q_0$. It follows that $\bar{\Pi}_0 = \Pi_0$, $\bar{B}_0 = \bar{B} = B - ARH(H^-R)'D$ and $\bar{B}_0\bar{Q}_0 = BQ_0 = B_0Q_0$, hence $\bar{G}_1 = \bar{G}_0 + \bar{B}_0\bar{Q}_0 = G_0 + B_0Q_0 = G_1 = G_1F_1$. Choose $\bar{Q}_1 = Q_1 = F_1^{-1}Q_1$ such that $\bar{\Pi}_1 = \Pi_1$, $\bar{\Pi}_0\bar{Q}_1 = \Pi_0Q_1$, $\bar{D}\bar{\Pi}_1\bar{D}^- = H^-D\Pi_1D^-H$, and further

$$\begin{aligned}
\bar{B}_1 &= \bar{B}_0\bar{P}_0 - \bar{G}_1\bar{D}^-(\bar{D}\bar{\Pi}_1\bar{D}^-)'\bar{D}\bar{\Pi}_0 \\
&= B_0P_0 - ARH(H^-R)'D - G_1D^-H(H^-D\Pi_1D^-H)'H^-D\Pi_0 \\
&= B_0P_0 - G_1D^-(D\Pi_1D^-)'D\Pi_0 + G_1D^-(D\Pi_1D^-)'D\Pi_0 \\
&\qquad - ARH(H^-R)'D - G_1D^-H(H^-RD\Pi_1D^-RH)'H^-D\Pi_0 \\
&= B_1 + G_1D^-(D\Pi_1D^-)'D\Pi_0 - ARH(H^-R)'D - G_1D^-H\{(H^-R)'D\Pi_1D^-RH \\
&\qquad + H^-R(D\Pi_1D^-)'RH + H^-RD\Pi_1D^-(RH)'\}H^-D \\
&= B_1 - ARH(H^-R)'D - G_1D^-H(H^-R)'D\Pi_1 - G_1\Pi_1D^-(RH)'H^-RD \\
&= B_1 - G_1D^-H(H^-R)'D\Pi_1 - ARH(H^-R)'D + G_1\Pi_1D^-RH(H^-R)'D.
\end{aligned}$$

In the last expression we have used that

$$D^-(RHH^-R)'D = D^-R'D = 0.$$

Compute $G_1\Pi_1D^-RH(H^-R)'D - ARH(H^-R)'D = G_1(\Pi_1 - I)D^-RH(H^-R)'D$ and

$$\begin{aligned}
G_1(\Pi_1 - I) &= G_1((I - Q_0)(I - Q_1) - I) = G_1(-Q_0 - Q_1 + Q_0Q_1) \\
&= G_1(-Q_0 + Q_0Q_1) = -G_1Q_0P_1.
\end{aligned}$$

This yields the required expression

$$\bar{B}_1 = B_1 - G_1D^-H(H^-R)'D\Pi_1 + G_1Q_0Z_{10}\Pi_0$$

with $Z_{10} := -Q_0P_1D^-RH(H^-R)'D$.
Next, supposing the relations (2.32)–(2.34) to be given up to i, we show their validity for $i+1$. Derive

$$\begin{aligned}
\bar{G}_{i+1} &= \bar{G}_i + \bar{B}_i\bar{Q}_i = \{G_i + \bar{B}_i(F_i \cdots F_1)^{-1}Q_i\}F_i \cdots F_1 \\
&= \{G_i + \bar{B}_i\Pi_{i-1}(F_i \cdots F_1)^{-1}Q_i\}F_i \cdots F_1,
\end{aligned}$$

and, because of $\Pi_{i-1}F_1^{-1} \cdots F_i^{-1} = \Pi_{i-1}$, we obtain further

$$\begin{aligned}\bar{G}_{i+1} &= \Big\{G_i + B_i Q_i - G_i D^- H (H^- R)' D \Pi_i Q_i + G_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1} Q_i\Big\} F_i \cdots F_1 \\ &= \Big\{G_{i+1} + G_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1} Q_i\Big\} F_i \cdots F_1 \\ &= G_{i+1}\Big\{I + P_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1} Q_i\Big\} F_i \cdots F_1 \\ &= G_{i+1} F_{i+1} F_i \cdots F_1,\end{aligned}$$

with nonsingular matrix functions

$$F_{i+1} = I + P_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1} Q_i, \quad F_{i+1}^{-1} = I - P_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1} Q_i.$$

Put $\bar{Q}_{i+1} := (F_{i+1} \cdots F_1)^{-1} Q_{i+1} F_{i+1} \cdots F_1$, and compute

$$\begin{aligned}\bar{\Pi}_i \bar{Q}_{i+1} &= \Pi_i \bar{Q}_{i+1} = \Pi_i F_1^{-1} \cdots F_{i+1}^{-1} Q_{i+1} F_{i+1} \cdots F_1 \\ &= \Pi_i Q_{i+1} F_{i+1} \cdots F_1 = \Pi_i Q_{i+1} \Pi_i F_{i+1} \cdots F_1 = \Pi_i Q_{i+1} \Pi_i = \Pi_i Q_{i+1},\end{aligned}$$

$$\bar{\Pi}_{i+1} = \bar{\Pi}_i - \bar{\Pi}_i \bar{Q}_{i+1} = \Pi_i - \Pi_i Q_{i+1} = \Pi_{i+1}.$$

It remains to verify the expression for $\bar{B}_{i+1}$. We derive

$$\begin{aligned}\bar{B}_{i+1} &= \bar{B}_i \bar{P}_i - \bar{G}_{i+1} \bar{D}^- (\bar{D} \bar{\Pi}_{i+1} \bar{D}^-)' \bar{D} \bar{\Pi}_i \\ &= \bar{B}_i \Pi_i - G_{i+1} F_{i+1} \cdots F_1 D^- H (H^- D \Pi_{i+1} D^- H)' H^- D \Pi_i,\end{aligned}$$

and

$$\begin{aligned}\bar{B}_{i+1} = \Big\{&B_i - G_i D^- H (H^- R)' D \Pi_i + G_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_{i-1}\Big\} \Pi_i \\ &- G_{i+1}(F_{i+1} \cdots F_1 - I) D^- H (H^- D \Pi_{i+1} D^- H)' H^- D \Pi_i \\ &- G_{i+1} D^- H \{(H^- R)' R D \Pi_{i+1} D^- R H + H^- R (D \Pi_{i+1} D^-)' R H \\ &+ H^- R D \Pi_{i+1} D^- (R H)'\} H^- D \Pi_i,\end{aligned}$$

and

$$\begin{aligned}\bar{B}_{i+1} = &B_i P_i - G_i D^- H (H^- R)' D \Pi_i + G_i \sum_{j=0}^{i-1} Q_j Z_{ij} \Pi_i \\ &- G_{i+1} D^- H (H^- R)' D \Pi_{i+1} - G_{i+1} D^- (D \Pi_{i+1} D^-)' D \Pi_i \\ &- G_{i+1} \Pi_{i+1} D^- (R H)' H^- R D \Pi_i \\ &- G_{i+1}(F_{i+1} \cdots F_1 - I) D^- H (H^- D \Pi_{i+1} D^- H)' H^- D \Pi_i,\end{aligned}$$

and

$$\begin{aligned}\bar{B}_{i+1} = B_{i+1} - G_{i+1}D^-H(H^-R)'D\Pi_{i+1} - G_{i+1}P_iD^-H(H^-R)'D\Pi_i\\ +G_{i+1}\Pi_{i+1}D^-H(H^-R)'D\Pi_i + G_{i+1}P_i\sum_{j=0}^{i-1}Q_jZ_{ij}\Pi_i\\ -G_{i+1}(F_{i+1}\cdots F_i - I)D^-H(H^-D\Pi_{i+1}D^-H)'H^-D\Pi_i.\end{aligned}$$

Finally, decomposing

$$P_i\sum_{j=0}^{i-1}Q_jZ_{ij}\Pi_i = \sum_{j=0}^{i-1}Q_jZ_{ij}\Pi_i - Q_i\sum_{j=0}^{i-1}Q_jZ_{ij}\Pi_i,$$

and expressing

$$F_{i+1}\cdots F_1 - I = \sum_{j=0}^{i}Q_j\mathfrak{A}_{i+1,j},$$

and taking into account that

$$G_{i+1}\{\Pi_{i+1} - P_i\}D^-H(H^-R)'D\Pi_i = G_{i+1}\sum_{j=0}^{i}Q_i\mathfrak{B}_{i+1,j}D^-H(H^-R)'D\Pi_i$$

we obtain

$$\bar{B}_{i+1} = B_{i+1} - G_{i+1}D^-H(H^-R)'D\Pi_{i+1} + \sum_{j=0}^{i}Q_jZ_{i+1,j}D\Pi_i.$$

□

By Theorem 2.21, the characteristic values and the tractability index are invariant under refactorizations of the leading term. In this way, the size of A and D may change or not (cf. Examples 2.4 and 2.19).

It is worth mentioning that also the associated function space accommodating the solutions of the DAE remains invariant under refactorizations as the next proposition shows.

Proposition 2.22. *Given the matrix function $D \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n))$ and the projector function $R \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n))$ onto $\operatorname{im} D$, let $H \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^s, \mathbb{R}^n))$ be given together with a generalized inverse $H^- \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^s))$ such that $H^-HH^- = H^-$, $HH^-H = H$, and $RHH^-R = R$. Then, for $\bar{D} = H^-D$, it holds that*

$$\mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m) = \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m).$$

Proof. For any $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ we find $\bar{D}x = H^-Dx \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^s)$, and hence $x \in \mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m)$. Conversely, for $x \in \mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m)$, we find $Dx = RDx = RHH^-Dx = RH\bar{D}x \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^s)$, and hence $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. □

2.4 Decoupling regular DAEs

The main objective of this section is the characterization of *regular DAEs* by means of admissible matrix function sequences and the projector based *structural decoupling* of each regular DAE (2.1) into an inherent regular ODE

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

and a triangular subsystem of several equations including differentiations

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ (Dv_1)' \\ \vdots \\ (Dv_{\mu-1})' \end{bmatrix}$$
$$+ \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} + \begin{bmatrix} \mathcal{H}_0 \\ \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_{\mu-1} \end{bmatrix} D^- u = \begin{bmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} q.$$

This structural decoupling is associated with the decomposition (see Theorem 2.30)

$$x = D^- u + v_0 + v_1 + \cdots + v_{\mu-1}.$$

2.4.1 Preliminary decoupling rearrangements

We apply admissible projector functions $Q_0, \ldots, Q_\kappa$ to rearrange terms within the DAE (2.1) in a similar way as done in Chapter 1 on constant coefficient DAEs for obtaining decoupled systems. The objective of the rearrangements is to place a matrix function G_κ in front of the derivative component $(D\Pi_\kappa x)'$, the rank of which is as large as possible, and at the same time to separate terms living in $N_0 + \cdots + N_\kappa$. We emphasize that we do not change the given DAE at all, and we do not transform the variables. We work just with the given DAE and its unknown. What we do are *rearrangements of terms and separations or decouplings of solution components* by means of projector functions. We proceed stepwise. Within this procedure, the special form of the matrix functions B_i appears to make good sense.
This part is valid for general DAEs with proper leading term, possibly with less or more variables than equations ($m \neq k$). The rearranged DAE versions serve then as the basis for further decouplings and solutions in the present chapter and also in Chapter 10.

First rewrite (2.1) as

$$G_0 D^- (Dx)' + B_0 x = q, \tag{2.35}$$

and then as

$$G_0 D^-(Dx)' + B_0(Q_0 x + P_0 x) = q$$

and rearrange this in order to increase the rank of the leading coefficient to

$$(G_0 + B_0 Q_0)(D^-(Dx)' + Q_0 x) + B_0 P_0 x = q,$$

or

$$G_1 D^-(Dx)' + B_0 P_0 x + G_1 Q_0 x = q. \tag{2.36}$$

Compute

$$\begin{aligned} P_1 D^-(Dx)' &= P_0 P_1 D^-(Dx)' + Q_0 P_1 D^-(Dx)' \\ &= D^- D P_0 P_1 D^-(Dx)' + Q_0 P_1 D^-(Dx)' \\ &= D^-(D P_0 P_1 x)' - D^-(D P_0 P_1 D^-)' Dx + Q_0 P_1 D^-(Dx)' \\ &= D^-(D P_0 P_1 x)' - D^-(D P_0 P_1 D^-)' Dx - (I - P_0) Q_1 D^-(Dx)' \\ &= D^-(D\Pi_1 x)' - D^-(D\Pi_1 D^-)' Dx - (I - \Pi_0) Q_1 D^-(D\Pi_0 x)', \end{aligned}$$

and hence

$$G_1 D^-(Dx)' = G_1 D^-(D\Pi_1 x)' - G_1 D^-(D\Pi_1 D^-)' D P_0 x - G_1 (I - \Pi_0) Q_1 D^-(D\Pi_0 x)'.$$

Inserting this into (2.36) yields

$$\begin{aligned} G_1 D^-(D\Pi_1 x)' &+ (B_0 P_0 - G_1 D^-(D\Pi_1 D^-)' D P_0) x \\ &+ G_1 \{Q_0 x - (I - \Pi_0) Q_1 D^-(Dx)'\} = q, \end{aligned}$$

and, regarding the definition of the matrix function B_1,

$$G_1 D^-(D\Pi_1 x)' + B_1 x + G_1 \{Q_0 x - (I - \Pi_0) Q_1 D^-(Dx)'\} = q. \tag{2.37}$$

Note that, if $N_0 \cap N_1 = 0$, then the derivative $(D\Pi_1 x)'$ is no longer involved in the term

$$Q_1 D^-(Dx)' = Q_1 D^- D P_0 Q_1 D^-(Dx)' = Q_1 D^-(D P_0 Q_1 x)' - Q_1 D^-(D P_0 Q_1 D^-)' Dx.$$

In the next step we move a part of the term $B_1 x$ in (2.37) to the leading term, and so on. Proposition 2.23 describes the result of these systematic rearrangements.

Proposition 2.23. *Let the DAE (2.1) with properly stated leading term have the admissible projectors $Q_0, \dots, Q_\kappa$, where $\kappa \in \mathbb{N} \cup \{0\}$.*

(1) *Then this DAE can be rewritten in the form*

$$G_\kappa D^-(D\Pi_\kappa x)' + B_\kappa x + G_\kappa \sum_{l=0}^{\kappa-1} \{Q_l x + (I - \Pi_l)(P_l - Q_{l+1} P_l) D^-(D\Pi_l x)'\} = q. \tag{2.38}$$

(2) *If, additionally, all intersections* $\widehat{N}_i$, $i = 1, \dots, \kappa$, *are trivial, then the DAE (2.1) can be rewritten as*

$$G_\kappa D^-(D\Pi_\kappa x)' + B_\kappa x + G_\kappa \sum_{l=0}^{\kappa-1} \{Q_l x - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)' + V_l D\Pi_l x\} = q, \tag{2.39}$$

with coefficients

$$V_l = (I-\Pi_l)\{P_l D^-(D\Pi_l D^-)' - Q_{l+1}D^-(D\Pi_{l+1}D^-)'\}D\Pi_l D^-,\ l = 0, \dots, \kappa-1.$$

Comparing with the rearranged DAE obtained in the constant coefficient case (cf. (1.35)), now we observe the extra terms V_l caused by time-dependent movements of certain subspaces. They disappear in the time-invariant case.

Proof (of Proposition 2.23). (1) In the case of $\kappa = 0$, equation (2.35) is just a trivial reformulation of (2.1). For $\kappa = 1$ we are done by considering (2.37). For applying induction, we suppose for $i+1 \le \kappa$, that (2.1) can be rewritten as

$$G_i D^-(D\Pi_i x)' + B_i x + G_i \sum_{l=0}^{i-1} \{Q_l x + (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l x)'\} = q. \tag{2.40}$$

Represent $B_i x = B_i P_i x + B_i Q_i x = B_i P_i x + G_{i+1}Q_i x$ and derive

$$\begin{aligned} G_i D^-(D\Pi_i x)' &= G_{i+1}P_{i+1}P_i D^-(D\Pi_i x)' \\ &= G_{i+1}\{\Pi_{i+1}P_i D^-(D\Pi_i x)' + (I-\Pi_i)P_{i+1}P_i D^-(D\Pi_i x)'\} \\ &= G_{i+1}\{D^-D\Pi_{i+1}D^-(D\Pi_i x)' + (I-\Pi_i)P_{i+1}P_i D^-(D\Pi_i x)'\} \\ &= G_{i+1}D^-(D\Pi_{i+1}x)' - G_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i x \\ &\qquad + G_{i+1}(I-\Pi_i)(P_i - Q_{i+1}P_i)D^-(D\Pi_i x)'). \end{aligned}$$

Taking into account that $(I-\Pi_i) = Q_0P_1\cdots P_i + \cdots + Q_{i-1}P_i + Q_i$ and $G_iQ_l = G_{i+1}Q_l$, $l = 0, \dots, i-1$, we realize that (2.40) can be reformulated to

$$\begin{aligned} &G_{i+1}D^-(D\Pi_{i+1}x)' + (B_iP_i - G_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i)x \\ &\quad + G_{i+1}Q_i x + G_{i+1}\sum_{l=0}^{i-1}\{Q_l x + (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l x)'\} \\ &\quad + G_{i+1}(I-\Pi_i)(P_i - Q_{i+1}P_i)D^-(D\Pi_i x)' = q. \end{aligned}$$

We obtain in fact

$$G_{i+1}D^-(D\Pi_{i+1}x)' + B_{i+1}x + G_{i+1}\sum_{l=0}^{i}\{Q_l x + (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l x)'\} = q$$

as we tried for.

(2) Finally assuming $\widehat{N}_i = \{0\}$, $i = 1, \dots, \kappa$, and taking into account Proposition 2.7,

we compute the part in question as

$$\mathcal{F} := \sum_{l=0}^{k-1}(I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l x)' = \sum_{l=0}^{k-1}(I-\Pi_l)(P_l - Q_{l+1})D^-(D\Pi_l x)'$$
$$= \sum_{l=0}^{k-1}(I-\Pi_l)\Big\{P_l D^-(D\Pi_l x)' - Q_{l+1}D^- D\Pi_l Q_{l+1}D^-(D\Pi_l x)'\Big\}.$$

Applying the relations

$$\begin{aligned}(D\Pi_l x)' &= (D\Pi_l D^-)'(D\Pi_l x) + D\Pi_l D^-(D\Pi_l x)',\\ (I-\Pi_l)P_l D^- D\Pi_l D^- &= (I-\Pi_l)P_l\Pi_l D^- = 0,\\ D\Pi_l Q_{l+1}D^-(D\Pi_l x)' &= (D\Pi_l Q_{l+1}x)' - (D\Pi_l Q_{l+1}D^-)'D\Pi_l x,\end{aligned}$$

$$\begin{aligned}Q_{l+1}(D\Pi_l Q_{l+1}D^-)'D\Pi_l &= Q_{l+1}(D\Pi_l D^-)'D\Pi_l - Q_{l+1}(D\Pi_{l+1}D^-)'D\Pi_l\\ &= -Q_{l+1}(D\Pi_{l+1}D^-)'D\Pi_l,\end{aligned}$$

we obtain, with the coefficients V_l described by the assertion,

$$\mathcal{F} = \sum_{l=0}^{k-1}(I-\Pi_l)\Big\{P_l D^-(D\Pi_l D^-)'D\Pi_l x + Q_{l+1}D^-(D\Pi_l Q_{l+1}D^-)'D\Pi_l x$$
$$- Q_{l+1}D^-(D\Pi_l Q_{l+1}x)'\Big\} = \sum_{l=0}^{k-1}\Big\{V_l D\Pi_l x - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)'\Big\},$$

and this completes the proof. □

How can one make use of the rearranged version of the DAE (2.1) and the structural information included in this version? We discuss this question in the next subsection for the case of regular DAEs, that is, if $m = k$ and a nonsingular G_μ exists. We study nonregular cases in Chapter 10.

For the moment, to gain a first impression, we cast a glance at the simplest situation, if G_0 already has maximal rank. Later on we assign the *tractability index 0* to each DAE whose matrix functions G_0 already have maximal rank. Then the DAE (2.35) splits into the two parts

$$G_0 D^-(Dx)' + G_0 G_0^- B_0 x = G_0 G_0^- q, \quad \mathcal{W}_0 B_0 x = \mathcal{W}_0 q. \tag{2.41}$$

Since $\operatorname{im} G_0$ is maximal, it holds that $\operatorname{im} B_0 Q_0 \subseteq \operatorname{im} G_1 = \operatorname{im} G_0$, hence $\mathcal{W}_0 B_0 = \mathcal{W}_0 B_0 P_0$. Further, since $DG_0^- G_0 = D$, we find the DAE (2.35) to be equivalent to the system

$$(Dx)' - R'Dx + DG_0^- B_0 D^- Dx + DG_0^- B_0 Q_0 x = DG_0^- q, \ \ \mathcal{W}_0 B_0 D^- Dx = \mathcal{W}_0 q, \tag{2.42}$$

the solution of which decomposes as $x = D^- Dx + Q_0 x$. It becomes clear that this DAE comprises an explicit ODE for Dx, that has an undetermined part $Q_0 x$ to be

chosen arbitrarily. The ODE for Dx is accompanied by a consistency condition applied to Dx and q. If G_0 is surjective, the consistency condition disappears. If G_0 is injective, then the undetermined component Q_0x disappears. If G_0 is nonsingular, which happens just for $m = k$, then the DAE is nothing other than a regular implicit ODE with respect to Dx.

Example 2.24 (Nonregular DAE). The DAE

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

leads to

$$G_0(t) = \begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix}, \quad Q_0(t) = \begin{bmatrix} 0 & t \\ 0 & 1 \end{bmatrix}, \quad B_0(t) = \begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix}, \quad G_1 = G_0.$$

Compute further

$$D^-(t) = \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \qquad R = 1, \quad G_0^-(t) = \begin{bmatrix} 0 & -1 \\ 0 & 0 \end{bmatrix}, \quad \mathcal{W}_0(t) = \begin{bmatrix} 1 & -t \\ 0 & 0 \end{bmatrix},$$
$$DG_0^- B_0 D^- = 0, \qquad B_0 Q_0 = 0, \qquad DG_0^- = \begin{bmatrix} 0 & 1 \end{bmatrix}.$$

For the second equation in formula (2.42) we obtain

$$\mathcal{W}_0 B_0 x = \mathcal{W}_0 q \Leftrightarrow -x_1 + tx_2 = q_1 - tq_2$$

and the inherent explicit ODE in formula (2.42) reads

$$(-x_1 + tx_2)' = q_2.$$

In this way the consistency condition $(q_1 - tq_2)' = q_2$ follows. The solution is

$$\begin{aligned} x(t) &= D^-(-x_1 + tx_2) + Q_0 x \\ &= \begin{bmatrix} x_1 - tx_2 \\ 0 \end{bmatrix} + \begin{bmatrix} tx_2 \\ x_2 \end{bmatrix}, \end{aligned}$$

with an arbitrary continuous function x_2. □

Of course, if the tractability index is greater than 0, things become much more subtle.

2.4.2 Regularity and basic decoupling of regular DAEs

We define regularity for DAEs after the model of classical ODE theory. The system

$$A(t)x'(t) + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.43}$$

with continuous coefficients, is named a regular implicit ODE or an ODE having solely regular line-elements, if the matrix $A(t) \in L(\mathbb{R}^m)$ remains nonsingular on the given interval. Then the homogeneous version of this ODE has a solution space of dimension m and the inhomogeneous ODE is solvable for each continuous excitation q. No question, these properties are maintained, if one turns to a subinterval. On the other hand, a point at which the full-rank condition of the matrix $A(t)$ becomes defective is a critical point, and different kinds of singularities are known to arise (e.g. [123]).
Roughly speaking, in our view, a regular DAE should have similar properties. It should be such that the homogeneous version has a finite-dimensional solution space and no consistency conditions related to the excitations q arise for inhomogeneous equations, which rules out DAEs with more or less unknowns than equations. Additionally, each restriction of a DAE to a subinterval should also inherit all characteristic values.

In the case of constant coefficients, regularity of DAEs is bound to regular pairs of square matrices. In turn, regularity of matrix pairs can be characterized by means of admissible matrix sequences and the associated characteristic values, as described in Section 1.2. A pair of $m \times m$ matrices is regular, if and only if an admissible matrix sequence shows a nonsingular matrix G_μ and the characteristic value $r_\mu = m$. Then the Kronecker index of the given matrix pair results as the smallest such index μ. The same idea applies now to DAEs with time-varying coefficients, too. However, we are now facing continuous *matrix functions* in distinction to the constant matrices in Chapter 1. While, in the case of constant coefficients, admissible projectors do always exist, their existence is now tied to several *rank conditions*. These rank conditions are indeed relevant to the problem. A point at which these rank conditions are defective is considered as a critical point.

We turn back to the DAE (2.1), i.e.,

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}. \tag{2.44}$$

We are looking for solutions in the function space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. Recall that the ranks $r_i = \operatorname{rank} G_i$ in admissible matrix function sequences (see Definitions 2.6, 2.9, Theorem 2.8) give the meaning of characteristics of the DAE on the given interval. The following regularity notion proves to meet the above expectations.

Definition 2.25. The DAE (2.44) with properly stated leading term and $m = k$ is said to be, on the given interval,

(1) *regular with tractability index 0*, if $r_0 = m$,
(2) *regular with tractability index* $\mu \in \mathbb{N}$, if there is an admissible matrix function sequence with characteristic values $r_{\mu-1} < r_\mu = m$,
(3) *regular*, if the DAE is regular with any tractability index μ (i.e., case (1) or (2) apply).

This regularity notion is well-defined in the sense that it is independent of the special choice of the admissible projector functions, which is guaranteed by Theorem 2.8.

Since for a regular DAE the matrix function G_μ is nonsingular, all intersections $\widehat{N}_i = N_i \cap (N_0 + \cdots + N_{i-1})$ are trivial, as a consequence of Proposition 2.7. Then it holds that

$$X_i = (N_0 + \cdots + N_{i-1}) \ominus \widehat{N}_i = N_0 + \cdots + N_{i-1} = N_0 \oplus \cdots \oplus N_{i-1} \subseteq \ker Q_i,$$

$i = 1, \ldots, \mu - 1$, thus $Q_i(I - \Pi_{i-1}) = 0$, and, equivalently,

$$Q_i Q_j = 0, \quad 0 \le j \le i-1, \ \ i = 1, \ldots, \mu - 1. \tag{2.45}$$

Additionally, Proposition 2.7 (4) yields $G_\mu Q_j = B_j Q_j$, thus

$$Q_j = G_\mu^{-1} B_j \Pi_{j-1} Q_j, \quad j = 1, \ldots, \mu - 1. \tag{2.46}$$

While, in the general Definition 2.6, only the part $\Pi_{j-1} Q_j = \Pi_{j-1} - \Pi_j$ of an admissible projector function Q_j is required to be continues, for a regular DAE, the admissible projector functions are continuous in all their components, as follows from the representation (2.46).
We emphasize once again that, for regular DAEs, the admissible projector functions are always *regular* admissible, and they are continuous in all components. At this place, we draw the readers attention to the fact that, in papers dealing exclusively with regular DAEs, the requirements for trivial intersections $\widehat{N}_i$ and the continuity of Q_i are usually already incorporated into the admissibility notion (e.g., [170]) or into the regularity notion (e.g., [167], [137]). Then, the relations (2.46) are constituent parts of the definitions (see also the recent monograph [194]).

Here is a further special quality of regular DAEs: The associated subspaces (cf. Theorem 2.8)

$$S_i = \ker \mathcal{W}_i B = \{z \in \mathbb{R}^m : B_i z \in \operatorname{im} G_i\} = S_{i-1} + N_{i-1}$$

are now $\mathcal{C}$-subspaces, too. They have the constant dimensions r_i. This can be immediately checked. By Lemma A.9, the nonsingularity of G_μ implies the decomposition $N_{\mu-1} \oplus S_{\mu-1} = \mathbb{R}^m$, thus $\dim S_{\mu-1} = r_{\mu-1}$. Regarding the relation $\ker(G_{\mu-2} + \mathcal{W}_{\mu-2} B_{\mu-2} Q_{\mu-2}) = N_{\mu-2} \cap S_{\mu-2}$, we conclude by Proposition 2.5 (3) that $N_{\mu-2} \cap S_{\mu-2}$ has the same dimension as $N_{\mu-1}$ has. This means $\dim N_{\mu-2} \cap S_{\mu-2} = m - r_{\mu-1}$. Next, the representation $S_{\mu-1} = S_{\mu-2} + N_{\mu-2}$ leads to $r_{\mu-1} = \dim S_{\mu-2} + (m - r_{\mu-2}) - (m - r_{\mu-1})$, therefore $\dim S_{\mu-2} = r_{\mu-2}$, and so on.

We decouple the regular DAE (2.44) into its characteristic components, in a similar way as we did with constant coefficient DAEs in Subsection 1.2.2. Since G_μ is nonsingular, by introducing $Q_\mu = 0$, $P_\mu = I$, $\Pi_\mu = \Pi_{\mu-1}$, the sequence $Q_0, \ldots, Q_{\mu-1}, Q_\mu$ is admissible, and we can apply Proposition 2.23. The DAE (2.44) can be rewritten as

$$G_\mu D^-(D\Pi_{\mu-1}x)' + B_\mu x \qquad (2.47)$$
$$+ G_\mu \sum_{l=0}^{\mu-1} \{Q_l x - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)' + V_l D\Pi_l x\} = q.$$

If the coefficients were constant, we would have $D^-(D\Pi_{\mu-1}x)' = (D^-D\Pi_{\mu-1}x)' = (\Pi_{\mu-1}x)'$, further $D^-(D\Pi_l Q_{l+1}x)' = (\Pi_l Q_{l+1}x)'$, and $V_l = 0$. This means that formula (2.47) precisely generalizes formula (1.35) obtained for constant coefficients. The new formula (2.47) contains the extra terms V_l which arise from subspaces moving with time. They disappear in the time-invariant case.

In Subsection 1.2.2, the decoupled version of the DAE is generated by the scaling with G_μ^{-1}, and then by the splitting by means of the projectors $\Pi_{\mu-1}$ and $I-\Pi_{\mu-1}$. Here we go a slightly different way and use $D\Pi_{\mu-1}$ instead of $\Pi_{\mu-1}$. Since $\Pi_{\mu-1}$ can be recovered from $D\Pi_{\mu-1}$ due to $\Pi_{\mu-1} = D^-D\Pi_{\mu-1}$, no information gets lost. Equation (2.47) scaled by G_μ^{-1} reads

$$D^-(D\Pi_{\mu-1}x)' + G_\mu^{-1}B_\mu x \qquad (2.48)$$
$$+ \sum_{l=0}^{\mu-1} \{Q_l x - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)' + V_l D\Pi_l x\} = G_\mu^{-1}q.$$

The detailed expression for V_l (Proposition 2.23) is

$$V_l = (I-\Pi_l)\{P_l D^-(D\Pi_l D^-)' - Q_{l+1}D^-(D\Pi_{l+1}D^-)'\}D\Pi_l D^-.$$

This yields $D\Pi_{\mu-1}V_l = 0$, $l = 0,\ldots,\mu-1$, and multiplying (2.48) by $D\Pi_{\mu-1}$ results in the equation

$$D\Pi_{\mu-1}D^-(D\Pi_{\mu-1}x)' + D\Pi_{\mu-1}G_\mu^{-1}B_\mu x = D\Pi_{\mu-1}G_\mu^{-1}. \qquad (2.49)$$

Applying the $\mathcal{C}^1$-property of the projector $D\Pi_{\mu-1}D^-$, and recognizing that $B_\mu = B_\mu\Pi_{\mu-1} = B_\mu D^-D\Pi_{\mu-1}$, we get

$$(D\Pi_{\mu-1}x)' - (D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}x + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^-D\Pi_{\mu-1}x = D\Pi_{\mu-1}G_\mu^{-1}q. \qquad (2.50)$$

Equation (2.50) is an explicit ODE with respect to the component $D\Pi_{\mu-1}x$. A similar ODE is described by formula (1.37) for the time-invariant case. Our new ODE (2.50) generalizes the ODE (1.37) in the sense that, due to $D^-D\Pi_{\mu-1} = \Pi_{\mu-1}$, equation (2.50) multiplied by D^- coincides with (1.37) for constant coefficients.

Definition 2.26. For the regular DAE (2.44) with tractability index μ, and admissible projector functions $Q_0,\ldots,Q_{\mu-1}$, the resulting explicit regular ODE

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}B_\mu D^-u = D\Pi_{\mu-1}G_\mu^{-1}q \qquad (2.51)$$

is called an *inherent explicit regular ODE* (IERODE) of the DAE.

It should be pointed out that there is a great variety of admissible projector functions. In consequence, there are various projector functions $\Pi_{\mu-1}$, and the IERODE (2.51) is not unique, except for the index-1 case. So far, we know the nullspace $N_0 + \cdots + N_{\mu-1}$ of the projector function $\Pi_{\mu-1}$ to be independent of the choice of the admissible projector functions $Q_0, \ldots, Q_{\mu-1}$, which means the subspace $N_0 + \cdots + N_{\mu-1}$ is unique; it is determined by the DAE coefficients only (Theorem 2.8). Later on we introduce advanced *fine decouplings* which make the corresponding IERODE unique.

Lemma 2.27. *If the DAE (2.44) is regular with index μ, and $Q_0, \ldots, Q_{\mu-1}$ are admissible, then the subspace* $\operatorname{im} D\Pi_{\mu-1}$ *is an invariant subspace for the IERODE (2.51), that is, the following assertion is valid for the solutions $u \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ of the ODE (2.51):*

$$u(t_*) \in \operatorname{im}(D\Pi_{\mu-1})(t_*), \text{ with a certain } t_* \in \mathcal{I} \;\Leftrightarrow\; u(t) \in \operatorname{im}(D\Pi_{\mu-1})(t)\ \forall\, t \in \mathcal{I}.$$

Proof. Let $\bar{u} \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ denote a solution of (2.51) with $\bar{u}(t_*) = (D\Pi_{\mu-1}D^-)(t_*)\bar{u}(t_*)$. We multiply the identity

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

by $I - D\Pi_{\mu-1}D^-$, and introduce the function $\bar{v} := (I - D\Pi_{\mu-1}D^-)\bar{u} \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$. This gives

$$(I - D\Pi_{\mu-1}D^-)\bar{u}' - (I - D\Pi_{\mu-1}D^-)(D\Pi_{\mu-1}D^-)'\bar{u} = 0,$$

further,

$$\bar{v}' - (I - D\Pi_{\mu-1}D^-)'\bar{u} - (I - D\Pi_{\mu-1}D^-)(D\Pi_{\mu-1}D^-)'\,\bar{u} = 0,$$

and

$$\bar{v}' - (I - D\Pi_{\mu-1}D^-)'\bar{v} = 0.$$

Because of $\bar{v}(t_*) = 0$, $\bar{v}$ must vanish identically, and hence $\bar{u} = D\Pi_{\mu-1}D^-\bar{u}$ holds true. □

We leave the IERODE for a while, and turn back to the scaled version (2.48) of the DAE (2.44). Now we consider the other part of this equation, which results from multiplication by the projector function $I - \Pi_{\mu-1}$. First we express

$$\begin{aligned}
&(I - \Pi_{\mu-1})D^-(D\Pi_{\mu-1}x)' + (I - \Pi_{\mu-1})G_\mu^{-1}B_\mu x \\
&\qquad = (I - \Pi_{\mu-1})G_\mu^{-1}\{G_\mu D^-(D\Pi_{\mu-1}x)' + B_{\mu-1}P_{\mu-1}x \\
&\qquad\qquad\qquad - G_\mu D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}x\} \\
&\qquad = (I - \Pi_{\mu-1})G_\mu^{-1}\{B_{\mu-1}P_{\mu-1}x + G_\mu D^- D\Pi_{\mu-1}D^-(D\Pi_{\mu-1}x)'\} \\
&\qquad = (I - \Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1}x,
\end{aligned}$$

and then obtain the equation

$$(I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1}x+\sum_{l=0}^{\mu-1}\{Q_lx+V_lD\Pi_lx\} \tag{2.52}$$
$$-\sum_{l=0}^{\mu-2}(I-\Pi_l)Q_{l+1}D^-(D\Pi_lQ_{l+1}x)'=(I-\Pi_{\mu-1})G_\mu^{-1}q,$$

which is the precise counterpart of equation (1.38). Again, the extra terms V_l comprise the time variation. By means of the decompositions

$$\begin{aligned}D\Pi_lx&=D\Pi_l(\Pi_{\mu-1}+I-\Pi_{\mu-1})x=D\Pi_{\mu-1}x+D\Pi_l(I-P_{l+1}\cdots P_{\mu-1})x\\&=D\Pi_{\mu-1}x+D\Pi_l(Q_{l+1}+P_{l+1}Q_{l+2}+\cdots+P_{l+1}\cdots P_{\mu-2}Q_{\mu-1})x\\&=D\Pi_{\mu-1}x+D\Pi_l(Q_{l+1}+\cdots+D\Pi_{\mu-2}Q_{\mu-1})x,\end{aligned}$$

we rearrange the terms in (2.52) once more as

$$\sum_{l=0}^{\mu-1}Q_lx-\sum_{l=0}^{\mu-2}(I-\Pi_l)Q_{l+1}D^-(D\Pi_lQ_{l+1}x)'+\sum_{l=0}^{\mu-2}\mathcal{M}_{l+1}D\Pi_lQ_{l+1}x \tag{2.53}$$
$$+\mathcal{K}\Pi_{\mu-1}x=(I-\Pi_{\mu-1})G_\mu^{-1}q,$$

with the continuous coefficients

$$\begin{aligned}\mathcal{K}:&=(I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1}+\sum_{l=0}^{\mu-1}V_lD\Pi_{\mu-1} \qquad (2.54)\\&=(I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1}+\sum_{l=0}^{\mu-1}(I-\Pi_l)\Big\{P_lD^-(D\Pi_lD^-)'\\&\qquad\qquad-Q_{l+1}D^-(D\Pi_{l+1}D^-)'\Big\}D\Pi_{\mu-1}\\&=(I-\Pi_{\mu-1})G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1}+\sum_{l=1}^{\mu-1}(I-\Pi_{l-1})(P_l-Q_l)D^-(D\Pi_lD^-)'D\Pi_{\mu-1}\end{aligned}$$

and

$$\begin{aligned}\mathcal{M}_{l+1}:&=\sum_{j=0}^{l}V_jD\Pi_lQ_{l+1}D^- \qquad (2.55)\\&=\sum_{j=0}^{l}(I-\Pi_j)\{P_jD^-(D\Pi_jD^-)'-Q_{j+1}D^-(D\Pi_{j+1}D^-)'\}D\Pi_lQ_{l+1}D^-,\\&\qquad\qquad l=0,\dots,\mu-2.\end{aligned}$$

The coefficients $\mathcal{M}_{l+1}$ vanish together with the V_j in the constant coefficient case.

Next we provide a further splitting of the subsystem (2.53) according to the decomposition

$$I - \Pi_{\mu-1} = Q_0 P_1 \cdots P_{\mu-1} + \cdots + Q_{\mu-2} P_{\mu-1} + Q_{\mu-1}$$

into μ parts. Notice that the products $Q_i P_{i+1} \cdots P_{\mu-1}$ are also continuous projectors. To prepare the further decoupling we provide some useful properties of our projectors and coefficients.

Lemma 2.28. *For the regular DAE (2.44) with tractability index μ, and admissible projector functions $Q_0, \ldots, Q_{\mu-1}$, the following relations become true:*

(1) $$Q_i P_{i+1} \cdots P_{\mu-1}(I - \Pi_l) = 0, \qquad l = 0, \ldots, i-1, \quad i = 1, \ldots, \mu-2,$$
$$Q_{\mu-1}(I - \Pi_l) = 0, \qquad l = 0, \ldots, \mu-2,$$

(2) $$Q_i P_{i+1} \cdots P_{\mu-1}(I - \Pi_i) = Q_i, \qquad i = 0, \ldots, \mu-2,$$
$$Q_{\mu-1}(I - \Pi_{\mu-1}) = Q_{\mu-1},$$

(3) $$Q_i P_{i+1} \cdots P_{\mu-1}(I - \Pi_{i+s}) = Q_i P_{i+1} \cdots P_{i+s}, \qquad s = 1, \ldots, \mu-1-i, \quad i = 0, \ldots, \mu-2,$$

(4) $$Q_i P_{i+1} \cdots P_{\mu-1} \mathcal{M}_{l+1} = 0, \qquad l = 0, \ldots, i-1, \quad i = 0, \ldots, \mu-2,$$
$$Q_{\mu-1} \mathcal{M}_{l+1} = 0, \qquad l = 0, \ldots, \mu-2,$$

(5) $$Q_i P_{i+1} \cdots P_{\mu-1} Q_s = 0 \text{ if } s \neq i, \qquad s = 0, \ldots, \mu-1,$$
$$Q_i P_{i+1} \cdots P_{\mu-1} Q_i = Q_i, \qquad i = 0, \ldots, \mu-2,$$

(6) $$\mathcal{M}_j = \sum_{l=1}^{j-1} (I - \Pi_{l-1})(P_l - Q_l) D^- (D \Pi_{j-1} Q_j D^-)' D \Pi_{j-1} Q_j D^-, \qquad j = 1, \ldots, \mu-1,$$

(7) $$\Pi_{\mu-1} G_\mu^{-1} B_\mu = \Pi_{\mu-1} G_\mu^{-1} B_0 \Pi_{\mu-1}, \text{ and hence}$$
$$D \Pi_{\mu-1} G_\mu^{-1} B_\mu D^- = D \Pi_{\mu-1} G_\mu^{-1} B D^-.$$

Proof. (1) The first part of the assertion results from the relation $Q_i P_{i+1} \cdots P_{\mu-1} = Q_i P_{i+1} \cdots P_{\mu-1} \Pi_{i-1}$, and the inclusion $\operatorname{im}(I - \Pi_l) \subseteq \ker \Pi_{i-1}$, $l \leq i-1$. The second part is a consequence of the inclusion $\operatorname{im}(I - \Pi_l) \subseteq \ker Q_{\mu-1}$, $l \leq \mu-2$.
(2) This is a consequence of the relations $P_{i+1} \cdots P_{\mu-1}(I - \Pi_i) = (I - \Pi_i)$ and $Q_i(I - \Pi_i) = Q_i$.
(3) We have

$$Q_i P_{i+1} \cdots P_{\mu-1} \Pi_{\mu-1} = 0, \quad \text{thus} \quad Q_i P_{i+1} \cdots P_{\mu-1}(I - \Pi_{\mu-1}) = Q_i P_{i+1} \cdots P_{\mu-1}.$$

Taking into account that $Q_j(I - \Pi_{i+s}) = 0$ for $j > i+s$, we find

$$\begin{aligned} Q_i P_{i+1} \cdots P_{\mu-1}(I-\Pi_{i+s}) &= Q_i P_{i+1} \cdots P_{i+s} P_{i+s+1} \cdots P_{\mu-1}(I-\Pi_{i+s}) \\ &= Q_i P_{i+1} \cdots P_{i+s} P_{i+s+1} \cdots P_{\mu-1}(I-\Pi_{i+s}) \\ &= Q_i P_{i+1} \cdots P_{i+s}(I-\Pi_{i+s}) = Q_i P_{i+1} \cdots P_{i+s}. \end{aligned}$$

(4) This is a consequence of (1).
(5) This is evident.
(6) We derive

$$\begin{aligned} \mathcal{M}_j &= \sum_{l=1}^{j-1} (I-\Pi_l) P_l D^- (D\Pi_l D^-)' D\Pi_{j-1} Q_j D^- \\ &\qquad - \sum_{l=0}^{j-2} (I-\Pi_l) Q_{l+1} D^- (D\Pi_{l+1} D^-)' D\Pi_{j-1} Q_j D^- \\ &= \sum_{l=1}^{j-1} (I-\Pi_l) P_l D^- \Big\{ (D\Pi_{j-1} Q_j D^-)' - D\Pi_l D^- (D\Pi_{j-1} Q_j D^-)' \Big\} D\Pi_{j-1} Q_j D^- \\ &\qquad - \sum_{l=0}^{j-2} (I-\Pi_l) Q_{l+1} D^- \Big\{ (D\Pi_{j-1} Q_j D^-)' \\ &\qquad\qquad - D\Pi_{l+1} D^- (D\Pi_{j-1} Q_j D^-)' \Big\} D\Pi_{j-1} Q_j D^- \\ &= \sum_{l=1}^{j-1} (I-\Pi_l) P_l D^- (D\Pi_{j-1} Q_j D^-)' D\Pi_{j-1} Q_j D^- \\ &\qquad - \sum_{l=0}^{j-2} (I-\Pi_l) Q_{l+1} D^- (D\Pi_{j-1} Q_j D^-)' D\Pi_{j-1} Q_j D^- \\ &= \sum_{l=1}^{j-1} (I-\Pi_{l-1}) P_l D^- (D\Pi_{j-1} Q_j D^-)' D\Pi_{j-1} Q_j D^- \\ &\qquad - \sum_{l=1}^{j-1} (I-\Pi_{l-1}) Q_l D^- (D\Pi_{j-1} Q_j D^-)' D\Pi_{j-1} Q_j D^-. \end{aligned}$$

(7) Owing to $P_\mu = I$, it holds that

$$\begin{aligned} B_\mu &= B_{\mu-1} P_{\mu-1} - G_\mu D^- (D\Pi_\mu D^-)' D\Pi_{\mu-1} \\ &= B_{\mu-1} P_{\mu-1} - G_\mu D^- (D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1}. \end{aligned}$$

We compute

$$\begin{aligned} \Pi_{\mu-1} G_\mu^{-1} B_\mu &= \Pi_{\mu-1} G_\mu^{-1} \{ B_{\mu-1} P_{\mu-1} - G_\mu D^- (D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1} \} \\ &= \Pi_{\mu-1} G_\mu^{-1} B_{\mu-1} \Pi_{\mu-1} - \underbrace{\Pi_{\mu-1} D^- (D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1}}_{=0}. \end{aligned}$$

The next step is

$$\begin{aligned}\Pi_{\mu-1}G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1} &= \Pi_{\mu-1}G_\mu^{-1}\{B_{\mu-2}P_{\mu-2} - G_{\mu-1}D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-2}\}\Pi_{\mu-1}\\ &= \Pi_{\mu-1}G_\mu^{-1}B_{\mu-2}\Pi_{\mu-1} - \underbrace{\Pi_{\mu-1}P_{\mu-1}D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}}_{=0},\end{aligned}$$

and so on. □

As announced before we split the subsystem (2.53) into μ parts. Multiplying by the projector functions $Q_iP_{i+1}\cdots P_{\mu-1}$, $i=0,\ldots,\mu-2$, and $Q_{\mu-1}$, and regarding Lemma 2.28 one attains the system

$$\begin{aligned}Q_ix - Q_iQ_{i+1}D^-(D\Pi_iQ_{i+1}x)' - \sum_{l=i+1}^{\mu-2} Q_iP_{i+1}\cdots P_lQ_{l+1}D^-(D\Pi_lQ_{l+1}x)' \qquad (2.56)\\ + \sum_{l=i}^{\mu-2} Q_iP_{i+1}\cdots P_{\mu-1}\mathcal{M}_{l+1}D\Pi_lQ_{l+1}x\\ = -Q_iP_{i+1}\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1}x + Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}q, \qquad i=0,\ldots,\mu-2,\end{aligned}$$

as well as

$$Q_{\mu-1}x = -Q_{\mu-1}\mathcal{K}\Pi_{\mu-1}x + Q_{\mu-1}G_\mu^{-1}q. \qquad (2.57)$$

Equation (2.57) determines $Q_{\mu-1}x$ in terms of q and $\Pi_{\mu-1}x$. The i-th equation in (2.56) determines Q_ix in terms of q, $\Pi_{\mu-1}x$, $Q_{\mu-1}x,\ldots,Q_{i+1}x$, and so on, that is, the system (2.56), (2.57) successively determines all components of $I-\Pi_{\mu-1} = Q_0+\Pi_0Q_1+\cdots+\Pi_{\mu-2}Q_{\mu-1}$ in a unique way. Comparing with the constant coefficient case, we recognize that, the system (2.56), (2.57) generalizes the system (1.40), (1.41).

So far, the regular DAE (2.44) decouples into the IERODE (2.51) and the subsystem (2.56), (2.57) by means of each arbitrary admissible matrix function sequence. The solutions of the DAE can be expressed as

$$x = \Pi_{\mu-1}x + (I-\Pi_{\mu-1}x) = D^-u + (I-\Pi_{\mu-1})x,$$

whereby $(I-\Pi_{\mu-1})x$ is determined by the subsystem (2.56), (2.57), and $u = D\Pi_{\mu-1}D^-u$ is a solution of the IERODE, which belongs to its invariant subspace.

The property

$$\ker Q_i = \ker \Pi_{i-1}Q_i, \quad i=1,\ldots,\mu-1, \qquad (2.58)$$

is valid, since we may represent $Q_i = (I+(I-\Pi_{i-1})Q_i)\Pi_{i-1}Q_i$ with the nonsingular factor $I+(I-\Pi_{i-1})Q_i$, $i=1,\ldots,\mu-1$. This allows us to compute Q_ix from $\Pi_{i-1}Q_ix$ and vice versa. We take advantage of this in the following rather cosmetic changes.

Denote (cf. (1.45))

$$v_0 := Q_0x, \quad v_i := \Pi_{i-1}Q_ix, \quad i=1,\ldots,\mu-1, \qquad (2.59)$$

$$u := D\Pi_{\mu-1}x, \qquad (2.60)$$

such that we have the solution expression

$$x = v_0 + v_1 + \cdots + v_{\mu-1} + D^- u. \tag{2.61}$$

Multiply equation (2.57) by $\Pi_{\mu-2}$, and, if $i \geq 1$, the i-th equation in (2.56) by Π_{i-1}. This yields the following system which determines the functions $v_{\mu-1}, \ldots, v_0$ in terms of q and u:

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \left(\mathcal{D} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} \right)' \tag{2.62}$$

$$+ \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} + \begin{bmatrix} \mathcal{H}_0 \\ \mathcal{H}_1 \\ \vdots \\ \mathcal{H}_{\mu-1} \end{bmatrix} D^- u = \begin{bmatrix} \mathcal{L}_0 \\ \mathcal{L}_1 \\ \vdots \\ \mathcal{L}_{\mu-1} \end{bmatrix} q.$$

The matrix function $\mathcal{D} := (\mathcal{D}_{ij})_{i,j=0}^{\mu-1}$ has as entries the blocks $\mathcal{D}_{ii} = D\Pi_{i-1}Q_i$, $i = 1, \ldots, \mu-1$, $\mathcal{D}_{00} = 0$, and $\mathcal{D}_{ij} = 0$, if $i \neq j$. This matrix function is block-diagonal if $n = m$. The further coefficients in (2.62) are also continuous, and their detailed form is

$$\begin{aligned}
\mathcal{N}_{01} &:= -Q_0 Q_1 D^-, \\
\mathcal{N}_{0j} &:= -Q_0 P_1 \cdots P_{j-1} Q_j D^-, && j = 2, \ldots, \mu-1, \\
\mathcal{N}_{i,i+1} &:= -\Pi_{i-1} Q_i Q_{i+1} D^-, \\
\mathcal{N}_{ij} &:= -\Pi_{i-1} Q_i P_{i+1} \cdots P_{j-1} Q_j D^-, && j = i+2, \ldots, \mu-1,\ i = 1, \ldots, \mu-2, \\
\mathcal{M}_{0j} &:= Q_0 P_1 \cdots P_{\mu-1} \mathcal{M}_j D \Pi_{j-1} Q_j, && j = 1, \ldots, \mu-1, \\
\mathcal{M}_{ij} &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} \mathcal{M}_j D \Pi_{j-1} Q_j, && j = i+1, \ldots, \mu-1,\ i = 1, \ldots, \mu-2, \\
\mathcal{L}_0 &:= Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1}, && i = 1, \ldots, \mu-2, \\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1},
\end{aligned}$$

$$\begin{aligned}
\mathcal{H}_0 &:= Q_0 P_1 \cdots P_{\mu-1} \mathcal{K} \Pi_{\mu-1}, \\
\mathcal{H}_i &:= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} \mathcal{K} \Pi_{\mu-1}, && i = 1, \ldots, \mu-2, \\
\mathcal{H}_{\mu-1} &:= \Pi_{\mu-2} Q_{\mu-1} \mathcal{K} \Pi_{\mu-1},
\end{aligned}$$

with $\mathcal{K}$ and $\mathcal{M}_j$ defined by formulas (2.54), (2.55). Introducing the matrix functions $\mathcal{N}$, $\mathcal{M}$, $\mathcal{H}$, $\mathcal{L}$ of appropriate sizes according to (2.62), we write this subsystem as

$$\mathcal{N}(\mathcal{D}v)' + \mathcal{M}v + \mathcal{H}D^- u = \mathcal{L}q, \tag{2.63}$$

whereby the vector function v contains the entries $v_0, \ldots, v_{\mu-1}$.

Again, we draw the reader's attention to the great consistency with (1.46). The difficulties caused by the time-variations are now hidden in the coefficients $\mathcal{M}_{ij}$ which disappear for constant coefficients.
We emphasize that the system (2.62) is nothing other than a more transparent reformulation of the former subsystem (2.56), (2.57). The next proposition records important properties.

Proposition 2.29. *Let the DAE (2.44) be regular with tractability index μ, and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions. Then the coefficient functions in (2.62) have the further properties:*

(1) $\mathcal{N}_{ij} = \mathcal{N}_{ij} D \Pi_{j-1} Q_j D^-$ *and* $\mathcal{N}_{ij} D = \mathcal{N}_{ij} D \Pi_{j-1} Q_j$, *for* $j = 1, \ldots, \mu-1$, $i = 0, \ldots, \mu-2$.
(2) $\operatorname{rank} \mathcal{N}_{i,i+1} = \operatorname{rank} \mathcal{N}_{i,i+1} D = m - r_{i+1}$, *for* $i = 0, \ldots, \mu-2$.
(3) $\ker \mathcal{N}_{i,i+1} = \ker D \Pi_i Q_{i+1} D^-$, *and* $\ker \mathcal{N}_{i,i+1} D = \ker \Pi_i Q_{i+1}$, *for* $i = 0, \ldots, \mu-2$.
(4) *The subsystem (2.62) is a DAE with properly stated leading term.*
(5) *The square matrix function $\mathcal{N}\mathcal{D}$ is pointwise nilpotent with index μ, more precisely,* $(\mathcal{N}\mathcal{D})^\mu = 0$ *and* $\operatorname{rank} (\mathcal{N}\mathcal{D})^{\mu-1} = m - r_{\mu-1} > 0$.
(6) $\mathcal{M}_{i,i+1} = 0, \quad i = 0, \ldots, \mu-2$.

Proof. (1) This is given by the construction.
(2) Because of $\mathcal{N}_{i,i+1} = \mathcal{N}_{i,i+1} D D^-$, the matrix functions $\mathcal{N}_{i,i+1}$ and $\mathcal{N}_{i,i+1} D$ have equal rank. To show that this is precisely $m - r_{i+1}$ we apply the same arguments as for Lemma 1.27. First we validate the relation

$$\operatorname{im} Q_i Q_{i+1} = N_i \cap S_i.$$

Namely, $z \in N_i \cap S_i$ implies $z = Q_i z$ and $B_i z = G_i w$, therefore, $(G_i + B_i Q_i)(P_i w + Q_i z) = 0$, further $(P_i w + Q_i z) = Q_{i+1}(P_i w + Q_i z) = Q_{i+1} w$, $Q_i z = Q_i Q_{i+1} w$, and hence $z = Q_i z = Q_i Q_{i+1} w$.
Conversely, $z \in \operatorname{im} Q_i Q_{i+1}$ yields $z = Q_i z$, $z = Q_i Q_{i+1} w$. Then the identity $(G_i + B_i Q_i) Q_{i+1} = 0$ leads to $B_i z = B_i Q_i Q_{i+1} w = -G_i Q_{i+1} w$, thus $z \in N_i \cap S_i$.
The intersection $N_i \cap S_i$ has the same dimension as N_{i+1}, so that we attain $\dim \operatorname{im} Q_i Q_{i+1} = \dim N_{i+1} = m - r_{i+1}$.
(3) From (1) we derive the inclusions

$$\ker D \Pi_i Q_{i+1} D^- \subseteq \ker \mathcal{N}_{i,i+1}, \quad \ker \Pi_i Q_{i+1} \subseteq \ker \mathcal{N}_{i,i+1} D.$$

Because of $\Pi_i Q_{i+1} = D^-(D \Pi_i Q_{i+1} D^-) D$, and $\ker \Pi_i Q_{i+1} = \ker Q_{i+1}$, the assertion becomes true for reasons of dimensions.
(4) We provide the subspaces

$$\ker \mathcal{N} = \left\{ z = \begin{bmatrix} z_0 \\ \vdots \\ z_{\mu-1} \end{bmatrix} \in \mathbb{R}^{n\mu} : z_i \in \ker \Pi_{i-1} Q_i,\ i = 1, \ldots, \mu - 1 \right\}$$

and

$$\operatorname{im} \mathcal{D} = \left\{ z = \begin{bmatrix} z_0 \\ \vdots \\ z_{\mu-1} \end{bmatrix} \in \mathbb{R}^{n\mu} : z_i \in \operatorname{im} \Pi_{i-1} Q_i,\ i = 1, \ldots, \mu - 1 \right\}$$

which obviously fulfill the condition $\ker \mathcal{N} \oplus \operatorname{im} \mathcal{D} = \mathbb{R}^{n\mu}$. The border projector is $\mathcal{R} = \operatorname{diag}(0, D\Pi_0 Q_1 D^-, \ldots, D\Pi_{\mu-2} Q_{\mu-1} D^-)$, and it is continuously differentiable.
(5) The matrix function $\mathcal{N}\mathcal{D}$ is by nature strictly block upper triangular, and its main entries $(\mathcal{N}\mathcal{D})_{i,i+1} = \mathcal{N}_{i,i+1} D$ have constant rank $m - r_{i+1}$, for $i = 0, \ldots, \mu - 2$.
The matrix function $(\mathcal{N}\mathcal{D})^2$ has zero-entries on the block positions $(i, i+1)$, and the dominating entries are

$$((\mathcal{N}\mathcal{D})^2)_{i,i+2} = \mathcal{N}_{i,i+1} D \mathcal{N}_{i+1,i+2} D = \Pi_{i-1} Q_i Q_{i+1} \Pi_i Q_{i+1} Q_{i+2} = \Pi_{i-1} Q_i Q_{i+1} Q_{i+2},$$

which have rank $m - r_{i+2}$, and so on.
In $(\mathcal{N}\mathcal{D})^{\mu-1}$ there remains exactly one nontrivial block in the upper right corner, $((\mathcal{N}\mathcal{D})^{\mu-1})_{0,\mu-1} = (-1)^{\mu-1} Q_0 Q_1 \cdots Q_{\mu-1}$, and it has rank $m - r_{\mu-1}$.
(6) This property is a direct consequence of the representation of $\mathcal{M}_{i+1}$ in Lemma 2.28 (6) and Lemma 2.28 (1). □

By this proposition, the subsystem (2.62) is in turn a regular DAE with tractability index μ and transparent structure. Property (6) slightly eases the structure of (2.62). We emphasize that the DAE (2.62) lives in $\mathbb{R}^{m\mu}$. The solutions belong to the function space $\mathcal{C}^1_{\mathcal{D}}(\mathcal{I}, \mathbb{R}^{m\mu})$. Owing to the special form of the matrix function $\mathcal{L}$ on the right-hand side, each solution of (2.62) satisfies the conditions $v_0 = Q_0 v_0$ and $v_i = \Pi_{i-1} Q_i v_i$, for $i = 1, \ldots, \mu - 1$.

We now formulate the main result concerning the basic decoupling:

Theorem 2.30. *Let the DAE (2.44) be regular with tractability index μ, and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions. Then the DAE is equivalent via (2.59)–(2.61) to the system consisting of the IERODE (2.51) related to its invariant subspace* $\operatorname{im} D\Pi_{\mu-1}$*, and the subsystem (2.62).*

Proof. If $x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$ is a solution of the DAE, then the component $u := D\Pi_{\mu-1} x \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$ satisfies the IERODE (2.51) and belongs to the invariant subspace $\operatorname{im} \Pi_{\mu-1}$. The functions $v_0 := Q_0 x \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$, $v_i := \Pi_{i-1} Q_i x \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m)$, $i = 1, \ldots, \mu - 1$, form the unique solution of the system (2.62) corresponding to u. Thereby, we recognize that $D\Pi_{\mu-1} x = D\Pi_{\mu-1} D^- Dx$, $Dv_i := D\Pi_{i-1} Q_i x = D\Pi_{i-1} Q_i D^- Dx$, $i = 1, \ldots, \mu - 1$, are continuously differentiable functions since Dx and the used projectors are so.
Conversely, let $u = D\Pi_{\mu-1} x$ denote a solution of the IERODE, and let $v_0, \ldots, v_{\mu-1}$ form a solution of the subsystem (2.62). Then, it holds that $v_i = \Pi_{i-1} Q_i v_i$, for $i = 1, \ldots, \mu - 1$, and $v_0 = Q_0 v_0$. The functions u and $Dv_i = D\Pi_{i-1} Q_i v_i$, $i = 1, \ldots,$

$\mu-1$, are continuously differentiable. The composed function $x := D^- u + v_0 + v_1 + \cdots + v_{\mu-1}$ is continuous and has a continuous part Dx. It remains to insert x into the DAE, and to recognize that x fulfills the DAE. □

The coefficients of the IERODE and the system (2.62) are determined in terms of the DAE coefficients and the admissible matrix function sequence resulting from these coefficients. We can make use of these equations unless we suppose that there is a solution of the DAE. Considering the IERODE (2.51) and the system (2.62) as equations with unknown functions $u \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)$, $v_0 \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$, $v_i \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, $i = 1,\ldots,\mu-1$, we may solve these equations and construct continuous functions $x := D^- u + v_0 + v_1 + \cdots + v_{\mu-1}$ with $Dx = DD^- u + Dv_1 + \cdots + Dv_{\mu-1}$ being continuously differentiable, such that x satisfies the DAE. In this way we restrict our interest to those solutions u of the IERODE that have the property $u = D\Pi_{\mu-1}D^- u$. In this way one can prove the existence of DAE solutions, supposing the excitation and the coefficients to be sufficiently smooth.

The following additional description of the coupling coefficients $\mathcal{H}_0,\ldots,\mathcal{H}_{\mu-1}$ in the subsystem (2.62), which tie the solution u of the IERODE into this subsystem, supports the idea of an advanced decoupling. We draw the reader's attention to the consistency with Theorem 1.22 which provides the easier time-invariant counterpart of a complete decoupling. This lemma plays its role when constructing fine decouplings. Further, we make use of the given special representation of the coefficient $\mathcal{H}_0$ when describing the canonical projector function associated to the space of consistent values for the homogeneous DAE in the next subsection.

Lemma 2.31. *Let the DAE (2.44) be regular with tractability index μ. Let $Q_0,\ldots,Q_{\mu-1}$ be admissible projector functions, and*

$$
\begin{aligned}
Q_{0*} &:= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}\{B_0 + G_0D^-(D\Pi_{\mu-1}D^-)'D\},\\
Q_{k*} &:= Q_kP_{k+1}\cdots P_{\mu-1}G_\mu^{-1}\{B_k + G_kD^-(D\Pi_{\mu-1}D^-)'D\Pi_{k-1}\},\ k = 1,\ldots,\mu-2,\\
Q_{\mu-1*} &:= Q_{\mu-1}G_\mu^{-1}B_{\mu-1}.
\end{aligned}
$$

(1) *Then the coupling coefficients of the subsystem (2.62) have the representations*

$$
\begin{aligned}
\mathcal{H}_0 &= Q_{0*}\Pi_{\mu-1},\\
\mathcal{H}_k &= \Pi_{k-1}Q_{k*}\Pi_{\mu-1},\ k = 1,\ldots,\mu-2,\\
\mathcal{H}_{\mu-1} &= \Pi_{\mu-2}Q_{\mu-1*}\Pi_{\mu-1}.
\end{aligned}
$$

(2) *The $Q_{0*},\ldots,Q_{\mu-1*}$ are also continuous projector functions onto the subspaces $N_0,\ldots,N_{\mu-1}$, and it holds that $Q_{k*} = Q_{k*}\Pi_{k-1}$ for $k = 1,\ldots,\mu-1$.*

Proof. (1) For $k = 0,\ldots,\mu-2$, we express

$$
\begin{aligned}
\mathcal{A}_k &:= Q_kP_{k+1}\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1} \quad \text{(cf. (2.54) for } \mathcal{K} \text{ and Prop. 2.23 for } V_l)\\
&= Q_kP_{k+1}\cdots P_{\mu-1}G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1} + Q_kP_{k+1}\cdots P_{\mu-1}\sum_{l=0}^{\mu-1}V_lD\Pi_{\mu-1}.
\end{aligned}
$$

Regarding the identity $\Pi_l D^-(D\Pi_l D^-)'D\Pi_l = 0$ we derive first

$$\begin{aligned}
&\Pi_{k-1}\sum_{l=0}^{\mu-1} V_l D\Pi_{\mu-1} = \Pi_{k-1}\sum_{l=k}^{\mu-1} V_l D\Pi_{\mu-1}\\
&= \Pi_{k-1}\sum_{l=k}^{\mu-1}\{\underbrace{(I-\Pi_l)P_l}_{P_l-\Pi_l} D^-(D\Pi_l D^-)'D\Pi_{\mu-1}\\
&\qquad\qquad - (I-\Pi_l)Q_{l+1}D^-(D\Pi_{l+1}D^-)'D\Pi_{\mu-1}\}\\
&= \Pi_{k-1}\sum_{l=k}^{\mu-1}\{P_l D^-(D\Pi_l D^-)' - (I-\Pi_l)Q_{l+1}D^-(D\Pi_{l+1}D^-)'D\Pi_{\mu-1}D^-\}D\Pi_{\mu-1}\\
&= \Pi_{k-1}\sum_{l=k}^{\mu-1}\{P_l D^-(D\Pi_l D^-)' - (I-\Pi_l)Q_{l+1}D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}D^-\}D\Pi_{\mu-1}.
\end{aligned}$$

Then, taking into account that $Q_\mu = 0$, as well as the properties

$$Q_k P_{k+1}\cdots P_{\mu-1} = Q_k P_{k+1}\cdots P_{\mu-1}\Pi_{k-1},\ \ Q_k P_{k+1}\cdots P_{\mu-1}P_k = Q_k P_{k+1}\cdots P_{\mu-1}\Pi_k,$$
$$Q_k P_{k+1}\cdots P_{\mu-1}Q_l = 0,\ \text{ if }\ l \geq k+1,$$

we compute

$$\begin{aligned}
&Q_k P_{k+1}\cdots P_{\mu-1}\sum_{l=0}^{\mu-1} V_l D\Pi_{\mu-1}\\
&= Q_k P_{k+1}\cdots P_{\mu-1}\sum_{l=k+1}^{\mu-1} D^-(D\Pi_l D^-)'D\Pi_{\mu-1}\\
&\qquad + Q_k P_{k+1}\cdots P_{\mu-1}\underbrace{\sum_{l=k}^{\mu-1}\Pi_l Q_{l+1}}_{\Pi_k-\Pi_{\mu-1}} D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}\\
&= Q_k P_{k+1}\cdots P_{\mu-1}\sum_{l=k+1}^{\mu-1} D^-(D\Pi_l D^-)'D\Pi_{\mu-1}\\
&\qquad + Q_k P_{k+1}\cdots P_{\mu-1}P_k(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}.
\end{aligned}$$

This leads to

$$\begin{aligned}
\mathcal{A}_k = Q_k P_{k+1}\cdots P_{\mu-1}G_\mu^{-1}&\Big\{B_k\Pi_{\mu-1} - \sum_{j=k+1}^{\mu-1} G_j D^-(D\Pi_j D^-)'D\Pi_{\mu-1}\Big\}\\
&+ Q_k P_{k+1}\cdots P_{\mu-1}\sum_{l=k+1}^{\mu-1} D^-(D\Pi_l D^-)'D\Pi_{\mu-1}\\
&+ Q_k P_{k+1}\cdots P_{\mu-1}P_k(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}.
\end{aligned}$$

Due to $Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} G_j = Q_k P_{k+1} \cdots P_{\mu-1}$, for $j \geq k+1$, it follows that

$$\begin{aligned}
\mathcal{A}_k &= Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} B_k \Pi_{\mu-1} - Q_k P_{k+1} \cdots P_{\mu-1} \sum_{j=k+1}^{\mu-1} D^-(D\Pi_j D^-)' D\Pi_{\mu-1} \\
&\quad + Q_k P_{k+1} \cdots P_{\mu-1} \sum_{l=k+1}^{\mu-1} D^-(D\Pi_l D^-)' D\Pi_{\mu-1} \\
&\quad + Q_k P_{k+1} \cdots P_{\mu-1} P_k (D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1} \\
&= Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} B_k \Pi_{\mu-1} + Q_k P_{k+1} \cdots P_{\mu-1} P_k D^-(D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1} \\
&= Q_{k*} \Pi_{\mu-1},
\end{aligned}$$

which proves the relations $\mathcal{H}_0 = Q_0 P_1 \cdots P_{\mu-1} \mathcal{K} \Pi_{\mu-1} = Q_{0*} \Pi_{\mu-1}$, and $\mathcal{H}_k = \Pi_{k-1} Q_k P_{k+1} \cdots P_{\mu-1} \mathcal{K} \Pi_{\mu-1} = \Pi_{k-1} \mathcal{A}_k \Pi_{\mu-1} = Q_{k*} \Pi_{\mu-1}$, $k = 1, \ldots, \mu-2$. Moreover, it holds that $\mathcal{H}_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1} \mathcal{K} = Q_{\mu-1} G_\mu^{-1} B_{\mu-1} \Pi_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1*} \Pi_{\mu-1}$.

(2) Derive

$$\begin{aligned}
Q_{k*} Q_k &= Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} \{B_k + G_k D^-(D\Pi_{\mu-1} D^-)' D\Pi_{k-1}\} Q_k \\
&= Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} B_k Q_k + Q_k P_{k+1} \cdots P_{\mu-1} P_k D^-(D\Pi_{\mu-1} D^-)' D\Pi_{k-1} Q_k \\
&= \underbrace{Q_k P_{k+1} \cdots P_{\mu-1} Q_k}_{=Q_k} - \underbrace{Q_k P_{k+1} \cdots P_{\mu-1} P_k D^-(D\Pi_{\mu-1} D^-)}_{=0} (D\Pi_{k-1} Q_k D^-)' D.
\end{aligned}$$

Then, $Q_{k*} Q_{k*} = Q_{k*}$ follows. The remaining part is evident. □

2.4.3 Fine and complete decouplings

Now we advance the decoupling of the subsystem (2.62) of the regular DAE (2.44). As benefits of such a refined decoupling we get further natural information on the DAE being independent of the choice of projectors in the given context. In particular, we fix a unique natural IERODE.

2.4.3.1 Index-1 case

Take a closer look at the special case of regular index-1 DAEs. Let the DAE (2.44) be regular with tractability index 1. The matrix function $G_0 = AD$ is singular with constant rank. We take an arbitrary continuous projector function Q_0. The resulting matrix function $G_1 = G_0 + BQ_0$ is nonsingular. It follows that $Q_1 = 0$, $\Pi_1 = \Pi_0$ and $V_0 = 0$ (cf. Proposition 2.23), further $B_1 = BP_0 - G_1 D^-(D\Pi_0 D^-)' D\Pi_0 = BP_0$. The DAE scaled by G_1^{-1} is (cf. (2.48)) now

$$D^-(D\Pi_0 x)' + G_1^{-1}BP_0x + Q_0x = G_1^{-1}q.$$

Multiplication by $D\Pi_0 = D$ and $I - \Pi_0 = Q_0$ leads to the system

$$(Dx)' - R'Dx + DG_1^{-1}BD^-Dx = DG_1^{-1}q, \tag{2.64}$$

$$Q_0x + Q_0G_1^{-1}BD^-Dx = Q_0G_1^{-1}q, \tag{2.65}$$

and the solution expression $x = D^-Dx + Q_0x$. Equation (2.65) stands for the subsystem (2.62), i.e., for

$$Q_0x + \mathcal{H}_0D^-Dx = \mathcal{L}_0q,$$

with $\mathcal{H}_0 = Q_0\mathcal{K}\Pi_0 = Q_0G_1^{-1}B\Pi_0 = Q_0G_1^{-1}BP_0$, $\mathcal{L}_0 = Q_0G_1^{-1}$.
The nonsingularity of G_1 implies the decomposition $S_0 \oplus N_0 = \mathbb{R}^m$ (cf. Lemma A.9), and the matrix function $Q_0G_1^{-1}B$ is a representation of the projector function onto N_0 along S_0.
We can choose Q_0 to be the special projector function onto N_0 along S_0 from the beginning. The benefit of this choice consists in the property $\mathcal{H}_0 = Q_0G_1^{-1}BP_0 = 0$, that is, the subsystems (2.65) uncouples from (2.64).

Example 2.32 (Decoupling of a semi-explicit index-1 DAE). We reconsider the semi-explicit DAE from Example 2.3

$$\begin{bmatrix} I \\ 0 \end{bmatrix} (\begin{bmatrix} I & 0 \end{bmatrix} x)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} x = q$$

with nonsingular B_{22}. Here we have the subspaces

$$N_0 = \{z \in \mathbb{R}^{m_1+m_2} : z_1 = 0\} \quad \text{and} \quad S_0 = \{z \in \mathbb{R}^{m_1+m_2} : B_{21}z_1 + B_{22}z_2 = 0\},$$

and the projector function onto N_0 along S_0 is given by

$$Q_0 = \begin{bmatrix} 0 & 0 \\ B_{22}^{-1}B_{21} & I \end{bmatrix}.$$

This projector is reasonable owing to the property $\mathcal{H}_0 = 0$, although it is far from being orthogonal. It yields

$$D^- = \begin{bmatrix} I \\ -B_{22}^{-1}B_{21} \end{bmatrix}, G_1 = \begin{bmatrix} I + B_{12}B_{22}^{-1}B_{21} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}, G_1^{-1} = \begin{bmatrix} I & -B_{12}B_{22}^{-1} \\ -B_{22}^{-1}B_{21} & (I + B_{22}^{-1}B_{21}B_{12}) \end{bmatrix},$$

and the IERODE

$$x_1' + (B_{11} - B_{12}B_{22}^{-1}B_{21})x_1 = q_1 - B_{12}B_{22}^{-1}q_2.$$

Notice that in Example 2.3, Q_0 is chosen to be the orthoprojector, but precisely the same IERODE results for this choice. □

The last observation reflects a general property of regular index-1 DAEs as the following proposition states.

Proposition 2.33. *Let the DAE (2.44) be regular with index 1. Then its IERODE*

$$u' - R'u + DG_1^{-1}BD^-u = DG_1^{-1}q$$

is actually independent of the special choice of the continuous projector function Q_0.

Proof. We compare the IERODEs built for two different projector functions Q_0 and $\bar{Q}_0$. It holds that $\bar{G}_1 = G_0 + B\bar{Q}_0 = G_0 + BQ_0\bar{Q}_0 = G_1(P_0 + \bar{Q}_0) = G_1(I + Q_0\bar{Q}_0P_0)$ and $\bar{D}^- = \bar{D}^-D\bar{D}^- = \bar{D}^-R = \bar{D}^-DD^- = \bar{P}_0D^-$, therefore $D\bar{G}_1^{-1} = DG_1^{-1}$, $D\bar{G}_1^{-1}B\bar{D}^- = DG_1^{-1}B(I - \bar{Q}_0)D^- = DG_1^{-1}B(I - Q_0\bar{Q}_0)D^- = DG_1^{-1}BD^-$. □

Regular index-1 DAEs are transparent and simple, and the coefficients of their IERODEs are *always* independent of the projector choice. However, higher index DAEs are different.

2.4.3.2 Index-2 case

We take a closer look at the simplest class among regular higher index DAEs, the DAEs with tractability index $\mu = 2$.

Let the DAE (2.44) be regular with tractability index $\mu = 2$. Then the IERODE (2.51) and the subsystem (2.62) reduce to

$$u' - (D\Pi_1D^-)'u + D\Pi_1G_2^{-1}B_1D^-u = D\Pi_1G_2^{-1}q,$$

and

$$\begin{bmatrix} 0 & -Q_0Q_1D^- \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & D\Pi_0Q_1 \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} \right)' + \begin{bmatrix} v_0 \\ v_1 \end{bmatrix} + \begin{bmatrix} \mathcal{H}_0 \\ \mathcal{H}_1 \end{bmatrix} D^-u = \begin{bmatrix} Q_0P_1G_2^{-1} \\ \Pi_0Q_1G_2^{-1} \end{bmatrix} q,$$

with

$$\begin{aligned} \mathcal{H}_0 &= Q_0P_1\mathcal{K}\Pi_1 = Q_0P_1G_2^{-1}B_1\Pi_1 + Q_0(P_1 - Q_1)D^-(D\Pi_1D^-)'D\Pi_1 \\ &= Q_0P_1G_2^{-1}B_0\Pi_1 + Q_0P_1D^-(D\Pi_1D^-)'D\Pi_1 \\ \mathcal{H}_1 &= \Pi_0Q_1\mathcal{K}\Pi_1 = \Pi_0Q_1G_2^{-1}B_1\Pi_1. \end{aligned}$$

Owing to the nonsingularity of G_2, the decomposition (cf. Lemma A.9)

$$N_1 \oplus S_1 = \mathbb{R}^m$$

is given, and the expression $Q_1G_2^{-1}B_1$ appearing in $\mathcal{H}_1$ reminds us of the representation of the special projector function onto N_1 along S_1 (cf. Lemma A.10) which is uniquely determined. In fact, $Q_1G_2^{-1}B_1$ is this projector function. The subspaces N_1 and S_1 are given before one has to choose the projector function Q_1, and hence one can settle on the projector function Q_1 onto N_1 along S_1 at the beginning.

Thereby, the necessary admissibility condition $N_0 \subseteq \ker Q_1$ is fulfilled because of $N_0 \subseteq S_1 = \ker Q_1$. It follows that

$$Q_1 G_2^{-1} B_1 \Pi_1 = Q_1 G_2^{-1} B_1 P_1 = Q_1 P_1 = 0, \quad \mathcal{H}_1 = \Pi_0 Q_1 G_2^{-1} B_1 \Pi_1 = 0.$$

Example 2.34 (Advanced decoupling of Hessenberg size-2 DAEs). Consider once again the so-called Hessenberg size-2 DAE

$$\begin{bmatrix} I \\ 0 \end{bmatrix} (\begin{bmatrix} I & 0 \end{bmatrix} x)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix} x = q, \tag{2.66}$$

with the nonsingular product $B_{21}B_{12}$. Suppose the subspaces $\operatorname{im} B_{12}$ and $\ker B_{21}$ to be $\mathcal{C}^1$-subspaces. In Example 2.3, admissible matrix functions are built. This DAE is regular with index 2, and the projector functions

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad Q_1 = \begin{bmatrix} \Omega & 0 \\ -B_{12}^- & 0 \end{bmatrix}, \quad \Omega := B_{12} B_{12}^-, \tag{2.67}$$

are admissible, for each arbitrary reflexive inverse B_{12}^- such that Ω is continuously differentiable. We have further $D\Pi_1 D^- = I - \Omega$ and

$$S_0 = S_1 = \{z \in \mathbb{R}^{m_1+m_2} : B_{21} z_1 = 0\}.$$

In contrast to Example 2.3, where widely orthogonal projectors are chosen and

$$\ker Q_1 = \{z \in \mathbb{R}^{m_1+m_2} : B_{12}^* z_1 = 0\} = (N_0 \oplus N_1)^\perp \oplus N_0,$$

now we set $B_{12}^- := (B_{21}B_{12})^{-1} B_{21}$ such that Ω projects $\mathbb{R}^{m_1}$ onto $\operatorname{im} B_{12}$ along $\ker B_{21}$, and Q_1 projects $\mathbb{R}^m$ onto N_1 along

$$\ker Q_1 = \{z \in \mathbb{R}^{m_1+m_2} : B_{21} z_1 = 0\} = S_1.$$

Except for the very special case, if $\ker B_{12}^* = \ker B_{21}$, a nonsymmetric projector function $D\Pi_1 D^- = I - \Omega = I - B_{12}(B_{21}B_{12})^{-1} B_{21}$ results. However, as we already know, this choice has the advantage of a vanishing coupling coefficient $\mathcal{H}_1$.
In contrast to the admissible projector functions (2.67), the projector functions

$$Q_0 = \begin{bmatrix} 0 & 0 \\ B_{12}^-(B_{11} - \Omega')(I - \Omega) & I \end{bmatrix}, \quad Q_1 = \begin{bmatrix} \Omega & 0 \\ -B_{12}^- & 0 \end{bmatrix}, \quad \Omega := B_{12} B_{12}^-, \tag{2.68}$$

form a further pair of admissible projector functions again yielding $D\Pi_1 D^- = I - \Omega$. With $B_{12}^- := (B_{21}B_{12})^{-1} B_{21}$, this choice forces both coefficients $\mathcal{H}_1$ and $\mathcal{H}_0$ to disappear, and the subsystem (2.62) uncouples from the IERODE. One can check that the resulting IERODE coincides with that from (2.67). □

As mentioned before, the index-2 case has the simplest higher index structure. The higher the index, the greater the variety of admissible projector functions. We

recall Example 1.26 which shows several completely decoupling projectors for a time-invariant regular matrix pair with Kronecker index 2.

2.4.3.3 General benefits from fine decouplings

Definition 2.35. Let the DAE (2.44) be regular with tractability index μ, and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions.

(1) If the coupling coefficients $\mathcal{H}_1, \ldots, \mathcal{H}_{\mu-1}$ of the subsystem (2.62) vanish, then we speak of *fine decoupling projector functions* $Q_0, \ldots, Q_{\mu-1}$, and of a *fine decoupling*.

(2) If all the coupling coefficients $\mathcal{H}_0, \ldots, \mathcal{H}_{\mu-1}$ of the subsystem (2.62) vanish, then we speak of *complete decoupling projector functions* $Q_0, \ldots, Q_{\mu-1}$, and of a *complete decoupling*.

Special fine and complete decoupling projector functions Q_0, Q_1 are built in Examples 2.34 and (2.32).

Owing to the linearity of the DAE (2.44) its homogeneous version

$$A(t)(D(t)x(t))' + B(t)x(t) = 0, \quad t \in \mathcal{J}, \tag{2.69}$$

plays its role, and in particular the subspace

$$S_{can}(t) := \{z \in \mathbb{R}^m : \exists x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m),\ A(Dx)' + Bx = 0,\ x(t) = z\},\ t \in \mathcal{I}.$$

The subspace $S_{can}(t)$ represents the geometric locus of all solution values of the homogeneous DAE (2.69) at time t. In other words, $S_{can}(t)$ is the linear *space of consistent initial values* at time t for the homogeneous DAE.

For implicit regular ODEs (2.43), $S_{can}(t) = \mathbb{R}^m$ is simply the entire time-invariant state space $\mathbb{R}^m$. In contrast, for intrinsic DAEs, the proper inclusion

$$S_{can}(t) \subseteq S_0(t)$$

is valid. While $S_0(t)$ represents the so-called *obvious constraint* associated with the DAE (2.69), the subspace $S_{can}(t)$ serves, so to say, as the complete final constraint which also incorporates all hidden constraints.

In particular, for the semi-explicit DAE in Example 2.3, we find the obvious constraint

$$S_0(t) = \{z \in \mathbb{R}^{m_1+m_2} : z_2 = -B_{22}(t)^{-1}B_{21}(t)z_1\},\ \dim S_0(t) = m_1,$$

and further

$$S_{can}(t) = \{z \in \mathbb{R}^{m_1+m_2} : z_2 = -B_{22}(t)^{-1}B_{21}(t)z_1\} = S_0(t),$$

supposing $B_{22}(t)$ remains nonsingular. However, if $B_{22}(t) \equiv 0$, but $B_{21}(t)B_{12}(t)$ remains nonsingular, then

$$S_{can}(t) = \{z \in \mathbb{R}^{m_1+m_2} : B_{21}(t)z_1 = 0, \\ z_2 = -[(B_{21}B_{21})^{-1}B_{21}(B_{11} - (B_{12}((B_{21}B_{21})^{-1}B_{21})')](t)z_1\}$$

is merely a proper subspace of the obvious constraint

$$S_0(t) = \{z \in \mathbb{R}^{m_1+m_2} : B_{21}(t)z_1 = 0\}.$$

Example 2.4 confronts us even with a zero-dimensional subspace $S_{can}(t) = \{0\}$.

Except for those simpler cases, the canonical subspace S_{can} is not easy to access. It coincides with the finite eigenspace of the matrix pencil for regular linear time-invariant DAEs. Theorem 2.39 below provides a description by means of fine decoupling projector functions.

Definition 2.36. For the regular DAE (2.44) the time-varying subspaces $S_{can}(t)$, $t \in \mathcal{I}$, and $N_{can}(t) := N_0(t) + \cdots + N_{\mu-1}(t)$, $t \in \mathcal{I}$, are said to be the *canonical subspaces* of the DAE.

By Theorem 2.8, N_{can} is known to be independent of the special choice of admissible projectors, which justifies the notion. The canonical subspaces of the linear DAE generalize the finite and infinite eigenspaces of matrix pencils.

Applying fine decoupling projector functions $Q_0, \ldots, Q_{\mu-1}$, the subsystem (2.62) corresponding to the homogeneous DAE simplifies to

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \left(\mathcal{D} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} \right)' + \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} v_0 \\ v_1 \\ \vdots \\ v_{\mu-1} \end{bmatrix} \\ + \begin{bmatrix} \mathcal{H}_0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} D^- u = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \tag{2.70}$$

For given u, its solution components are determined successively as

$$v_{\mu-1} = 0, \ \ldots, \ v_1 = 0, \ v_0 = -\mathcal{H}_0 D^- u,$$

and hence each solution $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ of the homogeneous DAE possesses the representation

$$x = D^- u + v_0 = (I - \mathcal{H}_0)D^- u = (I - Q_{0*}\Pi_{\mu-1})D^- D\Pi_{\mu-1}D^- u = (I - Q_{0*})\Pi_{\mu-1}D^- u,$$

whereby $u = D\Pi_{\mu-1}D^- u$ is a solution of the homogeneous IERODE

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}BD^-u = 0,$$

and Q_{0*} is defined in Lemma 2.31. Owing to the relations $P_0Q_{0*} = 0$, the continuous matrix function $(I - Q_{0*})\Pi_{\mu-1}$ is also a projector function, and the nullspace is easily checked to be

$$\ker (I - Q_{0*})\Pi_{\mu-1} = N_{can}.$$

Since each solution of the homogeneous DAE can be represented in this way, the inclusion

$$S_{can} \subseteq \operatorname{im} (I - Q_{0*})\Pi_{\mu-1}$$

is valid. On the other hand, through each element of $\operatorname{im}((I - Q_{0*}(t))\Pi_{\mu-1}(t))$, at time t, there passes a DAE solution, and we obtain

$$\operatorname{im} (I - Q_{0*})\Pi_{\mu-1} = S_{can}.$$

In fact, fixing an arbitrary pair $x_0 \in \operatorname{im}((I - Q_{0*}(t_0))\Pi_{\mu-1}(t_0))$, $t_0 \in \mathcal{I}$, we determine the unique solution u of the standard IVP

$$u' - (D\Pi_{\mu-1}D^-)'u + D\Pi_{\mu-1}G_\mu^{-1}BD^-u = 0, \quad u(t_0) = D(t_0)\Pi_{\mu-1}(t_0)x_0,$$

and then the DAE solution $x := (I - Q_{0*})\Pi_{\mu-1}D^-u$. It follows that $x(t_0) = (I - Q_{0*}(t_0))\Pi_{\mu-1}(t_0)x_0 = x_0$. In consequence, the DAE solution passes through $x_0 \in \operatorname{im}((I - Q_{0*}(t_0))\Pi_{\mu-1}(t_0))$.
Owing to the projector properties, the decomposition

$$N_{can}(t) \oplus S_{can}(t) = \mathbb{R}^m, \quad t \in \mathcal{I}, \tag{2.71}$$

becomes valid. Moreover, now we see that S_{can} is a $\mathcal{C}$-subspace of dimension $d = m - \sum_{i=0}^{\mu-1}(m - r_i)$.

Definition 2.37. For a regular DAE (2.44) with tractability index μ, which has a fine decoupling, the projector function $\Pi_{can} \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$ being uniquely determined by

$$\operatorname{im} \Pi_{can} = S_{can}, \quad \ker \Pi_{can} = N_{can}$$

is named *the canonical projector function* of the DAE.

We emphasize that both canonical subspaces S_{can} and N_{can}, and the canonical projector function Π_{can}, depend on the index μ. Sometimes it is reasonable to indicate this by writing $S_{can\,\mu}$, $N_{can\,\mu}$ and $\Pi_{can\,\mu}$.
The canonical projector plays the same role as the spectral projector does in the time-invariant case.

Remark 2.38. In earlier papers also the subspaces S_i (e.g., [159]) and the single projector functions $Q_0, \ldots, Q_{\mu-1}$ forming a fine decoupling (e.g., [157], [164]) are named canonical. This applies, in particular, to the projector function $Q_{\mu-1}$ onto $N_{\mu-1}$ along $S_{\mu-1}$. We do not use this notation. We know the canonical projector function Π_{can} in Definition 2.37 to be unique, however, for higher index cases, the

single factors P_i in the given representation by means of fine decoupling projectors are not uniquely determined as is demonstrated by Example 1.26.

Now we are in a position to gather the fruit of the construction.

Theorem 2.39. *Let the regular index-μ DAE (2.44) have a fine decoupling.*

(1) *Then the canonical subspaces S_{can} and N_{can} are $\mathcal{C}$-subspaces of dimensions $d = m - \sum_{i=0}^{\mu-1}(m - r_i)$ and $m - d$.*
(2) *The decomposition (2.71) is valid, and the canonical projector function has the representation*

$$\Pi_{can} = (I - Q_{0*})\Pi_{\mu-1},$$

with fine decoupling projector functions $Q_0, \ldots, Q_{\mu-1}$.
(3) *The coefficients of the IERODE (2.51) are independent of the special choice of the fine decoupling projector functions.*

Proof. It remains to verify (3). Let two sequences of fine decoupling projector functions $Q_0, \ldots, Q_{\mu-1}$ and $\bar{Q}_0, \ldots, \bar{Q}_{\mu-1}$ be given. Then the canonical projector function has the representations $\Pi_{can} = (I - Q_{0*})\Pi_{\mu-1}$ and $\Pi_{can} = (I - \bar{Q}_{0*})\bar{\Pi}_{\mu-1}$. Taking into account that $\bar{D}^- = \bar{P}_0 D^-$ we derive

$$D\Pi_{\mu-1}D^- = D\Pi_{can}D^- = D\bar{\Pi}_{\mu-1}D^- = D\bar{\Pi}_{\mu-1}\bar{D}^-.$$

Then, with the help of Lemma 2.12 yielding the relation $\bar{G}_\mu = G_\mu Z_\mu$, we arrive at

$$D\bar{\Pi}_{\mu-1}\bar{G}_\mu^{-1} = D\Pi_{\mu-1}D^- DZ_\mu^{-1}G_\mu^{-1} = D\Pi_{\mu-1}G_\mu^{-1},$$

$$D\bar{\Pi}_{\mu-1}\bar{G}_\mu^{-1}B\bar{D}^- = D\Pi_{\mu-1}G_\mu^{-1}B\bar{D}^- = D\Pi_{\mu-1}G_\mu^{-1}B(I - \bar{Q}_0)D^- = D\Pi_{\mu-1}G_\mu^{-1}BD^-,$$

and this proves the assertion. □

For regular index-1 DAEs, each continuous projector function Q_0 already generates a fine decoupling. Therefore, Proposition 2.33 is now a special case of Theorem 2.39 (3).

DAEs with fine decouplings, later on named *fine* DAEs, allow an intrinsic DAE theory in Section 2.6 addressing solvability, qualitative flow behavior and the characterization of admissible excitations.

2.4.3.4 Existence of fine and complete decouplings

For regular index-2 DAEs, the admissible pair Q_0, Q_1 provides a fine decoupling, if Q_1 is chosen such that $\ker Q_1 = S_1$. This is accompanied by the requirement that $\operatorname{im} D\Pi_1 D^- = DS_1$ is a $\mathcal{C}^1$-subspace. We point out that, for fine decouplings, we need some additional smoothness with respect to the regularity notion. While regularity with index 2 comprises the *existence* of an arbitrary $\mathcal{C}^1$ decomposition (i.e., the existence of a continuously differentiable projector function $D\Pi_1 D^-$)

$$\operatorname{im} D\Pi_1 D^- \oplus \underbrace{\operatorname{im} D\Pi_0 Q_1 D^-}_{=DN_1} \oplus \ker A = \mathbb{R}^n,$$

one needs for fine decouplings that the *special* decomposition

$$DS_1 \oplus DN_1 \oplus \ker A = \mathbb{R}^n,$$

consists of $\mathcal{C}^1$-subspaces. For instance, the semi-explicit DAE in Example 2.34 possesses fine decoupling projector functions, if both subspaces $\operatorname{im} B_{12}$ and $\ker B_{21}$ are continuously differentiable. However, for regularity, it is enough if $\operatorname{im} B_{12}$ is a $\mathcal{C}^1$-subspace, as demonstrated in Example 2.3.

Assuming the coefficients A, D, B to be $\mathcal{C}^1$, and choosing a continuously differentiable projector function Q_0, the resulting DN_1 and DS_1 are always $\mathcal{C}^1$-subspaces. However, we do not feel comfortable with such a generous sufficient smoothness assumption, though it is less demanding than that in derivative array approaches, where one naturally has to require $A, D, B \in \mathcal{C}^2$ for the treatment of an index-2 problem.

We emphasize that only certain continuous subspaces are additionally assumed to belong to the class $\mathcal{C}^1$. Since the precise description of these subspaces is somewhat cumbersome, we use instead the wording *the coefficients of the DAE are sufficiently smooth* just to indicate the smoothness problem.

In essence, the additional smoothness requirements are related to the coupling coefficients $\mathcal{H}_1, \ldots, \mathcal{H}_{\mu-1}$ in the subsystem (2.62), and in particular to the special projectors introduced in Lemma 2.31. It turns out that, for a fine decoupling of a regular index-μ DAE, certain parts of the coefficients A, D, B have to be continuously differentiable up to degree $\mu - 1$. This meets the common understanding of index μ DAEs, and it is closely related to solvability conditions. We present an example for more clarity.

Example 2.40 (Smoothness for a fine decoupling). Consider the DAE

$$\underbrace{\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&0 \end{bmatrix}}_{A} \left(\underbrace{\begin{bmatrix} 1&0&0&0\\ 0&1&0&0\\ 0&0&1&0\\ 0&0&0&0 \end{bmatrix}}_{D} x \right)' + \underbrace{\begin{bmatrix} 0&0&0&0\\ 0&0&0&-1\\ 0&-1&0&0\\ \alpha&0&-1&0 \end{bmatrix}}_{B} x = 0,$$

on the interval $\mathcal{I} = [0,1]$. According to the basic continuity assumption, B is continuous, that is, $\alpha \in \mathcal{C}([0,1])$. Taking a look at the solution satisfying the initial condition $x_1(0) = 1$, that is

$$x_1(t) = 1,\ x_3(t) = \alpha(t),\ x_2(t) = x_3'(t) = \alpha'(t),\ x_4(t) = x_3''(t) = \alpha''(t)$$

we recognize that we must more reasonably assume $\alpha \in \mathcal{C}^2([0,1])$. We demonstrate by constructing a fine decoupling sequence that this is precisely the smoothness needed.

The first elements of the matrix function sequence can be chosen, respectively, computed as

$$Q_0 = \begin{bmatrix} 0&0&0&0\\ 0&0&0&0\\ 0&0&0&0\\ 0&0&0&1 \end{bmatrix}, \; G_1 = \begin{bmatrix} 1&0&0&0\\ 0&1&0&-1\\ 0&0&1&0\\ 0&0&0&0 \end{bmatrix}, \; Q_1 = \begin{bmatrix} 0&0&0&0\\ 0&1&0&0\\ 0&0&0&0\\ 0&1&0&0 \end{bmatrix}, \; G_2 = \begin{bmatrix} 1&0&0&0\\ 0&1&0&-1\\ 0&-1&1&0\\ 0&0&0&0 \end{bmatrix}.$$

We could continue with

$$Q_2 = \begin{bmatrix} 0&0&0&0\\ 0&0&1&0\\ 0&0&1&0\\ 0&0&1&0 \end{bmatrix}, \; G_3 = \begin{bmatrix} 1&0&0&0\\ 0&1&0&-1\\ 0&-1&1&0\\ 0&0&-1&0 \end{bmatrix},$$

which shows the DAE to be regular with tractability index 3, and Q_0, Q_1, Q_2 to be admissible, if $\alpha \in \mathcal{C}([0,1])$. However, we dismiss this choice of Q_2 and compute it instead corresponding to the decomposition

$$N_2 \oplus S_2 = \{z \in \mathbb{R}^4 : z_1 = 0, z_2 = z_3 = z_4\} \oplus \{z \in \mathbb{R}^4 : \alpha z_1 = z_3\} = \mathbb{R}^4.$$

This leads to

$$Q_2 = \begin{bmatrix} 0&0&0&0\\ \alpha&0&1&0\\ \alpha&0&1&0\\ \alpha&0&1&0 \end{bmatrix}, \; D\Pi_2 D^- = \Pi_2 = \begin{bmatrix} 1&0&0&0\\ 0&0&0&0\\ -\alpha&0&0&0\\ 0&0&1&0 \end{bmatrix},$$

and hence, for these Q_0, Q_1, Q_2 to be admissible, the function α is required to be continuously differentiable. The coupling coefficients related to the present projector functions are

$$\mathcal{H}_1 = \begin{bmatrix} 0&0&0&0\\ \alpha'&0&0&0\\ 0&0&0&0\\ 0&0&0&0 \end{bmatrix}, \quad \mathcal{H}_2 = 0.$$

If α' does not vanish identically, we have not yet reached a fine decoupling. In the next round we set $\bar{Q}_0 = Q_0$ such that $\bar{G}_1 = G_1$, but then we put

$$\bar{Q}_1 := Q_{1*} := Q_1 P_2 G_3^{-1} \{B_1 + G_1 D^- (D\Pi_2 D^-)' D\Pi_0\} = \begin{bmatrix} 0&0&0&0\\ \alpha'&1&0&0\\ 0&0&0&0\\ \alpha'&1&0&0 \end{bmatrix},$$

in accordance with Lemma 2.31 (see also Lemma 2.41 below). It follows that

$$D\bar{\Pi}_1 D^- = \bar{\Pi}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\alpha' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \; \bar{G}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ -\alpha' & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

and we see that, to ensure that $D\bar{\Pi}_1 D^-$ becomes continuously differentiable, and $\bar{Q}_0, \bar{Q}_1$ admissible, we need a two times continuously differentiable function α. Then we have $\bar{N}_2 = N_2$, which allows for the choice $\bar{Q}_2 = Q_2$. The resulting $\bar{Q}_0, \bar{Q}_1, \bar{Q}_2$ are fine decoupling projector functions. □

In general, if the DAE (2.44) is regular with tractability index μ, and $Q_0, \ldots, Q_{\mu-1}$ are admissible projector functions, then the decomposition

$$N_{\mu-1} \oplus S_{\mu-1} = \mathbb{R}^m$$

holds true (cf. Lemma A.9). If the last projector function $Q_{\mu-1}$ is chosen such that the associated subspace $S_{\mu-1} \supseteq N_0 \oplus \cdots \oplus N_{\mu-2}$ becomes its nullspace, that is $\ker Q_{\mu-1} = S_{\mu-1}$, $\operatorname{im} Q_{\mu-1} = N_{\mu-1}$, then it follows (cf. Lemma A.10) that $Q_{\mu-1} = Q_{\mu-1} G_{\mu-1}^{-1} B_{\mu-1}$, and hence (cf. (2.54))

$$\begin{aligned} \mathcal{H}_{\mu-1} &:= \Pi_{\mu-2} Q_{\mu-1} \mathcal{K}_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1} \mathcal{K} \\ &= \Pi_{\mu-2} \underbrace{Q_{\mu-1}(I - \Pi_{\mu-1})}_{=Q_{\mu-1}} G_\mu^{-1} B_{\mu-1} \Pi_{\mu-1} \\ &\quad + \sum_{l=0}^{\mu-1} \Pi_{\mu-2} \underbrace{Q_{\mu-1}(I - \Pi_l)}_{=0} (P_l - Q_l)(D\Pi_l D^-)' D\Pi_{\mu-1} \\ &= \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1} B_{\mu-1} \Pi_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1} \Pi_{\mu-1} = 0. \end{aligned}$$

So far one can prevail on the coefficients $\mathcal{H}_{\mu-1}$ to vanish by determining $\ker Q_{\mu-1} = S_{\mu-1}$. This confirms the existence of complete decoupling projector functions for regular index-1 DAEs, and the existence of fine decoupling projector functions for regular index-2 DAEs.

Remember that, for regular constant coefficient DAEs with arbitrary index, complete decoupling projectors are provided by Theorem 1.22. We follow the lines of [169] to prove a similar result for general regular DAEs (2.44).

Having Lemma 2.31 we are well prepared to construct fine decoupling projector functions for the general regular DAE (2.44). As in Example 2.40, we successively improve the decoupling with the help of Lemma 2.31 in several rounds. We begin by forming arbitrary admissible projector functions $Q_0, \ldots, Q_{\mu-2}$ and $G_{\mu-1}$. Then we determine $Q_{\mu-1}$ by $\ker Q_{\mu-1} = S_{\mu-1}$ and $\operatorname{im} Q_{\mu-1} = N_{\mu-1}$. This yields $G_\mu = G_{\mu-1} + B_{\mu-1} Q_{\mu-1}$ as well as

$$\begin{aligned} Q_{\mu-1} &= Q_{\mu-1} G_\mu^{-1} B_{\mu-1} = Q_{\mu-1*}, \quad \text{and} \\ \mathcal{H}_{\mu-1} &= \Pi_{\mu-2} Q_{\mu-1*} \Pi_{\mu-1} = \Pi_{\mu-2} Q_{\mu-1} \Pi_{\mu-1} = 0. \end{aligned}$$

If $\mu = 2$ we already have a fine decoupling. If $\mu \geq 3$, we assume $D\Pi_{\mu-3}Q_{\mu-2*}D^-$, which is a priori continuous, to be even continuously differentiable, and compose a new sequence from the previous one. We set

$$\bar{Q}_0 := Q_0, \ldots, \bar{Q}_{\mu-3} = Q_{\mu-3}, \quad \text{and} \quad \bar{Q}_{\mu-2} = Q_{\mu-2*}.$$

$D\bar{\Pi}_{\mu-2}D^- = D\Pi_{\mu-3}D^- - D\Pi_{\mu-3}Q_{\mu-2*}D^-$ is continuously differentiable, and the projector functions $\bar{Q}_0, \ldots, \bar{Q}_{\mu-2}$ are admissible. Further, some technical calculations yield

$$\bar{G}_{\mu-1} = G_{\mu-1} \underbrace{\{I + \bar{Q}_{\mu-2}P_{\mu-2} + (I - \Pi_{\mu-3})Q_{\mu-2}D^-(D\bar{\Pi}_{\mu-2}D^-)'D\Pi_{\mu-3}\bar{Q}_{\mu-2}\}}_{Z_{\mu-1}}.$$

The matrix function $Z_{\mu-1}$ remains nonsingular; it has the pointwise inverse

$$Z_{\mu-1}^{-1} = I - \bar{Q}_{\mu-2}P_{\mu-2} - (I - \Pi_{\mu-3})Q_{\mu-2}D^-(D\bar{\Pi}_{\mu-2}D^-)'D\Pi_{\mu-3}Q_{\mu-2}.$$

We complete the current sequence by

$$\bar{Q}_{\mu-1} := Z_{\mu-1}^{-1}Q_{\mu-1}Z_{\mu-1} = Z_{\mu-1}^{-1}Q_{\mu-1}.$$

It follows that $\bar{Q}_{\mu-1}\bar{Q}_{\mu-2} = Z_{\mu-1}^{-1}Q_{\mu-1}Q_{\mu-2*} = 0$ and $\bar{Q}_{\mu-1}\bar{Q}_i = Z_{\mu-1}^{-1}Q_{\mu-1}Q_i = 0$ for $i = 0, \ldots, \mu-3$. Applying several basic properties (e.g., $\bar{\Pi}_{\mu-2} = \bar{\Pi}_{\mu-2}\Pi_{\mu-2}$) we find the representation $D\bar{\Pi}_{\mu-1}D^- = (D\bar{\Pi}_{\mu-2}D^-)(D\Pi_{\mu-1}D^-)$ which shows the continuous differentiability of $D\bar{\Pi}_{\mu-1}D^-$. Our new sequence $\bar{Q}_0, \ldots, \bar{Q}_{\mu-1}$ is admissible. We have further $\operatorname{im}\bar{G}_{\mu-1} = \operatorname{im}G_{\mu-1}$, thus

$$\bar{S}_{\mu-1} = S_{\mu-1} = \ker \mathcal{W}_{\mu-1}B = \ker \mathcal{W}_{\mu-1}BZ_{\mu-1} = Z_{\mu-1}^{-1}S_{\mu-1}.$$

This makes it clear that, $\bar{Q}_{\mu-1} = Z_{\mu-1}^{-1}Q_{\mu-1}$ projects onto $\bar{N}_{\mu-1} = Z_{\mu-1}^{-1}N_{\mu-1}$ along $\bar{S}_{\mu-1} = Z_{\mu-1}^{-1}S_{\mu-1}$, and therefore the new coupling coefficient satisfies $\bar{\mathcal{H}}_{\mu-1} = 0$. Additionally, making further technical efforts one attains $\bar{\mathcal{H}}_{\mu-2} = 0$.
If $\mu = 3$, a fine decoupling is reached. If $\mu \geq 4$, we build the next sequence analogously as

$$\bar{\bar{Q}}_0 := \bar{Q}_0, \ldots, \bar{\bar{Q}}_{\mu-4} := \bar{Q}_{\mu-4},\ \bar{\bar{Q}}_{\mu-3} := \bar{Q}_{\mu-3*},$$
$$\bar{\bar{Q}}_{\mu-2} := \bar{Z}_{\mu-2}^{-1}\bar{Q}_{\mu-2}\bar{Z}_{\mu-2},\ \bar{\bar{Q}}_{\mu-1} := \bar{Z}_{\mu-1}^{-1}\bar{Q}_{\mu-1}\bar{Z}_{\mu-1}.$$

Supposing $D\bar{\Pi}_{\mu-4}\bar{Q}_{\mu-3*}D^-$ to be continuously differentiable, we prove the new sequence to be admissible, and to generate the coupling coefficients

$$\bar{\bar{\mathcal{H}}}_{\mu-1} = 0, \quad \bar{\bar{\mathcal{H}}}_{\mu-2} = 0, \quad \bar{\bar{\mathcal{H}}}_{\mu-3} = 0.$$

And so on. Lemma 2.41 below guarantees the procedure reaches its goal.

Lemma 2.41. *Let the DAE (2.44) with sufficiently smooth coefficients be regular with tractability index $\mu \geq 3$, and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions.*
Let $k \in \{1, \ldots, \mu-2\}$ be fixed, and let $\bar{Q}_k$ be an additional continuous projector function onto $N_k = \ker G_k$ such that $D\Pi_{k-1}\bar{Q}_k D^-$ is continuously differentiable and the inclusion $N_0 + \cdots + N_{k-1} \subseteq \ker \bar{Q}_k$ is valid. Then the following becomes true:

(1) *The projector function sequence*

$$\begin{aligned}
&\bar{Q}_0 := Q_0, \ldots, \bar{Q}_{k-1} := Q_{k-1},\\
&\bar{Q}_k,\\
&\bar{Q}_{k+1} := Z_{k+1}^{-1} Q_{k+1} Z_{k+1}, \ldots, \bar{Q}_{\mu-1} := Z_{\mu-1}^{-1} Q_{\mu-1} Z_{\mu-1},
\end{aligned}$$

with the continuous nonsingular matrix functions $Z_{k+1}, \ldots, Z_{\mu-1}$ determined below, is also admissible.

(2) *If, additionally, the projector functions $Q_0, \ldots, Q_{\mu-1}$ provide an advanced decoupling in the sense that the conditions (cf. Lemma 2.31)*

$$Q_{\mu-1*}\Pi_{\mu-1} = 0, \ldots, \; Q_{k+1*}\Pi_{\mu-1} = 0$$

are given, then also the relations

$$\bar{Q}_{\mu-1*}\bar{\Pi}_{\mu-1} = 0, \ldots, \; \bar{Q}_{k+1*}\bar{\Pi}_{\mu-1} = 0, \tag{2.72}$$

are valid, and further

$$\bar{Q}_{k*}\bar{\Pi}_{\mu-1} = (Q_{k*} - \bar{Q}_k)\Pi_{\mu-1}. \tag{2.73}$$

The matrix functions Z_i are consistent with those given in Lemma 2.12; however, for easier reading we do not access this general lemma in the proof below. In the special case given here, Lemma 2.12 yields simply $Z_0 = I, Y_1 = Z_1 = I, \ldots,$ $Y_k = Z_k = I$, and further

$$Y_{k+1} = I + Q_k(\bar{Q}_k - Q_k) + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{kl} \bar{Q}_k = \left(I + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{kl} Q_k\right)\left(I + Q_k(\bar{Q}_k - Q_k)\right),$$

$$Z_{k+1} = Y_{k+1},$$

$$Y_j = I + \sum_{l=0}^{j-2} Q_l \mathfrak{A}_{j-1l} Q_{j-1}, \quad Z_j = Y_j Z_{j-1}, \quad j = k+2, \ldots, \mu.$$

Besides the general property $\ker \bar{\Pi}_j = \ker \Pi_j$, $j = 0, \ldots, \mu-1$, which follows from Lemma 2.12, now it additionally holds that

$$\operatorname{im} \bar{Q}_k = \operatorname{im} Q_k, \quad \text{but} \quad \ker \bar{Q}_j = \ker Q_j, \quad j = k+1, \ldots, \mu-1.$$

We refer to Appendix B for the extensive calculations proving this lemma.

Lemma 2.41 guarantees the existence of fine decoupling projector functions, and it confirms the procedure sketched above to be reasonable.
The following theorem is the time-varying counterpart of Theorem 1.22 on constant coefficient DAEs.

Theorem 2.42. *Let the DAE (2.44) be regular with tractability index μ.*

(1) *If the coefficients of the DAE are sufficiently smooth, then a fine decoupling exists.*
(2) *If there is a fine decoupling, then there is also a complete decoupling.*

Proof. (1) The first assertion is a consequence of Lemma 2.41 and the procedure described above.
(2) Let fine decoupling projectors $Q_0, \ldots, Q_{\mu-1}$ be given. We form the new sequence

$$\bar{Q}_0 := Q_{0*},\ \bar{Q}_1 := Z_1^{-1} Q_1 Z_1, \ldots,\ \bar{Q}_{\mu-1} := Z_{\mu-1}^{-1} Q_{\mu-1} Z_{\mu-1},$$

with the matrix functions Z_j from Lemma 2.12, in particular $Z_1 = I + \bar{Q}_0 P_0$. It holds that $\bar{D}^- = \bar{P}_0 D^-$. Owing to the special form of Z_j, the relations $\Pi_{j-1} Z_j = \Pi_{j-1}$, $\Pi_{j-1} Z_j^{-1} = \Pi_{j-1}$ are given for $j \le i-1$. This yields $\bar{Q}_i \bar{Q}_j = \bar{Q}_i Z_j^{-1} Q_j Z_j = \bar{Q}_i \underbrace{\Pi_{i-1} Z_j^{-1} Q_j}_{=0} Z_j = 0$.
Expressing $D\bar{\Pi}_1 \bar{D}^- = D\bar{P}_0 Z_1^{-1} P_1 Z_1 \bar{P}_0 D^- = D \underbrace{P_0 Z_1^{-1} P_1}_{\Pi_1} Z_1 \bar{P}_0 D^- = D\Pi_1 D^-$, and successively,

$$\begin{aligned} D\bar{\Pi}_i \bar{D}^- &= D\bar{\Pi}_{i-1} Z_i^{-1} P_i Z_i \bar{P}_D^- \\ &= D\bar{\Pi}_{i-1} \bar{D}^- D Z_i^{-1} P_i Z_i \bar{P}_D^- = D \underbrace{\Pi_{i-1} D^- D Z_i^{-1} P_i}_{\Pi_i} Z_i \bar{P}_D^- = D\Pi_i D^-, \end{aligned}$$

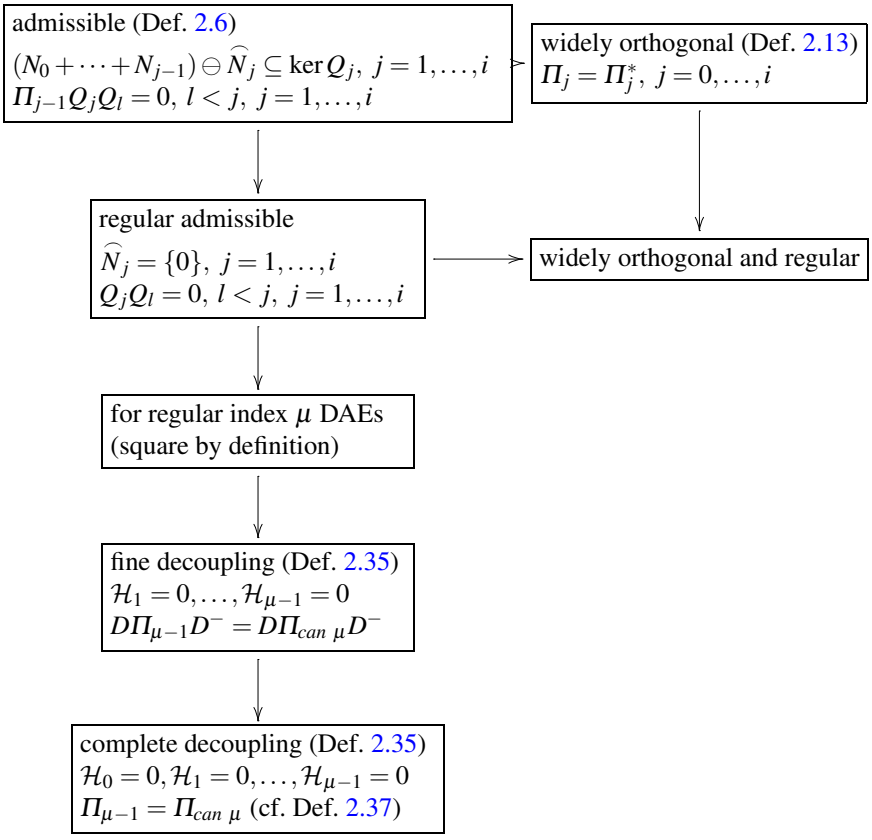
we see that the new sequence of projector functions $\bar{Q}_0, \ldots, \bar{Q}_{\mu-1}$ is admissible, too. Analogously to Lemma 2.41, one shows

$$\bar{\mathcal{H}}_{\mu-1} = 0, \ldots,\ \bar{\mathcal{H}}_1 = 0, \quad \bar{\mathcal{H}}_0 = (Q_{0*} - \bar{Q}_0)\Pi_{\mu-1},$$

and this completes the proof. □

2.5 Hierarchy of admissible projector function sequences for linear DAEs

The matrices $Q_0, \ldots, Q_i$ are admissible projectors, where Q_j projects onto $N_j = \ker G_j$, $j = 0, \ldots, i$, with $P_0 := I - Q_0$, $\Pi_0 := P_0$ and $P_j := I - Q_j$, $\Pi_j := \Pi_{j-1} P_j$, $\widehat{N}_j := (N_0 + \cdots + N_{j-1}) \cap N_j$, $j = 1, \ldots, i$.



2.6 Fine regular DAEs

Here we continue to investigate regular DAEs (2.44) which have tractability index μ and fine decoupling projector functions $Q_0,\ldots,Q_{\mu-1}$. It is worth emphasizing once more that Theorem 2.42 guarantees the existence of a fine decoupling for all regular DAEs with sufficiently smooth coefficients.

Definition 2.43. Equation (2.44) is said to be a *fine DAE* on the interval $\mathcal{I}$, if it is regular there and possesses a fine decoupling.

By Theorem 2.39 and Lemma 2.31,

$$\Pi_{can}=(I-Q_{0*})\Pi_{\mu-1}=(I-\mathcal{H}_0)\Pi_{\mu-1}$$

is the canonical projector function onto S_{can} along N_{can}, and hence $D\Pi_{can}=D\Pi_{\mu-1}$, and therefore $D\Pi_{can}D^-=D\Pi_{\mu-1}D^-$, and $\operatorname{im}D\Pi_{\mu-1}=\operatorname{im}D\Pi_{can}=DS_{can}$. Taking into account also Lemma 2.28 (7), the IERODE can now be written as

$$u' - (D\Pi_{can}D^-)'u + D\Pi_{can}G_\mu^{-1}BD^-u = D\Pi_{can}G_\mu^{-1}q, \tag{2.74}$$

and, by Lemma 2.27, the subspace DS_{can} is a time-varying invariant subspace for its solutions, which means $u(t_0) \in D(t_0)S_{can}(t_0)$ implies $u(t) \in D(t)S_{can}(t)$ for all $t \in \mathcal{I}$. This invariant subspace also applies to the homogeneous version of the IERODE. Here, the IERODE is unique, its coefficients are independent of the special choice of the fine decoupling projector functions, as pointed out in the previous subsection. With regard to the fine decoupling, Proposition 2.29 (6), and the fact that $v_i = \Pi_{i-1}Q_iv_i$ holds true for $i = 1, \ldots, \mu-1$, the subsystem (2.62) simplifies slightly to

$$v_0 = -\sum_{l=1}^{\mu-1} \mathcal{N}_{0l}(Dv_l)' - \sum_{l=2}^{\mu-1} \mathcal{M}_{0l}\, v_l - \mathcal{H}_0D^-u + \mathcal{L}_0q, \tag{2.75}$$

$$v_i = -\sum_{l=i+1}^{\mu-1} \mathcal{N}_{il}(Dv_l)' - \sum_{l=i+2}^{\mu-1} \mathcal{M}_{il}\, v_l + \mathcal{L}_iq, \quad i = 1, \ldots, \mu-3, \tag{2.76}$$

$$v_{\mu-2} = -\mathcal{N}_{\mu-2,\mu-1}(Dv_{\mu-1})' + \mathcal{L}_{\mu-2}q, \tag{2.77}$$

$$v_{\mu-1} = \mathcal{L}_{\mu-1}q. \tag{2.78}$$

By Theorem 2.30, the DAE (2.44) is equivalent to the system consisting of the IERODE and the subsystem (2.75)–(2.78).

2.6.1 Fundamental solution matrices

The following solvability assertion is a simple consequence of the above.

Theorem 2.44. *If the homogeneous DAE is fine, then,*

(1) *for each arbitrary $x^0 \in \mathbb{R}^m$, the IVP*

$$A(Dx)' + Bx = 0, \quad x(t_0) - x^0 \in N_{can}(t_0), \tag{2.79}$$

is uniquely solvable in $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$,

(2) *the homogeneous IVP*

$$A(Dx)' + Bx = 0, \quad x(t_0) \in N_{can}(t_0),$$

has the trivial solution only, and

(3) *through each $x_0 \in S_{can}(t_0)$ there passes exactly one solution.*

Remark 2.45. Sometimes is seems to be more comfortable to describe the initial condition in (2.79) by an equation, for instance, as

$$\Pi_{can}(t_0)(x(t_0) - x^0) = 0, \tag{2.80}$$

and as

$$C(x(t_0) - x^0) = 0, \tag{2.81}$$

by any matrix C such that $\ker C = \ker \Pi_{can}(t_0) = N_{can}(t_0)$. For instance, taking arbitrary admissible projector functions $\tilde{Q}_0, \ldots, \tilde{Q}_{\mu-1}$, one can choose C such that $C = C\tilde{\Pi}_{can}(t_0)$ (cf. Theorem 3.66).

Proof. (2) The initial condition yields $u(t_0) = D(t_0)\Pi_{can}(t_0)x(t_0) = 0$. Then, the resulting homogeneous IVP for the IERODE admits the trivial solution $u = 0$ only. Therefore, the DAE solution $x = \Pi_{can}D^- u$ vanishes identically, too.
(1) We provide the solution u of the homogeneous IERODE which satisfies the initial condition $u(t_0) = D(t_0)\Pi_{can}(t_0)x^0$. Then we form the DAE solution $x = \Pi_{can}D^- u$, and check that the initial condition is met:

$$\begin{aligned} x(t_0) - x^0 &= \Pi_{can}(t_0)D(t_0)^- u(t_0) - x^0 = \Pi_{can}(t_0)D(t_0)^- D(t_0)\Pi_{can}(t_0)x^0 - x^0 \\ &= -(I - \Pi_{can}(t_0))x^0 \in N_{can}(t_0). \end{aligned}$$

Owing to (2) this is the only solution of the IVP.
(3) We provide the IVP solution as in (1), with x^0 replaced by x_0. This leads to

$$x(t_0) = \Pi_{can}(t_0)D(t_0)^- u(t_0) = \Pi_{can}(t_0)D(t_0)^- D(t_0)\Pi_{can}(t_0)x_0 = \Pi_{can}(t_0)x_0 = 0.$$

The uniqueness is ensured by (2). □

By Theorem 2.44, regular homogeneous DAEs are close to regular homogeneous ODEs. This applies also to their fundamental solution matrices.
Denote by $U(t,t_0)$ the classical fundamental solution matrix of the IERODE, that is, of the explicit ODE (2.74), which is normalized at $t_0 \in \mathcal{I}$, i.e., $U(t_0,t_0) = I$.
For each arbitrary initial value $u_0 \in D(t_0)S_{can}(t_0)$, the solution of the homogeneous IERODE passing through remains for ever in this invariant subspace, which means $U(t,t_0)u_0 \in D(t)S_{can}(t)$ for all $t \in \mathcal{I}$, and hence

$$U(t,t_0)D(t_0)\Pi_{can}(t_0) = D(t)\Pi_{can}(t)D(t)^- U(t,t_0)D(t_0)\Pi_{can}(t_0), \quad t \in \mathcal{I}. \tag{2.82}$$

Each solution of the homogeneous DAE can now be expressed as

$$\begin{aligned} x(t) =& (I - \mathcal{H}_0(t))D(t)^- U(t,t_0)u_0 = \Pi_{can}(t)D(t)^- U(t,t_0)u_0, \\ & t \in \mathcal{I},\ u_0 \in D(t_0)S_{can}(t_0), \end{aligned} \tag{2.83}$$

and also as

$$x(t) = \underbrace{\Pi_{can}(t)D(t)^- U(t,t_0)D(t_0)\Pi_{can}(t_0)}_{X(t,t_0)}x^0, \quad t \in \mathcal{I}, \quad \text{with} \quad x^0 \in \mathbb{R}^m. \tag{2.84}$$

If $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ satisfies the homogeneous DAE, then there is exactly one $u_0 \in D(t_0)S_{can}(t_0)$ such that the expression (2.83) is valid, and there are elements $x^0 \in \mathbb{R}^m$ such that (2.84) applies. Except for the index-0 case, x^0 is not unique.

Conversely, for each arbitrary $x^0 \in \mathbb{R}^m$, formula (2.84) provides a solution of the homogeneous DAE. We know that the solution values of the homogeneous DAE lie in the d-dimensional canonical subspace S_{can}, in particular $x(t_0) \in S_{can}(t_0)$. Therefore, starting from an arbitrary $x^0 \in \mathbb{R}^m$, the consistency of $x(t_0)$ with x^0 cannot be expected. What we always attain is the relation

$$x(t_0) = \Pi_{can}(t_0)x^0,$$

but the condition $x(t_0) = x_0$ is exclusively reserved for x_0 belonging to $S_{can}(t_0)$.

The composed matrix function

$$X(t,t_0) := \Pi_{can}(t)D(t)^- U(t,t_0)D(t_0)\Pi_{can}(t_0), \quad t \in \mathcal{I}, \tag{2.85}$$

arising in the solution expression (2.84) plays the role of a fundamental solution matrix of the DAE (2.44). In comparison with the (regular) ODE theory, there are several differences to be considered. By construction, it holds that $X(t_0,t_0) = \Pi_{can}(t_0)$ and

$$\operatorname{im} X(t,t_0) \subseteq S_{can}(t), \quad N_{can}(t_0) \subseteq \ker X(t,t_0), \quad t \in \mathcal{I}, \tag{2.86}$$

so that $X(t,t_0)$ is a *singular* matrix, except for the case $\mu = 0$. $X(.,t_0)$ is continuous, and $DX(.,t_0) = D\Pi_{can}D^- U(.,t_0)D(t_0)\Pi_{can}(t_0)$ is continuously differentiable, thus the columns of $X(.,t_0)$ are functions belonging to $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$.
We show that $X(t,t_0)$ has constant rank d. Fix an arbitrary $t \neq t_0$ and investigate the nullspace of $X(t,t_0)$. $X(t,t_0)z = 0$ means $U(t,t_0)D(t_0)\Pi_{can}(t_0)z \in \ker \Pi_{can}(t)D(t)^-$, and with regard to (2.82) this yields $U(t,t_0)D(t_0)\Pi_{can}(t_0)z = 0$, thus $D(t_0)\Pi_{can}(t_0)z = 0$, and further $\Pi_{can}(t_0)z = 0$. Owing to (2.86), and for reasons of dimensions, it follows that

$$\operatorname{im} X(t,t_0) = S_{can}(t), \quad \ker X(t,t_0) = N_{can}(t_0), \quad \operatorname{rank} X(t,t_0) = d, \quad t \in \mathcal{I}. \tag{2.87}$$

Lemma 2.46. *The matrix function*

$$X(t,t_0)^- = \Pi_{can}(t_0)D(t_0)^- U(t,t_0)^{-1}D(t)\Pi_{can}(t), \quad t \in \mathcal{I},$$

is the reflexive generalized inverse of $X(t,t_0)$ determined by

$$XX^-X = X, \quad X^-XX^- = X^-, \quad X^-X = \Pi_{can}(t_0), \quad XX^- = \Pi_{can}.$$

Proof. Applying the invariance (2.82), we derive

$$\begin{aligned} X^-X &= \Pi_{can}(t_0)D(t_0)^- U^{-1}D\Pi_{can}\Pi_{can}D^- UD(t_0)\Pi_{can}(t_0) \\ &= \Pi_{can}(t_0)D(t_0)^- U^{-1}\underbrace{D\Pi_{can}D^- UD(t_0)\Pi_{can}(t_0)}_{UD(t_0)\Pi_{can}(t_0)} = \Pi_{can}(t_0), \end{aligned}$$

and $X^-XX^- = (X^-X)X^- = X^-, \quad XX^-X = X(X^-X) = X.$
Next we verify the relation

$$U^{-1}D\Pi_{can} = D(t_0)\Pi_{can}(t_0)D(t_0)^- U^{-1}D\Pi_{can}, \tag{2.88}$$

which in turn implies

$$\begin{aligned} XX^- &= \Pi_{can}D^- U D(t_0)\Pi_{can}(t_0)\Pi_{can}(t_0)D(t_0)^- U^{-1}D\Pi_{can} \\ &= \Pi_{can}D^- U \underbrace{D(t_0)\Pi_{can}(t_0)D(t_0)^- U^{-1}D\Pi_{can}}_{U^{-1}D\Pi_{can}} = \Pi_{can}. \end{aligned}$$

From

$$U' - (D\Pi_{can}D^-)'U + D\Pi_{can}G_\mu^{-1}BD^-U = 0, \quad U(t_0) = 0,$$

it follows that

$$U^{-1\prime} + U^{-1}(D\Pi_{can}D^-)' - U^{-1}D\Pi_{can}G_\mu^{-1}BD^- = 0.$$

Multiplication by $D\Pi_{can}D^-$ on the right results in the explicit ODE

$$V' = V(D\Pi_{can}D^-)' + VD\Pi_{can}G_\mu^{-1}BD^-$$

for the matrix function $V = U^{-1}D\Pi_{can}D^-$. Then, the matrix function $\tilde{V} := (I - D(t_0)\Pi_{can}(t_0)D(t_0)^-)V$ vanishes identically as the solution of the classical homogeneous IVP

$$\tilde{V}' = \tilde{V}(D\Pi_{can}D^-)' + \tilde{V}D\Pi_{can}G_\mu^{-1}BD^-, \quad \tilde{V}(t_0) = 0,$$

and this proves (2.88). □

The columns of $X(.,t_0)$ are solutions of the homogeneous DAE, and the matrix function $X(.,t_0)$ itself satisfies the equation

$$A(DX)' + BX = 0, \tag{2.89}$$

as well as the initial condition

$$X(t_0,t_0) = \Pi_{can}(t_0), \tag{2.90}$$

or, equivalently,

$$\Pi_{can}(t_0)(X(t_0,t_0) - I) = 0. \tag{2.91}$$

Definition 2.47. Let the DAE (2.44) be fine. Each matrix function $Y \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^s, \mathbb{R}^m))$, $d \le s \le m$, is said to be a *fundamental solution matrix* of the DAE, if its columns belong to $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, the equation

$$A(DY)' + BY = 0$$

is satisfied, and the condition $\operatorname{im} Y = S_{can}$ is valid.
A fundamental solution matrix is named *of minimal size*, if $s = d$, and *of maximal size*, if $s = m$.

A maximal size fundamental solution matrix Y is said to be *normalized at* t_0, if $\Pi_{can}(t_0)(Y(t_0) - I) = 0$.

In this sense, the above matrix function $X(.,t_0)$ (cf. (2.85)) is a maximal size fundamental solution normalized at t_0.

Remark 2.48. Concerning fundamental solution matrices of DAEs, there is no common agreement in the literature. Minimal and maximal size fundamental solution matrices, as well as relations among them, were first described in [9] for standard form index-1 DAEs. A comprehensive analysis for regular lower index DAEs, both in standard form and with properly stated leading term, is given in [7]. This analysis applies analogously to regular DAEs with arbitrary index.
Roughly speaking, minimal size fundamental solution matrices have a certain advantage in view of computational aspects, since they have full column rank. For instance, the Moore–Penrose inverse can be easily computed. In contrast, the benefits from maximal size fundamental solution matrices are a natural normalization and useful group properties as pointed out, e.g., in [11], [7].
If $X(t,t_0)$ is the maximal size fundamental solution matrix normalized at $t_0 \in \mathcal{I}$, and $X(t,t_0)^-$ is the generalized inverse described by Lemma 2.46, then it holds for all $t,t_0,t_1 \in \mathcal{I}$ that

$$X(t,t_1)X(t_1,t_0) = X(t,t_0), \quad \text{and} \quad X(t,t_0)^- = X(t_0,t),$$

as immediate consequences of the construction, and Lemma 2.46.

2.6.2 Consistent initial values and flow structure

Turning to inhomogeneous DAEs, first suppose the excitation to be such that a solution exists. Before long, we shall characterize the classes of admissible functions in detail.

Definition 2.49. The function $q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ is named an *admissible excitation* for the DAE (2.44), if the DAE is solvable for this q, i.e., if a solution $x_q \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ exists such that $A(Dx_q)' + Bx_q = q$.

Proposition 2.50. *Let the DAE (2.44) be fine with tractability index* μ.

(1) *Then,* $q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ *is an admissible excitation, if and only if the IVP*

$$A(Dx)' + Bx = q, \quad x(t_0) \in N_{can}(t_0), \tag{2.92}$$

admits a unique solution.

(2) *Each* $q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$, *which for* $\mu \geq 2$ *fulfills the condition* $q = G_\mu P_1 \cdots P_{\mu-1} G_\mu^{-1} q$, *is an admissible excitation.*

Proof. (1) Let q be admissible and x_q the associated solution. Then the function $\tilde{x}(t) := x_q(t) - X(t,t_0)x_q(t_0)$, $t \in \mathcal{I}$, satisfies the IVP (2.92). The uniqueness results from Theorem 2.44 (2). The reverse is trivial.
(2) From the condition $q = G_\mu P_1 \cdots P_{\mu-1} G_\mu^{-1} q$ it follows that

$$\begin{aligned}\mathcal{L}_i q &= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1} q \\ &= \Pi_{i-1} Q_i P_{i+1} \cdots P_{\mu-1} P_1 \cdots P_{\mu-1} G_\mu^{-1} q = 0, \quad i = 1, \ldots \mu - 2, \\ \mathcal{L}_{\mu-1} q &= \Pi_{\mu-2} Q_{\mu-1} G_\mu^{-1} q = \Pi_{\mu-2} Q_{\mu-1} P_1 \cdots P_{\mu-1} G_\mu^{-1} q = 0.\end{aligned}$$

In consequence, the subsystem (2.76)–(2.78) yields successively $v_{\mu-1}, \ldots, v_1 = 0$. The IERODE (2.74) is solvable for each arbitrary continuous excitation. Denote by u_* an arbitrary solution corresponding to q. Then, the function

$$v_0 = -\mathcal{H}_0 D^- u_* + \mathcal{L}_0 q = -\mathcal{H}_0 D^- u_* + Q_0 G_\mu^{-1} q$$

results from equation (2.75), and

$$x := D^- u_* + v_0 = \Pi_{can} D^- u_* + Q_0 G_\mu^{-1} q$$

is a solution of the DAE (2.44) corresponding to this excitation q. □

For a fine index-1 DAE, all continuous functions q are admissible. For fine higher index DAEs, the additional projector function $G_\mu P_1 \cdots P_{\mu-1} G_\mu^{-1}$ cuts away the "dangerous" parts of a function, and ensures that only the zero function is differentiated within the subsystem (2.75)–(2.78). For higher index DAEs, general admissible excitations have certain smoother components. We turn back to this problem later on.

Example 2.51 (A fine index-2 DAE). Consider the DAE

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1 & 0 \end{bmatrix} x \right)' + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{bmatrix} x = q.$$

Here, α is a continuous scalar function. Set and derive

$$D^- = \begin{bmatrix} 1 & -\alpha \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad G_0 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix},$$

and further

$$Q_1 = \begin{bmatrix} 0 & -\alpha & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad Q_1 Q_0 = 0, \quad D\Pi_1 D^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The projector functions Q_0, Q_1 are admissible, G_2 is nonsingular, and hence the DAE is regular with tractability index 2. The given property $\ker Q_1 = S_1 = \{z \in \mathbb{R}^3 :$

$z_2 = 0\}$ indicates that Q_0, Q_1 already provide a fine decoupling. The DAE is fine. Compute additionally

$$\Pi_{can} = \Pi_1 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad G_2^{-1} = \begin{bmatrix} 1 & 0 & -\alpha \\ 0 & 0 & 1 \\ 0 & -1 & 1 \end{bmatrix}, \quad G_2 P_1 G_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

A closer look at the detailed equations makes it clear that each admissible excitation q must have a continuously differentiable component q_3. By the condition $q = G_2 P_1 G_2^{-1} q$, the third component of q is put to be zero. □

Theorem 2.52. *Let the DAE (2.44) be fine. Let $q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$ be an admissible excitation, and let the matrix $C \in L(\mathbb{R}^m, \mathbb{R}^s)$ have the nullspace* $\ker C = N_{can}(t_0)$.

(1) *Then, for each $x^0 \in \mathbb{R}^m$, the IVP*

$$A(Dx)' + Bx = q, \quad C(x(t_0) - x^0) = 0, \tag{2.93}$$

admits exactly one solution.

(2) *The solution of the IVP (2.93) can be expressed as*

$$x(t, t_0, x^0) = X(t, t_0)x^0 + x_q(t),$$

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

$$A(Dx)' + Bx = q, \quad Cx(t_0) = 0. \tag{2.94}$$

Proof. (1) It holds that $C = C\Pi_{can}(t_0)$. Since q is admissible, by Proposition 2.50(1), the solution x_q exists and is unique. Then the function $x_* := X(., t_0)x^0 + x_q$ belongs to the function space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ and satisfies the DAE. Further, x_* meets the initial condition

$$C(x_*(t_0) - x^0) = C\Pi_{can}(t_0)(x_*(t_0) - x^0) = C\Pi_{can}(t_0)(\Pi_{can}(t_0)x^0 + x_q(t_0) - x^0) = 0,$$

and hence, x_* satisfies the IVP (2.93). By Theorem 2.44, x_* is the only IVP solution. This proves at the same time (2). □

We take a further look at the structure of the DAE solutions x_q and $x(., t_0, x^0)$. For the given admissible excitation q, we denote

$$v := v_1 + \cdots + v_{\mu-1} + \mathcal{L}_0 q - \sum_{l=1}^{\mu-1} \mathcal{N}_{0l}(Dv_l)' - \sum_{l=2}^{\mu-1} \mathcal{M}_{0l} v_l, \tag{2.95}$$

whereby $v_1, \ldots, v_{\mu-1} \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ are determined by equations (2.76)–(2.78), depending on q. All the required derivatives exist due to the admissibility of q. If q vanishes identically, so does v. By construction, $v(t) \in N_{can}(t)$, $t \in \mathcal{I}$, and $Dv = Dv_1 + \cdots + Dv_{\mu-1}$, thus $v \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. The function v is fully determined

by q and the coefficients of the subsystem (2.75)–(2.78). It does not depend either on the initial condition nor the IERODE solution.

Introduce further the continuously differentiable function u_q as

$$\begin{aligned} u_q(t) &:= \int_{t_0}^{t} U(t,t_0)U(s,t_0)^{-1}D(s)\Pi_{can}(s)G_\mu^{-1}(s)q(s)ds \\ &= U(t,t_0)\int_{t_0}^{t} X(s,t_0)^{-}G_\mu^{-1}(s)q(s)ds, \quad t\in\mathcal{I}, \end{aligned}$$

that is, as the solution of the inhomogeneous IERODE completed by the homogeneous initial condition $u(t_0)=0$. Now the solution x_q and, in particular, its value at t_0, can be expressed as

$$\begin{aligned} &x_q(t) = D(t)^{-}u_q(t) - \mathcal{H}_0(t)D(t)^{-}u_q(t) + v(t) = \Pi_{can}(t)D(t)^{-}u_q(t) + v(t), \\ &x_q(t_0) = v(t_0) \in N_{can}(t_0). \end{aligned}$$

The solution of the IVP (2.93) and its value at t_0 can be written in the form

$$x(t,t_0,x^0) = X(t,t_0)x^0 + \Pi_{can}(t)D(t)^{-}u_q(t) + v(t), \tag{2.96}$$

$$x(t_0,t_0,x^0) = \Pi_{can}(t_0)x^0 + v(t_0), \tag{2.97}$$

but also as

$$\begin{aligned} x(t,t_0,x^0) &= \Pi_{can}(t)D(t)^{-}U(t,t_0)D(t_0)\Pi_{can}(t_0)x^0 + \Pi_{can}(t)D(t)^{-}u_q(t) + v(t) \\ &= \Pi_{can}(t)D(t)^{-}\underbrace{\{U(t,t_0)D(t_0)\Pi_{can}(t_0)x^0 + u_q(t)\}}_{u(t,t_0,D(t_0)\Pi_{can}(t_0)x^0)} + v(t). \end{aligned}$$

The last representation

$$x(t,t_0,x^0) = \underbrace{\Pi_{can}(t)D(t)^{-}}_{\Uparrow} \underbrace{u(t,t_0,D(t_0)\Pi_{can}(t_0)x^0)}_{\Uparrow} + \underbrace{v(t)}_{\Uparrow}$$

wrapping — inherent flow — perturbation

unveils the general solution structure of fine DAEs to be the perturbed and wrapped flow of the IERODE along the invariant subspace DS_{can}. If the wrapping is thin (bounded) and the perturbation disappears, then the situation is close to regular ODEs. However, it may well happen that wrapping and perturbation dominate (cf. Example 2.57 below). In extreme cases, it holds that $S_{can} = \{0\}$, thus the inherent flow vanishes, and only the perturbation term remains (cf. Example 2.4).

From Theorem 2.52, and the representation (2.96), it follows that, for each given admissible excitation, the set

$$\mathcal{M}_{can,q}(t) := \{z + v(t) : z \in S_{can}(t)\}, \quad t\in\mathcal{I}, \tag{2.98}$$

is occupied with solution values at time t, and all solution values at time t belong to this set. In particular, for $x_0 \in \mathcal{M}_{can,q}(t_0)$ it follows that $x_0 = z_0 + v(t_0)$, $z_0 \in S_{can}(t_0)$; further $\Pi_{can}(t_0)x_0 = z_0$ and

$$x(t_0,t_0,x_0) = \Pi_{can}(t_0)x_0 + v(t_0) = z_0 + v(t_0) = x_0.$$

By construction, the inclusions

$$\begin{aligned} S_{can}(t) \subseteq S_0(t) &= \{z \in \mathbb{R}^m : B(t)z \in \operatorname{im} A(t)\} = \ker \mathcal{W}_0(t)B(t), \\ \mathcal{M}_{can,q}(t) \subseteq \mathcal{M}_0(t) &= \{x \in \mathbb{R}^m : B(t)x - q(t) \in \operatorname{im} A(t)\} \end{aligned}$$

are valid, whereby $\mathcal{W}_0(t)$ is again a projector along $\operatorname{im} A(t) = \operatorname{im} G_0(t)$. Recall that $S_{can}(t)$ and $S_0(t)$ have the dimensions $d = m - \sum_{j=0}^{\mu-1}(m - r_j) = r_0 - \sum_{j=1}^{\mu-1}(m - r_j)$ and r_0, respectively. Representing the obvious constraint set as

$$\begin{aligned} \mathcal{M}_0(t) &= \{x \in \mathbb{R}^m : \mathcal{W}_0(t)B(t)x = \mathcal{W}_0(t)q(t)\} \\ &= \{z + (\mathcal{W}_0(t)B(t))^- \mathcal{W}_0(t)q(t) : z \in S_0(t)\} \end{aligned}$$

we know that $\mathcal{M}_0(t)$, as an affine space, inherits its dimension from $S_0(t)$, while $\mathcal{M}_{can,q}(t)$ has the same dimension d as $S_{can}(t)$.

Since $d = r_0$ if $\mu = 1$, and $d < r_0$ if $\mu > 1$, $\mathcal{M}_{can,q}(t)$ coincides with $\mathcal{M}_0(t)$ for index-1 DAEs, however, for higher index DAEs, $\mathcal{M}_{can,q}(t)$ is merely a proper subset of $\mathcal{M}_0(t)$. $\mathcal{M}_{can,q}(t)$ is the set of consistent values at time t. Knowledge of this set gives rise to an adequate stability notion for DAEs. As pointed out in [7] for lower index cases, in general, $\mathcal{M}_{can,q}$ is a time-varying affine linear subspace of dimension d.

2.6.3 Stability issues

As for regular time-varying ODEs (e.g., [80]), we may consider the qualitative behavior of solutions of DAEs.

Definition 2.53. Let the fine DAE (2.44) with an admissible excitation q be given on the infinite interval $\mathcal{I} = [0,\infty)$. The DAE is said to be

(1) *stable*, if for every $\varepsilon > 0$, $t_0 \in \mathcal{I}$, a value $\delta(\varepsilon,t_0) > 0$ exists, such that the conditions $x_0, \bar{x}_0 \in \mathcal{M}_{can,q}(t_0)$, $|x_0 - \bar{x}_0| < \delta(\varepsilon,t_0)$ imply the existence of solutions $x(.,t_0,x_0)$, $x(.,t_0,\bar{x}_0) \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ as well as the inequality

$$|x(t,t_0,x_0) - x(t,t_0,\bar{x}_0)| < \varepsilon, \quad t_0 \le t,$$

(2) *uniformly stable*, if $\delta(\varepsilon,t_0)$ in (1) is independent of t_0,
(3) *asymptotically stable*, if (1) holds true, and

$$|x(t,t_0,x_0)-x(t,t_0,\bar{x}_0)| \xrightarrow[t\to\infty]{} 0 \quad \text{for all} \quad x_0,\bar{x}_0 \in \mathcal{M}_{can,q}(t_0),\ t_0 \in \mathcal{I},$$

(4) *uniformly asymptotically stable*, if the limit in (3) is uniform with respect to t_0.

Remark 2.54. We can dispense with the explicit use of the set $\mathcal{M}_{can,q}(t_0)$ within the stability notion by turning to appropriate IVPs (cf. Theorem 2.52). This might be more comfortable from the practical point of view.
Let $C \in L(\mathbb{R}^m,\mathbb{R}^s)$ denote a matrix that has precisely $N_{can}(t_0)$ as nullspace, for instance $C = \Pi_{\mu-1}(t_0)$ or $C = \Pi_{can}(t_0)$.
The DAE (2.44) is stable, if for every $\varepsilon > 0$, $t_0 \in \mathcal{I}$, there exists a value $\delta_C(\varepsilon,t_0) > 0$ such that the IVPs

$$A(Dx)'+Bx = q, \quad C(x(t_0)-x^0) = 0,$$
$$A(Dx)'+Bx = q, \quad C(x(t_0)-\bar{x}^0) = 0,$$

with $x^0,\ \bar{x}^0 \in \mathbb{R}^m$, $|C(x^0-\bar{x}^0)| < \delta_C(\varepsilon,t_0))$, have solutions $x(.,t_0,x^0),\ x(.,t_0,\bar{x}^0) \in C_D^1(\mathcal{I},\mathbb{R}^m)$, and it holds that $|x(.,t_0,x^0)-x(.,t_0,\bar{x}^0)| < \varepsilon$, for $t \geq t_0$.
This notion is equivalent to the previous one. Namely, denoting by C^- a generalized reflexive inverse of C such that $C^-C = \Pi_{can}(t_0)$, and considering the relation

$$\begin{aligned} C^-C(x^0-\bar{x}^0) &= \Pi_{can}(t_0)x^0 - \Pi_{can}(t_0)\bar{x}^0 \\ &= \underbrace{\Pi_{can}(t_0)x^0 + v(t_0)}_{=x_0\in\mathcal{M}_0(t_0)} - \underbrace{(\Pi_{can}(t_0)\bar{x}^0 + v(t_0))}_{=\bar{x}_0\in\mathcal{M}_0(t_0)} = x_0-\bar{x}_0, \end{aligned}$$

we know that the existence of $\delta(\varepsilon,t_0)$ in Definition 2.53 implies the existence of $\delta_C(\varepsilon,t_0) = |C|\delta(\varepsilon,t_0)$. Conversely, having $\delta_C(\varepsilon,t_0)$ we may put $\delta(\varepsilon,t_0) = |C^-|\delta_C(\varepsilon,t_0)$.

Making use of the linearity,

$$x(t,t_0,x_0)-x(t,t_0,\bar{x}_0) = X(t,t_0)(x_0-\bar{x}_0) \tag{2.99}$$

we trace back the stability questions to the growth behavior of the fundamental solution matrices. Applying normalized maximal size fundamental solution matrices we modify well-known results on flow properties of explicit ODEs (e.g., [80]) so that they can be considered for DAEs.

Theorem 2.55. *Let the DAE (2.44) be fine and the excitation q be admissible. Then the following assertions hold true, with positive constants K_{t_0}, K and α:*

(1) *If $|X(t,t_0)| \leq K_{t_0}$, $t \geq t_0$, then the DAE is stable.*
(2) *If $|X(t,t_0)| \xrightarrow[t\to\infty]{} 0$, then the DAE is asymptotically stable.*
(3) *If $|X(t,t_0)X(s,t_0)^-| \leq K$, $t_0 \leq s \leq t$, then the DAE is uniformly stable.*
(4) *If $|X(t,t_0)X(s,t_0)^-| \leq Ke^{-\alpha(t-s)}$, $t_0 \leq s \leq t$, then the DAE is uniformly asymptotically stable.*

Proof. (1) It suffices to put $\delta(t_0,\varepsilon)=\varepsilon/K_{t_0}$.
(2) This is now obvious.
(4) Take $x_0,\bar{x}_0\in\mathcal{M}_{can,q}(t_0)$, $z_0:=x_0-\bar{x}_0\neq 0$ such that $z_0\in S_{can}$ and $X(t,t_0)z_0$ has no zeros. For $t\geq s$, we compute

$$\frac{|X(t,t_0)z_0|}{|X(s,t_0)z_0|}=\frac{|X(t,t_0)\Pi_{can}z_0|}{|X(s,t_0)z_0|}=\frac{|X(t,t_0)X(s,t_0)^-X(s,t_0)z_0|}{|X(s,t_0)z_0|}$$
$$\leq |X(t,t_0)X(s,t_0)^-|\leq K\mathrm{e}^{-\alpha(t-s)}.$$

This implies

$$|x(t,t_0,x_0)-x(t,t_0,\bar{x}_0)|=|X(t,t_0)z_0|\leq K\mathrm{e}^{-\alpha(t-s)}|x(s,t_0,x_0)-x(s,t_0,\bar{x}_0)|.$$

(3) This is proved as (4) by letting $\alpha=0$. □

In the theory of explicit ODEs, for instance, in the context of boundary value problems, the notion of dichotomy plays its role. The flow of a dichotomic ODE accommodates both decreasing and increasing modes. The same can happen for DAEs. As for explicit ODEs, we relate dichotomy of DAEs to the flow of homogeneous equations. More precisely, we apply maximal size fundamental solution matrices $X(t,t_0)$ normalized at a reference point t_0. The following definition resembles that for ODEs.

Definition 2.56. The fine DAE (2.44) is said to be *dichotomic*, if there are constants $K,\alpha,\beta\geq 0$, and a nontrivial projector (not equal to the zero or identity matrix) $P_{dich}\in L(\mathbb{R}^m)$ such that $P_{dich}=\Pi_{can}(t_0)P_{dich}=P_{dich}\Pi_{can}(t_0)$, and the following inequalities apply for all $t,s\in\mathcal{I}$:

$$|X(t,t_0)P_{dich}X(s,t_0)^-|\leq Ke^{-\alpha(t-s)},\quad t\geq s,$$
$$|X(t,t_0)(I-P_{dich})X(s,t_0)^-|\leq Ke^{-\beta(s-t)},\quad t\leq s.$$

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

Sometimes it is reasonable to write the last inequality in the form

$$|X(t,t_0)(\Pi_{can}(t_0)-P_{dich})X(s,t_0)^-|\leq Ke^{-\beta(s-t)},\quad t\leq s.$$

It should be pointed out that dichotomy is actually independent of the reference point t_0. Namely, for $t_1\neq t_0$, with $P_{dich,t_1}:=X(t_1,t_0)P_{dich}X(t_1,t_0)^-$ we have a projector such that $P_{dich,t_1}=\Pi_{can}(t_1)P_{dich,t_1}=P_{dich,t_1}\Pi_{can}(t_1)$ and

$$|X(t,t_1)P_{dich,t_1}X(s,t_1)^-|\leq Ke^{-\alpha(t-s)},\quad t\geq s,$$
$$|X(t,t_1)(\Pi_{can}(t_1)-P_{dich,t_1})X(s,t_1)^-|\leq Ke^{-\beta(s-t)},\quad t\leq s.$$

Analogously to the ODE case, the flow of a dichotomic homogeneous DAE is divided into two parts, one containing in a certain sense a nonincreasing solution, the other with nondecreasing ones. More precisely, for a nontrivial $x_0\in\operatorname{im}P_{dich}\subseteq$

$S_{can}(t_0)$, the DAE solution $x(t,t_0,x_0) = X(t,t_0)x_0$ has no zeros, and it satisfies for $t \geq s$ the inequalities

$$\begin{aligned}\frac{|x(t,t_0,x_0)|}{|x(s,t_0,x_0)|} &= \frac{|X(t,t_0)x_0|}{|X(s,t_0)x_0|} = \frac{|X(t,t_0)P_{dich}\Pi_{can}(t_0)x_0|}{|X(s,t_0)x_0|} \\ &= \frac{|X(t,t_0)P_{dich}X(s,t_0)^- X(s,t_0)x_0|}{|X(s,t_0)x_0|} \\ &\leq |X(t,t_0)P_{dich}X(s,t_0)^-| \leq Ke^{-\alpha(t-s)}.\end{aligned}$$

For solutions $x(t,t_0,x_0) = X(t,t_0)x_0$ with $x_0 \in \operatorname{im}(I-P_{dich})\Pi_{can} \subseteq S_{can}(t_0)$ we show analogously, for $t \leq s$,

$$\begin{aligned}\frac{|x(t,t_0,x_0)|}{|x(s,t_0,x_0)|} &= \frac{|X(t,t_0)x_0|}{|X(s,t_0)x_0|} = \frac{|X(t,t_0)(I-P_{dich})\Pi_{can}(t_0)x_0|}{|X(s,t_0)x_0|} \\ &= \frac{|X(t,t_0)(I-P_{dich})X(s,t_0)^- X(s,t_0)x_0|}{|X(s,t_0)x_0|} \\ &\leq |X(t,t_0)(I-P_{dich})X(s,t_0)^-| \leq Ke^{-\beta(s-t)}.\end{aligned}$$

The canonical subspace of the dichotomic DAE decomposes into

$$S_{can}(t) = \operatorname{im}X(t,t_0) = \operatorname{im}X(t,t_0)P_{dich} \oplus \operatorname{im}X(t,t_0)(I-P_{dich}) =: S^-_{can}(t) \oplus S^+_{can}(t).$$

The following two inequalities result for $t \geq s$, and they characterize the subspaces S^-_{can} and S^+_{can} as those containing nonincreasing and nondecreasing solutions, respectively:

$$\begin{aligned}|x(t,t_0,x_0)| &\leq Ke^{-\alpha(t-s)}|x(s,t_0,x_0)|, \quad \text{if} \quad x_0 \in S^-_{can}, \\ \frac{1}{K}e^{\beta(t-s)}|x(s,t_0,x_0)| &\leq |x(t,t_0,x_0)|, \quad \text{if} \quad x_0 \in S^+_{can}.\end{aligned}$$

In particular, for $s = t_0$ it follows that

$$\begin{aligned}|x(t,t_0,x_0)| &\leq Ke^{-\alpha(t-t_0)}|x_0|, \quad \text{if} \quad x_0 \in S^-_{can}, \\ \frac{1}{K}e^{\beta(t-t_0)}|x_0| &\leq |x(t,t_0,x_0)|, \quad \text{if} \quad x_0 \in S^+_{can}.\end{aligned}$$

If $\alpha > 0$, and $\mathcal{I} = [t_0,\infty)$, then $|x(t,t_0,x_0)|$ tends to zero for t tending to ∞, if x_0 belongs to $S^-_{can}(t_0)$. If $\beta > 0$ and $x_0 \in S^+_{can}(t_0)$, then $x(t,t_0,x_0)$ growths unboundedly with increasing t.

As for explicit ODEs, dichotomy makes good sense on infinite intervals I. The growth behavior of fundamental solutions is also important for the condition of boundary value problems stated on compact intervals (e.g., [2] for explicit ODEs, also [146] for index-1 DAEs). Dealing with compact intervals one supposes a constant K of *moderate size*.

Example 2.57 (Dichotomic IERODE and dichotomic DAE). Consider the semi-explicit DAE

$$\begin{bmatrix} I \\ 0 \end{bmatrix} \left(\begin{bmatrix} I & 0 \end{bmatrix} x \right)' + \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} x = 0,$$

consisting of three equations, $m_1 = 2, m_2 = 1, n = 2$. Let B_{22} have no zeros, and let the coefficients be such that

$$B_{11} + B_{12} \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix} = \begin{bmatrix} \alpha & 0 \\ 0 & -\beta \end{bmatrix}, \quad \begin{bmatrix} \gamma_1 & \gamma_2 \end{bmatrix} := -B_{22}^{-1} B_{21},$$

with constants $\alpha, \beta \geq 0$. Then, the canonical projector function and the IERODE have the form (cf. Example 2.32)

$$\Pi_{can} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \gamma_1 & \gamma_2 & 0 \end{bmatrix}, \quad \text{and} \quad u' + \begin{bmatrix} \alpha & 0 \\ 0 & -\beta \end{bmatrix} u = 0.$$

The IERODE is obviously dichotomic. Compute the fundamental solution matrix of the DAE and its generalized inverse:

$$X(t,t_0) = \begin{bmatrix} e^{-\alpha(t-t_0)} & 0 & 0 \\ 0 & e^{\beta(t-t_0)} & 0 \\ \gamma_1(t) e^{-\alpha(t-t_0)} & \gamma_2(t) e^{\beta(t-t_0)} & 0 \end{bmatrix},$$

$$X(t,t_0)^- = \begin{bmatrix} e^{\alpha(t-t_0)} & 0 & 0 \\ 0 & e^{-\beta(t-t_0)} & 0 \\ \gamma_1(t_0) e^{\alpha(t-t_0)} & \gamma_2(t_0) e^{-\beta(t-t_0)} & 0 \end{bmatrix}.$$

The projector

$$P_{dich} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ \gamma_1(t_0) & 0 & 0 \end{bmatrix}, \quad \Pi_{can}(t_0) - P_{dich} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \gamma_2(t_0) & 0 \end{bmatrix},$$

meets the condition of Definition 2.56, and it follows that

$$X(t,t_0) P_{dich} X(t,t_0)^- = \mathrm{e}^{-\alpha(t-t_0)} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ \gamma_1(t) & 0 & 0 \end{bmatrix}, \text{ and } \quad S_{can}^-(t) = \operatorname{span} \begin{bmatrix} 1 \\ 0 \\ \gamma_1(t) \end{bmatrix},$$

$$X(t,t_0)(\Pi_{can}(t_0) - P_{dich}) X(t,t_0)^- = \mathrm{e}^{\beta(t-t_0)} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \gamma_2(t) & 0 \end{bmatrix}, \text{ and}$$

$$S_{can}^+(t) = \operatorname{span} \begin{bmatrix} 0 \\ 1 \\ \gamma_2(t) \end{bmatrix}.$$

If both γ_1 and γ_2 are bounded functions, then this DAE is dichotomic. If, additionally, α and β are positive, the DAE has an exponential dichotomy. We see that if the entries of the canonical projector remain bounded, then the dichotomy of the IERODE is passed over to the DAE. In contrast, if the functions γ_1, γ_2 grow unboundedly, the situation within the DAE may change. For instance, if $\alpha = 0$ and $\beta > 0$, then the fundamental solution

$$X(t,t_0) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \mathrm{e}^{\beta(t-t_0)} & 0 \\ \gamma_1(t) & \gamma_2(t)\mathrm{e}^{\beta(t-t_0)} & 0 \end{bmatrix}$$

indicates that each nontrivial solution will grow unboundedly though the IERODE is dichotomic. □

The last example is too simple in the sense that $DS_{can} = \operatorname{im} D = \mathbb{R}^n$ is valid, which happens only for regular index-1 DAEs, if A has full column rank, and D has full row rank. In general, DS_{can} is a time-varying subspace of $\operatorname{im} D$, and the IERODE at the whole does not comprise an exponential dichotomy. Here the question is whether the IERODE shows dichotomic behavior along its (time-varying) invariant subspace DS_{can}. We do not go into more details in this direction.

2.6.4 Characterizing admissible excitations and perturbation index

The fine decoupling of a regular DAE into the IERODE (2.74) and the subsystem (2.75)–(2.78) allows a precise and detailed description of admissible excitations. Remember that the equations (2.75)–(2.78), which means

$$v_0 = -\sum_{l=1}^{\mu-1} \mathcal{N}_{0l}(Dv_l)' - \sum_{l=2}^{\mu-1} \mathcal{M}_{0l}\, v_l - \mathcal{H}_0 D^- u + \mathcal{L}_0 q, \tag{2.100}$$

$$v_i = -\sum_{l=i+1}^{\mu-1} \mathcal{N}_{il}(Dv_l)' - \sum_{l=i+2}^{\mu-1} \mathcal{M}_{il}\, v_l + \mathcal{L}_i q, \quad i = 1,\ldots,\mu-3, \tag{2.101}$$

$$v_{\mu-2} = -\mathcal{N}_{\mu-2,\mu-1}(Dv_{\mu-1})' + \mathcal{L}_{\mu-2} q, \tag{2.102}$$

$$v_{\mu-1} = \mathcal{L}_{\mu-1} q, \tag{2.103}$$

constitute the subsystem (2.62) specified for fine decouplings. We quote once again the coefficients

$$
\begin{aligned}
\mathcal{N}_{01} &:= -Q_0Q_1D^-, \\
\mathcal{N}_{0j} &:= -Q_0P_1\cdots P_{j-1}Q_jD^-, && j=2,\dots,\mu-1,\\
\mathcal{N}_{i,i+1} &:= -\Pi_{i-1}Q_iQ_{i+1}D^-, \\
\mathcal{N}_{ij} &:= -\Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^-, && j=i+2,\dots,\mu-1,\ i=1,\dots,\mu-2,\\
\mathcal{M}_{0j} &:= Q_0P_1\cdots P_{\mu-1}\mathcal{M}_jD\Pi_{j-1}Q_j, && j=1,\dots,\mu-1,\\
\mathcal{M}_{ij} &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}\mathcal{M}_jD\Pi_{j-1}Q_j, && j=i+1,\dots,\mu-1,\ i=1,\dots,\mu-2,\\
\mathcal{L}_0 &:= Q_0P_1\cdots P_{\mu-1}G_\mu^{-1}, \\
\mathcal{L}_i &:= \Pi_{i-1}Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}, && i=1,\dots,\mu-2,\\
\mathcal{L}_{\mu-1} &:= \Pi_{\mu-2}Q_{\mu-1}G_\mu^{-1}, \\
\mathcal{H}_0 &:= Q_0P_1\cdots P_{\mu-1}\mathcal{K}\Pi_{\mu-1}.
\end{aligned}
$$

For the detailed form of $\mathcal{K}$ and $\mathcal{M}_j$ we refer to (2.54) and (2.55), respectively. All these coefficients are continuous by construction.

The IERODE is solvable for each arbitrary continuous inhomogeneity, therefore, additional smoothness requirements may occur only from the subsystem equations (2.100)–(2.102).

This causes us to introduce the following function space, if $\mu \geq 2$:

$$
\begin{aligned}
\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m) := \Big\{ & q\in\mathcal{C}(\mathcal{I},\mathbb{R}^m):\\
& v_{\mu-1} := \mathcal{L}_{\mu-1}q, && Dv_{\mu-1}\in\mathcal{C}^1(\mathcal{I},\mathbb{R}^n),\\
& v_{\mu-2} := -\mathcal{N}_{\mu-2,\mu-1}(Dv_{\mu-1})' + \mathcal{L}_{\mu-2}q, && Dv_{\mu-2}\in\mathcal{C}^1(\mathcal{I},\mathbb{R}^n),\\
& v_i := -\sum_{l=i+1}^{\mu-1}\mathcal{N}_{il}(Dv_l)' - \sum_{l=i+2}^{\mu-1}\mathcal{M}_{il}\,v_l + \mathcal{L}_iq, && Dv_i\in\mathcal{C}^1(\mathcal{I},\mathbb{R}^n),\\
& && i=1,\dots,\mu-3\Big\}. \qquad (2.104)
\end{aligned}
$$

Additionally we set for $\mu = 1 : \mathcal{C}^{ind\,1}(\mathcal{I},\mathbb{R}^m) := \mathcal{C}(\mathcal{I},\mathbb{R}^m)$.
The function space $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ makes sense unless there are further smoothness assumptions concerning the coefficients. It contains, in particular, all continuous functions q that satisfy the condition $q = G_\mu P_1\cdots P_{\mu-1}G_\mu^{-1}q$ (cf. Proposition 2.50), which implies $v_1 = 0,\dots,v_{\mu-1} = 0$.
The function space $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ is always a proper subset of the continuous function space $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$. The particular cases $\mu = 2$ and $\mu = 3$ are described in detail as

$$
\begin{aligned}
\mathcal{C}^{ind\,2}(\mathcal{I},\mathbb{R}^m) &:= \Big\{ q\in\mathcal{C}(\mathcal{I},\mathbb{R}^m) : v_1 := \mathcal{L}_1q,\ Dv_1\in\mathcal{C}^1(\mathcal{I},\mathbb{R}^n)\Big\} \qquad (2.105)\\
&= \Big\{ q\in\mathcal{C}(\mathcal{I},\mathbb{R}^m) : D\Pi_0Q_1G_2^{-1}q\in\mathcal{C}^1(\mathcal{I},\mathbb{R}^m)\Big\} = \mathcal{C}^1_{D\Pi_0Q_1G_2^{-1}}(\mathcal{I},\mathbb{R}^m),
\end{aligned}
$$

and

$$\mathcal{C}^{ind\,3}(\mathcal{I},\mathbb{R}^m) := \Big\{ q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : v_2 := \mathcal{L}_2 q,\, Dv_2 \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n),$$
$$v_1 := -\mathcal{N}_{12}(Dv_2)' + \mathcal{L}_1 q,\, Dv_1 \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n) \Big\} \tag{2.106}$$
$$= \Big\{ q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : v_2 := \Pi_1 Q_2 G_3^{-1} q,\, Dv_2 \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n),$$
$$v_1 := \Pi_0 Q_1 Q_2 D^- (Dv_2)' + \Pi_0 Q_1 P_2 G_3^{-1} q,\, Dv_1 \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n) \Big\}.$$

We now introduce the linear operator $L : \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ by means of

$$Lx := A(Dx)' + Bx, \quad x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m), \tag{2.107}$$

so that the DAE (2.44) is represented by the operator equation $Lx = q$, and an excitation q is admissible, exactly if it belongs to the range $\operatorname{im} L$ of the operator L.

Proposition 2.58. *If the DAE (2.44) is fine with tractability index $\mu \in \mathbb{N}$, then the linear operator L has the range*

$$\operatorname{im} L = \mathcal{C}(\mathcal{I},\mathbb{R}^m), \qquad \text{if} \quad \mu = 1,$$
$$\operatorname{im} L = \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m) \subset \mathcal{C}(\mathcal{I},\mathbb{R}^m), \qquad \text{if} \quad \mu \geq 2.$$

Proof. The index-1 case is already known from Proposition 2.50 and the definition of L. Turn to the case $\mu \geq 2$. By means of the decoupled version, to each excitation $q \in \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$, we find a solution $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ of the DAE, so that the inclusion $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m) \subseteq \operatorname{im} L$ follows. Namely, owing to the properties of q (cf. (2.104)), there is a solution $v_{\mu-1} \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ of equation (2.103), then a solution $v_{\mu-2} \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ of (2.102), and solutions $v_i \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ of (2.101), successively for $i = \mu-3,\ldots,1$. Furthermore, compute a solution u of the IERODE, and v_0 from equation (2.100). Finally put $x := D^- u + v_0 + \cdots + v_{\mu-1}$.
To show the reverse inclusion $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m) \supseteq \operatorname{im} L$ we fix an arbitrary $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ and investigate the resulting $q := A(Dx)' + Bx$. We again apply the decoupling. Denote $v_0 := Q_0 x$, and $v_i := \Pi_{i-1} Q_i x$, for $i = 1,\ldots,\mu-1$. Since the projector functions $D\Pi_{i-1}Q_i D^-$, $i = 1,\ldots,\mu-1$, and the function Dx are continuously differentiable, so are the functions $Dv_i = D\Pi_{i-1}Q_i D^- Dx$, $i = 1,\ldots,\mu-1$. Now equation (2.103) yields $v_{\mu-1} := \mathcal{L}_{\mu-1} q \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, equation (2.102) gives $v_{\mu-2} := -\mathcal{N}_{\mu-2\,\mu-1}(Dv_{\mu-1})' + \mathcal{L}_{\mu-2} q \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, and so on. □

At this point, the reader's attention should be directed to the fact that the linear function space $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ does not necessarily contain all continuously differentiable functions. For instance, if D is continuous, but fails to be continuously differentiable, then there are constant functions x_{const} such that Dx_{const} fails to be continuously differentiable, and hence x_{const} does not belong to $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$. In contrast, if D is continuously differentiable and its nullspace is nontrivial, then the proper inclusion

$$\mathcal{C}^1(\mathcal{I},\mathbb{R}^m) \subset \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$$

is valid. Similar aspects are to be considered if one deals with the space $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ comprising the admissible excitations. For $\mu \geq 2$, only if the involved coefficients $\mathcal{L}_i$, $\mathcal{N}_{ij}$ and $\mathcal{M}_{ij}$ are sufficiently smooth, does the inclusion

$$\mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m) \subset \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m),$$

hold true. Of course, the index-1 case is simple with

$$\mathcal{C}(\mathcal{I},\mathbb{R}^m) = \mathcal{C}^{ind\,1}(\mathcal{I},\mathbb{R}^m).$$

To achieve more transparent estimates we introduce, for each function w being continuous on $\mathcal{I}$ and $t_0, t_1 \in \mathcal{I}$, $t_0 < t_1$, the expression

$$\|w\|_{\infty}^{[t_0,t_1]} := \max_{t_0 \leq \tau \leq t_1} |w(\tau)|,$$

which is the maximum-norm related to the compact interval $[t_0,t_1]$. Moreover, for $q \in \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ and $t_0, t_1 \in \mathcal{I}$, $t_0 < t_1$, we introduce

$$\|q\|_{ind\,\mu}^{[t_0,t_1]} := \|q\|_{\infty}^{[t_0,t_1]} + \|(Dv_{\mu-1})'\|_{\infty}^{[t_0,t_1]} + \cdots + \|(Dv_1)'\|_{\infty}^{[t_0,t_1]},$$

which means for the special cases $\mu = 2$ and $\mu = 3$:

$$\begin{aligned}
\|q\|_{ind\,2}^{[t_0,t_1]} &:= \|q\|_{\infty}^{[t_0,t_1]} + \|(Dv_1)'\|_{\infty}^{[t_0,t_1]} = \|q\|_{\infty}^{[t_0,t_1]} + \|(D\Pi_0 Q_1 G_2^{-1} q)'\|_{\infty}^{[t_0,t_1]},\\
\|q\|_{ind\,3}^{[t_0,t_1]} &:= \|q\|_{\infty}^{[t_0,t_1]} + \|(Dv_2)'\|_{\infty}^{[t_0,t_1]} + \|(Dv_1)'\|_{\infty}^{[t_0,t_1]}\\
&= \|q\|_{\infty}^{[t_0,t_1]} + \|(D\Pi_1 Q_2 G_3^{-1} q)'\|_{\infty}^{[t_0,t_1]}\\
&\quad + \|(D\Pi_0 Q_1 Q_2 D^{-}(D\Pi_1 Q_2 G_3^{-1} q)' + D\Pi_0 Q_1 P_2 G_3^{-1} q)'\|_{\infty}^{[t_0,t_1]}.
\end{aligned}$$

Theorem 2.59. *Let the DAE (2.44) be fine with tractability index $\mu \in \mathbb{N}$. Let $t_0 \in \mathcal{I}$ and let C be a matrix such that* $\ker C = N_{can}(t_0)$. *Let the compact interval $[t_0,\bar{t}] \subseteq \mathcal{I}$ be fixed. Then the following assertions are true:*

(1) *The excitation q is admissible, if and only if it belongs to $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$.*

(2) *For each pair $q \in \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$, $x^0 \in \mathbb{R}^m$, the solution $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ of the IVP*

$$A(Dx)' + Bx = q, \quad C(x(t_0) - x^0) = 0, \tag{2.108}$$

satisfies the inequality

$$|x(t)| \leq \|x\|_{\infty}^{[t_0,t]} \leq c\left\{|\Pi_{can}(t_0)x^0| + \|q\|_{ind\,\mu}^{[t_0,t]}\right\}, \quad t_0 \leq t \leq \bar{t}, \tag{2.109}$$

whereby the constant c depends only on the interval.

(3) *If the DAE coefficients are so smooth that $\mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m) \subset \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$, and*

$$\|q\|_{ind\,\mu}^{[t_0,t]} \le c_0\Big\{\|q\|_\infty^{[t_0,t]} + \sum_{l=1}^{\mu-1}\|q^{(l)}\|_\infty^{[t_0,t]}\Big\},\ \text{for}\ q\in\mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m),$$

then, for each pair $q\in\mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m)$, $x^0\in\mathbb{R}^m$, *it holds that*

$$\|x\|_\infty^{[t_0,t]} \le K\Big\{|\Pi_{can}(t_0)x^0| + \|q\|_\infty^{[t_0,t]} + \sum_{l=1}^{\mu-1}\|q^{(l)}\|_\infty^{[t_0,t]}\Big\}. \tag{2.110}$$

Proof. (1) is a consequence of Proposition 2.58, and (3) results from (2). It remains to verify (2). We apply the solution representation (2.96). First we consider the function v defined by (2.95), for a given $q\in\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$. One has in detail

$$\begin{aligned} v_{\mu-1} &= \mathcal{L}_{\mu-1}q, \quad \text{thus} \quad \|v_{\mu-1}\|_\infty^{[t_0,t]} \le \bar{c}_{\mu-1}\|q\|_{ind\,\mu}^{[t_0,t]},\\ v_{\mu-2} &= \mathcal{L}_{\mu-2}q - \mathcal{N}_{\mu-2\,\mu-1}(Dv_{\mu-1})', \quad \text{thus} \quad \|v_{\mu-2}\|_\infty^{[t_0,t]} \le \bar{c}_{\mu-2}\|q\|_{ind\,\mu}^{[t_0,t]}, \end{aligned}$$

and so on, such that

$$\|v_i\|_\infty^{[t_0,t]} \le \bar{c}_i\|q\|_{ind\,\mu}^{[t_0,t]}, \quad i=\mu-3,\dots,1,$$

with certain constants $\bar{c}_i$. Then, with a suitable constant $\bar{c}$, it follows that

$$\|v\|_\infty^{[t_0,t]} \le \bar{c}\|q\|_{ind\,\mu}^{[t_0,t]}.$$

Now the representation (2.96) leads to the inequality

$$|x(t)| \le \|x\|_\infty^{[t_0,t]} \le c_1\,|\Pi_{can}(t_0)x^0| + c_2\,\|q\|_\infty^{[t_0,t]} + \|q\|_{ind\,\mu}^{[t_0,t]}, \quad t_0\le t\le\bar{t},$$

with c_1 being a bound of the fundamental solution matrix $X(t,t_0)$, $c_3 := \bar{c}$ and c_2 resulting as a bound of the term $X(t,t_0)X(s,t_0)^- G_\mu^{-1}(s)$, whereby s varies between t_0 and t. We finish the proof by letting $c := \max\{c_1, c_2+c_3\}$. □

The inequality (2.110) suggests that the DAE has so-called *perturbation index* μ (cf. [103, 105]). The concept of perturbation index interprets the index as a measure of sensitivity of the solution with respect to perturbations of the given problem. Applied to our DAE (2.44), the definition ([105, page 478]) becomes:

Definition 2.60. Equation (2.44) has *perturbation index* μ_p along a solution x_* on the interval $[t_0,\bar{t}]$, if μ_p is the smallest integer such that, for all functions $\tilde{x}$ having a defect

$$A(D\tilde{x})' + B\tilde{x} - q = \delta$$

there exists on $[t_0,\bar{t}]$ an estimate

$$|\tilde{x}(t) - x_*(t)| \le C\{|\tilde{x}(t_0) - x_*(t_0)| + \|\delta\|_\infty^{[t_0,t]} + \dots + \|\delta^{(\mu_p-1)}\|_\infty^{[t_0,t]}\},$$

whenever the expression on the right-hand side is sufficiently small.

Owing to the linearity, the DAE (2.44) has perturbation index μ_p (along each solution) on the interval $[t_0,\bar{t}]$, if for all functions $x = \tilde{x} - x_*$ having a defect $A(Dx)' + Bx = \delta$ an estimate

$$|x(t)| \le C\{|x(t_0)| + \|\delta\|_\infty^{[t_0,t]} + \cdots + \|\delta^{(\mu_p-1)}\|_\infty^{[t_0,t]}\}, \tag{2.111}$$

is valid.
The definition of the perturbation index does not specify function classes meant for the solutions and defects, but obviously one has to suppose $\delta \in \mathcal{C}^{\mu_p-1}$, such that the notion applies to sufficiently smooth problems only. In fact, the required estimate (2.111) corresponds to the inequality (2.110), which is available for smooth problems only. Therefore, we observe that a fine DAE with tractability index μ and sufficiently smooth coefficients has at the same time perturbation index μ.

All in all, the solution $x = x(x^0, q)$ of the IVP (2.108) depends on the value x^0 as well as on the function q. It is shown that x varies smoothly with x^0 such that, concerning this aspect, the DAE solutions are close to the ODE solutions. However, solutions of higher index DAEs show an ambivalent character. With respect to their variable q they are essentially ill-posed. More precisely, the linear operator $L : \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m)$ described in (2.107) has the range $\operatorname{im} L = \mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ which is a proper nonclosed subset in $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$, if $\mu \ge 2$. This makes the IVP (2.108) essentially ill-posed with respect to the excitations q. We recall of Example 1.5 which clearly shows this ill-posed character.

2.7 Specifications for regular standard form DAEs

At present, most of the literature on DAEs is devoted to standard form DAEs

$$E(t)x'(t) + F(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.112}$$

where E and F are smooth square matrix functions. Here we assume $E(t)$ to have constant rank on the given interval whereas points at which $E(t)$ change its rank are considered to be critical.
As proposed in [96], one can treat (2.112) as

$$E(t)(P(t)x(t))' + (F(t) - E(t)P'(t))x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.113}$$

by means of a continuously differentiable projector function P such that $\ker P = \ker E$. The DAE (2.113) has a properly stated leading term, and all results of the previous sections apply. In particular, we build the matrix function sequence beginning with

$$A := E,\ D := P,\ R = P,\ B := F - EP',\ G_0 = E,\ B_0 := B,$$

develop decouplings, etc. However, now the new question arises: which effects are caused by a change from one projector function P to another one? Clearly, the matrix function sequence depends on the projector function P.
Suppose P and $\tilde{P}$ to be two continuously differentiable projector functions such that

$$\ker E = \ker P = \ker \tilde{P}.$$

Besides (2.113) we consider

$$E(t)(\tilde{P}(t)x(t))' + (F(t) - E(t)\tilde{P}'(t))x(t) = q(t), \quad t \in \mathcal{I}. \tag{2.114}$$

Proposition 2.22 guarantees that the function spaces $\mathcal{C}_P^1(\mathcal{I}, \mathbb{R}^m)$ and $\mathcal{C}_{\tilde{P}}^1(\mathcal{I}, \mathbb{R}^m)$ coincide. Furthermore, the DAE (2.114) results from the DAE (2.113) by a *refactorization of the leading term*. Namely, set

$$A := E,\ D := P,\ R := P,\ B := F - EP', \quad \text{and} \quad H := \tilde{P},\ H^- := \tilde{P}.$$

Then, condition (2.27) is satisfied with $RHH^-R = P\tilde{P}P = P = R$, and the refactorized DAE (2.28) coincides with (2.114) because of (cf. (2.29))

$$\begin{aligned} \bar{A} &= AH = E\tilde{P} = E, & \bar{D} &= H^-D = \tilde{P}P = \tilde{P}, \\ \bar{B} &= B - ARH(H^-R)'D = F - EP' - E\tilde{P}'P \\ &= F - E\tilde{P}P' - E\tilde{P}'P = F - E(\tilde{P}P)' \\ &= F - E\tilde{P}'. \end{aligned}$$

In consequence, by Theorem 2.21 on refactorizations, the subspaces $\operatorname{im} G_i$, S_i, and $N_0 + \cdots + N_i$, as well as the characteristic values r_i, are independent of the special choice of P. This justifies the following regularity notion for standard form DAEs which traces the problem back to Definition 2.25 for DAEs with properly stated leading terms.

Definition 2.61. The standard form DAE (2.112) is *regular with tractability index* μ, if the properly stated version (2.113) is so for one (or, equivalently, for each) continuously differentiable projector function P with $\ker P = \ker E$.
The characteristic values of (2.113) are named *characteristic values* of (2.112).
The canonical subspaces S_{can} and N_{can} of (2.113) are called *canonical subspaces* of (2.112).

While the canonical subspaces S_{can} and N_{can} are independent of the special choice of P, the IERODE resulting from (2.113) obviously depends on P:

$$u' - (P\Pi_{\mu-1})'u + P\Pi_{\mu-1}G_\mu^{-1}Bu = P\Pi_{\mu-1}G_\mu^{-1}q, \quad u \in \operatorname{im} P\Pi_{\mu-1}. \tag{2.115}$$

This is a natural consequence of the standard formulation.

When dealing with standard form DAEs, the choice $P_0 := P$, $D^- = P$ suggests itself to begin the matrix function sequence with. In fact, this is done in the related previous work. Then the accordingly specialized sequence is

$$\begin{aligned} G_0 &= E, & B_0 &= F - EP_0' = F - G_0\Pi_0', \\ G_{i+1} &= G_i + B_iQ_i, & B_{i+1} &= B_iP_i - G_{i+1}P_0\Pi_{i+1}'\Pi_i, \quad i \geq 0. \end{aligned} \tag{2.116}$$

In this context, the projector functions $Q_0, \ldots, Q_\kappa$ are *regular admissible*, if

(a) the projector functions $G_0, \ldots, G_\kappa$ have constant ranks,
(b) the relations $Q_iQ_j = 0$ are valid for $j = 0, \ldots, i-1$, $i = 1, \ldots, \kappa$,
(c) and $\Pi_0, \ldots, \Pi_\kappa$ are continuously differentiable.

Then, it holds that $P\Pi_i = \Pi_i$, and the IERODE of a regular DAE (2.112) is

$$u' - \Pi_{\mu-1}'u + \Pi_{\mu-1}G_\mu^{-1}Bu = \Pi_{\mu-1}G_\mu^{-1}q, \quad u \in \operatorname{im}\Pi_{\mu-1}. \tag{2.117}$$

In previous papers exclusively devoted to regular DAEs, some higher smoothness is supposed for Q_i, and these projector functions are simply called admissible, without the addendum *regular*. A detailed description of the decoupling supported by the specialized matrix function (2.116) can be found in [194].

Remark 2.62. In earlier papers (e.g., [157], [159], [111], [160]) the matrix function sequence

$$G_{i+1} = G_i + B_iQ_i, \quad B_{i+1} = B_iP_i - G_{i+1}\Pi_{i+1}'\Pi_i, \quad i \geq 0, \tag{2.118}$$

is used, which is slightly different from (2.116). While [157], [159] provide solvability results and decouplings for regular index-2 and index-3 DAEs, [111] deserves attention in proving the invariance of the tractability index $\mu \in \mathbb{N}$ with respect to transformations (see also [160], but notice that, unfortunately, there is a misleading misprint in the sequence on page 158). In these earlier papers the famous role of the sum spaces $N_0 + \cdots + N_i$ was not yet discovered, so that the reasoning is less transparent and needs patient readers.
In [167, Remark 2.6] it is thought that the sequence (2.116) coincides with the sequence (2.118); however this is not fully correct. Because of

$$\begin{aligned} B_{i+1} &= B_iP_i - G_{i+1}P_0\Pi_{i+1}'\Pi_i = B_iP_i - G_{i+1}\Pi_{i+1}'\Pi_i + G_{i+1}Q_0 \underbrace{\Pi_{i+1}'}_{(P_0\Pi_{i+1})'}\Pi_i \\ &= B_iP_i - G_{i+1}\Pi_{i+1}'\Pi_i + G_{i+1}Q_0P_0'\Pi_{i+1}, \end{aligned}$$

both matrix function sequences in fact coincide, if $Q_0P_0' = 0$. One can always arrange that $Q_0P_0' = 0$ is locally valid. Namely, for each fixed $t_* \in \mathcal{I}$, we find a neighborhood $\mathcal{N}_{t_*}$ such that $\ker E(t) \oplus \ker E(t_*)^\perp = \mathbb{R}^m$ holds true for all $t \in \mathcal{N}_{t_*}$. The projector function Q_0 onto $\ker E(t)$ along $\ker E(t_*)^\perp$ has the required property

$$Q_0P_0' = Q_0(P_0(t_*)P_0)' = Q_0P_0(t_*)P_0' = 0.$$

Owing to the independence of the choice of the projector function $P_0 = P$, the regularity notions for (2.112), defined by means of (2.116) or by (2.118), are actually

consistent, and the sum subspaces, the canonical subspaces, and the characteristic values are precisely the same.
Several papers on lower index DAEs use subspace properties rather than rank conditions for the index definition. For instance, in [163], an index-2 tractable DAE is characterized by a constant-dimensional nontrivial nullspace N_1, together with the transversality condition $N_1 \oplus S_1 = \mathbb{R}^m$. Owing to Lemma A.9, this is equivalent to the condition for G_1 to have constant rank lower than m, and the requirement for G_2 to remain nonsingular.

Theorem 2.63. *Let the DAE (2.112) be regular with tractability index μ and fine. Let the matrix $C \in L(\mathbb{R}^m, \mathbb{R}^s)$ be such that* $\ker C = N_{can}(t_0)$.

(1) *Then, the IVP*
$$Ex' + Fx = 0, \quad Cx(t_0) = 0,$$
has the zero solution only.

(2) *For each admissible excitation q, and each $x^0 \in \mathbb{R}^m$, the IVP*
$$Ex' + Fx = q, \quad C(x(t_0) - x^0) = 0,$$
has exactly one solution in $\mathcal{C}_P^1(\mathcal{I}.\mathbb{R}^m)$.

(3) *For each given admissible excitation q, the set of consistent initial values at time t_0 is*
$$\mathcal{M}_{can,q}(t_0) = \{z + v(t_0) : z \in S_{can}(t_0)\},$$
whereby v is constructed as in (2.95) by means of fine decoupling projector functions.

(4) *If the coefficients of the DAE are sufficiently smooth, then each $q \in \mathcal{C}^{\mu-1}(\mathcal{I}, \mathbb{R}^m)$ is admissible. If the interval $\mathcal{I}$ is compact, then for the IVP solution from (2), the inequality*
$$\|x\| \le K\left(|\Pi_{can}(t_0)x^0| + \|q\|_\infty + \sum_{l=1}^{\mu-1} \|q^{(l)}\|_\infty\right) \tag{2.119}$$
is valid with a constant K independent of q and x^0.

Proof. (1) and (2) are consequences of Theorem 2.44(2) and Theorem 2.52(1), respectively. Assertion (4) follows from Theorem 2.59(3). Assertion (3) results from the representations (2.95) and (2.98), with $D = D^- = P$. □

The inequality (2.119) indicates that the DAE has *perturbation index* μ (cf. Definition 2.60).

2.8 The T-canonical form

Definition 2.64. The structured continuous coefficient DAE with properly stated leading term

$$\left[\begin{array}{c|ccc} I_d & & & \\ \hline & 0 & \tilde{\mathcal{N}}_{0,1} & \cdots & \tilde{\mathcal{N}}_{0,\mu-1} \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & \tilde{\mathcal{N}}_{\mu-2,\mu-1} \\ & & & & 0 \end{array}\right] \left(\left[\begin{array}{c|ccc} I_d & & & \\ \hline & 0 & & & \\ & & I_{m-r_1} & & \\ & & & \ddots & \\ & & & & I_{m-r_{\mu-1}} \end{array}\right] \tilde{x} \right)' \tag{2.120}$$

$$+ \left[\begin{array}{c|ccc} \tilde{\mathcal{W}} & & & \\ \hline \tilde{\mathcal{H}}_0 & I_{m-r_0} & & \\ \vdots & & \ddots & \\ \vdots & & & \ddots \\ \tilde{\mathcal{H}}_{\mu-1} & & & I_{m-r_{\mu-1}} \end{array}\right] \tilde{x} = \tilde{q},$$

$m = d + \sum_{j=0}^{\mu-1} (m - r_j)$, as well as its counterpart in standard form

$$\begin{bmatrix} I_d & 0 \\ 0 & \tilde{\mathcal{N}} \end{bmatrix} \tilde{x}' + \begin{bmatrix} \tilde{\mathcal{W}} & 0 \\ \tilde{\mathcal{H}} & I_{m-d} \end{bmatrix} \tilde{x} = \tilde{q}, \tag{2.121}$$

with

$$\tilde{\mathcal{N}} = \begin{bmatrix} 0 & \tilde{\mathcal{N}}_{0,1} & \cdots & \tilde{\mathcal{N}}_{0,\mu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \tilde{\mathcal{N}}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix},$$

are said to be in *T(ractability)-canonical form*, if the entries $\tilde{\mathcal{N}}_{0,1}, \ldots, \tilde{\mathcal{N}}_{\mu-2,\mu-1}$ are full column rank matrix functions, that is $\operatorname{rank} \mathcal{N}_{i-1,i} = m - r_i$, for $i = 1, \ldots, \mu - 1$.

The subscript μ indicates the tractability index μ, and at the same time the uniform nilpotency index of the upper block triangular matrix function $\tilde{\mathcal{N}}$. $\tilde{\mathcal{N}}^{\mu}$ vanishes identically, and $\tilde{\mathcal{N}}^{\mu-1}$ has the only nontrivial entry $\tilde{\mathcal{N}}_{0,1}\tilde{\mathcal{N}}_{1,2}\cdots\tilde{\mathcal{N}}_{\mu-2,\mu-1}$ of rank $m - r_{\mu-1}$ in the upper right corner. If the coefficients $\tilde{\mathcal{H}}_0, \ldots, \tilde{\mathcal{H}}_{\mu-1}$ vanish, the T-canonical form (2.121) looks precisely like the Weierstraß–Kronecker canonical form for constant matrix pencils.

Generalizing Proposition 1.28, we show that a DAE (2.44) is regular with tractability index μ if and only if it can be brought into T-canonical form by a regular multiplication, a regular transformations of the unknown function, and a refactorization of the leading term as described in Section 2.3. This justifies the attribute *canonical*. The structural sizes $r_0, \ldots, r_{\mu-1}$ coincide with the characteristic values from the tractability index framework.

Theorem 2.65. (1) *The DAE (2.44) is regular with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, if and only if there are pointwise*

regular matrix functions $L, K \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$, and a constant-rank refactorization matrix function $H \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^s, \mathbb{R}^n))$, $RHH^-R = R$, such that premultiplication by L, the transformation $x = K\tilde{x}$, and the refactorization of the leading term by H yield a DAE in T-canonical form, whereby the entry $\tilde{\mathcal{N}}_{i-1,i}$ has size $(m - r_{i-1}) \times (m - r_i)$ and

$$\operatorname{rank} \tilde{\mathcal{N}}_{i-1,i} = m - r_i, \quad \text{for} \quad i = 1, \ldots, \mu - 1.$$

(2) *If the DAE (2.44) is regular with tractability index μ, and its coefficients are smooth enough for the existence of completely decoupling projector functions, then the DAE is equivalent to a T-canonical form with zero coupling coefficients $\tilde{\mathcal{H}}_0, \ldots, \tilde{\mathcal{H}}_{\mu-1}$.*

Proof. (1) If the DAE has T-canonical form, one can construct a matrix function sequence and admissible projector functions in the same way as described in Subsection 1.2.6 for constant matrix pencils, and this shows regularity and confirms the characteristic values.
The reverse implication is more difficult. Let the DAE (2.44) be regular with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$. Let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions. As explained in Subsection 2.4.2, the DAE decomposes into equation (2.49) being a pre-version of the IERODE and subsystem (2.63) together

$$\underbrace{\begin{bmatrix} D\Pi_{\mu-1}D^- & 0 \\ 0 & \mathcal{N} \end{bmatrix}}_{\mathfrak{A}} \Big(\underbrace{\begin{bmatrix} D\Pi_{\mu-1}D^- & 0 \\ 0 & \mathcal{D} \end{bmatrix}}_{\mathfrak{D}} \begin{bmatrix} u \\ v \end{bmatrix} \Big)' + \underbrace{\begin{bmatrix} \mathcal{W} & 0 \\ \mathcal{H}D^- & \mathcal{M} \end{bmatrix}}_{\mathfrak{B}} \begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} \mathcal{L}_d \\ \mathcal{L} \end{bmatrix} q. \quad (2.122)$$

This is an inflated system in $\mathbb{R}^{m(\mu+1)}$, with $\mathcal{W} := D\Pi_{\mu-1}G_\mu^{-1}BD^-$, further coefficients given in Subsection 2.4.2, and the unknown functions

$$\begin{bmatrix} u \\ v \end{bmatrix} := \begin{bmatrix} u \\ v_0 \\ \vdots \\ v_{\mu-1} \end{bmatrix} := \begin{bmatrix} D\Pi_{\mu-1} \\ Q_0 \\ \Pi_0 Q_1 \\ \vdots \\ \Pi_{\mu-2}Q_{\mu-1} \end{bmatrix} x.$$

We condense this inflated system back to $\mathbb{R}^m$ in a similar way as in Proposition 1.28. The projector functions $D\Pi_{\mu-1}D^-$ and $D\Pi_{i-1}Q_iD^-$ are continuously differentiable, and so are their ranges and nullspaces. The $\mathcal{C}^1$-subspace $\operatorname{im}(D\Pi_{\mu-1}D^-)^*$ has dimension $d = m - \sum_{i=0}^{\mu-1}(m - r_i)$, and it is spanned by continuously differentiable basis functions, which means that there is a matrix function $\Gamma_d \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^d))$ such that

$$\operatorname{im}(D\Pi_{\mu-1}D^-)^* = \operatorname{im}\Gamma_d^*, \quad \ker\Gamma_d^* = \{0\},$$

and hence

$$\operatorname{im} \Gamma_d = \mathbb{R}^d, \quad \ker \Gamma_d = (\operatorname{im}(D\Pi_{\mu-1}D^-)^*)^\perp = \ker D\Pi_{\mu-1}D^-.$$

By Proposition A.17, there is a pointwise reflexive generalized inverse $\Gamma_d^- \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^d, \mathbb{R}^n))$ such that $\Gamma_d\Gamma_d^- = I_d$ and $\Gamma_d^-\Gamma_d = D\Pi_{\mu-1}D^-$. Analogously we find $\Gamma_i \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^{m-r_i}))$ and $\Gamma_i^- \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^{m-r_i}, \mathbb{R}^n))$ such that for $i = 1, \ldots, \mu-1$

$$\operatorname{im} \Gamma_i = \mathbb{R}^{m-r_i}, \quad \ker \Gamma_i = \ker D\Pi_{i-1}Q_iD^-, \quad \Gamma_i\Gamma_i^- = I_{m-r_i}, \quad \Gamma_i^-\Gamma_i = D\Pi_{i-1}Q_iD^-.$$

This implies

$$\Gamma_i D = \Gamma_i D\Pi_{i-1}Q_i, \quad D^-\Gamma_i^- = \Pi_{i-1}Q_iD^-\Gamma_i^-, \quad \Gamma_i DD^-\Gamma_i^- = \Gamma_i\Gamma_i^- = I_{m-r_i}.$$

Finally we provide $\Gamma_0 \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^{m-r_0}))$ and $\Gamma_0^- \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^{m-r_0}, \mathbb{R}^m))$ such that

$$\operatorname{im} \Gamma_0 = \mathbb{R}^{m-r_0}, \quad \ker \Gamma_0 = \ker Q_0, \quad \Gamma_0\Gamma_0^- = I_{m-r_0}, \quad \Gamma_0^-\Gamma_0 = Q_0.$$

Then we compose

$$\Gamma := \begin{bmatrix} \Gamma_d & \\ & \Gamma_{sub} \end{bmatrix}, \quad \Gamma^- := \begin{bmatrix} \Gamma_d^- & \\ & \Gamma_{sub}^- \end{bmatrix},$$

$$\Gamma_{sub} := \begin{bmatrix} \Gamma_0 & & & \\ & \Gamma_1 D & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1}D \end{bmatrix}, \quad \Gamma_{sub}^- := \begin{bmatrix} \Gamma_0^- & & & \\ & D^-\Gamma_1^- & & \\ & & \ddots & \\ & & & D^-\Gamma_{\mu-1}^- \end{bmatrix}$$

such that $\Gamma\Gamma^- = I_m$, $\Gamma_{sub}\Gamma_{sub}^- = I_{m-d}$, and

$$\Gamma^-\Gamma = \begin{bmatrix} D\Pi_{\mu-1}D^- & & & & \\ & Q_0 & & & \\ & & \Pi_0Q_1 & & \\ & & & \ddots & \\ & & & & \Pi_{\mu-2}Q_{\mu-1} \end{bmatrix},$$

$$\Gamma_{sub}^-\Gamma_{sub} = \begin{bmatrix} Q_0 & & & \\ & \Pi_0Q_1 & & \\ & & \ddots & \\ & & & \Pi_{\mu-2}Q_{\mu-1} \end{bmatrix}.$$

Additionally we introduce

$$\Omega := \begin{bmatrix} 0 & & & \\ & \Gamma_1 & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1} \end{bmatrix}, \quad \Omega^- := \begin{bmatrix} 0 & & & \\ & \Gamma_1^- & & \\ & & \ddots & \\ & & & \Gamma_{\mu-1}^- \end{bmatrix},$$

such that

$$\Omega^-\Omega = \begin{bmatrix} 0 & & & \\ & D\Pi_0 Q_1 D^- & & \\ & & \ddots & \\ & & & D\Pi_{\mu-2}Q_{\mu-1}D^- \end{bmatrix}, \quad \Omega\Omega^- = \begin{bmatrix} 0 & & & \\ & I_{m-r_1} & & \\ & & \ddots & \\ & & & I_{m-r_{\mu-1}} \end{bmatrix}.$$

For the coefficients of the inflated system (2.122) it follows that

$$\Gamma_{sub}^-\Gamma_{sub}\,\mathcal{N} = \mathcal{N}\Omega^-\Omega = \mathcal{N}, \quad \Gamma_{sub}^-\Gamma_{sub}\mathcal{M} = \mathcal{M}\Gamma_{sub}^-\Gamma_{sub}, \quad \mathcal{D} = \Omega^-\Gamma_{sub},$$

and further

$$\Gamma\mathfrak{A} = \begin{bmatrix} \Gamma_d D\Pi_{\mu-1}D^- & \\ & \Gamma_{sub}\mathcal{N} \end{bmatrix} = \begin{bmatrix} \Gamma_d & \\ & \Gamma_{sub}\mathcal{N}\Omega^-\Omega \end{bmatrix} = \begin{bmatrix} I_d & \\ & \Gamma_{sub}\mathcal{N}\Omega^- \end{bmatrix} \begin{bmatrix} \Gamma_d & \\ & \Omega \end{bmatrix},$$

$$\Gamma\mathfrak{B} = \begin{bmatrix} \Gamma_d\mathcal{W} & 0 \\ \Gamma_{sub}\mathcal{H}D^- & \Gamma_{sub}\mathcal{M} \end{bmatrix} = \begin{bmatrix} \Gamma_d\mathcal{W}\Gamma_d^-\Gamma_d & 0 \\ \Gamma_{sub}\mathcal{H}D^-\Gamma_d^-\Gamma_d & \Gamma_{sub}\mathcal{M}\Gamma_{sub}^-\Gamma_{sub} \end{bmatrix}$$

$$= \begin{bmatrix} \Gamma_d\mathcal{W}\Gamma_d^- & 0 \\ \Gamma_{sub}\mathcal{H}D^-\Gamma^- & \Gamma_{sub}\mathcal{M}\Gamma_{sub} \end{bmatrix} \begin{bmatrix} \Gamma_d & 0 \\ 0 & \Gamma_{sub} \end{bmatrix},$$

$$\mathfrak{D} = \begin{bmatrix} \Gamma_d^-\Gamma_d & 0 \\ 0 & \Omega^-\Gamma_{sub} \end{bmatrix} = \begin{bmatrix} \Gamma_d^- & 0 \\ 0 & \Omega^- \end{bmatrix} \begin{bmatrix} \Gamma_d & 0 \\ 0 & \Gamma_{sub} \end{bmatrix}.$$

Multiplying the inflated system (2.122) by the condensing matrix function Γ and introducing the new variables

$$\tilde{x} := \begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} := \begin{bmatrix} \Gamma_d & 0 \\ 0 & \Gamma_{sub} \end{bmatrix} \begin{bmatrix} u \\ v \end{bmatrix}$$

gives

$$\underbrace{\begin{bmatrix} I & 0 \\ 0 & \Gamma_{sub}\mathcal{N}\Omega^- \end{bmatrix}\begin{bmatrix} \Gamma_d & 0 \\ 0 & \Omega \end{bmatrix}}_{\bar{A}}\left(\underbrace{\begin{bmatrix} \Gamma_d^- & 0 \\ 0 & \Omega^- \end{bmatrix}}_{\bar{D}}\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix}\right)'$$
$$+\underbrace{\begin{bmatrix} \Gamma_d\mathcal{W}\Gamma_d^- & 0 \\ \Gamma_{sub}\mathcal{H}D^-\Gamma_d^- & \Gamma_{sub}\mathcal{M}\Gamma_{sub}^- \end{bmatrix}}_{\bar{B}}\begin{bmatrix} \tilde{u} \\ \tilde{v} \end{bmatrix} = \underbrace{\Gamma\begin{bmatrix} \mathcal{L}_d \\ \mathcal{L} \end{bmatrix}}_{\bar{L}}q.$$

This last DAE lives in $\mathbb{R}^m$, but the border space of its leading term is $\mathbb{R}^{n(\mu+1)}$. Because of

$$\ker\bar{A} = \ker\begin{bmatrix} \Gamma_d & 0 \\ 0 & \Omega \end{bmatrix} = \ker\bar{R}, \quad \operatorname{im}\bar{D} = \operatorname{im}\bar{R},$$

with the border projector $\bar{R} = \begin{bmatrix} D\Pi_{\mu-1}D^- & 0 \\ 0 & \Omega^-\Omega \end{bmatrix}$ the refactorization of the leading term (cf. Section 2.3) by means of

$$H := \begin{bmatrix} \Gamma_d^- & 0 \\ 0 & \Omega^- \end{bmatrix}, \quad H^- = \begin{bmatrix} \Gamma_d & 0 \\ 0 & \Omega \end{bmatrix}$$

suggests itself. H has constant rank d, and H^- is the reflexive generalized inverse with

$$H^-H = \begin{bmatrix} I_d & 0 \\ 0 & \Omega\Omega^- \end{bmatrix},\ HH^- = \begin{bmatrix} D\Pi_{\mu-1}D^- & 0 \\ 0 & \Omega^-\Omega \end{bmatrix} = \bar{R},\ \bar{R}HH^-\bar{R} = \bar{R}.$$

This way we arrive at the DAE

$$\tilde{A}(\tilde{D}\tilde{x})' + \tilde{B}\tilde{x} = \bar{L}q,$$
$$\tilde{A} := \begin{bmatrix} I & 0 \\ 0 & \Gamma_{sub}\mathcal{N}\Omega^- \end{bmatrix}, \tilde{D} := \begin{bmatrix} I & 0 \\ 0 & \Omega\Omega^- \end{bmatrix}, \tilde{B} := \begin{bmatrix} \Gamma_d\mathcal{W}\Gamma_d^- - \Gamma_d'\Gamma_d^- & 0 \\ \Gamma_{sub}\mathcal{H}D^-\Gamma_d^- & \tilde{B}_{22} \end{bmatrix}.$$

The entry

$$\begin{aligned}\tilde{B}_{22} &:= \Gamma_{sub}\mathcal{M}\Gamma_{sub}^- - \Gamma_{sub}\mathcal{N}\Omega^-\Omega'\Omega^- \\ &= \Gamma_{sub}\Gamma_{sub}^- + \Gamma_{sub}(\mathcal{M}-I)\Gamma_{sub} - \Gamma_{sub}\mathcal{N}\Omega^-\Omega'\Omega^- =: I + \tilde{\mathcal{M}}\end{aligned}$$

has upper block triangular form, with identity diagonal blocks. $\tilde{\mathcal{M}}$ is strictly upper block triangular, and $I+\tilde{\mathcal{M}}$ remains nonsingular. Scaling the DAE by $\operatorname{diag}(I, (I+\tilde{\mathcal{M}})^{-1})$ yields

$$\begin{bmatrix} I & 0 \\ 0 & \tilde{\mathcal{N}} \end{bmatrix}\left(\begin{bmatrix} I & 0 \\ 0 & \Omega\Omega^- \end{bmatrix}\tilde{x}\right)' + \begin{bmatrix} \tilde{\mathcal{W}} & 0 \\ \tilde{\mathcal{H}} & I \end{bmatrix}\tilde{x} = \begin{bmatrix} I & 0 \\ 0 & (I+\tilde{\mathcal{M}})^{-1} \end{bmatrix}\bar{L}q, \tag{2.123}$$

with coefficients

$$\tilde{\mathcal{N}} := (I+\tilde{\mathcal{M}})^{-1}\Gamma_{sub}\mathcal{N}\Omega^{-}, \quad \tilde{\mathcal{H}} := (I+\tilde{\mathcal{M}})^{-1}\Gamma_{sub}\mathcal{H}D^{-}\Gamma_d^{-},$$
$$\tilde{\mathcal{W}} := \Gamma_d\mathcal{W}\Gamma_d^{-} - \Gamma_d'\Gamma_d^{-}.$$

The DAE (2.123) has T-canonical form, if the entries $\tilde{\mathcal{N}}_{i,i+1}$ have full column rank. Therefore, we take a closer look at these entries. Having in mind that $\tilde{\mathcal{M}}$ is strictly upper block triangular, we derive

$$\begin{aligned}\tilde{\mathcal{N}}_{i,i+1} &= (\Gamma_{sub}\mathcal{N}\Omega)_{i,i+1} = \Gamma_i D\mathcal{N}_{i,i+1}\Gamma_{i+1}^{-} = -\Gamma_i D\Pi_{i-1}Q_iQ_{i+1}D^{-}\Gamma_{i+1}^{-}\\ &= -\Gamma_i\Gamma_i^{-}\Gamma_i DQ_{i+1}D^{-}\Gamma_{i+1}^{-} = -\Gamma_i DQ_{i+1}D^{-}\Gamma_{i+1}^{-}.\end{aligned}$$

Then, $\tilde{\mathcal{N}}_{i,i+1}z = 0$ means $\Gamma_i D\mathcal{N}_{i,i+1}\Gamma_{i+1}^{-}z = 0$, thus $\mathcal{N}_{i,i+1}\Gamma_{i+1}^{-}z = 0$. Applying Proposition 2.29 (3) we find that $D\Pi_iQ_{i+1}D^{-}\Gamma_{i+1}^{-}z = \Gamma_{i+1}^{-}z \in \ker D\Pi_iQ_{i+1}D^{-}$, and hence $\Gamma_{i+1}^{-}z = 0$, therefore $z = 0$. This shows that $\tilde{\mathcal{N}}_{i,i+1}$ is injective for $i = 1,\ldots,\mu-2$. The injectivity of $\tilde{\mathcal{N}}_{0,1}$ follows analogously. We obtain in fact a T-canonical form. The resulting transformations are

$$L = \begin{bmatrix} I & 0 \\ 0 & (I+\tilde{\mathcal{M}})^{-1}\end{bmatrix}\Gamma\begin{bmatrix}\mathcal{L}_d \\ \mathcal{L}\end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & (I+\tilde{\mathcal{M}})^{-1}\end{bmatrix}\begin{bmatrix} \Gamma_d D\Pi_{\mu-1} \\ \Gamma_0Q_0 \\ \Gamma_1D\Pi_0Q_1 \\ \vdots \\ \Gamma_{\mu-1}D\Pi_{\mu-2}Q_{\mu-1}\end{bmatrix}G_\mu^{-1}$$

and

$$K = \Gamma\begin{bmatrix} D\Pi_{\mu-1} \\ Q_0 \\ \Pi_0Q_1 \\ \vdots \\ \Pi_{\mu-2}Q_{\mu-1}\end{bmatrix} = \begin{bmatrix} \Gamma_d D\Pi_{\mu-1} \\ \Gamma_0Q_0 \\ \Gamma_1D\Pi_0Q_1 \\ \vdots \\ \Gamma_{\mu-1}D\Pi_{\mu-2}Q_{\mu-1}\end{bmatrix}.$$

Both matrix functions K and L are continuous and pointwise nonsingular. This completes the proof of (1).
The assertion (2) now follows immediately, since $\mathcal{H} = 0$ implies $\tilde{\mathcal{H}} = 0$. □

2.9 Regularity intervals and critical points

Critical points *per se* attract much special interest and effort. In particular, to find out whether the ODE with a so-called singularity of the first kind (e.g. [123])

$$x'(t) = \frac{1}{t}M(t)x(t) + q(t),$$

has bounded solutions, standard ODE theory is of no avail, and one is in need of smarter tools using the eigenstructure of the matrix $M(0)$.

In the case of DAEs, the inherent ODE might be affected by singularities. For instance, the DAEs in [124] show inherent ODEs having a singularity of the first kind. The following example is taken from [124].

Example 2.66 (Rank drop in G_1 causes a singular inherent ODE). The DAE

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

has a properly stated leading term on $[0,1]$. It is accompanied by the matrix functions

$$G_0(t) = \begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix}, \quad Q_0(t) = \frac{1}{2}\begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix}, \quad G_1(t) = \begin{bmatrix}2 & 0\\ 2+\frac{t}{2} & \frac{t}{2}\end{bmatrix},$$

such that the DAE is regular with tractability index 1 just on the interval $(0,1]$. The inherent ODE resulting there applies to $u(t) = x_1(t) - x_2(t)$, and it reads

$$u'(t) = -\frac{2}{t}(t+2)u(t) + \frac{1}{t}((t+2)q_1(t) - 2q_2(t)).$$

Observe that, in view of the closed interval $[0,1]$, this is no longer a regular ODE but an inherent explicit *singular* ODE (IESODE). Given a solution $u(\cdot)$ of the IESODE, a DAE solution is formed by

$$x(t) = \frac{1}{t}\begin{bmatrix}t+2\\ 2\end{bmatrix}u(t) + \frac{1}{t}\begin{bmatrix}-q_1(t)+q_2(t)\\ -q_1(t)+q_2(t)\end{bmatrix}.$$

We refer to [124] for the specification of bounded solutions by means of boundary conditions as well as for collocation approximations. □

One could presume that rank changes in G_1 would always lead to singular inherent ODEs, but the situation is much more intricate. A rank drop of the matrix function G_1 is not necessarily accompanied by a singular inherent ODE, as the next example shows.

Example 2.67 (Rank drop in G_1 does not necessarily cause a singular inherent ODE). The DAE

$$\begin{bmatrix}1\\0\end{bmatrix}\left(\begin{bmatrix}t & 1\end{bmatrix}x(t)\right)' + \begin{bmatrix}\beta(t) & 0\\ 0 & 1\end{bmatrix}x(t) = q(t),$$

with an arbitrary continuous real function β, has a properly stated leading term on $(-\infty,\infty)$. Put

$$G_0(t) = \begin{bmatrix}t & 1\\ 0 & 0\end{bmatrix}, \quad D(t)^- = \frac{1}{1+t^2}\begin{bmatrix}t\\ 1\end{bmatrix}, \quad Q_0(t) = \frac{1}{1+t^2}\begin{bmatrix}1 & -t\\ -t & t^2\end{bmatrix},$$

and compute

$$G_1(t) = \frac{1}{1+t^2}\begin{bmatrix}\beta(t)+t+t^3 & 1+t^2-t\beta(t)\\ -t & t^2\end{bmatrix}, \quad \omega_1(t) := \det G_1(t) = t(1+t^2).$$

This DAE is regular with index 1 on the intervals $(-\infty,0)$ and $(0,\infty)$. The point $t_* = 0$ is a critical one. The inherent ODE reads, with $u(t) = tx_1(t) + x_2(t)$,

$$u'(t) = -\frac{\beta(t)}{t}u(t) + q_1(t) + \frac{\beta(t)}{t}q_2(t).$$

All DAE solutions have the form

$$x(t) = \frac{1}{t}\begin{bmatrix}1\\0\end{bmatrix}u(t) + \frac{1}{t}\begin{bmatrix}-q_2(t)\\ tq_2(t)\end{bmatrix}.$$

Obviously, if the function β has a zero at $t_* = 0$, or if it actually vanishes identically, then there is no singularity within the inherent ODE, even though the matrix $G_1(t_*)$ becomes singular. Remember that the determinant ω_1 does not at all depend on the coefficient β.
We turn to a special case. Set q identically zero, $\beta(t) = t^\gamma$, with an integer $\gamma \geq 0$. The inherent ODE simplifies to

$$u'(t) = -t^{\gamma-1}u(t).$$

If $\gamma = 0$, this is a singular ODE, and its solutions have the form $u(t) = \frac{1}{t}c$. All nontrivial solutions grow unboundedly, if t approaches zero. In contrast, if $\gamma \geq 1$, the ODE is regular, and it has the solutions $u(t) = e^{-\frac{1}{\gamma}t^\gamma}u(0)$ which remain bounded. However, among the resulting nontrivial DAE solutions

$$x(t) = \frac{1}{t}\begin{bmatrix}1\\0\end{bmatrix}u(t)$$

there is no bounded one, even if $\gamma \geq 1$. □

As adumbrated by the above example, apart from the singularities concerning the inherent ODE, DAEs involve further sources of critical points which are unacquainted at all in explicit ODEs. In DAEs, not only the inherent ODE but also the associated subsystem (2.62) which constitutes the wrapping up, and which in higher index cases includes the differentiated parts, might be hit by singularities. In the previous two examples which show DAEs being almost overall index 1, a look at the solution representations supports this idea. The next example provides a first impression of a higher index case.

Example 2.68 (Rank drop in G_2). The DAE with properly stated leading term

$$\begin{bmatrix}1&0\\0&1\\0&0\end{bmatrix}\left(\begin{bmatrix}1&0&0\\0&1&0\end{bmatrix}x(t)\right)' + \begin{bmatrix}0&0&\beta(t)\\1&1&0\\ \gamma(t)&0&0\end{bmatrix}x(t) = q(t)$$

yields

$$G_0(t)=\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\end{bmatrix},\quad Q_0(t)=\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix},\quad G_1(t)=\begin{bmatrix}1&0&\beta(t)\\0&1&0\\0&0&0\end{bmatrix},\quad \Pi_0(t)=\begin{bmatrix}1&0&0\\0&1&0\\0&0&0\end{bmatrix},$$

and further $\widehat{N}_1(t)=N_1(t)\cap N_0(t)=\{z\in\mathbb{R}^3 : z_1=0, z_2=0, \beta(t)z_3=0\}$. Supposing $\beta(t)\neq 0$, for all t, we derive

$$Q_1(t)=\begin{bmatrix}1&0&0\\0&0&0\\-\frac{1}{\beta(t)}&0&0\end{bmatrix},\quad \Pi_0(t)Q_1(t)=\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix},\quad G_2(t)=\begin{bmatrix}1&0&\beta(t)\\1&1&0\\\gamma(t)&0&0\end{bmatrix},$$

and $\omega_2(t):=\det G_2(t)=-\beta(t)\gamma(t)$. The projector functions Q_0, Q_1 are the widely orthogonal ones. Taking a look at the following equivalent formulation of the DAE,

$$\begin{aligned}x_1(t)&=\frac{1}{\gamma(t)}q_3(t),\\ x_2'(t)+x_2(t)&=q_2(t)-\frac{1}{\gamma(t)}q_3(t),\\ x_3(t)&=\frac{1}{\beta(t)}(q_1(t)-(\frac{1}{\gamma(t)}q_3(t))'),\end{aligned}$$

we see the correspondence of zeros of the function γ to rank drops in G_2, and to critical solution behavior.

Observe also that if we dispense with the demand that the function β has no zeros, and allow a zero at a certain point t_*, then the intersection $\widehat{N}_1(t_*)$ is nontrivial, $\widehat{N}_1(t_*)=N_0(t_*)$, and the above projector function $Q_1(t)$ grows unboundedly, if t approaches t_*. Nevertheless, since by construction G_2 depends just on the product $\Pi_1 Q_2$, we can continue forming the next matrix function G_2 considering the product $\Pi_0 Q_1$ that has a continuous extension. Then a zero of the function β also leads to a zero of $\det G_2$.

Apart from critical points, the resulting IERODE applies to

$$u=D\Pi_1 x=\begin{bmatrix}0\\x_2\end{bmatrix},$$

and it reads

$$u'+\underbrace{\begin{bmatrix}0&0\\0&1\end{bmatrix}}_{D\Pi_1 G_2^{-1}B_1 D^-}u=\underbrace{\begin{bmatrix}0\\q_2-\frac{1}{\gamma}q_3\end{bmatrix}}_{D\Pi_1 G_2^{-1}q}.$$

Observe the coefficient $D\Pi_1 G_2^{-1}BD^-$ to be independent of the functions β and γ, while $D\Pi_1 G_2^{-1}$ does not depend on β. Therefore, the IERODE does not at all suffer from zeros of β.

Notice that, if one restricts interest to homogeneous DAEs only, then one cannot see the singular solution behavior in this example. □

Next we consider DAEs which fail to have a proper leading term on the entire given interval.

Example 2.69 (Rank drop in A causes a singular inherent ODE). Consider the DAE

$$\underbrace{\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}}_{A} \Big(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}}_{D} x \Big)' + \underbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}}_{B} x = q, \tag{2.124}$$

given on the interval $\mathcal{I} = [-1,1]$. The function α is continuous. Let b_{21} have no zeros. This system yields

$$\begin{aligned} x_1 &= \frac{1}{b_{21}}(q_2 - b_{22}x_2) \\ \alpha x_2' &= \underbrace{(\frac{b_{11}}{b_{21}}b_{11}b_{22} - b_{12})}_{=M} x_2 + q_1 - b_{11}q_2. \end{aligned}$$

For any $t_* \in \mathcal{I}$ with $\alpha(t_*) \neq 0$, there is an interval $\mathcal{I}_*$ around t_* such that α has no zeros on $\mathcal{I}_*$, and the DAE has a proper leading term there. On intervals where the function α has no zeros, one can write the ODE for x_2 in explicit form as

$$x_2' = \frac{1}{\alpha} M x_2 + \frac{1}{\alpha}(q_1 - b_{11}q_2). \tag{2.125}$$

Then, equation (2.125) is a well-defined continuous coefficient ODE on this interval so that standard solvability arguments apply.
In contrast, if $\alpha(t_*) = 0$, but $\alpha(t) \neq 0$, for $t \in \mathcal{I}$, $t \neq t_*$, then equation (2.125) becomes a singular ODE, more precisely, an explicit ODE with a singularity at t_*. For those kinds of equations special treatment is advisable. We have to expect a singular flow behavior of the component $x_2(t)$. The component $x_1(t)$ may inherit properties of $x_2(t)$ depending on the coefficient function b_{22}. Let us glance over typical situations.
Let $t_* = 0$, M be a constant function, $\alpha(t) = t$, $q(t) = 0$ on $\mathcal{I}$. Then the solutions of the singular ODE (2.125) are $x_2(t) = ct^M$, with a real constant c. The behavior of these solutions heavily depends on the sign of M. Figure 2.1 shows the different flow behavior of the component $x_2(t)$ in the cases $M = 2$, $M = -2$, $M = 0$, respectively:
If $M = 2$, then all solutions cross the origin, while no solution satisfies an initial condition $x_2(0) \neq 0$.
If $M = -2$, just the trivial solution passes the origin, and all other solutions grow unboundedly if t tends to zero. Again, there is no solution with $x_2(0) \neq 0$.
If $M = 0$, then every constant function solves the ODE, and every IVP is uniquely solvable. Derive, for $t \in [-1,0)$ and $t \in (0,1]$,

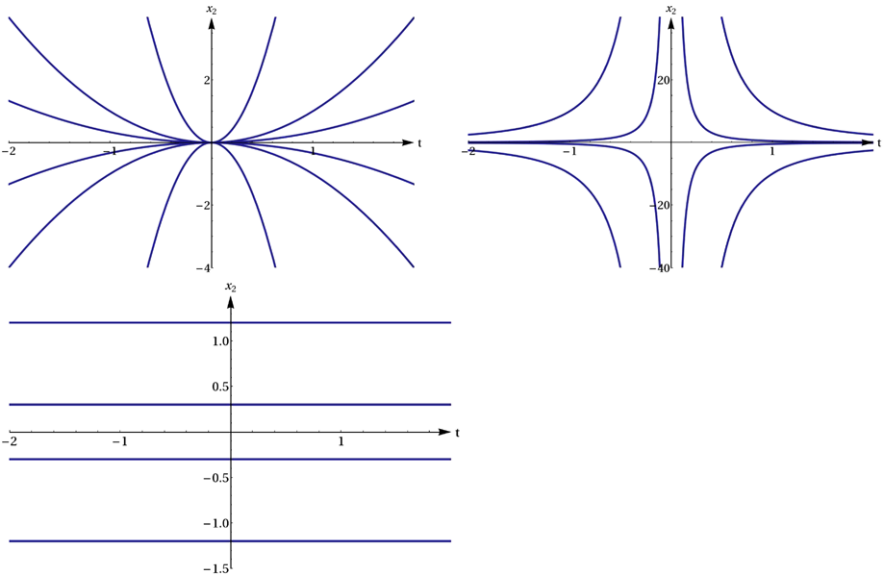


Fig. 2.1 $M=2$, $M=-2$, $M=0$

$$G_0(t)=\begin{bmatrix}0 & t\\ 0 & 0\end{bmatrix},\ Q_0(t)=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix},\ G_1(t)=\begin{bmatrix}b_{11}(t) & t\\ b_{21}(t) & 0\end{bmatrix}.$$

This shows the DAE to be regular with index 1 on both intervals $[-1,0)$ and $(0,1]$. □

Example 2.70 (Rank change in A causes an index change). Now we put the continuous entry (cf. Figure 2.2)

$$\alpha(t)=\begin{cases}0 & \text{for } t\in[-1,0]\\ t^{\frac{1}{3}} & \text{for } t\in(0,1]\end{cases}$$

into the DAE

$$\begin{bmatrix}0 & \alpha\\ 0 & 0\end{bmatrix}\left(\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}x\right)'+\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}x=q, \tag{2.126}$$

which has a properly stated leading term merely on the subinterval $(0,1]$.
The admissible matrix function sequence

$$G_0=\begin{bmatrix}0 & \alpha\\ 0 & 0\end{bmatrix},\ Q_0=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix},\ G_1=\begin{bmatrix}1 & \alpha\\ 0 & 0\end{bmatrix},\ Q_1=\begin{bmatrix}0 & -\alpha\\ 0 & 1\end{bmatrix},\ G_1=\begin{bmatrix}1 & \alpha\\ 0 & 1\end{bmatrix},$$

indicates the characteristic values $r_0=1, r_1=1$ and $r_2=2$ on $(0,1]$. The DAE is regular with index 2 there.
For every $q_1\in\mathcal{C}((0,1],\mathbb{R})$, $q_2\in\mathcal{C}^1((0,1],\mathbb{R})$, there is a unique solution $x\in\mathcal{C}_D^1((0,1],\mathbb{R}^2)$. The particular solution corresponding to $q_1(t)=0$, $q_2(t)=t^{\frac{1}{3}}$,

$t \in (0,1]$, reads $x_1(t) = -\frac{1}{3}t^{-\frac{1}{3}}$, $x_2(t) = t^{\frac{1}{3}}$.
On the subinterval $[-1,0]$ the leading term is no longer properly stated, but we may turn to a proper reformulation if we replace D by $\bar{D} = 0$. Then, for $t \in [-1,0]$, it follows that

$$G_0(t) = 0,\ Q_0(t) = I,\ G_1(t) = I,\ r_0 = 0,\ r_1 = 2,$$

and the DAE is regular with index 1 on the interval $[-1,0]$. On this subinterval, for every continuous q, the solution is simply $x = q$. In particular, for $q_1(t) = 0$, $q_2(t) = -|t|^{\frac{1}{3}}$, the solution on this subinterval is $x_1(t) = 0$, $x_2(t) = -|t|^{\frac{1}{3}}$.
Altogether, combining now the segments, we have an excitation q that is continuous on the entire interval $\mathcal{I}$, and its second component is continuously differentiable on $(0,1]$. We have two solution segments. Can these segments be glued together to form a solution on the entire interval? While the second component has a continuous extension, the first has not, as shown in Figure 2.3.
Relaxing the minimal smoothness requirements for the excitation on the subintervals, and assuming more generously q to be continuous with a continuously differentiable second component on the whole interval $\mathcal{I}$, then, for every such q, there exists a unique solution $x \in C_D^1(\mathcal{I}, \mathbb{R}^2)$. This means that, in the smoother setting, the critical point does not matter. Those kinds of critical points which can be healed by higher smoothness are said to be harmless. However, we stress once more that in a setting with minimal smoothness, these points are in fact critical. Written as

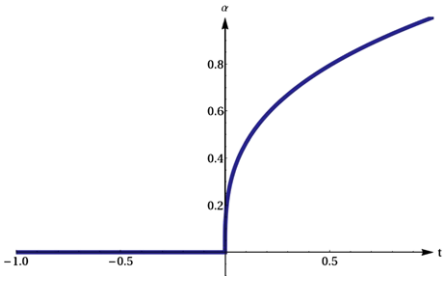


Fig. 2.2 Continuous function α of Example 2.70

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} x' + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x = q, \tag{2.127}$$

the DAE (2.126) yields a special DAE in standard canonical form (SCF). To ensure continuously differentiable solutions on the entire interval, one now has to suppose not only that q is continuously differentiable, but also that αq_2 is so. □

Example 2.71 (Index drop in A yielding a harmless critical point). We replace the function α in (2.126) by a different one and turn to

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} x \right)' + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x = q, \tag{2.128}$$

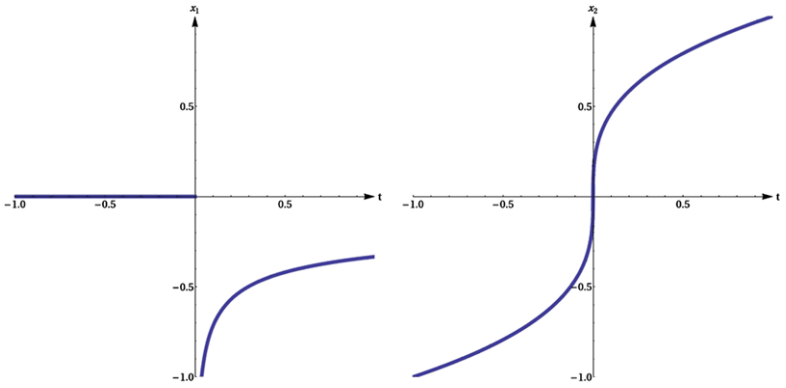


Fig. 2.3 Solution segments of x_1, x_2 in Example 2.70

with

$$\alpha(t) = \begin{cases} -|t|^{\frac{1}{3}} & \text{for } t \in [-1,0) \\ t^{\frac{1}{3}} & \text{for } t \in [0,1]. \end{cases}$$

The DAE (2.128) has a properly stated leading term on the two subintervals $[-1,0)$ and $(0,1]$, but on the entire interval $[-1,1]$ the leading term fails to be properly stated. The point $t_* = 0$ is a critical one.
The matrix function sequence

$$G_0 = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \; G_1 = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}, \; Q_1 = \begin{bmatrix} 0 & -\alpha \\ 0 & 1 \end{bmatrix}, \; G_1 = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix},$$

is admissible with characteristic values $r_0 = 1, r_1 = 1$ and $r_2 = 2$ on the intervals $[-1,0)$ and $(0,1]$ which indicates the DAE to be regular with index 2 there.
We apply the excitation $q_1 \equiv 0$, $q_2 = \alpha$ on both subintervals. As in the previous example, the first components of the solution segments cannot be glued together, as is sketched in Figure 2.4. For smoother excitations, we again obtain solutions belonging to $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^2)$. Furthermore, if q and αq_2 are smooth, then the SCF version (cf. 2.127) has $\mathcal{C}^1$-solutions. Those critical points which disappear in smoother settings are said to be harmless. □

Equations (2.124), (2.126) and (2.128) possess the following property independently of the behavior of the function α: There is a subspace $N_A \subseteq \mathbb{R}^2$, such that

$$N_A \oplus \operatorname{im} D = \mathbb{R}^2, \quad N_A = \operatorname{span}\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix} \right\} \subseteq \ker A.$$

This property motivates the following generalization of proper leading terms for DAEs (2.1) having continuous coefficients as before.

Definition 2.72. For the DAE (2.1), let the time-varying subspace $\operatorname{im} D$ be a $\mathcal{C}^1$-subspace on $\mathcal{I}$, and let a further $\mathcal{C}^1$-subspace N_A exist such that

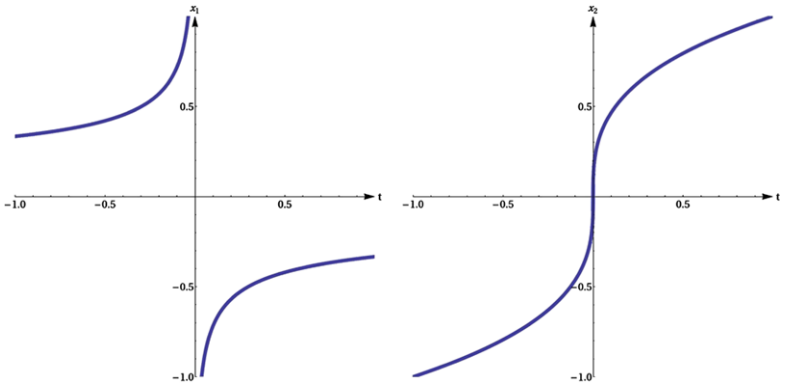


Fig. 2.4 Solution segments of x_1, x_2 in Example 2.71

$$N_A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad N_A(t) \subseteq \ker A(t), \quad t \in \mathcal{I}. \tag{2.129}$$

(1) If $N_A(t) = \ker A(t)$ for all t from a dense subset of the interval $\mathcal{I}$, then we speak of a DAE with an *almost proper leading term*.
(2) If $\dim \ker D(t) \geqslant 1$, then equation (2.1) is called a DAE with a *quasi-proper leading term* on $\mathcal{I}$.

DAEs with proper leading terms constitute a particular case of DAEs with almost-proper leading terms, if $N_A(t) = \ker A(t)$ holds true for all $t \in \mathcal{I}$.
The DAEs in Examples 2.69–2.71 have quasi-proper leading terms on the entire given interval $\mathcal{I}$. Examples 2.69 and 2.71 show even almost proper leading terms.
Example 2.70 represents a simple case of a DAE in SCF. Large classes of DAEs with those quasi-proper leading term, including the DAEs in SCF, are treated in detail in Chapter 9.
To some extent, the quasi-proper DAE form is quite comfortable. However, we should be aware that, in the general setting of quasi-proper DAEs, there is no way of indicating basic level critical points as in Examples 2.69, 2.70, and 2.71. This is why we prefer properly stated leading terms.

Our examples clearly account for the correspondence between singular solution behavior and points at which the matrix function sequence loses one of the required constant-rank properties. Roughly speaking, at all points where the matrix function sequence determining regularity cannot be built, we expect a critical (in some sense) solution behavior. We refer to [194] for a closer view of the relevant literature. As in [194], we consider critical (in [194] named *singular*) points to be the counterparts of regular points. Therefore, in this section, we deal with square DAEs (2.1) the coefficients A of which do not necessarily show constant rank. We recall Examples 2.69, 2.70, and 2.71 once more, which demonstrate critical solution behavior corresponding to the rank changes of A.
The DAE in Example 2.70 fails to have a proper leading term on the subinterval $[-1,0)$. On this subinterval, the special matrix function $A = 0$ has constant rank 0 and $\ker A = \mathbb{R}^2$ is a C^1-subspace on $[-1,0)$. For this reason, in this special case,

we could do with a proper refactorization to a proper leading term. Such proper refactorizations apply also in general cases as the next proposition says.

Proposition 2.73. *Let the DAE (2.1) have a quasi-proper leading term on the given interval $\mathcal{I}$. Let $\tilde{\mathcal{I}} \subseteq \mathcal{I}$ be a subinterval such that* $\ker A$ *is a $\mathcal{C}^1$-subspace on $\tilde{\mathcal{I}}$. Then $\tilde{R} := A^+A$ is continuously differentiable on $\tilde{\mathcal{I}}$ and the DAE has there the reformulation with a proper leading term*

$$A(\tilde{R}Dx)' + (B - A\tilde{R}'D)x = q, \quad t \in \tilde{\mathcal{I}}. \tag{2.130}$$

Proof. The matrix function $\tilde{R}$ is continuously differentiable as an orthoprojector function along a $\mathcal{C}^1$-subspace. We rewrite the leading term in the DAE (2.1) on the subinterval as

$$A(Dx)' = A\tilde{R}(Dx)' = A(\tilde{R}Dx)' - A\tilde{R}'Dx,$$

which leads to (2.130).
Introduce R as the projector function onto $\operatorname{im} D$ along N_A so that $\operatorname{im} D = \operatorname{im} R$ and $N_A = \operatorname{im}(I - R)$. Owing to condition (2.129), R is well defined. Additionally, it holds that $\operatorname{im}(I - R) = N_A \subseteq \ker A = \ker \tilde{R}$, and hence $\tilde{R}(I - R) = 0$, thus $\tilde{R} = \tilde{R}R$. This implies $\operatorname{im} \tilde{R}D = \operatorname{im} \tilde{R}R = \operatorname{im} \tilde{R}$, which proves the spaces $\ker A = \ker \tilde{R}$ and $\operatorname{im} \tilde{R}D = \operatorname{im} \tilde{R}$ to be transversal $\mathcal{C}^1$-subspaces on $\tilde{\mathcal{I}}$. Therefore, the DAE (2.130) has a proper leading term. □

Definition 2.74. Let the DAE (2.1), with $m = k$, have a quasi-proper leading term. Then, $t_* \in \mathcal{I}$ is said to be a *regular point* of the DAE, if there is an open interval $\mathcal{I}_*$ containing t_* such that either the original DAE is regular on $\tilde{\mathcal{I}} := \mathcal{I} \cap \mathcal{I}_*$ or $\ker A$ is a $\mathcal{C}^1$-subspace on $\tilde{\mathcal{I}}$ and the proper reformulation (2.130) is a regular DAE on $\tilde{\mathcal{I}}$.
Otherwise, t_* is said to be a *critical point*.
Each open interval on which the DAE is regular is called a *regularity interval*. Denote by $\mathcal{I}_{reg}$ the set of all $t \in \mathcal{I}$ being regular points of the DAE.

In this sense, $t_* = 0$ is the only critical point of the DAEs in Examples 2.66, 2.67, 2.69, 2.70, and 2.71, while in Example 2.68 the set of critical points is formed by the zeros of the functions β and γ. The left boundary point in Example 2.66 is a critical point while the right boundary point is regular.
By definition, Example 2.66 shows the regularity interval $(0,1)$ but $\mathcal{I}_{reg} = (0,1]$. We find the regularity intervals $(-\infty,0)$ and $(0,\infty)$ in Example 2.67, whereby the characteristic values are on both sides $r_0 = 1, r_1 = 2$ and $\mu = 2$.
In Example 2.68, regularity intervals exist around all inner points t of the given interval where $\beta(t)\gamma(t) \neq 0$, with uniform characteristics $r_0 = 2, r_1 = 2, r_2 = 3$ and $\mu = 2$.
The peculiarity of Example 2.70 consists of the different characteristic values on the regularity intervals $(-1,0)$ and $(0,1)$.

Each regularity interval consists of regular points, exclusively. All subintervals of a regularity interval inherit the characteristic values. If there are intersecting regularity intervals, then the DAE has common characteristic values on these intervals, and the union of regularity intervals is a regularity interval, again ([173], applying

widely orthogonal projector functions one can simplify the proof given there).
The set $\mathcal{I}_{reg} \subseteq \mathcal{I}$ may be described as the union of disjoint regularity intervals, eventually completed by the regular boundary points. By definition, $\mathcal{I} \setminus \mathcal{I}_{reg}$ is the set of critical points of the DAE (2.1).

The regularity notion (cf. Definitions 2.6 and 2.25) involves several constant-rank conditions. In particular, the proper leading term brings the matrix function $G_0 = AD$ with constant rank $r_0 = r$. Further, the existence of regular admissible matrix functions includes that, at each level $k = 1, \ldots, \mu - 1$,

(A) the matrix function G_k has constant rank r_k, and

(B) the intersection $\widehat{N}_k$ is trivial, i.e., $\widehat{N}_k = \{0\}$.

Owing to Proposition 2.7 we have $\ker \Pi_{k-1} = N_0 + \cdots + N_{k-1}$, and hence

$$\widehat{N}_k = N_k \cap (N_0 + \cdots + N_{k-1}) = \ker G_k \cap \ker \Pi_{k-1}.$$

Then, the intersection $\widehat{N}_k$ is trivial, exactly if the matrix function

$$\begin{bmatrix} G_k \\ \Pi_{k-1} \end{bmatrix} \tag{2.131}$$

has full column rank m. This means that condition (B) also represents a rank condition.

Suppose the coefficients A, D and B of the DAE are sufficiently smooth (at most class C^{m-1} will do). Then, if the algebraic *rank conditions* are fulfilled, the requirements for the projector functions Π_k and $D\Pi_k D^-$ to be continuous respectively continuously differentiable, can be satisfied at one level after the other. In consequence (cf. [173, 174, 194]), a critical point can be formally characterized as the location where the coefficient A has a rank drop, or where one of the constant-rank conditions type (A) or type (B), at a level $k \geq 1$, is violated first.

Definition 2.75. Let the DAE (2.44) have a quasi-proper leading term, and t_* be a critical point. Then, t_* is called

(1) a *critical point of type* 0, if $\operatorname{rank} G_0(t_*) < r := \operatorname{rank} D(t_*)$,
(2) a *critical point of type A at level* $k \geq 1$ (briefly, type k-A), if there are admissible projectors functions $Q_0, \ldots, Q_{k-1}$, and G_k changes its rank at t_*,
(3) a *critical point of type B at level* $k \geq 1$ (briefly, type k-B), if there are admissible projector functions $Q_0, \ldots, Q_{k-1}$, the matrix function G_k has constant rank, but the full-rank condition for the matrix function (2.131) is violated at t_*.

It is worth emphasizing that the proposed typification of critical points remains invariant with respect to transformations and refactorizations (Section 2.3), and also with respect to the choice of admissible projector functions (Subsection 2.2.2).

The DAEs in Examples 2.66 and 2.67 have the type 1-A critical point $t_* = 0$. In Example 2.68, the zeros of the function γ are type 2-A critical points, while the zeros of the function β yield type 1-B critical points. Examples 2.69, 2.70 and 2.71 show different cases of type 0 critical points.

While the zero of the function α in Example 2.71 yields a harmless critical point, in contrast, in Example 2.69, the zero of α causes a singular inherent ODE.
How do harmless critical points differ from the other critical points? As suggested by Example 2.71, we prove the nonsingularity of the matrix function G_μ to indicate harmless critical points in general.

Let the DAE (2.44) have an almost proper leading term. For simplicity, let DD^* be continuously differentiable such that the widely orthogonal projector functions can be used. Assume the set of regular points $\mathcal{I}_{reg}$ to be dense in $\mathcal{I}$.
Let Q_0 be the orthogonal projector function onto $\ker D =: N_0$, which is continuous on the entire interval $\mathcal{I}$, since D has constant rank r there. Set $G_0 = AD$, $B_0 = B$, $G_1 = G_0 + BQ_0$. These functions are also continuous on $\mathcal{I}$. For all $t \in \mathcal{I}_{reg}$ it holds further that $\operatorname{rank} G_0(t) = r$. On each regularity interval, which is a regularity region, we construct the matrix function sequence by means of widely orthogonal projector functions up to G_μ, whereby μ denotes the lowest index such that $G_\mu(t)$ is nonsingular for all $t \in \mathcal{I}_{reg}$. In particular, $\Pi_1, \ldots, \Pi_{\mu-1}$ are defined and continuous on each part of $\mathcal{I}_{reg}$. Assume now that

$$\Pi_1, \ldots, \Pi_{\mu-1} \quad \text{have continuous extensions on } \mathcal{I}, \tag{2.132}$$

and we keep the same denotation for the extensions. Additionally, suppose

$$D\Pi_1 D^-, \ldots, D\Pi_{\mu-1} D^- \quad \text{are continuously differentiable on } \mathcal{I}.$$

Then, the projector functions $\Pi_{i-1} Q_i = \Pi_{i-1} - \Pi_i$, $i = 1, \ldots, \mu - 1$, have continuous extensions, too, and the matrix function sequence (cf. (2.5)–(2.8), and Proposition 2.7)

$$\begin{aligned} B_i &= B_{i-1}\Pi_{i-1} - G_i D^- (D\Pi_i D^-)' D\Pi_{i-1}, \\ G_{i+1} &= G_i + B_i \Pi_{i-1} Q_i, \quad i = 1, \ldots, \mu - 1, \end{aligned}$$

is defined and continuous on the entire interval $\mathcal{I}$. In contrast to the regular case, where the matrix functions G_j have constant rank on the entire interval $\mathcal{I}$, now, for the time being, the projector functions Q_j are given on $\mathcal{I}_{reg}$ only, and

$$N_i(t) = \operatorname{im} Q_i(t) = \ker G_i(t), \quad \text{for all} \quad t \in \mathcal{I}_{reg}.$$

The projector function $\Pi_0 = P_0$ inherits the constant rank $r = \operatorname{rank} D$ from D. On each of the regularity intervals, the rank r_0 of G_0 coincides with the rank of D, and hence we are aware of the uniform characteristic value $r_0 = r$ on all regularity intervals, that is on $\mathcal{I}_{reg}$.
Owing to its continuity, the projector function Π_1 has constant rank on $\mathcal{I}$. Taking into account the relations

$$\ker \Pi_1(t) = N_0(t) \oplus N_1(t), \quad \dim N_0(t) = m - r_0, \quad \dim N_1(t) = m - r_1, \ t \in \mathcal{I}_{reg}$$

we recognize the characteristic value $r_1 = \operatorname{rank} G_1$ to be also uniform on $\mathcal{I}_{reg}$, and so on. In this way we find out that all characteristics

$$r_0 \le \cdots \le r_{\mu-1} < r_\mu = m \quad \text{are uniform on } \mathcal{I}_{reg}.$$

In particular, the DAE has index μ on $\mathcal{I}_{reg}$.
Denote by $G_\mu(t)^{adj}$ the matrix of cofactors to $G_\mu(t)$, and introduce the determinant $\omega_\mu(t) := \det G_\mu(t)$, such that

$$\omega_\mu(t) G_\mu(t)^{-1} = G_\mu(t)^{adj}, \quad t \in \mathcal{I}_{reg}.$$

By construction, it results that $G_\mu Q_i = B_i Q_i = B_i \Pi_{i-1} Q_i$, for $i = 1, \ldots, \mu - 1$, thus

$$\omega_\mu(t) Q_i(t) = G_\mu(t)^{adj} B_i(t) \Pi_{i-1}(t) Q_i(t), \quad i = 1, \ldots, \mu - 1, \quad t \in \mathcal{I}_{reg}. \quad (2.133)$$

The last expression possesses a continuous extension, and hence $\omega_\mu Q_i = G_\mu^{adj} B_i \Pi_{i-1} Q_i$ is valid on $\mathcal{I}$.
Observe that a nonsingular $G_\mu(t_*)$ also indicates that the projector functions $Q_1, \ldots, Q_{\mu-1}$ have continuous extensions over the critical point t_*. In this case, the decoupling formulas (2.51), (2.62) keep their value for the continuous extensions, and it is evident that the critical point is a harmless one.
In contrast, if G_μ has a rank drop at the critical point t_*, then the decoupling formulas actually indicate different but singular solution phenomena. Additionally, several projector functions Q_j may suffer discontinuities, as is the case in Example 2.68.

Next, by means of the widely orthogonal projector functions, on each regularity interval, we apply the basic decoupling (see Subsection 2.4.2, Theorem 2.30) of a regular DAE into the IERODE (2.51) and the subsystem (2.62). In order to safely obtain coefficients that are continuous on the entire interval $\mathcal{I}$, we multiply the IERODE (2.51) by ω_μ, the first row of (2.62) by ω_μ^μ, the second by $\omega_\mu^{\mu-1}$, and so on up to the last line which we multiply by ω_μ. With regard to assumption (2.132) and relation (2.133), the expressions $\omega_\mu G_\mu^{-1}$ and $\omega_\mu \mathcal{K}$, $\omega_\mu \mathcal{M}_{l+1}$ (cf. (2.54), (2.55)) are continuous on $\mathcal{I}$, and so are all the coefficients of the subsystem resulting from (2.62). Instead of the IERODE (2.51) we are now confronted with the equation

$$\omega_\mu u' - \omega_\mu (D\Pi_{\mu-1} D^-)' u + D\Pi_{\mu-1} G_\mu^{adj} B_\mu D^- u = D\Pi_{\mu-1} G_\mu^{adj} q, \quad (2.134)$$

which is rather a scalarly implicit inherent ODE or an inherent explicit singular ODE (IESODE). As is proved for regular DAEs by Theorem 2.30, the equivalence of the DAE and the system decoupled in this way is given. We refer to [194, Subsection 4.2.2] for a detailed description in a slightly different way. Here we take a look at the simplest lower index cases only.
The case $\mu = 1$ corresponds to the solution decomposition $x = D^- u + Q_0 x$, the inherent ODE

$$\omega_1 u' - \omega_1 R' u + DG_1^{adj} B_1 D^- u = DG_1^{adj} q, \tag{2.135}$$

and the subsystem

$$\omega_1 Q_0 x = -Q_0 G_1^{adj} B_1 D^- u + Q_0 G_1^{adj} q. \tag{2.136}$$

For $\mu = 2$, we apply the solution decomposition $x = D^- u + \Pi_0 Q_1 x + Q_0 x$. The inherent ODE reads

$$\omega_2 u' - \omega_2 (D\Pi_1 D^-)' u + D\Pi_1 G_2^{adj} B_1 D^- u = D\Pi_1 G_2^{adj} q, \tag{2.137}$$

and we have to add the subsystem

$$\begin{bmatrix} -\omega_2 Q_0 \omega_2 Q_1 D^- (D\Pi_0 Q_1 x)' \\ 0 \end{bmatrix} + \begin{bmatrix} \omega_2^2 Q_0 x \\ \omega_2 \Pi_0 Q_1 x \end{bmatrix} + \begin{bmatrix} Q_0 \omega_2 P_1 \omega_2 \mathcal{K} \Pi_1 \\ \Pi_0 Q_1 \omega_2 \mathcal{K} \Pi_1 \end{bmatrix} D^- u = \begin{bmatrix} Q_0 \omega_2 P_1 G_2^{adj} \\ \Pi_0 Q_1 G_2^{adj} \end{bmatrix} q. \tag{2.138}$$

A careful inspection of our examples proves that these formulas comprise a worst case scenario. For instance, in Example 2.68, not only is $D\Pi_1 G_2^{adj} B_1 D^-$ continuous but already $D\Pi_1 G_2^{-1} B_1 D^-$ can be extended continuously. However, as in Example 2.66, the worst case can well happen.

Proposition 2.76. *Let the DAE (2.1) have an almost proper leading term, and DD^* be continuously differentiable. Let the set of regular points $\mathcal{I}_{reg}$ be dense in $\mathcal{I}$. If the projector functions $\Pi_1, \ldots, \Pi_{\mu-1}$ associated with the widely orthogonal projector functions have continuous extensions on the entire interval $\mathcal{I}$, and $D\Pi_1 D^-, \ldots, D\Pi_{\mu-1} D^-$ are continuously differentiable, then the following holds true:*

(1) *The DAE has on $\mathcal{I}_{reg}$ uniform characteristics $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$.*
(2) *If $G_\mu(t_*)$ is nonsingular at the critical point t_*, then the widely orthogonal projector functions $Q_0, \ldots, Q_{\mu-1}$ themselves have continuous extensions over t_*. If the coefficients A, D, and B are sufficiently smooth, then t_* is a harmless critical point.*
(3) *If $G_\mu(t_*)$ is nonsingular at the critical point t_*, then $G_{\mu-1}(t)$ has necessarily constant rank $r_{\mu-1}$ on a neighborhood including t_*.*
(4) *If the DAE has index 1 on $\mathcal{I}_{reg}$, then its critical points fail to be harmless.*
(5) *A critical point of type B leads necessarily to a singular G_μ, and hence it can never been harmless.*

Proof. Assertion (1) is already verified. Assertion (2) follows immediately by making use of the decoupling. If A, D, B are smooth, then the coefficients of the subsystem (2.62) are also sufficiently smooth, and allow for the respective solutions.
Turn to (3). Owing to (2), $Q_{\mu-1}$ is continuous, and $\operatorname{rank} Q_{\mu-1}(t_*) = m - r_{\mu-1}$, $G_{\mu-1}(t_*) Q_{\mu-1}(t_*) = 0$ are valid, thus $\operatorname{rank} G_{\mu-1}(t_*) \le r_{\mu-1}$. The existence of a $z \in \ker G_{\mu-1}(t_*)$, $P_{\mu-1}(t_*) z = z \ne 0$, would imply $G_{\mu-1}(t_*) z = 0$, and hence would

contradict the nonsingularity of $G_{\mu-1}(t_*)$.
(4) is a direct consequence of (3).
For proving Assertion (5) we remember the relation

$$\Pi_{j-1}(t)Q_j(t) = \Pi_{j-1}(t)Q_j(t)\Pi_{j-1}(t), \quad t \in \mathcal{I}_{reg}.$$

These relations remain valid for the continuous extensions, that is, for $t \in \mathcal{I}$. Consider a type $k-B$ critical point t_*, and a nontrivial $z \in N_k(t_*) \cap (N_0(t_*) + \cdots + N_{\mu-1}(t_*))$, which means $G_k(t_*)z = 0$, $\Pi_{k-1}(t_*)z = 0$. This yields

$$\begin{aligned} G_\mu(t_*)z = G_k(t_*)z + B_k(t_*)Q_k(t_*)\Pi_{k-1}(t_*)z \\ + \cdots + B_{\mu-1}(t_*)\Pi_{\mu-2}(t_*)Q_{\mu-1}(t_*)\Pi_{k-1}(t_*)z = 0, \end{aligned}$$

and hence, $G_\mu(t_*)$ is singular. □

2.10 Strangeness versus tractability

2.10.1 Canonical forms

Among the traditional goals of the theory of linear time-varying DAEs are appropriate generalizations of the Weierstraß–Kronecker canonical form and equivalence transformations into these canonical forms. So far, except for the T-canonical form which applies to both standard form DAEs and DAEs with properly stated leading term (cf. Subsection 2.8), reduction to canonical forms is developed for standard form DAEs (e.g. [39], [25], [127]).
While equivalence transformations for DAEs with properly stated leading term include transformations K of the unknown, scalings L and refactorizations H of the leading term (cf. Section 2.3), equivalence transformations for standard form DAEs combine only the transformations K of the unknowns and the scalings L.
Transforming the unknown function by $x = K\tilde{x}$ and scaling the standard form DAE (2.112) by L yields the equivalent DAE

$$\underbrace{LEK}_{\tilde{E}}\tilde{x}' + \underbrace{(LFK + LEK')}_{\tilde{F}}\tilde{x} = Lq.$$

Therefore the transformation matrix functions K must be continuously differentiable.

In the remaining part of this subsection we use the letters K and H also for special entries in the matrix functions describing the coefficients of the canonical forms below. No confusion will arise from this.

Definition 2.77. The structured DAE with continuous coefficients

$$\begin{bmatrix} I_{m-l} & K \\ 0 & N \end{bmatrix} x' + \begin{bmatrix} W & 0 \\ H & I_l \end{bmatrix} x = q, \tag{2.139}$$

$0 \le l \le m$, is said to be in

(1) *standard canonical form* (SCF), if $H = 0,\ K = 0$, and N is strictly upper triangular,

(2) *strong standard canonical form* (SSCF), if $H = 0,\ K = 0$, and N is a constant, strictly upper triangular matrix,

(3) *S-canonical form*, if $H = 0$, $K = [0\ K_1 \dots K_\kappa]$, and

$$N = \begin{bmatrix} 0 & N_{1,2} & \cdots & N_{1,\kappa} \\ & \ddots & & \vdots \\ & & \ddots & N_{\kappa-1,\kappa} \\ & & & 0 \end{bmatrix} \begin{matrix} \}l_1 \\ \\ \}l_{\kappa-1} \\ \}l_\kappa \end{matrix},$$

is strictly upper block triangular with full row rank entries $N_{i,i+1},\ i = 1, \dots, \kappa - 1$,

(4) *T-canonical form*, if $K = 0$ and N is strictly upper block triangular with full column rank entries $N_{i,i+1},\ i = 1, \dots, \kappa - 1$.

In the case of time-invariant coefficients, these four canonical forms are obviously equivalent. However, this is no longer true for time-varying coefficients.
The matrix function N is nilpotent in all four canonical forms, and N has uniform nilpotency index κ in (3) and (4). N and all its powers N^k have constant rank in (2), (3) and (4). In contrast, in (1), the nilpotency index and the rank of N may vary with time. The S-canonical form is associated with DAEs with regular strangeness index $\zeta = \kappa - 1$ (cf. [127]), while the T-canonical form is associated with regular DAEs with tractability index $\mu = \kappa$ (cf. Subsection 2.8). The classification into SCF and SSCF goes back to [39] (cf. also [25]). We treat DAEs being transformable into SCF as quasi-regular DAEs in Chapter 9. Here we concentrate on the S-canonical form. We prove that each DAE being transformable into S-canonical form is regular with tractability index $\mu = \kappa$, and hence, each DAE with well-defined regular strangeness index ζ is a regular DAE with tractability index $\mu = \zeta + 1$. All the above canonical forms are given in standard form. For the T-canonical form, a version with properly stated leading term is straightforward (cf. Definition 2.64).

The strangeness index concept applies to standard form DAEs (2.112) with sufficiently smooth coefficients. A reader who is not familiar with this concept will find a short introduction in the next subsection. For the moment, *we interpret DAEs with regular strangeness index as those being transformable into S-canonical form*. This is justified by an equivalence result of [127], which is reflected by Theorem 2.78 below.

The regular strangeness index ζ is supported by a sequence of *characteristic values* $\bar{r}_i, \bar{a}_i, \bar{s}_i,\ i = 0, \dots, \zeta$, which are associated with constant-rank conditions for matrix functions, and which describe the detailed size of the S-canonical form. By definition, $s_\zeta = 0$ (cf. Subsection 2.10.2). These characteristic values are invariant

with respect to the equivalence transformations, however, they are not independent of each other.

Theorem 2.78. *Each DAE (2.112) with smooth coefficients, well-defined strangeness index ζ and characteristic values $\bar{r}_i, \bar{a}_i, \bar{s}_i$, $i = 0, \ldots, \zeta$, is equivalent to a DAE in S-canonical form with $\kappa = \zeta + 1$, $l = l_1 + \cdots + l_\kappa$, $m - l = \bar{r}_\zeta$, and*

$$l_1 \le \cdots \le l_\kappa, \quad l_1 = \bar{s}_{\kappa-2} = \bar{s}_{\zeta-1},\ l_2 = \bar{s}_{\kappa-3}, \ldots, l_{\kappa-1} = \bar{s}_0,\ l_\kappa = \bar{s}_0 + \bar{a}_0.$$

Proof. This assertion comprises the regular case of [127, Theorem 12] which considers more general equations having also underdetermined parts (indicated by nontrivial further characteristic values $\bar{u}_i$). □

By the next assertion, which represents the main result of this subsection, we prove each DAE with regular strangeness index ζ to be at the same time a regular DAE with tractability index $\mu = \zeta + 1$. Therefore, the tractability index concept applies at least to the entire class of DAEs which are accessible by the strangeness index concept. Both concepts are associated with characteristic values being invariant under equivalence transformations, and, of course, we would like to know how these characteristic values are related to each other. In particular, the question arises whether the constant-rank conditions supporting the strangeness index coincide with the constant-rank conditions supporting the tractability index.

Theorem 2.79. (1) *Let the standard form DAE (2.112) have smooth coefficients, regular strangeness index ζ and characteristic values $\bar{r}_i$, $\bar{a}_i$, $\bar{s}_i$, $i = 0, \ldots, \zeta$. Then this DAE is regular with tractability index $\mu = \zeta + 1$ and associated characteristic values*

$$r_0 = \bar{r}_0, \quad r_j = m - \bar{s}_{j-1}, \quad j = 1, \ldots, \mu.$$

(2) *Each DAE in S-canonical form with smooth coefficients can be transformed into T-canonical form with $H = 0$.*

Proof. (1) We prove the assertion by constructing a matrix function sequence and admissible projector functions associated with the tractability index framework for the resulting S-canonical form described by Theorem 2.78.
The matrix function N within the S-canonical form has constant rank $l - l_\kappa$. Exploiting the structure of N we compose a projector function $Q_0^{[N]}$ onto $\ker N$, which is upper block triangular, too. Then we set

$$P_0 := \begin{bmatrix} I_{m-l} & K Q_0^{[N]} \\ 0 & P_0^{[N]} \end{bmatrix}, \quad \text{such that} \quad \ker P_0 = \ker \begin{bmatrix} I_{m-l} & K \\ 0 & N \end{bmatrix}.$$

P_0 is a projector function. The DAE coefficients are supposed to be smooth enough so that P_0 is continuously differentiable. Then we can turn to the following properly stated version of the S-canonical form:

$$\begin{bmatrix} I_{m-l} & K \\ 0 & N \end{bmatrix} (P_0 x)' + \Big(\underbrace{\begin{bmatrix} W & 0 \\ 0 & I_l \end{bmatrix} - \begin{bmatrix} I_{m-l} & K \\ 0 & N \end{bmatrix} P_0'}_{\begin{bmatrix} W & -K' Q_0^{[N]} \\ 0 & I_l - N(P_0^{[N]})' \end{bmatrix}} \Big) x = q. \tag{2.140}$$

The product $NP_0^{[N]'}$ is again strictly upper block triangular, and $I_l - N(P_0^{[N]})'$ is nonsingular. Scaling the DAE by

$$\begin{bmatrix} I_{m-l} & 0 \\ 0 & (I_l - N(P_0^{[N]})')^{-1} \end{bmatrix}$$

yields

$$\begin{bmatrix} I_{m-l} & K \\ 0 & M_0 \end{bmatrix} (P_0 x)' + \begin{bmatrix} W & -K' Q_0^{[N]} \\ & I_l \end{bmatrix} x = q. \tag{2.141}$$

The matrix function M_0 has the same structure as N, and $\ker M_0 = \ker N$. For the subsystem corresponding to the second line of (2.141)

$$M_0 (P_0^{[N]} v)' + v = q_2.$$

Proposition B.2 in Appendix B provides a matrix function sequence $G_j^{[N]}$, $j = 0, \ldots, \kappa$, and admissible projector functions $Q_0^{[N]}, \ldots, Q_{\kappa-1}^{[N]}$ such that this subsystem is a regular DAE with tractability index $\mu^{[N]} = \kappa$ and characteristic values

$$r_i^{[N]} = l - l_{\kappa - i}, \quad i = 0, \ldots, \kappa - 1, \quad r_\kappa^{[N]} = l.$$

Now we compose a matrix function sequence and admissible projector functions for the DAE (2.141). We begin with $D = D^- = R = P_0$, and build successively for $i = 0, \ldots, \kappa$

$$G_i = \begin{bmatrix} I_{m-l} & * \\ 0 & G_i^{[N]} \end{bmatrix}, \quad Q_i = \begin{bmatrix} 0 & * \\ 0 & Q_i^{[N]} \end{bmatrix}, \quad \Pi_i = \begin{bmatrix} I_{m-l} & * \\ 0 & \Pi_i^{[N]} \end{bmatrix}, \quad B_i = \begin{bmatrix} W & * \\ 0 & B_i^{[N]} \end{bmatrix}.$$

The coefficients are supposed to be smooth enough so that the Π_i are continuously differentiable. It follows that the matrix functions G_i have constant ranks

$$r_i = m - l + r_i^{[N]} = m - l + l - l_{\kappa - i} = m - l_{\kappa - i}, \ i = 0, \ldots, \kappa - 1, \ r_\kappa = m - l + r_\kappa^{[N]} = m.$$

This confirms that the DAE is regular with tractability index $\mu = \kappa$. Applying again Theorem 2.78, we express $r_i = m - l_{\kappa - i} = \bar{s}_{i-1}$ for $i = 1, \ldots, \kappa - 1$, further $r_0 = m - (\bar{s}_0 + \bar{a}_0) = \bar{r}_0$, and this completes the proof of (1).
(2) This is a consequence of assertion (1), and the fact that each regular DAE with tractability index μ can be transformed into T-canonical form (with $\kappa = \mu$, cf. Theorem 2.65). □

2.10.2 Strangeness reduction

The original strangeness index concept is a special reduction technique for standard form DAEs (2.112)

$$E(t)x'(t) + F(t)x(t) = q(t)$$

with sufficiently smooth coefficients on a compact interval $\mathcal{I}$. We repeat the basic reduction step from [127]. For more details and a comprehensive discussion of reduction techniques we refer to [130] and [189].

As mentioned before, the strangeness index is supported by several constant-rank conditions. In particular, the matrix E in (2.112) is assumed to have constant rank $\bar{r}$. This allows us to construct continuous injective matrix functions T, Z, and $\bar{T}$ such that

$$\operatorname{im} T = \ker E, \quad \operatorname{im} \bar{T} = (\ker E)^{\perp}, \quad \operatorname{im} Z = (\operatorname{im} E)^{\perp}.$$

The columns of $T, \bar{T}$, and Z are basis functions of the corresponding subspaces. Supposing Z^*FT to have constant rank $\bar{a}$, we find a continuous injective matrix function V such that

$$\operatorname{im} V = (\operatorname{im} Z^*FT)^{\perp}.$$

If, additionally, $V^*Z^*F\bar{T}$ has constant rank $\bar{s}$, then one can construct pointwise nonsingular matrix functions K and L, such that the transformation $x = K\tilde{x}$, and scaling the DAE (2.112) by L leads to

$$\begin{bmatrix} I_{\bar{s}} & & & & \\ & I_{\bar{d}} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix} \tilde{x}' + \begin{bmatrix} 0 & \tilde{F}_{1,2} & 0 & \tilde{F}_{1,4} & \tilde{F}_{1,5} \\ 0 & 0 & 0 & \tilde{F}_{2,4} & \tilde{F}_{2,5} \\ 0 & 0 & I_{\bar{a}} & 0 & 0 \\ I_{\bar{s}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tilde{x} = Lq, \tag{2.142}$$

with $\bar{d} := \bar{r} - \bar{s}$.
The system (2.142) consists of $m = \bar{s} + \bar{d} + \bar{a} + \bar{s} + \bar{u}$ equations, $\bar{u} := m - \bar{r} - \bar{a} - \bar{s}$. The construction of K and L involves three smooth factorizations of matrix functions and the solution of a classical linear IVP (see [130]).
The fourth equation in (2.142) is simply $\tilde{x}_1 = (Lq)_4$, which gives rise to replacement of the derivative $\tilde{x}_1'$ in the first line by $(Lq)_4'$. Doing so we attain the new DAE

$$\underbrace{\begin{bmatrix} 0 & & & & \\ & I_{\bar{d}} & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{bmatrix}}_{E_{new}} \tilde{x}' + \underbrace{\begin{bmatrix} 0 & \tilde{F}_{1,2} & 0 & \tilde{F}_{1,4} & \tilde{F}_{1,5} \\ 0 & 0 & 0 & \tilde{F}_{2,4} & \tilde{F}_{2,5} \\ 0 & 0 & I_{\bar{a}} & 0 & 0 \\ I_{\bar{s}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{F_{new}} \tilde{x} = Lq - \begin{bmatrix} (Lq)_4' \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{2.143}$$

which is expected to have a lower index since the mentioned differentiation of $\tilde{x}_1$ is carried out analytically.

This *reduction step* is supported by the three rank conditions

$$\operatorname{rank} E = \bar{r}, \quad \operatorname{rank} Z^* F T = \bar{a}, \quad \operatorname{rank} V^* Z^* F \bar{T} = \bar{s}. \tag{2.144}$$

The following proposition guarantees these constant-rank conditions to be valid, if the DAE under consideration is regular in the tractability sense.

Proposition 2.80. *Let the DAE (2.112) be regular with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu$. Then the constant-rank conditions (2.144) are valid,*

$$\bar{r} = r_0, \quad \bar{a} = r_1 - r_0, \quad \bar{s} = m - r_1,$$

so that the reduction step is feasible.

Proof. We choose symmetric projector functions $\mathcal{W}_0$, Q_0 and $\mathcal{W}_1$, and verify the relations

$$\operatorname{rank} Z^* B T = \operatorname{rank} \mathcal{W}_0 B Q_0 = r_1 - r_0, \quad \operatorname{rank} V^* Z^* F \bar{T} = \operatorname{rank} \mathcal{W}_1 B = m - r_1.$$

□

The reduction from $\{E, F\}$ to $\{E_{new}, F_{new}\}$ can be repeated as long as the constant-rank conditions are given. This leads to an iterative reduction procedure. One starts with $\{E_0, F_0\}$:=$\{E, F\}$ and forms, for each $i \ge 0$, a new pair $\{E_{i+1}, F_{i+1}\}$ to $\{E_i, F_i\}$. This works as long as the three constant-rank conditions

$$\bar{r}_i = \operatorname{rank} E_i, \quad \bar{a}_i = \operatorname{rank} Z_i^* F_i T_i, \quad \bar{s}_i = \operatorname{rank} V_i^* Z_i^* F_i \bar{T}_i, \tag{2.145}$$

hold true.
The *strangeness index* $\zeta \in \mathbb{N} \cup \{0\}$ is defined to be

$$\zeta := \min\{i \in \mathbb{N} \cup \{0\} : \bar{s}_i = 0\}.$$

The strangeness index is the minimal index such that the so-called strangeness disappears. ζ is named the *regular strangeness index*, if there are no so-called underdetermined parts during the iteration such that $\bar{u}_i = 0$ and $\bar{r}_i + \bar{a}_i + \bar{s}_i = m$ for all $i = 0, \ldots, \zeta$.
The values $\bar{r}_i$, $\bar{a}_i$, $\bar{s}_i$, $i \ge 0$, and several additional ones, are called *characteristic values* associated with the strangeness index concept.
If the original DAE (2.112) has regular strangeness index ζ, then the reduction procedure ends up with the DAE

$$\begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix} \tilde{\tilde{x}}' + \begin{bmatrix} 0 & 0 \\ 0 & I_a \end{bmatrix} \tilde{\tilde{x}} = \tilde{\tilde{q}},$$

with $d = \bar{d}_\zeta$, $a = \bar{a}_\zeta$.

Remark 2.81. Turn for a moment back to time-invariant DAEs and constant matrix pairs. If the matrix pair $\{E, F\}$ is regular with Kronecker index μ (which is the same

as the tractability index μ), and characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, then this pair has the regular strangeness index $\zeta = \mu - 1$. The characteristic values associated with the strangeness index can then be obtained from the $r_0, \ldots, r_\mu$ by means of the formulas

$$\begin{aligned} \bar{r}_i &= m - \sum_{j=0}^{i} (m - r_j), \\ \bar{a}_i &= \sum_{j=0}^{i} (m - r_j) - (m - r_{i+1}), \\ \bar{s}_i &= m - r_{i+1}, \quad i = 0, \ldots, \zeta. \end{aligned}$$

The same relations apply to DAEs with time-varying coefficients, too (cf. [139]).

2.10.3 Projector based reduction

Although linear regular higher index DAEs are well understood, they are not accessible to direct numerical integration as pointed out in Chapter 8. Especially for this reason, different kinds of index reduction have their meaning.
We formulate a reduction step for the DAE (2.44) with properly stated leading term, i.e.,

$$A(Dx)' + Bx = q,$$

by applying the projector function $\mathcal{W}_1$ associated with the first terms of the matrix function sequence. $\mathcal{W}_1$ projects along $\operatorname{im} G_1 = \operatorname{im} G_0 \oplus \operatorname{im} \mathcal{W}_0 B Q_0$, and, because of $\operatorname{im} A \subseteq \operatorname{im} G_0 \subseteq \operatorname{im} G_1$, multiplication of the DAE by $\mathcal{W}_1$ leads to the derivative-free equations

$$\mathcal{W}_1 Bx = \mathcal{W}_1 q. \tag{2.146}$$

We emphasize that these equations are just a part of the derivative-free equations, except for the case $\mathcal{W}_0 = \mathcal{W}_1$, which is given in Hessenberg systems, and in Example 2.82 below. The complete set is described by

$$\mathcal{W}_0 Bx = \mathcal{W}_0 q. \tag{2.147}$$

We suppose the matrix function $\mathcal{W}_1$ to have constant rank $m - r_1$, which is at least ensured in regular DAEs. For regular DAEs the subspace

$$S_1 = \ker \mathcal{W}_1 B$$

is known to have dimension r_1.
Introduce a continuous reflexive generalized inverse $(\mathcal{W}_1 B)^-$, and put

$$Z_1 := I - (\mathcal{W}_1 B)^- \mathcal{W}_1 B.$$

Z_1 is a continuous projector function onto S_1. Because of $\mathcal{W}_1 B Q_0 = 0$ the following properties hold true:

$$\begin{aligned} Z_1 Q_0 &= Q_0 \\ DZ_1 &= DZ_1 P_0 = DZ_1 D^- D \\ DZ_1 D^- &= DZ_1 D^- DZ_1 D^- \\ \operatorname{im} DZ_1 D^- &= \operatorname{im} DZ_1 = DS_1 = DS_0. \end{aligned}$$

$DZ_1 D^-$ is a priori a continuous projector function. Assuming the DAE coefficients to be sufficiently smooth, it becomes continuously differentiable, and we do so. In consequence, for each function $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ it follows that

$$DZ_1 x = DZ_1 D^- Dx \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n), \quad D(I - Z_1)x = Dx - DZ_1 x \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n),$$

which allows us to write the DAE as

$$A(DZ_1 x)' + A(D(I - Z_1)x)' + Bx = q. \tag{2.148}$$

Equation (2.146) is consistent, since, for reasons of dimensions, $\operatorname{im} \mathcal{W}_1 B = \operatorname{im} \mathcal{W}_1$. It follows that

$$(I - Z_1)x = (\mathcal{W}_1 B)^- \mathcal{W}_1 q. \tag{2.149}$$

This allows us to remove the derivative $(D(I - Z_1)x)'$ from the DAE, and to replace it by the exact solution part derived from (2.146). The resulting new DAE

$$A(DZ_1 x)' + Bx = q - A(D(\mathcal{W}_1 B)^- \mathcal{W}_1 q)'$$

has no properly stated lading term. This is why we express $A(DZ_1 x)' = A\{DZ_1 D^-(DZ_1 x)' + (DZ_1 D^-)' DZ_1 x\}$, and turn to the new DAE with a properly stated leading term

$$\underbrace{ADZ_1 D^-}_{A_{new}} \underbrace{(DZ_1}_{D_{new}} x)' + \underbrace{(A(DZ_1 D^-)' DZ_1 + B)}_{B_{new}} x = q - A(D(\mathcal{W}_1 B)^- \mathcal{W}_1 q)' \tag{2.150}$$

which has the same solutions as the original DAE (2.44) has, and which is expected to have a lower index (cf. [138]).

Example 2.82 (Index reduction step). We reconsider the DAE (2.10) from Example 2.4,

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & -t & 1 \\ 0 & 0 & 0 \end{bmatrix}}_{A(t)} \left(\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D} x(t) \right)' + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -t & 1 \end{bmatrix}}_{B(t)} x(t) = q(t), \quad t \in \mathbb{R},$$

where an admissible matrix function sequence for this DAE is generated. This DAE is regular with tractability index 3. Now compute

$$\mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{W}_1 B(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -t & 1 \end{bmatrix}.$$

Since $\mathcal{W}_1 B$ is already a projector function, we can set $(\mathcal{W}_1 B)^- = \mathcal{W}_1 B$. This implies

$$Z_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t & 0 \end{bmatrix}, \quad D(t)Z_1(t) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t & 0 \end{bmatrix},$$

and finally the special DAE (2.150)

$$\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{A_{new}(t)} \Bigg(\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t & 0 \end{bmatrix}}_{D_{new}(t)} x(t) \Bigg)' + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -t & 1 \end{bmatrix}}_{B_{new}(t)} x(t) = \begin{bmatrix} q_1(t) \\ q_2(t) - q_3'(t) \\ q_3(t) \end{bmatrix}, \quad t \in \mathbb{R},$$

which is indeed regular with tractability index 2. □

For the special choice $(\mathcal{W}_1 B)^- = (\mathcal{W}_1 B)^+$, the resulting Z_1 is the orthoprojector function onto S_1. This version is the counterpart to the strangeness reduction step from Subsection 2.10.2.

At first glance it seems to be somewhat arbitrary to figure out just the equations (2.146) for reduction. However, after the explanations below it will be seen as a nice option.

An analogous reduction step can be arranged by choosing the complete set of derivative-free equations (2.147) as a candidate. For regular DAEs, the subspace $\ker \mathcal{W}_0 B = S_0$ has dimension r_0, and we again obtain consistency, as well as the projector $Z_0 := I - (\mathcal{W}_0 B)^- \mathcal{W}_0 B$ onto S_0. From (2.147) it follows that

$$(I - Z_0)x = (\mathcal{W}_0 B)^- \mathcal{W}_0 q.$$

Now we need a smoother solution x to be able to differentiate this expression. To be more transparent we assume at least D and Z_0, as well as the solution x, to be continuously differentiable, and turn to the standard form

$$\underbrace{AD}_{E} x' + \underbrace{(B - AD')}_{F} x = q.$$

Here we express

$$x' = (Z_0 x)' + ((\mathcal{W}_0 B)^- \mathcal{W}_0 q)' = Z_0 x' + Z_0' x + ((\mathcal{W}_0 B)^- \mathcal{W}_0 q)',$$

such that we arrive at the new DAE

$$\underbrace{EZ_0}_{E_{new}} x' + \underbrace{(F + EZ_0')}_{F_{new}} x = q - E((\mathcal{W}_0 B)^- \mathcal{W}_0 q)'. \tag{2.151}$$

This kind of reduction is in essence the procedure described in [189]. The description in [189] concentrates on the coefficient pairs, and one turns to a condensed version of the pair $\{EZ_0, (I - \mathcal{W}_0)(F + EZ_0')\}$.

In the following we do not provide a precise proof of the index reduction, but explain the idea behind it. Assume the DAE (2.44) to be regular with tractability index μ and characteristic values $r_0 \le \cdots \le r_{\mu-1} = r_\mu = m$, and take a further look to the completely decoupled version consisting of the IERODE (2.51) and the subsystem (cf. (2.63))

$$\mathcal{N}(\mathcal{D}v)' + \mathcal{M}v = \mathcal{L}q. \tag{2.152}$$

This subsystem comprises the inherent differentiations. It reads in detail

$$\begin{bmatrix} 0 & \mathcal{N}_{0,1} & \cdots & \mathcal{N}_{0,\mu-1} \\ & 0 & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{bmatrix} \begin{bmatrix} 0 \\ (D\Pi_0 Q_1 x)' \\ \vdots \\ (D\Pi_{\mu-2} Q_{\mu-1} x)' \end{bmatrix} \tag{2.153}$$

$$+ \begin{bmatrix} I & \mathcal{M}_{0,1} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} Q_0 x \\ \Pi_0 Q_1 x \\ \vdots \\ \Pi_{\mu-2} Q_{\mu-1} x \end{bmatrix} = \begin{bmatrix} \mathcal{L}_0 q \\ \mathcal{L}_1 q \\ \vdots \\ \mathcal{L}_{\mu-1} q \end{bmatrix}.$$

We see that if we replace the derivative term $(D\Pi_{\mu-2} Q_{\mu-1} x)'$ by its exact solution part $(D\mathcal{L}_{\mu-1} q)'$ we arrive at the system

$$\mathcal{N}_{new} \begin{bmatrix} 0 \\ (D\Pi_0 Q_1 x)' \\ \vdots \\ (D\Pi_{\mu-3} Q_{\mu-2} x)' \\ 0 \end{bmatrix} + \begin{bmatrix} I & \mathcal{M}_{0,1} & \cdots & \mathcal{M}_{0,\mu-1} \\ & I & \ddots & \vdots \\ & & \ddots & \mathcal{M}_{\mu-2,\mu-1} \\ & & & I \end{bmatrix} \begin{bmatrix} Q_0 x \\ \Pi_0 Q_1 x \\ \vdots \\ \Pi_{\mu-2} Q_{\mu-1} x \end{bmatrix} \tag{2.154}$$

$$= \begin{bmatrix} \mathcal{L}_0 q - \mathcal{N}_{0,\mu-1} (\mathcal{L}_{\mu-1} q)' \\ \mathcal{L}_1 q - \mathcal{N}_{1,\mu-1} (\mathcal{L}_{\mu-1} q)' \\ \vdots \\ \mathcal{L}_{\mu-2} q - \mathcal{N}_{\mu-2,\mu-1} (\mathcal{L}_{\mu-1} q)' \\ \mathcal{L}_{\mu-1} q \end{bmatrix}.$$

While the matrix function $\mathcal{N}$ has nilpotency index μ, the new matrix function

$$\mathcal{N}_{new} = \begin{bmatrix} 0 & \mathcal{N}_{0,1} & \cdots & \mathcal{N}_{0,\mu-2} & 0 \\ & 0 & \ddots & \vdots & 0 \\ & & \ddots & \mathcal{N}_{\mu-3,\mu-2} & 0 \\ & & & 0 & 0 \\ & & & & 0 \end{bmatrix}$$

has nilpotency index $\mu - 1$ (cf. Proposition 2.29). That means, replacing the derivative $(D\Pi_{\mu-2}Q_{\mu-1}x)'$ by the true solution term reduces the index by one. Clearly, replacing further derivatives and successively solving the subsystem for $(I-\Pi_{\mu-1})x = Q_0x+\Pi_0Q_1x+\cdots+\Pi_{\mu-2}Q_{\mu-1}x$ reduces the index up to one. We keep in mind that, replacing at least the derivative $(D\Pi_{\mu-2}Q_{\mu-1}x)'$ reduces the index by at least one. However, in practice, we are not given the decoupled system. How can we otherwise make sure that this derivative is replaced?
Consider for a moment the equation

$$\mathcal{W}_{\mu-1}Bx = \mathcal{W}_{\mu-1}q \tag{2.155}$$

that is also a part of the derivative-free equations of our DAE. Since the subspace $S_{\mu-1} = \ker \mathcal{W}_{\mu-1}$ has dimension $r_{\mu-1}$, the matrix function $\mathcal{W}_{\mu-1}B$ has constant rank $m - r_{\mu-1}$, and equation (2.155) is consistent, we obtain with $Z_{\mu-1} := I - (\mathcal{W}_{\mu-1}B)^-\mathcal{W}_{\mu-1}B$ a continuous projector function onto $S_{\mu-1}$, and it follows that

$$(I - Z_{\mu-1})x = (\mathcal{W}_{\mu-1}B)^-\mathcal{W}_{\mu-1}q.$$

Since we use completely decoupling projector functions $Q_0, \ldots, Q_{\mu-1}$, we know that $\Pi_{\mu-2}Q_{\mu-1}$ is the projector function onto $\operatorname{im}\Pi_{\mu-2}Q_{\mu-1}$ along $S_{\mu-1}$. Therefore, with $I - Z_{\mu-1}$ and $\Pi_{\mu-2}Q_{\mu-1}$ we have two projector functions along $S_{\mu-1}$. This yields

$$I - Z_{\mu-1} = (I - Z_{\mu-1})\Pi_{\mu-2}Q_{\mu-1}, \quad \Pi_{\mu-2}Q_{\mu-1} = \Pi_{\mu-2}Q_{\mu-1}(I - Z_{\mu-1}),$$

and therefore, by replacing $(D(I - Z_{\mu-1})x)'$ we replace at the same time $(D\Pi_{\mu-2}Q_{\mu-1}x)'$. This means that turning from the original DAE (2.44) to

$$ADZ_{\mu-1}D^-(DZ_{\mu-1}x)' + (A(DZ_{\mu-1}D^-)'DZ_{\mu-1} + B)x = q - A(D(\mathcal{W}_{\mu-1}B)^-\mathcal{W}_{\mu-1}q)'$$

indeed reduces the index by one. However, the use of $Z_{\mu-1}$ is rather a theoretical option, since $\mathcal{W}_{\mu-1}$ is not easy to obtain. The point is that working instead with (2.146) and Z_1 as described above, and differentiating the extra components $D(I - Z_1)x$, includes the differentiation of the component $D(I - Z_{\mu-1})x$ as part of it. In this way, the reduction step from (2.44) to (2.150) seems to be a reasonable compromise from both theoretical and practical viewpoints.
At this point we emphasize that there are various possibilities to compose special reduction techniques.

2.11 Generalized solutions

We continue to consider linear DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{2.156}$$

with coefficients $A \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^m))$, $D \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n))$, $B \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$. $\mathcal{I} \subseteq \mathbb{R}$ denotes an interval. Here we focus on IVPs. Let $t_0 \in \mathcal{I}$ be fixed. We state the initial condition in the form

$$Cx(t_0) = z, \tag{2.157}$$

by means of a matrix $C \in L(\mathbb{R}^m, \mathbb{R}^d)$ which satisfies the condition

$$C = CD(t_0)^- D(t_0), \tag{2.158}$$

and which is further specified below (cf. Theorem 2.52) where appropriate.
A classical solution of the IVP (2.156), (2.157) is a continuous function x which possesses a continuously differentiable component Dx and satisfies the initial condition as well as pointwise the DAE. Excitations corresponding to classical solutions are at least continuous.

2.11.1 Measurable solutions

A straightforward generalization is now to turn to measurable solution functions x such that the part Dx is absolutely continuous, the initial condition makes sense owing to condition (2.158), and the DAE is satisfied for almost every $t \in \mathcal{I}$. The corresponding right-hand sides q are also measurable functions.
DAEs with excitations $q \in L_2(\mathcal{I}, \mathbb{R}^m)$ result, e.g., from Galerkin approximations of PDAEs (cf. [207], [203]). Furthermore, in optimization problems one usually applies measurable control functions.

We point out that regularity of the DAE, its characteristic values and the tractability index are determined from the coefficient functions A, D and B alone. Also the decoupling procedure is given in terms of these coefficient functions. Therefore, the regularity notion, the tractability index, characteristic values and the decoupling procedure retain their meaning also if we change the nature of the solutions and excitations.

We use the function space

$$H_D^1(\mathcal{I}, \mathbb{R}^m) := \{x \in L_2(\mathcal{I}, \mathbb{R}^m) : Dx \in H^1(\mathcal{I}, \mathbb{R}^n)\}$$

to accommodate the generalized solutions. For $x \in H_D^1(\mathcal{I}, \mathbb{R}^m)$ the resulting defect $q := A(Dx)' + Bx$ belongs to $L_2(\mathcal{I}, \mathbb{R}^m)$. Conversely, given an excitation $q \in L_2(\mathcal{I}, \mathbb{R}^m)$, it seems to make sense if we ask for IVP solutions from $H_D^1(\mathcal{I}, \mathbb{R}^m)$. The following proposition is a counterpart of Proposition 2.50. Roughly speaking,

in higher index cases, excitations are directed to the inherent regular ODE and the nullspace component only, which is ensured by means of the filtering projector $G_\mu P_1 \cdots P_{\mu-1} G_\mu^{-1}$.

Proposition 2.83. *Let the DAE (2.156) be fine with tractability index μ and characteristic values $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$. Put $d = m - \sum_{j=1}^{\mu}(m - r_{j-1})$, Let $Q_0, \ldots, Q_{\mu-1}$ be completely decoupling projectors. Set $V_1 := I$ and $V_\mu := G_\mu P_1 \cdots P_{\mu-1} G_\mu^{-1}$ if $\mu > 1$.*
Let the matrix C in the initial condition 2.157 have the property $\ker C = N_{can}(t_0)$.
Then, for every $z \in \operatorname{im} C$ and $q = V_\mu p$, $p \in L_2(\mathcal{I}, \mathbb{R}^m)$, the IVP (2.156), (2.157) has exactly one solution $x \in H_D^1(\mathcal{I}, \mathbb{R}^m)$.
If, additionally, the component $Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1} q = Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1} p$ is continuous, then the solution x is continuous.

Proof. We refer to Section 2.6 for details. Applying the decoupling and regarding condition $q = V_\mu p$, we arrive at the inherent regular ODE

$$u' - (D\Pi_{can} D^-)' u + D\Pi_{can} G_\mu^{-1} B D^- u = D\Pi_{can} G_\mu^{-1} p \tag{2.159}$$

and the solution expression

$$x = D^- u + Q_0 P_1 \cdots P_{\mu-1} G_\mu^{-1} p. \tag{2.160}$$

The initial value problem for (2.159) with the initial condition

$$u(t_0) = D(t_0) C^- z$$

has (cf. [79, pp. 166–167]) a continuous solution u with u' in $L_2(\mathcal{I}, \mathbb{R}^n)$, and which satisfies the ODE for almost all t. Then, by (2.160), x is in $H_D^1(\mathcal{I}, \mathbb{R}^m)$, and the initial condition (2.157) is fulfilled:

$$Cx(t_0) = CD(t_0)^- u(t_0) = CD(t_0)^- D(t_0) C^- z = CC^- z = z.$$

The second part of the assertion follows from the solution representation (2.160). □

If the tractability index equals 1, then $V_1 = I$ is valid, and hence the operator

$$L : H_D^1(\mathcal{I}, \mathbb{R}^m) \to L_2(\mathcal{I}, \mathbb{R}^m), \quad Lx := A(Dx)' + Bx,$$

is surjective. The corresponding IVP operator $\mathcal{L}$ is then a bijection.

If the DAE (2.156) is regular with tractability index 2, then by means of completely decoupling projectors we obtain the solution expression

$$x = D^- u + \Pi_0 Q_1 G_2^{-1} q + Q_0 P_1 G_2^{-1} q + Q_0 Q_1 D^- (D\Pi_0 Q_1 G_2^{-1} q)',$$

where u satisfies

$$u' - (D\Pi_1 D^-)' u + D\Pi_1 G_2^{-1} B D^- u = D\Pi_1 G_2^{-1} q, \quad u(t_0) = D(t_0) C^- z,$$

for the IVP (2.156), (2.157). In this way, the excitation q is supposed to be from the function space

$$\{q \in L_2(\mathcal{I}, \mathbb{R}^m) : D\Pi_0 Q_1 G_2^{-1} q \in H^1(\mathcal{I}, \mathbb{R}^n)\}.$$

Because of $V_2 = G_2 P_1 G_2^{-1}$ and $D\Pi_0 Q_1 G_2^{-1} V_2 = 0$, the excitation $q = V_2 p$ in the above proposition belongs to this space for trivial reasons.
Dealing with piecewise smooth excitations q, the solution expression shows how jumps are passed onto the solution.
We refer to Chapter 12 for a discussion of abstract differential equations, which also includes related results.
In all the above cases, suitably posed initial conditions play their role. If one replaces the initial condition (2.157) by the condition $x(t_0) = x_0$ we used to apply for regular ODEs, and which makes sense only for solutions being continuous, then the value x_0 must be consistent. Otherwise solvability is lost. As discussed in Subsection 2.6.2, a consistent value depends on the excitation.

2.11.2 Distributional solutions

The theory of *distributional solutions* allows us to elude the problem with inconsistent initial values and to consider discontinuous excitations q. We briefly address DAEs having $\mathcal{C}^\infty$-coefficients and a distributional excitation. For facts on generalized functions we refer to [201], [215].

Let $\mathfrak{D}$ denote the space of functions from $\mathcal{C}^\infty(\mathcal{I}, \mathbb{R})$ with compact support in $\mathcal{I}$, and $\mathfrak{D}'$ its dual space. The elements of $\mathfrak{D}'$ are said to be generalized functions or distributions. Denote by $\langle \cdot, \cdot \rangle$ the dual pairing between $\mathfrak{D}'$ and $\mathfrak{D}$.
For $y \in [\mathfrak{D}']^k$ and $\varphi \in [\mathfrak{D}]^k$, $k \in \mathbb{N}$, we define

$$\langle y, \varphi \rangle := \sum_{j=1}^{k} \langle y_j, \varphi_j \rangle.$$

For a matrix function $M \in \mathcal{C}^\infty(\mathcal{I}, L(\mathbb{R}^k, \mathbb{R}^l))$, $l, k \in \mathbb{N}$, and $y \in [\mathfrak{D}']^k$ we define the product $My \in [\mathfrak{D}']^l$ by

$$\langle My, \varphi \rangle = \langle y, M^* \varphi \rangle, \quad \forall \varphi \in [\mathfrak{D}]^l.$$

This product is well defined since $M^* \varphi$ belongs to $[\mathfrak{D}]^k$. For this, the $\mathcal{C}^\infty$ property of M is crucial.
Any distribution $y \in [\mathfrak{D}']^k$ possesses the distributional derivative $y' \in [\mathfrak{D}']^k$ defined by means of

$$\langle y', \varphi \rangle = -\langle y, \varphi' \rangle, \quad \forall \varphi \in [\mathfrak{D}]^k.$$

The product rule $(My)' = M'y + My'$ is valid.

Now we are prepared to consider distributional solutions of the given DAE (2.156) supposing its coefficient functions A, D, B have all entries belonging to $\mathcal{C}^\infty(\mathcal{I}, \mathbb{R})$.
Given a distributional excitation $q \in [\mathfrak{D}']^m$, a distribution $x \in [\mathfrak{D}']^m$ is said to be a *distributional solution* of the (generalized) DAE (2.156) if

$$\langle A(Dx)' + Bx, \varphi \rangle = \langle q, \varphi \rangle, \quad \forall \varphi \in [\mathfrak{D}]^m, \tag{2.161}$$

or, equivalently,

$$\langle x, -D^*(A^*\varphi)' + B^*\varphi \rangle = \langle q, \varphi \rangle, \quad \forall \varphi \in [\mathfrak{D}]^m. \tag{2.162}$$

Since the entries of A, D, B belong to $\mathcal{C}^\infty(\mathcal{I}, \mathbb{R})$, for regular DAEs, all admissible matrix functions and admissible projector functions have those entries, too. And hence the decoupling procedure described in Section 2.4 keeps its value also for the distributional solution. Every regular DAE possesses distributional solutions.

2.12 Notes and references

(1) For constant coefficient DAEs

$$\bar{E}\bar{x}'(t) + \bar{F}\bar{x}(t) = \bar{q}(t), \tag{2.163}$$

the Kronecker index and regularity are well defined via the properties of the matrix pencil $\{\bar{E}, \bar{F}\}$, and these characteristics are of particular importance in view of an appropriate numerical treatment. From about 1970, challenged by circuit simulation problems, numerical analysts and experts in circuit simulation begun to devote much work to the numerical integration of larger systems of implicit ODEs and DAEs (e.g., [86], [64], [202], [89]). In particular, linear variable coefficient DAEs

$$\bar{E}(t)\bar{x}'(t) + \bar{F}(t)\bar{x}(t) = \bar{q}(t) \tag{2.164}$$

were tackled by the implicit Euler method

$$\bar{E}(t_l)\frac{1}{h}(\bar{x}_l - \bar{x}_{l-1}) + \bar{F}(t_l)\bar{x}_l = \bar{q}(t_l).$$

Obviously, for the method to be just feasible, the matrix $\frac{1}{h}\bar{E}(t_l) + \bar{F}(t_l)$ must be nonsingular, but this can be guaranteed for all steps t_l and all sufficiently small stepsizes h, if one requires the so-called *local matrix pencils* $\{\bar{E}(t), \bar{F}(t)\}$ to be regular on the given interval (we mention at this point, that feasibility is by far not sufficient for a numerical integration method to work well). However, it was already discovered in [84] that the local pencils are not at all relevant characteristics of more general DAEs. Except for regular index-1 DAEs, local matrix pencils may change

their index and lose their regularity under smooth regular transformations of the variables. That means that the local matrix pencils $\{E(t),F(t)\}$ of the DAE

$$E(t)x'(t)+F(t)x(t)=q(t),\quad t\in\mathcal{I}, \tag{2.165}$$

which result from transforming $\bar{x}(t)=K(t)x(t)$ in the DAE (2.164), with a pointwise nonsingular continuously differentiable matrix function K, may have completely different characteristics from the local pencils $\{\bar{E}(t),\bar{F}(t)\}$. Nevertheless, the DAEs are equivalent, and hence, the local matrix pencils are irrelevant for determining the characteristics of a DAE. The coefficients of equivalent DAEs (2.164) and (2.165) are related by the formulas $E(t)=\bar{E}(t)K(t)$, $F(t)=\bar{F}(t)K(t)+\bar{E}(t)K'(t)$, which gives the impression that one can manipulate the resulting local pencil almost arbitrarily by choosing different transforms K.
In DAEs of the form

$$\bar{A}(t)(\bar{D}(t)\bar{x}(t))'+\bar{B}(t)\bar{x}(t)=\bar{q}(t), \tag{2.166}$$

the transformation $\bar{x}(t)=K(t)x(t)$ leads to the equivalent DAE

$$A(t)(D(t)x(t))'+B(t)x(t)=q(t). \tag{2.167}$$

The coefficients are related by $A(t)=\bar{A}(t)$, $D(t)=\bar{D}(t)K(t)$ and $B(t)=\bar{B}(t)K(t)$, and the local pencils $\{\bar{A}(t)\bar{D}(t),\bar{B}(t)\}$ and $\{A(t)D(t),B(t)\}=\{\bar{A}(t)\bar{D}(t)K(t),\bar{B}(t)K(t)\}$ are now equivalent. However, we do not consider this to justify the local pencils of the DAE (2.167) as relevant carriers of DAE essentials. For the DAE (2.167), also so-called *refactorizations of the leading term* yield equivalent DAEs, and any serious concept incorporates this fact. For instance, inserting $(Dx)'=(DD^+Dx)'=D(D^+Dx)'+D'D^+Dx$ does not really change the DAE (2.167), however, the local matrix pencils may change their nature as the following example demonstrates. This rules out the local pencils again. The DAE

$$\underbrace{\begin{bmatrix}0&1&0\\0&0&1\\0&0&0\end{bmatrix}}_{\tilde{A}}\Bigg(\underbrace{\begin{bmatrix}0&0&0\\0&1&0\\0&-t&1\end{bmatrix}}_{\tilde{D}(t)}x(t)\Bigg)'+\underbrace{\begin{bmatrix}1&0&0\\0&1&0\\0&-t&1\end{bmatrix}}_{\tilde{B}(t)}x(t)=q(t),\quad t\in\mathbb{R}, \tag{2.168}$$

has the local pencil $\{\tilde{A}\tilde{D}(t),\tilde{B}(t)\}$ which is regular with index 3. However, deriving

$$(\tilde{D}(t)x(t))'=(\tilde{D}(t)\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}x(t))'=\tilde{D}(t)\Bigg(\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}x(t)\Bigg)'+\tilde{D}'(t)\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix}x(t)$$

yields the equivalent DAE

$$\underbrace{\begin{bmatrix}0 & 1 & 0\\ 0 & -t & 1\\ 0 & 0 & 0\end{bmatrix}}_{A(t)}\Big(\underbrace{\begin{bmatrix}0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{bmatrix}}_{D}x(t)\Big)'+\underbrace{\begin{bmatrix}1 & 0 & 0\\ 0 & 0 & 0\\ 0 & -t & 1\end{bmatrix}}_{B(t)}x(t)=q(t),\quad t\in\mathbb{R},\tag{2.169}$$

the local matrix pencils $\{A(t)D,B(t)\}=\{E(t),F(t)\}$ of which are singular for all $t\in\mathbb{R}$.
We see, aiming for the characterization of a variable coefficient DAE, that it does not make sense to check regularity and index of the local pencils, neither for standard form DAEs nor for DAEs with properly stated leading term.

(2) Although in applications one commonly already has DAEs with proper leading term or standard form DAEs (2.165) the leading coefficient E of which has constant rank, there might be a different view of the constant-rank requirement for E seeing it as a drawback. In the early work on DAEs (cf. [39, 25]), the *standard canonical form* (SCF) of a DAE plays its role. By definition, the DAE (2.165) is in SCF, if it is in the form

$$\begin{bmatrix}I & 0\\ 0 & N(t)\end{bmatrix}x'(t)+\begin{bmatrix}W(t) & 0\\ 0 & I\end{bmatrix}x(t)=q(t),\tag{2.170}$$

where $N(t)$ is strictly lower (or upper) triangular. We emphasize that $N(t)$, and consequently $E(t)$, need not have constant rank on the given interval. Supposing the excitation q and the matrix function N are sufficiently smooth, this DAE has continuously differentiable solutions, and the flow does not show critical behavior.

In contrast, in our analysis, several constant-rank conditions play their role, in particular, each rank-changing point of $A(t)D(t)$ or $E(t)$ is considered as a critical point, that is, as a candidate for a point where something extraordinary with the solutions may happen. We motivate this opinion by Examples 2.69–2.71, among them also DAEs in SCF.

(3) The ambition to allow for matrix coefficients E with variable rank in a more general DAE theory is closely related to the SCF as well as to the derivative array approach (cf. [41]).
Given a DAE (2.165) with $m=k$ and coefficients $E,F\in\mathcal{C}^{2m}(\mathcal{I},L(\mathbb{R}^m))$, one considers the derivative array system (also, the prolongated or expanded system)

$$\underbrace{\begin{bmatrix}E(t) & 0 & . & . & . & 0\\ E'(t)+F(t) & E(t) & 0 & . & . & .\\ * & * & E(t) & . & . & .\\ . & . & . & . & . & .\\ * & * & * & * & * & E(t)\end{bmatrix}}_{\mathcal{J}_\kappa(t)}\begin{bmatrix}x^1\\ x^2\\ .\\ .\\ x^{\kappa+1}\end{bmatrix}=-\begin{bmatrix}F(t)\\ F'(t)\\ .\\ .\\ F^{(\kappa)}(t)\end{bmatrix}x+\begin{bmatrix}q(t)\\ q'(t)\\ .\\ .\\ q^{(\kappa)}(t)\end{bmatrix},\tag{2.171}$$

which results from (2.165) by formally differentiating this equation κ times, collecting all these equations, and replacing the derivative values $x^{(j)}(t)$ by jet variables x^j. The $(\kappa+1)m\times(\kappa+1)m$ matrix function $\mathcal{J}_\kappa$ is said to be *smoothly 1-full* on $\mathcal{I}$ ([25,

Definition 2.4.7]), if there is a smooth nonsingular matrix function $\mathcal{R}_\kappa$ such that

$$\mathcal{R}_\kappa(t)\mathcal{J}_\kappa(t) = \begin{bmatrix} I_m & 0 \\ 0 & \mathcal{K}(t) \end{bmatrix}.$$

If $\mathcal{J}_\kappa$ is smoothly 1-full, then an explicit vector field can be extracted from the derivative array system (2.171), say

$$x^1 = \mathcal{C}(t)x + \sum_{j=0}^{\kappa} \mathcal{D}_j(t)q^{(j)}(t).$$

The solution set of the DAE (2.165) is embedded into the solution set of the explicit ODE

$$x'(t) = \mathcal{C}(t)x(t) + \sum_{j=0}^{\kappa} \mathcal{D}_j(t)q^{(j)}(t), \tag{2.172}$$

which is called a *completion ODE* associated with the DAE, often also the *underlying ODE*.

In this context, one speaks (cf. [25, 41]) about *solvable systems* (2.165), if for every $q \in \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$ there exists at least one solution $x \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$, which is uniquely determined by its value at any $t \in \mathcal{I}$. Any DAE that is transformable into SCF is solvable in this sense. For every such solvable system, there is an index $\kappa \le m$ such that the derivative array matrix function $\mathcal{J}_\kappa$ has constant rank and is smoothly 1-full. The reverse statement is true under additional assumptions (cf. [25, 41]).

If N in the SCF (2.170) is the zero matrix, then the leading coefficient of this DAE has constant rank. Correspondingly, if the matrix function $\mathcal{J}_1$ has constant rank and is smoothly 1-full on $\mathcal{I}$, then E has constant rank. Namely, we have here

$$\mathcal{R}\mathcal{J}_1 = \mathcal{R}\begin{bmatrix} E & 0 \\ E'+F & E \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & \mathcal{K} \end{bmatrix}.$$

The block $\mathcal{K}$ has constant rank since $\mathcal{J}_1$ has. Now, E has constant rank because of

$$\mathcal{R}\begin{bmatrix} 0 \\ E \end{bmatrix} = \begin{bmatrix} 0 \\ \mathcal{K} \end{bmatrix}.$$

It becomes clear that the leading coefficient E of a *solvable system* (2.165) may undergo rank changes only in so-called higher index cases, that is, if $\kappa \ge 2$ is the lowest index such that $\mathcal{J}_\kappa$ has constant rank and is smoothly 1-full.

To illustrate what is going on we revisit the simple SCF-DAE

$$\begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix} x' + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x = q$$

given on the interval $\mathcal{I} = [-1, 1]$. The function α is a strictly positive on $(0, 1]$ and vanishes identically on $[-1, 0]$. Suppose α to be four times continuously differentiable. Notice that, in contrast, in Example 2.70 we only need continuous α. We

form the derivative array functions

$$\mathcal{J}_1 = \begin{bmatrix} 0 & \alpha & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & \alpha' & 0 & \alpha \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathcal{J}_2 = \begin{bmatrix} 0 & \alpha & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & \alpha' & 0 & \alpha & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & \alpha'' & 1 & \alpha' + \alpha'' & 0 & \alpha \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

These matrix functions have constant-rank two, respectively four. Multiplication by the smooth nonsingular matrix functions

$$\mathcal{R}_1 = \begin{bmatrix} 0 & 0 & 1 & -\alpha' \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & -\alpha \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \mathcal{R}_2 = \begin{bmatrix} 0 & 0 & 1 & -\alpha' & 0 & -\alpha \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & -\alpha & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\alpha'' & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

yields

$$\mathcal{R}_1\mathcal{J}_1 = \begin{bmatrix} 1 & 0 & 0 & \alpha \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{R}_2\mathcal{J}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & \alpha' + \alpha'' & 0 & \alpha \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}.$$

The derivative array function $\mathcal{J}_2$ is smoothly 1-full on the entire interval $\mathcal{I} = [-1,1]$ independently of the behavior of the function α. On the other hand, 1-fullness on $\mathcal{I}$ does not apply to $\mathcal{J}_1$. However, $\mathcal{J}_1$ is smoothly 1-full on the subinterval $[-1,0)$, where α vanishes identically. It becomes clear that the restriction of the DAE onto a subinterval does not necessarily show the same characteristic. A more rigorous characterization of the DAE would depend on the interval.

We stress once again that we aim for a DAE analysis including a regularity notion, which meets the lowest possible smoothness demands, and that we have good reasons for treating the rank-changing points of the leading coefficient as critical points and for regarding regularity intervals.
From our point of view, *regularity* of linear DAEs comprises the following three main aspects (cf. Definition 2.25):

(a) The homogeneous equation has a solution space of finite dimension d.
(b) Equations restricted to subintervals inherit property (a) with the same d.
(c) Equations restricted to subintervals inherit the further characteristics $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$.

(4) Regularity is an often applied notion in mathematics to characterize quite diverse features. Also, different regularity notions are already known for linear DAEs.

They refer to different intentions and are not consistent with each other. We pick up some of them.

Repeatedly (e.g., [129, 130]) regularity of linear DAEs is bound to the unique solvability of initial value problems for *every sufficiently smooth excitation* and consistent initial conditions. Note that this property is named *solvability*, e.g., in [41, 25].

In [25] the linear DAE is said to be regular, if the local matrix pencils remain regular, a property that is helpful for numerical integration.

In [189] the ordered matrix function pair $\{E,F\}$ is said to be regular, if $E(t)$ has constant rank $r < m$ and $E(t)E(t)^* + F(t)F(t)^*$ is nonsingular, a property that is useful for the reduction procedure. A DAE showing a regular coefficient pair is then named *reducible*. So, for instance, the constant coefficient pair

$$\left\{ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} \right\}$$

is regular in [189], but fails to be regular in [25].

Moreover, apart from higher smoothness demands, *complete reducibility* of the DAE and *complete regularity* of the pair $\{E,F\}$ in [189] are consistent with our regularity notion. Namely, it is proved in [189] by a careful inspection of all involved constant-rank requirements that complete reducibility is in full agreement with the conditions yielding a *well-defined regular strangeness index* (cf. Section 2.10). In turn, as shown in Section 2.10, the rank conditions supporting regularity in the tractability index context are also in agreement with those from the regular strangeness index.

(5) Distributional solutions of linear DAEs with constant coefficients were studied very early, e.g., [53]. Generalized (distributional) solutions of linear DAEs with smooth variable coefficients have been worked out in [187] (also [189, Chapter III]), whereby so-called impulsive-smooth distributions play a central role. Recently, also time-varying DAEs with discontinuous coefficients are on the agenda, see [208].

(6) Further solution generalizations such as *weak solutions* result from the special settings of partial differential-algebraic equations (PDAEs) and abstract differential-algebraic equations (ADAEs), see, e.g., [207, 191] and references therein.

(7) There are simple interrelations between standard form DAEs (2.165) and DAEs (2.167) with proper leading term.

If D is continuously differentiable, we rewrite the DAE (2.167) as

$$A(t)D(t)x'(t) + (B(t) + A(t)D'(t))x(t) = q(t), \tag{2.173}$$

which has standard form. If equation (2.167) has a properly stated leading term, the resulting matrix function $E = AD$ has constant-rank and the variable subspace $\ker E = \ker D$ is a $\mathcal{C}^1$-subspace.

Conversely, if a standard form DAE (2.165) with a constant rank matrix function E is given and, additionally, $\ker E$ is a $\mathcal{C}^1$-subspace, then, taking a continuously differentiable projector valued function P, $\ker P = \ker E$, we may write

$Ex' = EPx' = E(Px)' - EP'x$. In this way we obtain

$$E(t)(P(t)x(t))' + (F(t) - E(t)P'(t))x(t) = q(t), \tag{2.174}$$

which is a DAE with properly stated leading term. Now it is evident that any DAE (2.167) with a properly stated leading term and a continuously differentiable matrix function D yields a standard form DAE (2.165) such that the leading coefficient E has constant rank and $\ker E$ is a $\mathcal{C}^1$-subspace, and vice versa.

Moreover, there are various possibilities to factorize a given matrix function E, and to rewrite a standard form DAE as a DAE with proper leading term.
If the matrix function E itself is continuously differentiable and has constant rank, then its nullspace is necessarily a $\mathcal{C}^1$-subspace, so that we may use equation (2.174). Additionally in this case, by taking any continuously differentiable generalized inverse E^- and by writing $Ex' = EE^-Ex' = EE^-(Ex)' - EE^-E'x$ we form

$$E(t)E(t)^-(E(t)x(t))' + (F(t) - E(t)E(t)^-E'(t))x(t) = q(t)$$

which is also a DAE with properly stated leading term.
Furthermore, very often the original DAE consists of two kinds of equations, those containing derivatives and those which are derivative-free. Then, the matrix function E has the special form

$$E(t) = \begin{bmatrix} E_1(t) \\ 0 \end{bmatrix}, \quad \operatorname{rank} E_1(t) = \operatorname{rank} E(t),$$

or can be easily brought into this form. In this case, we can simply turn to

$$\begin{bmatrix} I \\ 0 \end{bmatrix} \big(E_1(t)x(t)\big)' + \Big(F(t) - \begin{bmatrix} E_1'(t) \\ 0 \end{bmatrix}\Big)x(t) = q(t).$$

We also point out the following full-rank factorization on a compact interval $\mathcal{I}$, which is provided by a continuous singular value decomposition (e.g. [49]),

$$E(t) = \begin{bmatrix} U_{11}(t) & U_{12}(t) \\ U_{21}(t) & U_{22}(t) \end{bmatrix} \begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} V_{11}(t) & V_{12}(t) \\ V_{21}(t) & V_{22}(t) \end{bmatrix}^* = \underbrace{\begin{bmatrix} U_{11}(t) \\ U_{21}(t) \end{bmatrix} \Sigma(t)}_{A(t)} \underbrace{\big[V_{11}^*(t) \; V_{21}^*(t)\big]}_{D(t)},$$

$\operatorname{rank} \Sigma(t) = \operatorname{rank} E(t) =: r$, $n = r$. The factors U, Σ and V are continuously differentiable, supposing E is so. Then, $A(t)$ has full column rank n and $D(t)$ has full row rank n.

As in the constant coefficient case, the special form of the factorization does not matter for the nature of the solution. Only the nullspace $\ker D = \ker E$ specifies what a solution is. Namely, for every two matrix functions $D \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^n)$ and $\bar{D} \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^{\bar{n}})$ with constant rank and the common nullspace $N := \ker D = \ker \bar{D}$, it holds that

$$\mathcal{C}_{\bar{D}}^1(\mathcal{I}, \mathbb{R}^m) = \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m).$$

Since the Moore–Penrose inverses D^+ and $\bar{D}^+$ are continuously differentiable, too, for any $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ we find $\bar{D}x = \bar{D}\bar{D}^+\bar{D}x = \bar{D}D^+Dx \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^{\bar{n}})$, and hence $x \in \mathcal{C}_{\bar{D}}^1(\mathcal{I},\mathbb{R}^m)$.

(8) Our stability analysis of linear DAEs as well as the stability issues in Part II concerning nonlinear index-1 DAEs carry forward the early ideas of [146, 96] as well as fruitful discussions on a series of special conferences on stability issues. The basic tool of our analysis is the explicit description of the unique IERODE defined by fine decouplings. We emphasize that we do not transform the given DAE, but work with the originally given data. In [17] they try another way of considering DAEs via transformation into SCF and proposing Lyapunov functions.

(9) In Section 2.2 we provide the admissible matrix function sequences and admissible projector functions together with their main properties. This part generalizes the ideas of [167], [170]. While [167], [170] are restricted to regular DAEs, we now give an adequate generalization for systems of k equations for m unknowns. The new preliminary rearrangement of the DAE terms for better structural insight in Subsection 2.4.1 is also valid for nonregular DAEs. We discuss this topic in Chapter 10 for over- and underdetermined DAEs. We emphasize once again that we only rearrange terms in the given setting, but we do not transform the DAE at all.

The discussion of regular and fine DAEs renews the ideas of [170] and [169], while the discussion of critical points reflects parts of [173, 174, 194], but we apply a relaxed notion of regular points by the introduction of quasi-proper leading terms. [194] is the first monograph offering a comprehensive introduction to the projector based decoupling of regular linear DAEs, both in standard form and with proper leading term. Moreover, this book considers critical points in the context of DAEs having almost overall uniform characteristics.

(10) In the present chapter we describe harmless critical points somewhat loosely as those which disappear in a smoother setting. A precise investigation on the background of the concept of *quasi-regular* DAEs (cf. Chapter 9) can be found in [59].

Chapter 3
Nonlinear DAEs

The objective of this chapter is a rigorous analysis of a large class of DAEs

$$f((d(x(t),t))',x(t),t) = 0$$

on a low smoothness level by means of admissible projector functions. In contrast to the usually applied derivative array approaches and reduction procedures, we do without those derivative arrays. We also do without providing solutions prior to and involving them into the characterization of the equation.

The chapter is organized as follows. We describe the basic assumptions, the setting of DAEs with properly involved derivative, their constraints and what we consider to be a linearization in Section 3.1. Section 3.2 provides admissible matrix function sequences and admissible projector functions, as well as their essential properties, as pointwise generalizations of the admissible matrix function sequences and admissible projector functions already defined in Chapter 2 on linear DAEs. In Section 3.3 we introduce *regularity regions* and provide necessary and sufficient regularity conditions via linearizations. We consider this to be the main result of the present chapter. It says that a DAE is regular with tractability index μ and characteristic values $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, if all its corresponding linearizations are regular with these characteristics, and vice versa. Characteristic values and regularity regions are shown to be invariant under transformations in Section 3.4. In this context, a *DAE having a well-defined index*, which is commonly supposed in the literature, appears to be a DAE comprising just one single regularity region. Therefore, the class of DAEs showing several regularity regions appears to be quite large. Nevertheless, Section 3.8 addresses the need for a further generalization of the DAE class by an advanced localization of regularity regarding the jet variables.
Section 3.5 deals with the special structure of Hessenberg form DAEs and verifies the full agreement of the tractability index concept with the trusted knowledge on Hessenberg DAEs. For DAEs arising in circuit simulation which are studied in Section 3.6, it is shown how the structure of the circuit can be exploited to reach useful information and to build the admissible projector functions, and then to provide the DAE characteristics.

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_3,

We prove strong local solvability assertions in Section 3.7. In Section 3.9 we derive perturbation results for DAEs having a linear derivative part by means of an operator setting. Section 3.11 offers hints to ease models. We discuss relations to the differentiation index in Section 3.10.

3.1 Basic assumptions and notions

3.1.1 Properly involved derivative

In this chapter we investigate general nonlinear equations

$$f((d(x(t),t))',x(t),t) = 0, \tag{3.1}$$

which satisfy the following assumption.

Assumption 3.1. *The function $f : \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f \longrightarrow \mathbb{R}^k$ is continuous on the open set $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f \subseteq \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$ and has continuous partial derivatives f_y, f_x with respect to the first two variables $y \in \mathbb{R}^n$, $x \in \mathcal{D}_f$.*
The function $d : \mathcal{D}_f \times \mathcal{I}_f \longrightarrow \mathbb{R}^n$ is continuously differentiable.

DAEs in the form (3.1) arise, for instance, in circuit simulation by means of the modified nodal analysis on a big scale (cf. Section 3.6).
Involving the derivative by means of an extra function into the DAE brings benefits in view of solvability.

Definition 3.2. A *solution x_* of equation (3.1)* is a continuous function defined on an interval $\mathcal{I}_* \subseteq \mathcal{I}_f$, with values $x_*(t) \in \mathcal{D}_f$, $t \in \mathcal{I}_*$, such that the function $u_*(.) := d(x_*(.),.)$ is continuously differentiable, and x_* satisfies the DAE (3.1) pointwise on $\mathcal{I}_*$.

In our context, the wording *the DAE is solvable* simply means the existence of a solution in this sense. In contrast, mostly in the literature on DAEs, solvability of a DAE means the existence of a continuously differentiable function satisfying the DAE pointwise. As for linear DAEs, one can expect lower smoothness solvability results also for nonlinear DAEs (3.1).

Example 3.3 (Solvability benefit). The DAE

$$\begin{aligned} (x_1(t) + x_2(t)x_3(t))' - q_1(t) &= 0,\\ x_2(t) - q_2(t) &= 0,\\ x_3(t) - q_3(t) &= 0, \quad t \in \mathcal{I}, \end{aligned}$$

has the form (3.1) with $k = m = 3$, $n = 1$,

$$f(y,x,t) := \begin{bmatrix} 1\\0\\0 \end{bmatrix} y + \begin{bmatrix} 0\\x_2\\x_3 \end{bmatrix} - q(t), \quad d(x,t) := x_1 + x_2x_3, \quad x \in \mathbb{R}^3,\ t \in \mathcal{I},\ y \in \mathbb{R}.$$

For each given continuous q and fixed $\bar{t} \in \mathcal{I}$, $\bar{c} \in \mathbb{R}$, this DAE has the solution

$$\begin{aligned} x_1(t) &= -q_2(t)q_3(t) + \bar{c} + q_2(\bar{t})q_3(\bar{t}) + \int_{\bar{t}}^{t} q_1(s)\mathrm{d}s, \\ x_2(t) &= q_2(t), \\ x_3(t) &= q_3(t), \quad t \in \mathcal{I}, \end{aligned}$$

which satisfies $x(\bar{t}) = \bar{x}$, $\bar{x}_1 := \bar{c}$, $\bar{x}_i := q_i(\bar{t})$, $i = 2,3$. Obviously, the second and third solution components are not necessarily continuously differentiable. Later on we refer to such a system as a regular index-1 DAE. Observe that $\ker f_y = \{0\}$, $\operatorname{im} d_x = \mathbb{R}$.
In contrast, rewriting this DAE in standard form as

$$\begin{aligned} x_1'(t) + x_3(t)\, x_2'(t) + x_2(t)\, x_3'(t) - q_1(t) &= 0, \\ x_2(t) - q_2(t) &= 0, \\ x_3(t) - q_3(t) &= 0, \quad t \in \mathcal{I}, \end{aligned}$$

we are led to solvability only for at least continuously differentiable q_2, q_3. □

Equation (3.1) covers linear DAEs (2.1) via

$$f(y,x,t) = A(t)y + B(t)x - q(t), \quad d(x,t) = D(t)x.$$

Semi-explicit systems

$$x_1'(t) + b_1(x_1(t), x_2(t), t) = 0, \tag{3.2}$$
$$b_2(x_1(t), x_2(t), t) = 0, \tag{3.3}$$

where the unknown function is partitioned into the two components $x_1(.)$, $x_2(.)$, and the derivative of the second component is absent, and, additionally, a part of the equations is derivative-free, represent the special case with

$$f(y,x,t) = \begin{bmatrix} I \\ 0 \end{bmatrix} y + b(x,t), \quad d(x,t) = x_1, \quad y \in \mathbb{R}^n,\ x \in \mathcal{D}_b =: \mathcal{D}_f,\ t \in \mathcal{I}_b =: \mathcal{I}_f.$$

Many authors restrict themselves to semi-explicit DAEs; often one deals with so-called Hessenberg form DAEs which are particular cases of semi-explicit DAEs.
Semi-explicit systems yield

$$f_y(y,x,t) = \begin{bmatrix} I \\ 0 \end{bmatrix} \in L(\mathbb{R}^n, \mathbb{R}^m), \quad d_x(x,t) = \begin{bmatrix} I & 0 \end{bmatrix} \in L(\mathbb{R}^m, \mathbb{R}^n),$$

$$\ker f_y(y,x,t) \oplus \operatorname{im} d_x(x,t) = \{0\} \oplus \mathbb{R}^n = \mathbb{R}^n, \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f,$$

which is typical for properly stated terms.
The following notion of a properly involved derivative generalizes this property.

Definition 3.4. Let the DAE (3.1) satisfy Assumption 3.1. The DAE (3.1) has on $\mathcal{D}_f \times \mathcal{I}_f$ a *properly involved derivative*, also called a *properly stated leading term*, if $\operatorname{im} d_x$ and $\ker f_y$ are $\mathcal{C}^1$-subspaces in $\mathbb{R}^n$, and the transversality condition

$$\ker f_y(y,x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^n, \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f, \tag{3.4}$$

holds.

Variable subspaces moving in $\mathbb{R}^n$, in particular $\mathcal{C}^1$-subspaces, are described in Appendix A.4. Any $\mathcal{C}^1$-subspace necessarily has constant dimension. Therefore, if the DAE has a properly involved derivative, then the partial derivatives $f_y(y,x,t)$ and $d_x(x,t)$ have constant rank on their definition domain. Denote

$$r := \operatorname{rank} d_x(x,t). \tag{3.5}$$

Due to the transversality condition (3.4), $f_y(y,x,t)$ has rank r, too.

From our point of view it makes good sense to figure out the term housing the derivative in a rigorous way, i.e., to use an improved model (3.1) compared with a so-called standard DAE

$$\mathfrak{f}(x'(t),x(t),t) = 0 \tag{3.6}$$

that leaves it undecided which derivatives are actually involved. Example 3.3 shows a DAE with properly stated leading term and also any semi-explicit system has a properly stated leading term, too. Both cases show trivial decomposition of $\mathbb{R}^n$.

The general first-order form of the equation describing the motion of a constrained multibody system is actually a special semi-explicit DAE. The circuit models described in Section 3.6 also yield DAEs with properly stated leading term. It seems that properly involved derivatives in the mathematical sense reflect the physical nature in basic applications well.

Whereas in properly stated leading terms both matrix functions $f_y(y,x,t)$ and $d_x(x,t)$ have necessarily constant rank, the matrix function $f_y(y,x,t)$ is allowed to change its rank in so-called quasi-proper leading terms in Chapter 9. Already in Chapter 2 concerning linear DAEs, we threw some light on different views concerning the constant-rank requirements, and of course, all arguments keep their value for nonlinear DAEs, too. At this place we emphasize once more the role of properly stated leading terms in detecting critical points on early stages. We demonstrate this by the next two examples.

Example 3.5 (Critical point detection). The system

$$\begin{aligned} x_1(t)\, x_1'(t) - x_2(t) &= 0, \\ x_1(t) - x_2(t) &= 0, \end{aligned}$$

possesses the solutions $x_{*1}(t) = x_{*2}(t) = t + c$, where c denotes an arbitrary real constant. Additionally, the identically vanishing function $\bar{x}_*(t) = 0$ is also a solution. Through every point on the diagonal line $x_1 = x_2$, except for the origin, there passes exactly one solution. However, two different solutions satisfy the initial condition

$x(0) = 0$, which characterizes the origin as a critical point. Writing the DAE in the form (3.1) with $n = 1$, $m = k = 2$, $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = \mathbb{R}$,

$$f(y,x,t) = \begin{bmatrix} x_1 y - x_2 \\ x_1 - x_2 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}, \quad d(x,t) = x_1, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

one arrives at a DAE which fails to have a properly stated leading term on the given definition domain $\mathcal{D}_f \times \mathcal{I}_f$. However, the leading term is stated properly on the open set

$$\{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x_1 \neq 0\},$$

where f_y keeps constant rank. This emphasizes the constant-rank condition (and the proper leading term setting) to be helpful in indicating critical points. □

So-called *quasi-linear equations*

$$A(x(t),t)(d(x(t),t))' + b(x(t),t) = 0 \tag{3.7}$$

are accommodated in (3.1) with $f(y,x,t) = A(x,t)y + b(x,t)$, $d(x,t) = D(t)x$. Quasi-linear DAEs have an extra term housing the derivative, and formally *leading* the equation. This justifies the name *properly stated leading term*. In a more general equation, we might not have a leading term. Then the notion *properly involved derivative* is more appropriate. However, we also keep using the more traditional notions *properly stated leading term* and *proper leading term* for the general nonlinear case.

A generally quite favorable version of a properly involved derivative term is given if

$$\ker f_y(y,x,t) = \{0\}, \quad \operatorname{im} d_x(x,t) = \mathbb{R}^n,$$

which means that the partial derivatives $f_y(y,x,t)$ and $d_x(x,t)$ have full column rank and full row rank, respectively, as it is the case for semi-explicit systems.

3.1.2 Constraints and consistent initial values

In Example 3.3, all solution values at time t must belong to the set

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^2 : x_2 = q_2(t), x_3 = q_3(t)\}.$$

This is a typical feature of DAEs, which consists in the fact that the flow determined by the DAE is restricted to certain lower-dimensional sets, the *constraint sets* (constrained manifold, if a manifold structure is given). This differs essentially from the situation given for regular ODEs where no such restrictions are met. In DAEs, certain components are completely fixed by others and the excitations. Just for some part of the components, if any, free integration constants allow for additional initial or boundary conditions and an actual flow. If an IVP of the form

$$f((d(x(t),t))',x(t),t)=0, \quad x(t_0)=x_0 \tag{3.8}$$

should be solvable, the initial value x_0 must meet these constraints. For instance, for getting an integration routine started, one needs suitable initial values. In general, as we shall see later, one needs a deep knowledge of the DAE structure to be able to formulate, e.g., initial conditions such that the IVP becomes solvable.

Definition 3.6. For a given DAE (3.1) and given $t_0 \in \mathcal{I}_f$, the value $x_0 \in \mathcal{D}_f$ is said to be a *consistent initial value* if the IVP (3.8) possesses a solution.

Except for transparent academic examples and very special situations, one can provide consistent initial values just approximately. This is a difficult task that should not be underrated.
For linear DAEs (2.1), all solution values at time $t \in \mathcal{I}$ obviously belong to the constraint set

$$\mathcal{M}_0(t)=\{x\in\mathbb{R}^m : B(t)x-q(t)\in \mathrm{im}A(t)\}.$$

However, the precise set of consistent values at time t is given by (2.98), that is

$$\mathcal{M}_{can,q}(t)=\{z+v(t) : z\in S_{can}(t)\}\subseteq \mathcal{M}_0(t).$$

Recall that the function v vanishes identically if q does, which implies $\mathcal{M}_{can,q}(t)=S_{can}(t)$.
We already know that $\mathcal{M}_{can,q}(t)=\mathcal{M}_0(t)$ exactly if the linear DAE is regular with index 1. In the case of higher index DAEs, $\mathcal{M}_{can,q}(t)$ is a lower-dimensional subset of $\mathcal{M}_0(t)$.
Not surprisingly, the situation is quite involved in nonlinear equations.

Example 3.7 (All values in $\mathcal{M}_0(t)$ are consistent). Consider the semi-explicit DAE

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

with two equations and two unknown functions on $\mathcal{D}_f=\{x\in\mathbb{R}^2 : x_2>0\}$, $\mathcal{I}_f=\mathbb{R}$. The real function γ is continuous on $\mathcal{I}_f$. We may write this DAE in the form (3.1) with $n=1$, $m=k=2$,

$$f(y,x,t)=\begin{bmatrix} y+x_1 \\ x_1^2+x_2^2-\gamma(t)-1 \end{bmatrix}, \quad f_y(y,x,t)=\begin{bmatrix}1\\0\end{bmatrix}, \quad d(x,t)=x_1, \quad d_x(x,t)=\begin{bmatrix}1 & 0\end{bmatrix},$$

yielding a DAE with properly stated leading term. Every solution value $x_*(t)$ must lie in the set

$$\mathcal{M}_0(t)=\{x\in\mathcal{D}_f : (x_1)^2+(x_2)^2-1-\gamma(t)=0\}.$$

Therefore, this set must be nonempty, which means the function γ satisfies $1+\gamma(t)>0$. Otherwise, the DAE has no solutions on those parts. The solution passing through $x_1(0)=x_{0,1}$, $x_2(0)=\sqrt{1-x_{0,1}^2+\gamma(0)}$ can be expressed as

$$x_{*1}(t) = e^{-t}x_{0,1}, \quad x_{*2}(t) = \sqrt{1 - e^{-2t}x_{0,1}^2 + \gamma(t)}.$$

Through each point of the set $\mathcal{M}_0(0)$ there passes exactly one solution at time $t = 0$, hence, the points of $\mathcal{M}_0(0)$ are consistent. Furthermore, the component $x_{0,1}$ serving as an integration constant can be chosen freely as long as $1 - x_{0,1}^2 + \gamma(0) > 0$ holds. Figure 3.1 shows, for $\gamma(t) \equiv 0$ and for $\gamma(t) = t^2$, the path of the solution $(x_1(t), x_2(t))$, $t \in [0,1]$, for the initial value $x_{0,1} = 0.98$, and the sets $\mathcal{M}_0(0)$ and $\mathcal{M}_0(1)$.

□

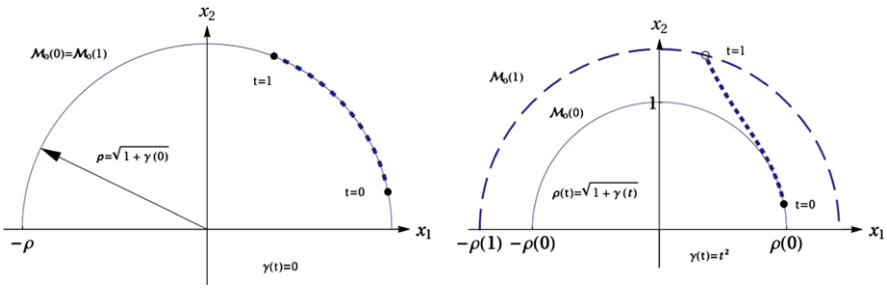


Fig. 3.1 Solution for $\gamma(t) \equiv 0$ and $\gamma(t) = t^2$ in Example 3.7

Example 3.8 (Hidden constraint). The DAE

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

given on $\mathcal{D}_f = \{x \in \mathbb{R}^3 : x_2 > 0\}$, $\mathcal{I}_f = \mathbb{R}$, with γ being continuously differentiable and satisfying $1 + \gamma(t) > 0$, looks quite similar to the previous one. We may write this DAE in the form (3.1), where $n = 2$, $m = k = 3$,

$$f(y,x,t) = \begin{bmatrix} y_1 + x_1 \\ x_2 y_2 - x_3 \\ x_1^2 + x_2^2 - \gamma(t) - 1 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 1 & 0 \\ 0 & x_2 \\ 0 & 0 \end{bmatrix},$$

$$d(x,t) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

yielding again a DAE with properly stated leading term. The solution values must belong to the set

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

However, a closer look at this DAE makes it clear that there is another set the solution values have to belong to. Namely, for any solution $x_*(.)$, differentiating the

identity $x_{*1}(t)^2 + x_{*2}(t)^2 - 1 = \gamma(t)$ and replacing the expressions for the derivatives we obtain the new identity

$$-2x_{*1}(t)^2 + 2x_{*3}(t) = \gamma'(t).$$

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

$$\mathcal{H}(t) := \{x \in \mathcal{D}_f : -2x_1^2 + 2x_3 - \gamma'(t) = 0\}.$$

In consequence, the obvious constraint set $\mathcal{M}_0(t)$ contains points which are no longer consistent, but the proper subset $\mathcal{M}_1(t) := \mathcal{M}_0(t) \cap \mathcal{H}(t) \subset \mathcal{M}_0(t)$ consists of consistent points. Figure 3.2 shows $M_1(t)$ for $\gamma(t) = -\frac{1}{2}\cos \pi t$ for $t = 0$ and $t = \frac{1}{2}$. □

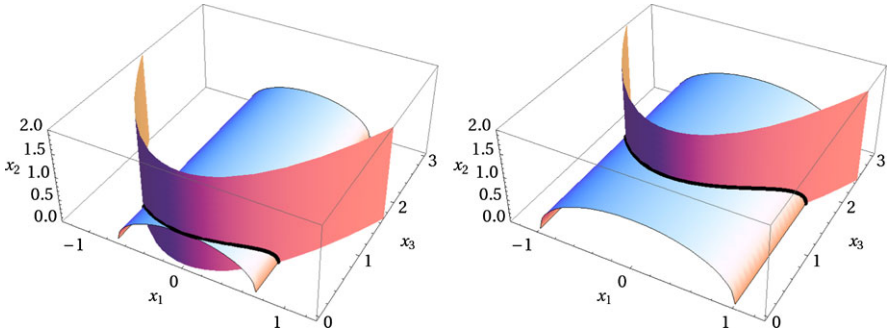


Fig. 3.2 $\mathcal{M}_1$ at $t = 0$ and $t = \frac{1}{2}$ in Example 3.8

Turn back to general DAEs (3.1) with properly involved derivative. The function d is continuously differentiable, and d_x has constant rank. By definition, a solution x_* is a continuous function on a certain interval $\mathcal{I}_*$ which has there a continuously differentiable part $u_*(.) := d(x_*(.),.)$. If x_* were to belong to class $\mathcal{C}^1$, then the identity

$$u_*'(t) - d_t(x_*(t),t) = d_x(x_*(t),t)x_*'(t),\ t \in \mathcal{I}_*,$$

would be given. Although we do not suppose the solutions to be from $\mathcal{C}^1$, the inclusion

$$u_*'(t) - d_t(x_*(t),t) \in \operatorname{im} d_x(x_*(t),t),\ t \in \mathcal{I}_*, \tag{3.9}$$

remains to be valid for all solutions (cf. Proposition C.1), and there is a continuous function (not necessarily unique) w_* such that

$$u_*'(t) = d_x(x_*(t),t)w_*(t) + d_t(x_*(t),t),\ t \in \mathcal{I}_*.$$

The inclusion (3.9) holds trivially, if $d_x(x,t)$ has full row rank. In general, for every solution of the DAE (3.1) the two identities

$$f(u_*'(t),x_*(t),t)=0, \quad t\in\mathcal{I}_*,$$

and

$$f(d_x(x_*(t),t)w_*(t)+d_t(x_*(t),t),x_*(t),t)=0, \quad t\in\mathcal{I}_*,$$

are valid, and hence, for all solutions, their values $x_*(t)$ must belong to the sets

$$\widetilde{\mathcal{M}}_0(t):=\{x\in\mathcal{D}_f : \exists y\in\mathbb{R}^n : f(y,x,t)=0\} \tag{3.10}$$

and

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

The sets $\mathcal{M}_0(t)$ and $\widetilde{\mathcal{M}}_0(t)$ are defined for all $t\in\mathcal{I}_f$. Eventually, they might be empty. The inclusion $\mathcal{M}_0(t)\subseteq\widetilde{\mathcal{M}}_0(t)$ is evident. For DAEs yielding a full row rank matrix function d_x, as is the case in Examples 3.7 and 3.8, these sets coincide. Both sets are obviously restriction sets or constraints for the DAE solutions.

Definition 3.9. For a DAE (3.1) with proper leading term, the set

$$\mathcal{M}_0(t):=\{x\in\mathcal{D}_f : \exists y\in\mathbb{R}^n : y-d_t(x,t)\in\operatorname{im} d_x(x,t),\ f(y,x,t)=0\}$$

is called the *obvious restriction set* or the *obvious constraint* of the DAE at $t\in\mathcal{I}_f$.

Turn for a moment to the special, but large class of quasi-linear DAEs (3.7). Remember that this class covers at least DAEs arising in circuit simulation and multibody systems. Denote by $\mathcal{W}_0(x,t)$ a projector matrix such that $\ker\mathcal{W}_0(x,t)=\operatorname{im}A(x,t)$. We represent

$$\begin{aligned}\widetilde{\mathcal{M}}_0(t)&=\{x\in\mathcal{D}_f : \exists y\in\mathbb{R}^n : A(x,t)y+b(x,t)=0\}\\ &=\{x\in\mathcal{D}_f : b(x,t)\in\operatorname{im}A(x,t)\}=\{x\in\mathcal{D}_f : \mathcal{W}_0(x,t)b(x,t)=0\},\end{aligned}$$

and

$$\begin{aligned}\mathcal{M}_0(t)&=\{x\in\mathcal{D}_f : \exists w\in\mathbb{R}^m : A(x,t)(d_x(x,t)w+d_t(x,t))+b(x,t)=0\}\\ &=\{x\in\mathcal{D}_f : b(x,t)\in\operatorname{im}A(x,t)\}=\{x\in\mathcal{D}_f : \mathcal{W}_0(x,t)b(x,t)=0\}=\widetilde{\mathcal{M}}_0(t).\end{aligned}$$

The sets $\mathcal{M}_0(t)$ and $\widetilde{\mathcal{M}}_0(t)$ coincide, and they are described by means of an equation which is more comfortable in some sense.

Observe that, for given t and $x\in\mathcal{M}_0(t)$, the corresponding y is uniquely determined, if the leading term is properly stated. Namely, for fixed $t\in\mathcal{I}_f, x\in\mathcal{M}_0(t)$, and $w,\bar{w}\in\mathbb{R}^m$ with

$$A(x,t)\underbrace{\big(d_x(x,t)w+d_t(x,t)\big)}_{=y}+b(x,t)=0,\quad A(x,t)\underbrace{\big(d_x(x,t)\bar{w}+d_t(x,t)\big)}_{=\bar{y}}+b(x,t)=0,$$

we derive $A(x,t)d_x(x,t)(w-\bar{w})=0$, and hence $y-\bar{y}=d_x(x,t)(w-\bar{w})=0$ owing to the property $\ker Ad_x=\ker d_x$.
The latter property holds true in general as the next proposition states.

Proposition 3.10. *If the DAE (3.1) has a properly involved derivative, then, for each $x\in\mathcal{M}_0(t)$ there exists exactly one $y\in\mathbb{R}^n$ such that $y-d_t(x,t)\in\operatorname{im}d_x(x,t)$, $f(y,x,t)=0$, which means,*

$$\mathcal{M}_0(t)=\{x\in\mathcal{D}_f:\exists!y\in\mathbb{R}^n:y-d_t(x,t)\in\operatorname{im}d_x(x,t),\ f(y,x,t)=0\},\ t\in\mathcal{I}_f.$$

Proof. Let $\bar{t}\in\mathcal{I}_f$ and $\bar{x}\in\mathcal{M}_0(\bar{t})$ be fixed. Suppose there are two different values $\bar{y},\bar{\bar{y}}\in\mathbb{R}^n$ such that

$$\begin{aligned}\bar{y}-d_t(\bar{x},\bar{t})&\in\operatorname{im}d_x(\bar{x},\bar{t}),\ \ f(\bar{y},\bar{x},\bar{t})=0,\\ \bar{\bar{y}}-d_t(\bar{x},\bar{t})&\in\operatorname{im}d_x(\bar{x},\bar{t}),\ \ f(\bar{\bar{y}},\bar{x},\bar{t})=0.\end{aligned}$$

Denote $\bar{N}:=\ker d_x(\bar{x},\bar{t})$, and introduce the vectors

$$\bar{w}:=d_x(\bar{x},\bar{t})^+(\bar{y}-d_t(\bar{x},\bar{t})),\quad \bar{\bar{w}}:=d_x(\bar{x},\bar{t})^+(\bar{\bar{y}}-d_t(\bar{x},\bar{t})),$$

thus $\bar{w},\bar{\bar{w}}\in\bar{N}^\perp=\operatorname{im}d_x(\bar{x},\bar{t})^+$. It follows that $\bar{y}-\bar{\bar{y}}=d_x(\bar{x},\bar{t})(\bar{w}-\bar{\bar{w}})$, as well as

$$f(d_x(\bar{x},\bar{t})\bar{w}+d_t(\bar{x},\bar{t}),\bar{x},\bar{t})=0,\quad f(d_x(\bar{x},\bar{t})\bar{\bar{w}}+d_t(\bar{x},\bar{t}),\bar{x},\bar{t})=0,$$

hence

$$\begin{aligned}0&=f\big(d_x(\bar{x},\bar{t})\bar{w}+d_t(\bar{x},\bar{t}),\bar{x},\bar{t}\big)-f\big(d_x(\bar{x},\bar{t})\bar{\bar{w}}+d_t(\bar{x},\bar{t}),\bar{x},\bar{t}\big)\\ &=\int_0^1\underbrace{f_y\big(s(d_x(\bar{x},\bar{t})\bar{w}+d_t(\bar{x},\bar{t}))+(1-s)(d_x(\bar{x},\bar{t})\bar{\bar{w}}+d_t(\bar{x},\bar{t})),\bar{x},\bar{t}\big)d_x(\bar{x},\bar{t})}_{M(s)}\,\mathrm{d}s(\bar{w}-\bar{\bar{w}}).\end{aligned}$$

The matrix $M(s)$ depends continuously on s. Since the DAE has a properly stated leading term, the matrix functions f_yd_x and d_x have common constant rank r, hence $\operatorname{rank}M(s)=r$, $s\in[0,1]$. The inclusion $\bar{N}\subseteq\ker M(s)$, $s\in[0,1]$, together with reasons of dimensions lead to the property $\bar{N}=\ker M(s)$, $s\in[0,1]$.
Then the inclusion $\bar{N}\subseteq\int_0^1\ker M(s)\mathrm{d}s$ is evident. Moreover, by applying a sufficiently fine decomposition $0=\tau_0<\tau_1<\cdots<\tau_S=1$ and considering the continuity of the matrix function M, one can successively verify that the rank of the matrix $\int_0^{\tau_i}\ker M(s)\mathrm{d}s$ is greater than or equal to r, $i=1,\ldots,S$. Again, for reasons of dimensions, it holds that $\bar{N}=\int_0^1\ker M(s)\mathrm{d}s$.
Now it follows that $\bar{w}-\bar{\bar{w}}\in\bar{N}$, thus $\bar{w}-\bar{\bar{w}}=0$, and finally $\bar{y}-\bar{\bar{y}}=0$. □

As we will see in Section 3.7, in the regular index-1 case, through each (x,t), $t \in \mathcal{I}_f$, $x \in \mathcal{M}_0(t)$, there passes exactly one solution. This is what we intended to obtain for index-1 DAEs.

The question whether the set $\mathcal{M}_0(t)$ might be a proper subset of $\widetilde{\mathcal{M}}_0(t)$ remains unsolved. In most cases the sets coincide as the next lemma shows.

Lemma 3.11. *Let the DAE (3.1) have a properly stated leading term and let the nullspace* $\ker f_y(y,x,t)$ *be independent of* y. *Let* $R(x,t) \in L(\mathbb{R}^n)$ *denote the projector matrix defined by*

$$\operatorname{im} R(x,t) = \operatorname{im} d_x(x,t), \quad \ker R(x,t) = \ker f_y(y,x,t), \quad \text{for } x \in \mathcal{D}_f, t \in \mathcal{I}_f.$$

(1) *Then the identities*

$$f(y,x,t) \equiv f(R(x,t)y,x,t), \quad f_y(y,x,t) \equiv f_y(R(x,t)y,x,t) \equiv f_y(y,x,t)R(x,t)$$

become true,

(2) R *is continuously differentiable on* $\mathcal{D}_f \times \mathcal{I}_f$,

(3) *and the set* $\mathcal{M}_0(t)$ *coincides with* $\widetilde{\mathcal{M}}_0(t)$ *for* $t \in \mathcal{I}_f$.

Proof. For $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, $y \in \mathbb{R}^n$, $\eta := (I - R(x,t))y$, it holds that

$$f(y,x,t) - f(R(x,t)y,x,t) = \int_0^1 f_y(sy + (1-s)R(x,t)y,x,t)\eta \, \mathrm{d}s = 0,$$

since $\eta \in \operatorname{im}(I - R(x,t)) = \ker f_y(sy + (1-s)R(x,t)y,x,t)$ independently of s, so the identities in the first assertion are validated.
The function R is continuously differentiable as a projector function defined from $\mathcal{C}^1$-subspaces.
For each fixed $t \in \mathcal{I}_f$, $x \in \widetilde{\mathcal{M}}_0(t)$, and a corresponding $\tilde{y}$, such that $0 = f(\tilde{y},x,t) = f(R(x,t)\tilde{y},x,t)$, we define $y := R(x,t)\tilde{y} + (I - R(x,t))d_t(x,t)$. It follows that

$$y - d_t(x,t) = R(x,t)(\tilde{y} - R(x,t)d_t(x,t)) \in \operatorname{im} d_x(x,t),$$

and

$$f(y,x,t) = f(R(x,t)y,x,t) = f(R(x,t)\tilde{y},x,t) = f(\tilde{y},x,t) = 0,$$

and hence x belongs to $\mathcal{M}_0(t)$. □

In the fully implicit case, if $\ker f_y(y,x,t)$ actually depends on y, the situation is less transparent.

For solvability, the obvious constraint set and all hidden constraint sets must be nonempty. Consistent values necessarily must belong to these sets.
Supposing the obvious constraint set is well accessible, all the following requirements concerning smoothness and constant ranks can be restricted to an open set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ being a neighborhood of the set $\{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x \in \mathcal{M}_0(t),\ t \in \mathcal{I}_f\}$. Also, $\mathcal{D}_f \times \mathcal{I}_f$ can be initially interpreted to be such a neighborhood.

The constraint sets, the obvious as well as the hidden ones, are strongly fixed by the given DAE. Providing a description of the set of consistent values of a given DAE is close to generating the solutions of this DAE. Perturbations of the right-hand side can essentially change the constraint sets, as we can observe already in the easier case of linear DAEs. This motivates us not to place primary emphasis on the constraints. We try to find another way to characterize DAEs and for persistence under perturbation.

We close this subsection by introducing the two additional subspaces

$$\begin{aligned} S(y,x,t) &:= \{z \in \mathbb{R}^m : f_x(y,x,t)z \in \operatorname{im} f_y(y,x,t)\},\\ S_0(x^1,x,t) &:= \{z \in \mathbb{R}^m : f_x(d_x(x,t)x^1 + d_t(x,t),x,t)z \\ &\qquad \in \operatorname{im} f_y(d_x(x,t)x^1 + d_t(x,t),x,t)\}, \end{aligned}$$

associated with the DAE (3.1). The subspaces are defined for $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, $y \in \mathbb{R}^n$ and $x^1 \in \mathbb{R}^m$, and they are closely related to each other. In most cases they coincide. By definition, one has

$$S_0(x^1,x,t) = S\big(d_x(x,t)x^1 + d_t(x,t),x,t\big) \quad \text{for all } x \in \mathcal{D}_f,\ t \in \mathcal{I}_f,\ x^1 \in \mathbb{R}^m,$$

and, on the other hand,

$$\begin{aligned} &S(y,x,t) = S_0(x^1,x,t) \quad \text{for } x \in \mathcal{D}_f,\ t \in \mathcal{I}_f,\\ &\text{and those } y \in \mathbb{R}^n,\ x^1 \in \mathbb{R}^m, \text{ which are related by } y = d_x(x,t)x^1 + d_t(x,t). \end{aligned}$$

It is evident that, if the partial derivative $f_x(y,x,t)$ and the subspace $\operatorname{im} f_y(y,x,t)$ are independent of y, then $S(y,x,t)$ is independent of y and $S_0(x^1,x,t)$ is independent of x^1, and both subspaces coincide, that is,

$$S(y,x,t) = S_0(x^1,x,t), \quad x \in \mathcal{D}_f,\ t \in \mathcal{I}_f,\ y \in \mathbb{R}^n,\ x^1 \in \mathbb{R}^m.$$

This property reflects the special situation given in linear DAEs and in all semi-explicit DAEs, and we write at times $S(x,t) := S(0,x,t) = S_0(0,x,t) =: S_0(x,t)$. For linear DAEs only $S_0(t)$ is used.

Turn once again to quasi-linear DAEs (3.7) and their obvious constraint

$$\mathcal{M}_0(t) = \{x \in \mathcal{D}_f : \mathcal{W}_0(x,t)b(x,t) = 0\}.$$

For these DAEs with $f(y,x,t) = A(x,t)y + b(x,t)$ it follows that

$$\begin{aligned} S(y,x,t) &= \ker \mathcal{W}_0(x,t)f_x(y,x,t),\\ S_0(x^1,x,t) &= \ker \mathcal{W}_0(x,t)f_x(d_x(x,t)x^1 + d_t(x,t),x,t). \end{aligned}$$

If, additionally, the matrix function $A(x,t)$ has a constant range and $\mathcal{W}_0$ is a constant projection along this range, then it holds that

$$\mathcal{M}_0(t) = \{x \in \mathcal{D}_f : \mathcal{W}_0 b(x,t) = 0\},$$
$$S(y,x,t) = \ker \mathcal{W}_0 b_x(x,t) = S_0(x^1,x,t),$$

which indicates a certain affinity of the subspaces $S(y,x,t)$, $S_0(x^1,x,t)$ and the tangent space $T_x\mathcal{M}_0(t)$, if $x \in \mathcal{M}_0(t)$ and the tangent space is well defined.

3.1.3 Linearization

Linearization plays an important role in various fields of nonlinear analysis. It is a standard tool for obtaining information on smooth nonlinear problems. Here we apply linearization for index determination in nonlinear DAEs. Roughly speaking, below, we introduce *regularity regions* of a nonlinear DAE so that all linearizations along sufficiently smooth reference functions residing in that region are regular linear DAEs with uniform characteristics.

For any *reference function* $x_* \in \mathcal{C}(\mathcal{I}_*,\mathbb{R}^m)$, $\mathcal{I}_* \subseteq \mathcal{I}_f$, with values in $\mathcal{D}_f$, i.e. $x_*(t) \in \mathcal{D}_f$, $t \in \mathcal{I}_*$, such that $d(x_*(\cdot),\cdot) \in \mathcal{C}^1(\mathcal{I}_*,\mathbb{R}^n)$, we consider the linear DAE

$$A_*(t)(D_*(t)x(t))' + B_*(t)x(t) = q(t), \quad t \in \mathcal{I}_*, \tag{3.11}$$

the coefficients of which are given by

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

The reference function $x_* \in \mathcal{C}(\mathcal{I}_*,\mathbb{R}^m)$ is not necessary a solution of the DAE (3.1).

Definition 3.12. The linear DAE (3.11) is said to be a *linearization of the nonlinear DAE (3.1) along the reference function* x_*.

We stress that here the linearization of the nonlinear DAE (3.1) along x_* represents the linear DAE (3.11) with unassigned general right-hand side q. In contrast, at times in a different context one restricts a linearization to the equation with specific right-hand side

$$A_*(t)(D_*(t)x(t))' + B_*(t)x(t) = -f((d(x_*(t),t))',x(t),t), \quad t \in \mathcal{I}_*,$$

supposing x_* to be close to a solution of the original equation.

The smoothness demands (3.1) for the nonlinear DAE (3.1) ensure that the linear DAE (3.11) is equipped with continuous coefficients. Moreover, if the DAE (3.1) has a properly involved derivative, the decomposition

$$\ker A_*(t) \oplus \operatorname{im} D_*(t) = \mathbb{R}^n, \quad t \in \mathcal{I}_*, \tag{3.12}$$

holds true with $\ker A_*$ and $\operatorname{im} D_*$ being $\mathcal{C}$-subspaces, but not necessarily $\mathcal{C}^1$-subspaces. This is a direct consequence of the construction.
To be able to apply the linear theory from Chapter 2, we look for additional conditions ensuring $\mathcal{C}^1$-subspaces and a continuously differentiable border projector associated to the decomposition (3.12).
If the subspace $\ker f_y(y,x,t)$ does not depend on y at all, owing to Lemma 3.11, we obtain $\mathcal{C}^1$-subspaces $\ker A_*$ and D_* by taking the linearization along a smoother reference function $x_* \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^m)$.
If $\ker f_y(y,x,t)$ depends on all its arguments, we can make do with $x_* \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^m)$, and, additionally, $d(x_*(.),.) \in \mathcal{C}^2(\mathcal{I}_*, \mathbb{R}^n)$. As a sufficient requirement for the latter we suppose d to have continuous second partial derivatives, and $x_* \in \mathcal{C}^2(\mathcal{I}_*, \mathbb{R}^m)$.

Before we extend the tractability index concept to general nonlinear DAEs, we introduce some convenient denotations and consider some basic properties. We start with

$$D(x,t) := d_x(x,t), \tag{3.13}$$

$$A(x^1,x,t) := f_y(D(x,t)x^1 + d_t(x,t),x,t), \tag{3.14}$$

$$B(x^1,x,t) := f_x(D(x,t)x^1 + d_t(x,t),x,t), \tag{3.15}$$

for $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, to be used throughout the whole chapter. D,A and B are continuous matrix functions. The coefficients of the DAE (3.11) linearized at a continuously differentiable reference function x_* now look like

$$\begin{aligned}
A_*(t) &= A(x_*'(t),x_*(t),t) = f_y(D(x_*(t),t)x_*'(t) + d_t(x_*(t),t),x_*(t),t),\\
D_*(t) &= D(x_*(t),t),\\
B_*(t) &= B(x_*'(t),x_*(t),t) = f_x(D(x_*(t),t)x_*'(t) + d_t(x_*(t),t),x_*(t),t), \quad t \in \mathcal{I}_*.
\end{aligned}$$

Lemma 3.13. *For a DAE (3.1) with a properly involved derivative, the decomposition*

$$\ker A(x^1,x,t) \oplus \operatorname{im} D(x,t) = \mathbb{R}^n, \quad \forall x^1 \in \mathbb{R}^m,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{3.16}$$

is true, and the subspaces $\ker A$ *and* $\operatorname{im} D$ *are at least* $\mathcal{C}$*-subspaces.*

Proof. Because of the assumption, the transversality condition (3.4) is valid. For each triple $(\bar{x}^1,\bar{x},\bar{t}) \in \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f$ we set $\bar{y} := D(\bar{x},\bar{t})\bar{x}^1 + d_t(\bar{x},\bar{t})$, which leads to

$$A(\bar{x}^1,\bar{x},\bar{t}) = f_y(\bar{y},\bar{x},\bar{t}), \quad D(\bar{x},\bar{t}) = d_x(\bar{x},\bar{t}),$$

hence

$$\ker A(\bar{x}^1,\bar{x},\bar{t}) \oplus \operatorname{im} D(\bar{x},\bar{t}) = \ker f_y(\bar{y},\bar{x},\bar{t}) \oplus \operatorname{im} d_x(\bar{x},\bar{t}) = \mathbb{R}^n.$$

The subspaces $\ker A$ and $\operatorname{im} D$ are at least $\mathcal{C}$-subspaces, since A and D are continuous matrix functions which have constant rank. □

In general we have to expect the subspace $\ker A(x^1,x,t)$ to depend on all its arguments x^1, x and t. If the subspace $\ker f_y(y,x,t)$ is independent of y, then $\ker A(x^1,x,t)$ is independent of x^1. We emphasize the importance of the class of DAEs (3.1) whose associate subspace $\ker f_y(y,x,t)$ is independent of y. This class covers quasi-linear DAEs (3.7) and all applications we know.

Lemma 3.14. *Let the DAE (3.1) have a properly involved derivative, and let* $\ker f_y(y,x,t)$ *be independent of* y. *Then the transversality conditions (3.4) and (3.16) are equivalent,*

Proof. Owing to Lemma 3.13 it remains to verify that (3.16) implies (3.4). Let (3.16) be valid. Let t and x be fixed. For each $y \in \mathbb{R}^n$ we find a $x^1 \in \mathbb{R}^m$ such that $R(x,t)(y - d_t(x,t)) = D(x,t)x^1$, thus $R(x,t)y = R(x,t)(D(x,t)x^1 + d_t(x,t))$, and hence

$$\begin{aligned} f_y(y,x,t) &= f_y(R(x,t)y,x,t) = f_y(R(x,t)(D(x,t)x^1 + d_t(x,t)),x,t) \\ &= f_y(D(x,t)x^1 + d_t(x,t)),x,t) = A(x^1,x,t). \end{aligned}$$

□

We see that if the subspace $\ker f_y(y,x,t)$ is independent of y then $\ker A(x^1,x,t) = \ker R(x,t)$ is independent of x^1, and both decompositions (3.4) and (3.16) de facto coincide.
If $\ker f_y(y,x,t)$ depends on y, then supposing condition (3.16) to be given, the definition of the *obvious constraint* accompanying the DAE (cf. Definition 3.9) leads to

$$\ker f_y(y,x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^n, \quad \forall y \in \mathbb{R}^n,\ x \in \mathcal{M}_0(t),\ t \in \mathcal{I}_f. \tag{3.17}$$

Altogether, we have

$$\begin{array}{l} \ker f_y(y,x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^n \\ \forall y \in \mathbb{R}^m,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f \end{array} \quad \begin{array}{c} \longrightarrow \\ \longleftarrow \\ \text{indep. of } y \\ \text{or } x \in \mathcal{M}_0(t) \end{array} \quad \begin{array}{l} \ker A(x^1,x,t) \oplus \operatorname{im} D(x,t) = \mathbb{R}^n \\ \forall x^1 \in \mathbb{R}^m,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f. \end{array}$$

The projector function R introduced in Lemma 3.11 plays its role in the further analysis. The following definition of the *border projector function* generalizes this projector function.

Definition 3.15. For a DAE (3.1) with properly involved derivative, the projector valued function R defined by

$$\operatorname{im} R(x^1,x,t) = \operatorname{im} D(x,t), \quad \ker R(x^1,x,t) = \ker A(x^1,x,t)$$

for $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, is named the *border projector function* or briefly the *border projector* of the DAE.
If $\ker A(x^1,x,t)$ is independent at all of x^1, we set $R(0,x,t) =: R(x,t)$.
If $\ker A(x^1,x,t)$ is independent at all of x^1 and x, and $\operatorname{im} D(x,t)$ does not depend on x, we write $R(t)$.

Under the smoothness Assumption 3.1 for f and d we have agreed upon by now, the border projector function R is continuous, but not necessarily continuously differentiable. For the analysis later on, we will need this function R to be continuously differentiable. Therefore, we want to provide assumptions about the original system leading to a continuously differentiable border projector R. We know, for a DAE with properly involved derivative, and $\ker f_y$ being independent of y, the border projector $R = R(x,t)$ is a priori continuously differentiable on $\mathcal{D}_f \times \mathcal{I}_f$ (cf. Lemma 3.11). On the other hand, if the subspace $\ker f_y(y,x,t)$ depends on y then $\ker A(x^1,x,t)$ depends on x^1, and so does $R(x^1,x,t)$. Then, if we require d to have continuous second partial derivatives, then R is continuously differentiable on $\mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f$.

We summarize the appropriate assumptions to be used later on.

Assumption 3.16. (Basic assumption for (3.1))

(a) *The function f is continuous on $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$ together with its first partial derivatives f_y, f_x. The function d is continuously differentiable on $\mathcal{D}_f \times \mathcal{I}_f$.*
(b) *The DAE (3.1) has a properly involved derivative.*
(c) *If $\ker f_y(y,x,t)$ depends on y, then d is supposed to have additional continuous second partial derivatives.*

Having a continuously differentiable border projector R, we only need to choose $\mathcal{C}^2$-functions x_* as reference functions for the linearization in order to obtain linear DAEs (3.11) with $\mathcal{C}^1$-subspaces $\ker A_*$, $\operatorname{im} D_*$, and hence with a properly stated leading term.

Definition 3.17. (Reference function set) Let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open. Denote by $\mathcal{C}_*^\nu(\mathcal{G})$ the set of all $\mathcal{C}^{\max(2,\nu)}$ functions with graph in $\mathcal{G}$. In other words, $x_* \in \mathcal{C}_*^\nu(\mathcal{G})$ if and only if $x_* \in \mathcal{C}^{\max(2,\nu)}(\mathcal{I}_*, \mathbb{R}^m)$, with $(x_*(t),t) \in \mathcal{G}$, $t \in \mathcal{I}_*$.

Under Assumption 3.16, all linearizations (3.11) of the general nonlinear DAE (3.1) along reference functions $x_* \in \mathcal{C}_*^2(\mathcal{G})$ have a properly stated leading term. This provides the basis to applying the ideas from Chapter 2.

3.2 Admissible matrix function sequences and admissible projector functions

In this section we construct an admissible matrix function sequence and associated admissible projector functions for the general nonlinear DAE (3.1) emulating the model of admissible matrix function sequences built for linear DAEs in Chapter 2. The DAE (3.1) is supposed to satisfy Assumption 3.16.
We start with the matrix functions A, D, B defined in (3.13)–(3.15),

$$A(x^1,x,t) \in L(\mathbb{R}^n, \mathbb{R}^k), \quad D(x,t) \in L(\mathbb{R}^m, \mathbb{R}^n), \quad B(x^1,x,t) \in L(\mathbb{R}^m, \mathbb{R}^k)$$

for $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$. Assumption 3.16 ensures that the matrix functions A, D, B are continuous, and the border projector function R associated with the decomposition (3.16) is continuously differentiable. Denote

$$N_0(x,t) := \ker D(x,t) \qquad \text{for } x \in \mathcal{D}_f,\ t \in \mathcal{I}_f,$$

and introduce $Q_0(x,t) \in L(\mathbb{R}^m)$ to be a projector onto $N_0(x,t)$, i.e.,

$$Q_0(x,t)^2 = Q_0(x,t), \quad \operatorname{im} Q_0(x,t) = N_0(x,t).$$

Set the complementary projector to be $P_0(x,t) := I - Q_0(x,t)$. Since $D(x,t)$ has constant rank r, its nullspace has constant dimension $m-r$. This allows for choosing Q_0, P_0 to be continuous, and we do so (see Lemma A.15). At this point we emphasize the advantage of projector functions against bases. While globally defined smooth bases might not exist, smooth projector functions do exist (see Remark A.16).

Next we introduce the generalized inverse $D(x^1,x,t)^- \in L(\mathbb{R}^n,\mathbb{R}^m)$ by means of the four conditions

$$\begin{aligned} D(x,t)D(x^1,x,t)^-D(x,t) &= D(x,t),\\ D(x^1,x,t)^-D(x,t)D(x^1,x,t)^- &= D(x^1,x,t)^-,\\ D(x,t)D(x^1,x,t)^- &= R(x^1,x,t),\\ D(x^1,x,t)^-D(x,t) &= P_0(x,t), \end{aligned} \tag{3.18}$$

for $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$. By (3.18), $D(x^1,x,t)^-$ is uniquely determined, and it is a continuous function (cf. Proposition A.17). Notice that D^-, in general, depends not only on (x,t) but also on x^1 since $R = R(x^1,x,t)$.

Denote (cf. Section 2.2)

$$G_0(x^1,x,t) := A(x^1,x,t)D(x,t),\ \Pi_0(x,t) := P_0(x,t),\ B_0(x^1,x,t) := B(x^1,x,t). \tag{3.19}$$

Since the derivative is properly involved, it holds that $\ker G_0(x^1,x,t) = \ker D(x,t) = N_0(x,t)$.
Next we form

$$\begin{aligned} G_1(x^1,x,t) &:= G_0(x^1,x,t) + B_0(x^1,x,t)Q_0(x,t),\\ N_1(x^1,x,t) &:= \ker G_1(x^1,x,t),\\ \Pi_1(x^1,x,t) &:= \Pi_0(x,t)P_1(x^1,x,t), \end{aligned} \tag{3.20}$$

with $Q_1(x^1,x,t)$ being a projector onto $N_1(x^1,x,t)$ and $P_1(x^1,x,t) := I - Q_1(x^1,x,t)$.

From the case of linear DAEs in Section 2.1 we know of the necessity to incorporate the derivative of $D\Pi_1 D^-$ into the expression for B_1. Now, since $D\Pi_1 D^-$ may depend on x^1, x, t, we use the total derivative in jet variables, which means that

$$(D\Pi_1 D^-)'(x^2,x^1,x,t) =: \operatorname{Diff}_1(x^2,x^1,x,t)$$

is defined to be the function of the variables $(x^2,x^1,x,t) \in \mathbb{R}^m \times \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f$ given by

$$\begin{aligned}\mathrm{Diff}_1(x^2,x^1,x,t) = (D\Pi_1 D^-)_{x^1}(x^1,x,t)x^2 + (D\Pi_1 D^-)_x(x^1,x,t)x^1 \\ + (D\Pi_1 D^-)_t(x^1,x,t).\end{aligned}$$

The new jet variable x^2 can be considered as a place holder for x''. Indeed, for the special choice $x := x(t)$, $x^1 := x'(t)$ and $x^2 := x''(t)$, we get

$$\frac{\mathrm{d}}{\mathrm{d}t}((D\Pi_1 D^-)(x'(t),x(t),t)) = \mathrm{Diff}_1(x''(t),x'(t),x(t),t).$$

In the following, we mostly drop the arguments. The given relations are then meant pointwise for all arguments.

Next, we introduce

$$\begin{aligned}B_1 &:= B_0P_0 - G_1D^-(D\Pi_1D^-)'D\Pi_0,\\ G_2 &:= G_1 + B_1Q_1,\\ N_2 &:= \ker G_2,\\ \Pi_2 &:= \Pi_1P_2,\end{aligned}$$

with Q_2 being a projector function onto N_2 and $P_2 := I - Q_2$. Now, $D\Pi_2D^-$ is a function of x^2, x^1, x, t, and, by taking the total derivative, a new jet variable x^3 appears, standing as a place holder for x'''.

As long as the expressions exist, we form the sequence for $i \geq 0$,

$$\begin{aligned}G_{i+1} &:= G_i + B_iQ_i,\\ N_{i+1} &:= \ker G_{i+1},\\ \Pi_{i+1} &:= \Pi_iP_{i+1},\\ B_{i+1} &:= B_iP_i - G_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i,\end{aligned} \tag{3.21}$$

with Q_{i+1} being a projector function onto N_{i+1}, $P_{i+1} := I - Q_{i+1}$. Here, $(D\Pi_{i+1}D^-)' =: \mathrm{Diff}_{i+1}$ is a function of $x^{i+2},\ldots,x^1,x,t$ satisfying

$$\begin{aligned}\mathrm{Diff}_{i+1}(x^{i+2},\ldots,x^1,x,t) = \sum_{j=1}^{i+1}(D\Pi_{i+1}D^-)_{x^j}(x^{i+1},\ldots,x^1,x,t)x^{j+1} \\ +(D\Pi_{i+1}D^-)_x(x^{i+1},\ldots,x^1,x,t)x^1 + (D\Pi_{i+1}D^-)_t(x^{i+1},\ldots,x^1,x,t).\end{aligned}$$

On each level i, a new jet variable x^i appears as a place holder for the i-th derivative $x^{(i)}$. In this way, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}((D\Pi_iD^-)(x^{(i)}(t),\ldots,x'(t),x(t),t)) = \mathrm{Diff}_i(x^{(i+1)}(t),\ldots,x'(t),x(t),t)$$

for the special choice $x := x(t), x^1 := x'(t), \ldots, x^{i+1} := x^{(i+1)}(t)$.

If the DAE (3.1) is linear with

$$f(y,x,t) = A(t)y + B(t)x - q(t), \quad d(x,t) = D(t)x,$$

the total derivatives $\mathrm{Diff}_i(x^{(i+1)}(t), \ldots, x'(t), x(t), t)$ simplify to time derivatives $(D\Pi_i D^-)'(t)$, and the matrix function sequence (3.19)–(3.21) coincides with that given in Section 2.2.

Example 3.18 (Sequence terminates at level 1). We continue to consider the semi-explicit DAE from Example 3.7

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

given on $\mathcal{D}_f = \{x \in \mathbb{R}^2 : x_2 > 0\}$, $\mathcal{I}_f = \mathbb{R}$. The real function γ is continuous on $\mathcal{I}_f$. We write the DAE in the form (3.1) with $n = 1$, $m = k = 2$,

$$f(y,x,t) = \begin{bmatrix} y + x_1 \\ x_1^2 + x_2^2 - \gamma(t) - 1 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad d(x,t) = x_1, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

yielding a DAE with properly stated leading term and

$$G_0 = AD = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad B_0 = f_x = \begin{bmatrix} 1 & 0 \\ 2x_1 & 2x_2 \end{bmatrix}.$$

Letting

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{yields} \quad G_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2x_2 \end{bmatrix}.$$

The matrix function G_1 remains nonsingular on the given definition domain, therefore, the matrix function sequence terminates at level 1. □

Example 3.19 (Sequence terminates at level 2). The DAE from Example 3.8

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

is given on $\mathcal{D}_f = \{x \in \mathbb{R}^3 : x_2 > 0\}$, $\mathcal{I}_f = \mathbb{R}$. We write this DAE in the form (3.1), where $n = 2$, $m = k = 3$,

$$f(y,x,t) = \begin{bmatrix} y_1 + x_1 \\ x_2 y_2 - x_3 \\ x_1^2 + x_2^2 - \gamma(t) - 1 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 1 & 0 \\ 0 & x_2 \\ 0 & 0 \end{bmatrix},$$

$$d(x,t) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix},$$

yielding a DAE with properly stated leading term and

$$G_0 = AD = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2^1 & -1 \\ 2x_1 & 2x_2 & 0. \end{bmatrix}.$$

Letting

$$Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \text{yields} \quad G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

G_1 is singular but has constant rank. Since $N_0 \cap N_1 = \{0\}$ we find a projector function Q_1 such that $N_0 \subseteq \ker Q_1$. Later on those projector functions are named admissible. We choose

$$Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{x_2} & 0 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -\frac{1}{x_2} & 1 \end{bmatrix}, \quad \Pi_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad D\Pi_1 D^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}.$$

We obtain $B_1 = B_0 P_0 Q_1$ and then

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 + x_2^1 & -1 \\ 0 & 2x_2 & 0 \end{bmatrix}$$

which is nonsingular on the given definition domain such that the matrix function sequence terminates. □

Not surprisingly, the matrix function sequence composed for nonlinear DAEs keeps the algebraic properties stated for linear DAEs in Section 2.2. For instance, the consecutive inclusions

$$\operatorname{im} G_0 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq G_{i+1}$$

remain valid. We are again interested in reaching a G_κ that is nonsingular or showing at least maximal possible rank.
With exactly the same arguments as used for Proposition 2.5, we obtain

Proposition 3.20. *Let the DAE (3.1) satisfy Assumption 3.16. Let a matrix function sequence* (3.19)–(3.21) *be given and, additionally, a projector valued function $\mathcal{W}_j$ such that pointwise* $\ker \mathcal{W}_j = \operatorname{im} G_j$, $j \geq 0$. *Then, the following relations become true:*

(1) $\ker \Pi_i \subseteq \ker B_{i+1}$,
(2) $\mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_i = \cdots = \mathcal{W}_{i+1} B_0 = \mathcal{W}_{i+1} B, \quad \mathcal{W}_{i+1} B_{i+1} = \mathcal{W}_{i+1} B_0 \Pi_i$,
(3) $\operatorname{im} G_{i+1} = \operatorname{im} G_i \oplus \operatorname{im} \mathcal{W}_i B Q_i$,
(4) $N_i \cap \ker B_i = N_i \cap N_{i+1} \subseteq N_{i+1} \cap \ker B_{i+1}$,
(5) $N_{i-1} \cap N_i \subseteq N_i \cap N_{i+1}$.

We keep in mind that the matrix function G_j and the projector functions Q_j, P_j and $\mathcal{W}_j$ may depend on the variables $x^j, \ldots, x^1, x, t$, and that all the above relations

are meant pointwise. While the original variables (x,t) vary in $\mathcal{D}_f \times \mathcal{I}_f$, each jet variable x^i varies in $\mathbb{R}^m$.
Although G_{i+1} may depend on the variables $x^{i+1}, x^i, \ldots, x^1, x, t$ its rank r_{i+1} depends at most on $x^i, \ldots, x^1, x, t$. This is a consequence of Proposition 3.20 (3).

As in Section 2.2 we turn to a smarter choice of the projector functions ensuring continuous matrix functions G_j and discovering certain invariants of the DAE.

Definition 3.21. Let the DAE (3.1) satisfy the basic Assumption 3.16. Let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open connected.
Let the projector function Q_0 onto $\ker D$ be continuous on $\mathcal{G}$, $P_0 = I - Q_0$. Let D^- be determined on $\mathbb{R}^m \times \mathcal{G}$ by (3.18). For the given level $\kappa \in \mathbb{N}$, we call the sequence $G_0, \ldots, G_\kappa$ an *admissible matrix function sequence* associated to the DAE on the set $\mathcal{G}$, if it is built by the rule

Set $G_0 := AD$, $B_0 := B$, $N_0 := \ker G_0$.
For $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1}Q_{i-1},$$
$$B_i := B_{i-1}P_{i-1} - G_i D^-(D\Pi_i D^-)' D\Pi_{i-1}$$
$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$
fix a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,
choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,
set $P_i := I - Q_i$, $\Pi_i := \Pi_{i-1}P_i$

and, additionally,

(a) the matrix function G_i has constant rank r_i on $\mathbb{R}^{mi} \times \mathcal{G}$, $i = 0, \ldots, \kappa$,
(b) the intersection $\widehat{N}_i$ has constant dimension $u_i := \dim \widehat{N}_i$ there,
(c) the product function Π_i is continuous and $D\Pi_i D^-$ is continuously differentiable on $\mathbb{R}^{mi} \times \mathcal{G}$, $i = 0, \ldots, \kappa$.

The projector functions $Q_0, \ldots, Q_\kappa$ in an admissible matrix function sequence are said to be *admissible* themselves.
An admissible matrix function sequence $G_0, \ldots, G_\kappa$ is said to be *regular admissible*, if

$$\widehat{N}_i = \{0\} \quad \forall\, i = 1, \ldots, \kappa.$$

Then, also the projector functions $Q_0, \ldots, Q_\kappa$ are called *regular admissible*.
The numbers $r_0 := \operatorname{rank} G_0, \ldots, r_\kappa := \operatorname{rank} G_\kappa$ and $u_1, \ldots, u_\kappa$ are named *characteristic values* of the DAE on $\mathcal{G}$.

The notion of *characteristic values* makes sense, since these values are independent of the special choice of admissible projector functions (cf. Theorem 3.23), and invariant under regular transformations (cf. Section 3.4).
To shorten the wording we often speak simply of *admissible projector functions* having in mind the admissible matrix function sequence built with these admissible

projector functions. Admissible projector functions are always cross-linked with their matrix function sequence. Changing a projector function yields a new matrix function sequence.

The following proposition gathers benefits of the smarter construction. We emphasize that now the products of projector functions Π_i are also projector valued.

Proposition 3.22. *If $Q_0, \dots, Q_\kappa$ are admissible on $\mathcal{G}$, then the following relations become true (on $\mathcal{G}$) for $i = 1, \dots, \kappa$:*

(1) $\ker \Pi_i = N_0 + \cdots + N_i$,
(2) *the products $\Pi_i = P_0 \cdots P_i$, $\Pi_{i-1}Q_i$, $D\Pi_i D^-$, $D\Pi_{i-1}Q_i D^-$ are projectors again,*
(3) $N_0 + \cdots + N_{i-1} \subseteq \ker \Pi_{i-1}Q_i$,
(4) $B_i = B_i \Pi_{i-1}$,
(5) $N_i \cap (N_0 + \cdots + N_{i-1}) \subseteq N_i \cap N_{i+1}$,
(6) $G_{i+1}Q_j = B_j Q_j$, $0 \le j \le i$,
(7) $D(N_0 + \cdots + N_i) = \operatorname{im} D\Pi_{i-1}Q_i \oplus \operatorname{im} D\Pi_{i-2}Q_{i-1} \oplus \cdots \oplus \operatorname{im} D\Pi_0 Q_1$.

Proof. All relations are considered pointwise, and the same arguments as in the proof of Proposition 2.7 are used. □

There is a great variety of admissible matrix function sequences left. Of course, also the admissible matrix function sequences strongly depend on the special choice of the involved projectors. However, fortunately, there are invariants.

Theorem 3.23. *Let the DAE (3.1) satisfy the basic assumption (3.16). Let, for the given $\kappa \in \mathbb{N}$, an admissible matrix function sequence $G_0, \dots, G_\kappa$ associated to the DAE exist. Then the subspaces*

$$\operatorname{im} G_j, \quad N_0 + \cdots + N_j, \quad S_j := \ker \mathcal{W}_j B, \quad \text{with } j = 0, \dots, \kappa + 1,$$

the numbers $r_0, \dots, r_\kappa$ and $u_1, \dots, u_\kappa$ as well as the functions

$$r_{\kappa+1} := \operatorname{rank} G_{\kappa+1}, \quad u_{\kappa+1} := \dim \widehat{N}_{\kappa+1},$$

are independent of the special choice of the involved admissible projector functions.

Proof. We repeat the arguments from Theorem 2.8. Basically, the assertions of Lemma 2.12 remain valid in our case. Inspecting the proof of this lemma we see that all properties used there are now guaranteed by Proposition 3.22. Since the product rule for the derivative

$$(D\bar{\Pi}_i \bar{D}^- \times D\Pi_i D^-)' = (D\bar{\Pi}_i \bar{D}^-)' D\Pi_i D^- + D\bar{\Pi}_i \bar{D}^- (D\Pi_i D^-)'$$

used in the proof of Lemma 2.12 is also valid for the total derivative we may adopt this proof. □

If the projector functions $Q_0, \dots, Q_\kappa$ are admissible on $\mathcal{G}$, then the nullspaces $N_0, \dots, N_\kappa$ and the sum spaces $N_0 + N_1, \dots, N_0 + \cdots + N_\kappa$ are $\mathcal{C}$-subspaces on $\mathcal{G}$, since they represent the ranges of the continuous projector functions $I - G_j^+ G_j$, $j = 0, \dots, \kappa$, and $I - \Pi_j$, $j = 1, \dots, \kappa$.
If all projector functions $Q_0, \dots, Q_\kappa$ are also continuous, then the intersection spaces $\widehat{N}_1, \dots, \widehat{N}_\kappa$, as well as the complement spaces $X_1, \dots, X_\kappa$, are $\mathcal{C}$-subspaces on $\mathcal{G}$, too, owing to $\widehat{N}_j = \operatorname{im} Q_j(I - \Pi_{j-1})$ and $X_j = \operatorname{im}(I - Q_j)(I - \Pi_{j-1})$, $j = 1, \dots, \kappa$.

There is a comfortable flexibility left within admissible projectors. We can fix special projectors by choosing them to be orthogonal as far as possible. We start with orthoprojectors $Q_0 = Q_0^*$, and choose, for $i \geq 1$,

$$\operatorname{im} Q_i = N_i, \quad \ker Q_i = (N_0 + \cdots + N_i)^\perp \oplus X_i, \tag{3.22}$$

with

$$X_i := (N_0 + \cdots + N_{i-1}) \cap \widehat{N}_i^\perp. \tag{3.23}$$

Definition 3.24. Admissible projector functions $Q_0, \dots, Q_\kappa$ are called *widely orthogonal*, if $Q_0 = Q_0^*$, and the conditions (3.22), (3.23) are valid for $i = 1, \dots, \kappa$.

Proposition 3.25. *In the case of widely orthogonal projector functions $Q_0, \dots, Q_\kappa$, the projectors Π_i, $\Pi_{i-1} Q_i$, $i = 1, \dots, \kappa$, are symmetric.*

Proof. The same arguments as in Proposition 2.15 apply. □

Widely orthogonal projector functions which are uniquely determined provide the associated matrix function sequence to be unique. This appears to be useful in the theory below and is helpful for ensuring the required smoothness of practically calculated projector functions.
The question whether the smoothness demands in Definition 3.21(c) are in agreement with the orthogonality requirement has a positive answer supposing the matrix function DD^* is continuously differentiable. This reflects the situation in Proposition 2.16.

Proposition 3.26. *Let an admissible matrix function sequence up to the level κ associated to the DAE (3.1) exist. Let, additionally to the given basic assumptions, the matrix function DD^* be continuously differentiable. Then, the matrix function sequence which meets the conditions (3.22), (3.23) is also admissible up to level κ.*

Proof. We show that if admissible projector functions $Q_0, \dots, Q_\kappa$ are given, then we can construct widely orthogonal ones, too. Let $r_0, \dots, r_\kappa$ and $u_1, \dots, u_\kappa$ denote the associate characteristic values of the DAE.
First, we choose the orthogonal projector $\bar{Q}_0 = \bar{Q}_0^*$ onto $\ker G_0$ and form $\bar{G}_1 = G_0 + B_0 \bar{Q}_0$. With the same arguments as in Proposition 2.16 we realize that

$$\dim \widehat{\bar{N}}_1 = u_1, \quad \text{for } \widehat{\bar{N}}_1 := \bar{N}_1 \cap \bar{N}_0.$$

Put

$$\bar{X}_1 = \bar{N}_0 \cap \widehat{\bar{N}_1}^{\perp}, \quad \operatorname{im}\bar{Q}_1 = \bar{N}_1, \quad \ker\bar{Q}_1 = (\bar{N}_0 + \bar{N}_1)^{\perp} \oplus \bar{X}_1.$$

Since $\bar{\Pi}_0 := \bar{P}_0$ and $\bar{\Pi}_1 := \bar{\Pi}_0\bar{P}_1$ are continuous, the projectors $\bar{Q}_0$, $\bar{Q}_1$ are admissible, supposing $D\bar{\Pi}_1\bar{D}^-$ is continuously differentiable. Next, we show that $D\bar{\Pi}_1\bar{D}^-$ is indeed continuously differentiable. As in Proposition 2.16,

$$\ker D\bar{\Pi}_1\bar{D}^- = \ker D\Pi_1 D^- = D(N_0 + N_1) \oplus \ker R$$

is already a $\mathcal{C}^1$-subspace. Denote $M_1 := (D(N_0 + N_1))^{\perp}$. Then, M_1 is a $\mathcal{C}^1$-subspace since $D(N_0 + N_1)$ is so. We have to verify that

$$\begin{aligned}\operatorname{im} D\bar{\Pi}_1\bar{D}^- &= \operatorname{im} D\bar{\Pi}_1 = D(\bar{N}_0 + \bar{N}_1)^{\perp} = D(N_0 + N_1)^{\perp} \\ &= DD^*(D(N_0 + N_1))^{\perp} = DD^*M_1\end{aligned}$$

is also a $\mathcal{C}^1$-subspace. Derive

$$\begin{aligned}M_1^{\perp} &= D(N_0 + N_1) = \operatorname{im} D\Pi_0 Q_1 D^- = \operatorname{im}(R - D\Pi_1 D^-) \\ &= \ker(I - R + D\Pi_1 D^-) = \operatorname{im}(I - R^* + (D\Pi_1 D^-)^*)^{\perp},\end{aligned}$$

thus $M_1 = \operatorname{im}(I - R^* + (D\Pi_1 D^-)^*)$. Because of

$$\ker DD^* = \ker D^* = \ker R^*$$

it follows that

$$DD^*M_1 = \operatorname{im} DD^*(D\Pi_1 D^-)^* = DD^*\operatorname{im}(D\Pi_1 D^-)^*.$$

The subspace $\operatorname{im}(D\Pi_1 D^-)^*$ is a $\mathcal{C}^1$-subspace, too. Since

$$\ker DD^* \cap \operatorname{im}(D\Pi_1 D^-)^* = \ker R^* \cap \operatorname{im} R^*(D\Pi_1 D^-)^* = 0,$$

a local $\mathcal{C}^1$-basis of $\operatorname{im}(D\Pi_1 D^-)^*$ multiplied by DD^* yields a local $\mathcal{C}^1$-basis of DD^*M_1, i.e., DD^*M_1 is in fact a $\mathcal{C}^1$-subspace (cf. Appendix A.4). Consequently, $\operatorname{im}(D\bar{\Pi}_1\bar{D}^-)^*$ and $\ker(D\bar{\Pi}_1\bar{D}^-)^*$ are $\mathcal{C}^1$-subspaces, which implies that $D\bar{\Pi}_1\bar{D}^-$ is continuously differentiable and $\bar{Q}_0$, $\bar{Q}_1$ are admissible.

On the further levels we proceed analogously (cf. Proposition 2.16), using for

$$M_i := (D(N_0 + \cdots + N_i))^{\perp}$$

the representation

$$\begin{aligned}M_i^{\perp} &= \operatorname{im}(R - D\Pi_i D^-) = \ker(I - R + D\Pi_i D^-), \\ M_i &= \operatorname{im}(I - R^* + (D\Pi_i D^-)^*), \qquad \text{and} \\ DD^*M_i &= DD^*\operatorname{im}(D\Pi_i D^-)^* = DD^*\operatorname{im} R^*(D\Pi_i D^-)^*.\end{aligned}$$

□

We close the present section with an assertion which reflects the main idea behind the construction of admissible matrix function sequences: we take the opportunity to make use of linearizations.

Lemma 3.27. *Let Assumption 3.16 be given for the DAE (3.1), and let $Q_0,\dots,Q_\kappa$ be admissible projector functions for the DAE (3.1) on the open connected set $\mathcal{G}\subseteq\mathcal{D}_f\times\mathcal{I}_f$.*
Then, for each reference function $x_*\in C_*^\kappa(\mathcal{G})$ (see Definition (3.17)), the resulting linearization (3.11) is a DAE with properly stated leading term, and $Q_{*0},\dots,Q_{*\kappa}$ defined by

$$\begin{aligned} Q_{*0}(t) &:= Q_0(x_*(t),t),\\ Q_{*i}(t) &:= Q_i(x_*^{(i)}(t),\dots,x_*'(t),x_*(t),t), \quad t\in\mathcal{I}_*,\ i=1,\dots,\kappa, \end{aligned}$$

are admissible projector functions for the linear DAE 3.11). If $Q_0,\dots,Q_\kappa$ are widely orthogonal, then so are $Q_{*0},\dots,Q_{*\kappa}$.

Proof. The properly stated leading term of the linear DAE (3.11) results from Assumption 3.16 and the smoothness of x_*. Denote

$$\begin{aligned} N_{*0}(t) &:= \ker D_*(t), \quad D_*(t) := D(x_*(t),t),\\ R_*(t) &:= R(x_*'(t),x_*(t),t), \quad D_*(t)^- := D(x_*'(t),x_*(t),t)^- \end{aligned}$$

and

$$G_{*0}(t) := A_*(t)D_*(t) = A(x_*'(t),x_*(t),t)D(x_*(t),t) = G_0(x_*'(t),x_*(t),t), \quad t\in\mathcal{I}_*.$$

It is evident that $G_{*0}(t)$ has constant rank r_0, and Q_{*0} is admissible. We construct a matrix function sequence for the linearized DAE (3.11), and indicate it by an asterix index. The matrix function

$$\begin{aligned} G_{*1}(t) &:= G_{*0}(t)+B_*(t)Q_{*0}(t)\\ &= G_0(x_*'(t),x_*(t),t)+B_0(x_*'(t),x_*(t),t)Q_0(x_*(t),t) \;=\; G_1(x_*'(t),x_*(t),t) \end{aligned}$$

is continuous and has constant rank r_1 on $\mathcal{I}_*$, and

$$N_{*1}(t) := \ker G_{*1}(t) = \ker G_1(x_*'(t),x_*(t),t) = N_1(x_*'(t),x_*(t),t)$$

has constant dimension $m-r_1$, while the intersection

$$\begin{aligned} \widehat{N}_{*1}(t) &:= N_{*1}(t)\cap N_{*0}(t)\\ &= N_1(x_*'(t),x_*(t),t)\cap N_0(x_*(t),t) \;=\; \widehat{N}_1(x_*'(t),x_*(t),t) \end{aligned}$$

has the constant dimension u_1 on $\mathcal{I}_*$. With

$$X_{*1}(t) := \operatorname{im} P_1(x_*'(t), x_*(t), t) Q_0(x_*(t), t)$$

we find the decomposition

$$N_{*0}(t) = \widehat{N}_{*1}(t) \oplus X_{*1}(t)$$

such that

$$Q_{*1}(t) X_{*1}(t) = Q_1(x_*'(t), x_*(t), t) X_{*1}(t) = 0, \quad t \in \mathcal{I}_*.$$

Finally for this stage,

$$\Pi_{*1}(t) = \Pi_1(x_*'(t), x_*(t), t) \quad \text{and} \quad (D_* \Pi_{*1} D_*^-)(t) = (D \Pi_1 D^-)(x_*'(t), x_*(t), t)$$

are, as composed functions, continuously differentiable on $\mathcal{I}_*$. Thus, Q_{*0}, Q_{*1} are admissible, and

$$(D_* \Pi_{*1} D_*^-)'(t) = \operatorname{Diff}_1(x_*''(t), x_*'(t), x_*(t), t).$$

We proceed analogously on the next stages, whereby we put

$$X_{*i}(t) := \operatorname{im} P_i\big(x_*^{(i)}(t), \ldots, x_*'(t), x_*(t), t\big)\big(I - \Pi_{i-1}(x_*^{(i-1)}(t), \ldots, x_*(t), t)\big).$$

□

3.3 Regularity regions

The regularity notion for linear DAEs in Section 2.6 is supported by several constant-rank conditions and comprises the following three main aspects:

(a) The solution space of the homogeneous equation has dimension $d < \infty$.
(b) Equations restricted to subintervals inherit property (a) with the same d.
(c) Equations restricted to subintervals inherit the characteristic values r_j, $j \geq 0$.

This feature is expressed in terms of admissible matrix functions and admissible projector functions by Definition 2.25. Linear time-varying DAEs are considered to be regular, if the time-dependent matrix functions G_i have constant rank r_i, and there is a nonsingular G_μ.
Now the matrix functions G_i not only depend on time but may also depend on x and on the jet variables $x^1, \ldots, x^i$. As in the linear case, we require constant rank r_i of the matrix functions G_i. Points where these rank conditions fail will be handled as critical ones.
The following regularity notion for the nonlinear DAE (3.1) comprises the above three regularity aspects for all corresponding linearizations (3.11).

Definition 3.28. Let the DAE (3.1) satisfy Assumption 3.16 with $k = m$, and let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be an open connected set. Then the DAE (3.1) is said to be

(1) *regular with tractability index 0* , if $r_0 = m$,
(2) *regular with tractability index* μ *on* $\mathcal{G}$, if on $\mathcal{G}$ an admissible matrix function sequence exists such that $r_{\mu-1} < r_\mu = m$,
(3) *regular on* $\mathcal{G}$, if it is, on $\mathcal{G}$, regular with any index (i.e., case (1) or (2) apply).

The constants $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu$ are named *characteristic values* of the regular DAE.
The open connected subset $\mathcal{G}$ is called a *regularity region* or *regularity domain*.
A point $(\bar{x},\bar{t}) \in \mathcal{D}_f \times \mathcal{I}_f$ is a *regular point*, if there is a regularity region $\mathcal{G} \ni (\bar{x},\bar{t})$.

By Theorem 3.23, regularity itself as well as the particular values μ and $r_0,\ldots,r_\mu$ are independent of the special choice of the admissible projectors, although the matrix functions $G_1,\ldots,G_\mu$ depend on it. In regular DAEs, all intersections $\widehat{N}_i$ are trivial ones, thus $u_i = 0$, $i \ge 1$. Namely, because of the inclusions (Propositions 3.22 (5), 3.20 (5))

$$\widehat{N}_i \subseteq N_i \cap N_{i+1} \subseteq N_{i+1} \cap N_{i+2} \subseteq \cdots \subseteq N_{\mu-1} \cap N_\mu,$$

for reaching a nonsingular G_μ, which means $N_\mu = \{0\}$, it is necessary to have $\widehat{N}_i = \{0\}$, $i \ge 1$. This is a useful condition for checking regularity in practice.
Definition 2.25 concerning regular linear DAEs and Definition 4.3 characterizing special regular DAEs with tractability index 1 are in agreement with Definition 3.28. Regularity intervals represent the specification of regularity regions for linear DAEs.

By definition, all points belonging to a regularity region are regular points, and they must show uniform characteristics.

The union of regularity regions is, if it is connected, a regularity region, too. Each open connected subset of a regularity region is again a regularity region, and it inherits all characteristics.

Further, the nonempty intersection of two regularity regions is also a regularity region. Only regularity regions with uniform characteristics yield nonempty intersections.

Maximal regularity regions are bordered by critical points. To characterize a DAE it is important to describe the maximal regularity regions with their characteristics. It should be emphasized that, for this aim there is no need to compute solutions.

Example 3.29 (A regular index-1 DAE). We reconsider the DAE from Example 3.3

$$\begin{aligned} (x_1(t) + x_3(t)x_2(t))' &= q_1(t), \\ x_2(t) &= q_2(t), \\ x_3(t) &= q_3(t), \quad t \in \mathcal{I}, \end{aligned}$$

that is (3.1) with $k = m = 3$, $n = 1$,

$$f(y,x,t) := \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} y + \begin{bmatrix} 0 \\ x_2 \\ x_3 \end{bmatrix} - q(t), \quad d(x,t) := x_1 + x_2 x_3, \quad x \in \mathbb{R}^3,\ t \in \mathcal{I},\ y \in \mathbb{R}.$$

The leading term is properly stated because of $\ker f_y = \{0\}$, $\operatorname{im} d_x = \mathbb{R}$. Observe that $\ker D(x,t) = \ker G_0(x,t)$ varies with x. G_0 has rank $r_0 = 1$. The obvious constraint is

$$\mathcal{M}_0(t) := \{x \in \mathbb{R}^3 : x_2 = q_2(t),\ x_3 = q_3(t)\}, \quad t \in \mathcal{I}.$$

Choosing the projector function

$$Q_0(x) := \begin{bmatrix} 0 & -x_3 & -x_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

we find

$$G_1(x) = \begin{bmatrix} 1 & x_3 & x_2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad r_1 = m,$$

that is, this DAE is regular with index 1 on the entire definition domain $\mathbb{R}^3 \times \mathcal{I}$, that is, there is a single maximal regularity region which coincides with the definition domain. For each given continuous q and fixed $\bar{t} \in \mathcal{I}$, $\bar{c} \in \mathbb{R}$, the DAE has the solution

$$\begin{aligned} x_{*1}(t) &= -q_2(t)q_3(t) + \bar{c} + q_2(\bar{t})q_3(\bar{t}) + \int_{\bar{t}}^{t} q_1(s)\mathrm{d}s, \\ x_{*2}(t) &= q_2(t), \\ x_{*3}(t) &= q_3(t), \quad t \in \mathcal{I}, \end{aligned}$$

which satisfies $x_*(\bar{t}) = \bar{x}$, $\bar{x}_1 := \bar{c}$, $\bar{x}_i := q_i(\bar{t})$, $i = 2,3$, $\bar{x} \in \mathcal{M}_0(\bar{t})$. It is evident that there is exactly one solution passing through each given $(\bar{t},\bar{x})$, $\bar{x} \in \mathcal{M}_0(\bar{t})$. Theorem 3.53 below confirms and generalizes this property. We observe that the solution x_* is continuous with a continuously differentiable part $x_{*1} + x_{*2}x_{*3}$, but the second and third solution components are not necessarily continuously differentiable. From this point of view the notation of this DAE in standard form is so to speak misleading.

Taking the identically vanishing function $x_{**}(t) \equiv 0$ as a fixed solution on the compact interval $[\bar{t},T]$ and considering q as a perturbation, $K := \max\{|q(t)| : t \in [\bar{t},T]\}$, we derive the inequality

$$|x_*(t) - x_{**}(t)| \le |x_*(\bar{t})| + \{(T-\bar{t}) + 2K\} \max_{\bar{t} \le s \le t} |q(s)|,$$

hence the problem possesses perturbation index 1 along the solution x_{**} (cf. [103]). □

Example 3.30 (Two maximal regularity regions with index 1). We consider the semi-explicit DAE from Example 3.7 (see also Example 3.18)

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

but now we suppose the definition domain $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = \mathbb{R}$. As in Example 3.18 we compute

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \text{yields} \quad G_1 = \begin{bmatrix} 1 & 0 \\ 0 & 2x_2 \end{bmatrix}.$$

The matrix function G_1 remains nonsingular for $x_2 \neq 0$, but this leads to the two maximal regularity regions, one region associated to $x_2 > 0$, the other one to $x_2 < 0$. The border points between these regularity regions are those with $x_2 = 0$. A closer look at the possible solutions confirms the critical behavior at these points.

Example 3.31 (Two maximal regularity regions with index 2). The DAE from Example 3.8

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

is now given on $\mathcal{D}_f = \mathbb{R}^3$, $\mathcal{I}_f = \mathbb{R}$. We proceed as in Example 3.19 to obtain

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 & -1 \\ 0 & 0 & 0 \end{bmatrix}.$$

G_1 is singular but has constant rank. We have

$$N_0 \cap N_1 = \{z \in \mathbb{R}^3 : z_1 = 0, x_2 z_2 = 0, z_3 = 0\}.$$

If $x_2 > 0$ or $x_2 < 0$ it holds that $N_0 \cap N_1 = \{0\}$, and we find an admissible projector function Q_1 such that $N_0 \subseteq \ker Q_1$. We choose the same Q_1 as in Example 3.19 and arrive in both cases at

$$G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2x_2 + x_2^1 & -1 \\ 0 & 2x_2 & 0 \end{bmatrix}$$

which is nonsingular for all $x_2 > 0$ and $x_2 < 0$. In this way we find two maximal regularity regions bordered by critical points with $x_2 = 0$. □

An admissible matrix function sequence incorporates by definition the existence of the first derivative of the projectors $D\Pi_i D^-$. This might need some additional smoothness demands for the functions f and d besides Assumption 3.16. Consider the following example to illustrate this fact.

Example 3.32 (Smoothness for Hessenberg size-2 DAEs).

$$\begin{aligned} x_1'(t) + b_1(x_1(t), x_2(t), t) &= 0, \quad \}m_1 \\ b_2(x_1(t), t) &= 0, \quad \}m_2 \end{aligned} \tag{3.24}$$

with the product $B_{21}B_{12}$ being nonsingular on the domain $\mathcal{D}_b \times \mathcal{I}_b$, $B_{ij} := b_{i,x_j}$. This is a so-called Hessenberg size-2 DAE. With

$$f(y,x,t) := Ay + b(x,t), \quad d(x,t) = Dx, \quad n = m_1, \quad k = m = m_1 + m_2$$

and

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} I \\ 0 \end{bmatrix}, \quad R = I,$$

the DAE (3.24) appears to be a DAE with a very simple properly stated leading term. We form

$$G_0 = AD = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \; B_0 = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & 0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & \\ & I \end{bmatrix}, \; \Pi_0 = P_0 = I - Q_0 = \begin{bmatrix} I & \\ & 0 \end{bmatrix}$$

and

$$G_1 = G_0 + B_0 Q_0 = \begin{bmatrix} I & B_{12} \\ 0 & 0 \end{bmatrix}, \quad r_0 = m_1, \quad r_1 = m_1.$$

This implies

$$N_1 = \{z \in \mathbb{R}^m : z_1 + B_{12} z_2 = 0\}, \quad N_1 \cap N_0 = \{z \in \mathbb{R}^m : z_1 = 0,\, B_{12} z_2 = 0\}.$$

From the nonsingularity of the product $B_{21}B_{12}$ it follows that $\ker B_{12} = \{0\}$ and, consequently, $N_0 \cap N_1 = 0$.
Choose a generalized inverse B_{12}^- to B_{12} (pointwise on $\mathcal{D}_b \times \mathcal{I}_b$) such that $B_{12}B_{12}^-B_{12} = B_{12}$, and

$$Q_1 = \begin{bmatrix} B_{12}B_{12}^- & 0 \\ -B_{12}^- & 0 \end{bmatrix}, \quad \Pi_0 Q_1 = \begin{bmatrix} B_{12}B_{12}^- & 0 \\ 0 & 0 \end{bmatrix}, \quad \Pi_1 = \Pi_0 P_1 = \begin{bmatrix} I - B_{12}B_{12}^- & \\ & 0 \end{bmatrix}.$$

Note that Q_1 is, except for the smoothness of $D\Pi_1 D^-$, an admissible projector function since Q_1 is a projector function onto N_1 and we have

$$X_1 := \{z \in \mathbb{R}^m : z_1 = 0\} = N_0 \ominus \widehat{N_1} = N_0 \ominus (N_0 \cap N_1) = N_0 \subseteq \ker Q_1.$$

This leads to $D\Pi_1 D^- = I - B_{12}B_{12}^-$. The matrix B_{12} has constant rank m_2 so that B_{12}^- can be chosen continuously. For a continuously differentiable $D\Pi_1 D^-$, $\operatorname{im} B_{12}$ must be a $\mathcal{C}^1$-subspace, but this is not necessarily guaranteed by the general Assumption 3.16. A *sufficient* condition for that is the additional existence of continuous second partial derivatives b_{1,x_1x_2}, b_{1,x_2x_2}, b_{1,x_2t}. However, if the subspace $\operatorname{im} B_{12}$ is a constant one, the projector $D\Pi_1 D^-$ can be chosen to be constant without any further smoothness, and Assumption 3.16 is enough. Next we form

$$B_1 = B_0 P_0 - G_1 D^- (D\Pi_1 D^-)' D\Pi_0 = \begin{bmatrix} B_{11} + (B_{12}B_{12}^-)' & 0 \\ B_{21} & 0 \end{bmatrix}$$

and

$$G_2 = G_1 + B_1 Q_1 = \begin{bmatrix} I + (B_{11} + (B_{12}B_{12}^-)')B_{12}B_{12}^- & B_{12} \\ B_{21}B_{12}B_{12}^- & 0 \end{bmatrix}.$$

Consider the homogeneous equation $G_2 z = 0$, i.e.,

$$z_1 + (B_{11} + (B_{12}B_{12}^-)')B_{12}B_{12}^- z_1 + B_{12}z_2 = 0, \tag{3.25}$$

$$B_{21}B_{12}B_{12}^- z_1 = 0. \tag{3.26}$$

Since $B_{21}B_{12}$ is nonsingular, (3.26) yields $B_{12}^- z_1 = 0$, and (3.25) reduces to

$$z_1 + B_{12}z_2 = 0.$$

Multiplying this equation by $(I - B_{12}B_{12}^-)$ implies $(I - B_{12}B_{12}^-)z_1 = 0$, hence $z_1 = 0$, and further $B_{12}z_2 = 0$, thus $z_2 = 0$.

Consequently, the Hessenberg size-2 system is a regular DAE with tractability index 2 on $\mathcal{D}_b \times \mathcal{I}_b$. Its characteristics are $r_0 = r_1 = m_1$, $r_2 = m$. □

The demand for the projector functions $D\Pi_i D^-$ to be continuously differentiable corresponds to the consecutive decomposition of the $\mathcal{C}^1$-subspace $\operatorname{im} R = \operatorname{im} D$ into further $\mathcal{C}^1$-subspaces by

$$\begin{aligned} R = DD^- &= D\Pi_0 D^- = D\Pi_1 D^- + D\Pi_0 Q_1 D^- \\ &= D\Pi_i D^- + D\Pi_{i-1} Q_i D^- + D\Pi_{i-2} Q_{i-1} D^- + \cdots + D\Pi_0 Q_1 D^-. \end{aligned}$$

Example 3.32 which is structurally very simple shows that, in the case of a constant subspace $\operatorname{im} B_{12}$, Assumption 3.16 is sufficient. For varying $\operatorname{im} B_{12}$, using special knowledge of the structure, we have specified somewhat mild sufficient conditions for the $\mathcal{C}^1$-property of $D\Pi_1 D^-$. From this point of view the requirement for b to belong to $\mathcal{C}^2$ or $\mathcal{C}^m$ looks much too generous. However, to figure out the milder sufficient smoothness conditions for more general DAEs needs hard technical work and it does not seem to allow for better insights. This is why we do not go into those details. Instead we use the phrasing *f and d satisfy Assumption 3.16, and they are sufficiently smooth*. Let us stress that, owing to the structural properties, it may happen that Assumption 3.16 is sufficient. On the other hand, in this context it would be greatly generous assuming f and d to belong to $\mathcal{C}^{\bar{\mu}}$, if $\bar{\mu} < m$ is a known upper bound of the index, or even to be from $\mathcal{C}^m$. In contrast, applying derivative array approaches, one has to suppose at least $\mathcal{C}^{\bar{\mu}+1}$ functions to be able to form the derivative array function $\mathcal{E}_{\bar{\mu}}$ on its own and to compute its Jacobian (cf. Section 3.10).

Theorem 3.33. (Necessary and sufficient regularity conditions)
Let the DAE (3.1) satisfy the Assumption 3.16, with $k = m$, and $DD^* \in \mathcal{C}^1(\mathcal{D}_f \times \mathcal{I}_f, L(\mathbb{R}^n))$. Let f and d be sufficiently smooth on the open connected subset $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$.

(1) *Then, the DAE (3.1) is regular on $\mathcal{G}$ if all linearizations (3.11) along reference functions $x_* \in \mathcal{C}_*^m(\mathcal{G})$ are regular linear DAEs, and vice versa.*
(2) *If (3.1) is regular with tractability index μ, and characteristics $r_0, \ldots, r_\mu$, then all linearizations (3.11) along reference functions $x_* \in \mathcal{C}_*^\mu(\mathcal{G})$ are regular linear DAEs with uniform index μ and uniform characteristics $r_0, \ldots, r_\mu$.*
(3) *If all linearizations (3.11) along $x_* \in \mathcal{C}_*^m(\mathcal{G})$ are regular linear DAEs, then they have a uniform index μ, and uniform characteristics $r_0, \ldots, r_\mu$, and the nonlinear DAE (3.1) is regular on $\mathcal{G}$ with these characteristics and index.*

Proof. Assertion (1) is a consequence of assertions (2) and (3). Assertion (2) follows immediately from Lemma 3.27. It remains to verify assertion (3).
If D is nonsingular, there is nothing to prove. So we suppose that $r_0 < m$. Let all linearizations (3.11) along functions $x_* \in \mathcal{C}_*^m(\mathcal{G})$ be regular. Introduce the matrix functions

$$G_0(x^1,x,t) := A(x^1,x,t)D(x,t),\ \ B_0(x^1,x,t) := B(x^1,x,t),\ \ N_0(x,t) := \ker D(x,t),$$

and choose $Q_0(x,t)$ to be the orthoprojector onto $N_0(x,t)$. $D(x^1,x,t)^-$ denotes the corresponding generalized inverse (cf. Section 3.2). Due to Assumption 3.16, the matrix function G_0 is continuous on $\mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f$ and has constant rank r_0, and hence Q_0 is continuous and so is D^-. Compute further

$$G_1(x^1,x,t) = G_0(x^1,x,t) + B_0(x^1,x,t)Q_0(x,t), \quad N_1(x^1,x,t) = \ker G_1(x^1,x,t).$$

Obviously, G_1 is also a continuous matrix function.
We show the intersection $N_1(x^1,x,t) \cap N_0(x,t)$ to be trivial for all $x^1 \in \mathbb{R}^m$, $(x,t) \in \mathcal{G}$, and $G_1(x^1,x,t)$ to have constant rank. Assume that there is a point $(\bar{x}^1,\bar{x},\bar{t}) \in \mathbb{R}^m \times \mathcal{G}$ such that $N_1(\bar{x}^1,\bar{x},\bar{t}) \cap N_0(\bar{x},\bar{t}) \neq \{0\}$. Consider the function x_*,

$$x_*(t) := \bar{x} + (t-\bar{t})\bar{x}^1, \quad t \in \mathcal{I}_* = (\bar{t}-\varepsilon,\bar{t}+\varepsilon),$$

with $\varepsilon > 0$ small enough to ensure $x_* \in \mathcal{C}_*^m(\mathcal{G})$. The linearization along this function x_* is regular because of the assumptions, and hence there are $Q_{*0},\ldots,Q_{*\mu_*-1}$ being admissible for (3.11), and G_{μ_*} is nonsingular. Since $D(x_*(t),t)D(x_*(t),t)^*$ is continuously differentiable with respect to t, we may consider $Q_{*0},\ldots,Q_{*\mu_*-1}$ to be widely orthogonal (cf. Proposition 2.16). In this way we arrive at

$$\begin{aligned} N_{*1}(\bar{t}) \cap N_{*0}(\bar{t}) &= N_1(x_*'(\bar{t}),x_*(\bar{t}),\bar{t}) \cap N_0(x_*(\bar{t}),\bar{t}) \\ &= N_1(\bar{x}^1,\bar{x},\bar{t}) \cap N_0(\bar{x},\bar{t}) \ \neq\ \{0\}, \end{aligned}$$

but this contradicts the property of regular linear DAEs to have those intersections just trivial (cf. also Section 2.6).
We turn to the rank of $G_1(x^1,x,t)$. Assume that there exist two points

$$\mathcal{P}_i := (x_i^1,x_i,t_i) \in \mathbb{R}^m \times \mathcal{G}, \quad i = 1,2,$$

with $\operatorname{rank} G_1(\mathcal{P}_1) > \operatorname{rank} G_1(\mathcal{P}_2)$. We connect $\mathcal{P}_1$ and $\mathcal{P}_2$ by a continuous curve lying in $\mathbb{R}^m \times \mathcal{G}$, and move along this curve starting at $\mathcal{P}_1$. We necessarily meet a point $\mathcal{P}_3$ where the rank changes. This means that $\operatorname{rank} G_1(\mathcal{P}_3) < \operatorname{rank} G_1(\mathcal{P}_1)$, and each neighborhood of $\mathcal{P}_3$ contains points $\mathcal{P}_4$ with $\operatorname{rank} G_1(\mathcal{P}_4) = \operatorname{rank} G_1(\mathcal{P}_1)$. Construct a function x_* passing through $\mathcal{P}_3$ and $\mathcal{P}_4$, i.e.,

$$x_*(t_i) = x_i, \quad x_*'(t_i) = x_i^1, \quad i = 3,4.$$

We may use the interpolation polynomial. Choosing $\mathcal{P}_4$ close enough to $\mathcal{P}_3$, we make sure that x_* belongs to $\mathcal{C}_*^m(\mathcal{G})$. In this way, for the DAE (3.11) linearized

along x_*, it holds that

$$\operatorname{rank} G_{*1}(t_4) = \operatorname{rank} G_1(\mathcal{P}_4) > \operatorname{rank} G_1(\mathcal{P}_3) = \operatorname{rank} G_{*1}(t_3),$$

but this contradicts the regularity of (3.11).
Next, since G_1 is continuous with constant rank r_1, we may construct Q_1 (pointwise on $\mathbb{R}^m \times \mathcal{G}$) to be the projector onto N_1 along $(N_0+N_1)^\perp \oplus N_0$. Q_1 is continuous since the involved subspaces are $\mathcal{C}^1$-subspaces (cf. Appendix A.4). It is justified by Proposition 3.26 that we can do this by considering widely orthogonal projectors only. If $D\Pi_1 D^-$ actually varies with its arguments, due to the smoothness of f and d, $D\Pi_1 D^-$ is $\mathcal{C}^1$ and Q_0, Q_1 are admissible (widely orthogonal).
We continue to construct the matrix function sequence for (3.1) with widely orthogonal projectors. Let $Q_0, \ldots, Q_\kappa$ be already shown to be widely orthogonal which includes admissibility. Form $G_{\kappa+1} = G_\kappa + B_\kappa Q_\kappa$ (pointwise for $x^i \in \mathbb{R}^m$, $i = 1, \ldots, \kappa+1$, $(x,t) \in \mathcal{G}$), and consider its nullspace.
The existence of a point $\bar{\mathcal{P}} := (\bar{x}^{\kappa+1}, \ldots, \bar{x}^1, \bar{x}, \bar{t})$ in $\mathbb{R}^{m(\kappa+1)} \times \mathcal{G}$ where the intersection

$$N_{\kappa+1}(\bar{\mathcal{P}}) \cap (N_0 + \cdots + N_\kappa)(\bar{x}^\kappa, \ldots, \bar{x}^1, \bar{x}, \bar{t})$$

is nontrivial would contradict the regularity of the linearization along x_*, with

$$x_*(\bar{t}) = \bar{x}, \quad x_*^{(i)}(t) = \bar{x}^i, \quad i = 1, \ldots, \kappa+1, \quad t \in (\bar{t}-\varepsilon, \bar{t}+\varepsilon),$$

$\varepsilon > 0$ small enough. Similarly as for G_1, we show that $G_{\kappa+1}$ has constant rank $r_{\kappa+1}$ on $\mathbb{R}^{m(\kappa+1)} \times \mathcal{G}$. The next step is the construction of $Q_{\kappa+1}$ such that

$$\operatorname{im} Q_{\kappa+1} = N_{\kappa+1}, \quad \ker Q_{\kappa+1} = (N_0 + \cdots + N_{\kappa+1})^\perp \oplus (N_0 + \cdots + N_\kappa).$$

Again, the involved subspaces are $\mathcal{C}^1$-subspaces, hence $Q_{\kappa+1}$ is continuous, and so are $P_{\kappa+1} = I - Q_{\kappa+1}$, $\Pi_{\kappa+1} = \Pi_\kappa P_{\kappa+1}$. The smoothness of f and d makes $D\Pi_{\kappa+1}D^-$ continuously differentiable, thus, $Q_0, \ldots, Q_{\kappa+1}$ are admissible (widely orthogonal).

It follows also that all linearizations must have uniform characteristics $r_0, \ldots, r_{\kappa+1}$ (cf. Lemma 3.27). We continue to construct the admissible matrix function sequence for the nonlinear DAE up to level μ using widely orthogonal projectors. It turns out that there must be a uniform index μ such that $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$. □

The necessary and sufficient regularity conditions provided by Theorem 3.33 represent the main goal and result of the present chapter. We had this in mind when we started to create the admissible matrix function sequence for the nonlinear DAE. Now, various questions can be traced back to linearizations. The next example shows that rank changes in the matrix functions G_1 indicate in fact points where somewhat unusual things happen with the solutions such that we have good reason for calling these points *critical*.

Example 3.34 (Singularities at rank drop points of G_1). The system

$$\begin{aligned} x_1'(t) - x_3(t) &= 0, \\ x_2(t)(1-x_2(t)) - \gamma(t) &= 0, \\ x_1(t)x_2(t) + x_3(t)(1-x_2(t)) - t &= 0, \end{aligned} \tag{3.27}$$

written as (3.1) with $k = m = 3$, $n = 1$, $f(y,x,t) = Ay + b(x,t)$,

$$A = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad b(x,t) = \begin{bmatrix} -x_3 \\ x_2(1-x_2) - \gamma(t) \\ x_1x_2 + x_3(1-x_2) - t \end{bmatrix}, \quad x \in \mathbb{R}^3,\ t \in \mathbb{R},$$

and $d(x,t) = x_1$ satisfies Assumption 3.16. The function γ is supposed to be continuous, $\gamma(t) \le \frac{1}{4}$.
The obvious constraint is

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^3 : x_2(1-x_2) = \gamma(t),\ x_1x_2 + x_3(1-x_2) = t\}.$$

Compute

$$D = \begin{bmatrix} 1\ 0\ 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0\ 0\ 0 \\ 0\ 1\ 0 \\ 0\ 0\ 1 \end{bmatrix}, \quad R = 1,$$

$$B_0(x,t) = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 1-2x_2 & 0 \\ x_2 & x_1 - x_3 & 1-x_2 \end{bmatrix}, \quad G_1(x,t) = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1-2x_2 & 0 \\ 0 & x_1 - x_3 & 1-x_2 \end{bmatrix}.$$

Then, $\det G_1(x,t) = (1-2x_2)(1-x_2)$ has the zeros $x_2 = \frac{1}{2}$ and $x_2 = 1$. This splits the definition domain $\mathcal{D}_f \times \mathcal{I}_f = \mathbb{R}^3 \times \mathbb{R}$ into the open sets

$$\begin{aligned} \mathcal{G}_1 &:= \left\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x_2 < \frac{1}{2}\right\}, \\ \mathcal{G}_2 &:= \left\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : \frac{1}{2} < x_2 < 1\right\}, \\ \mathcal{G}_3 &:= \{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : 1 < x_2\}, \end{aligned}$$

such that $\mathcal{D}_f \times \mathcal{I}_f$ is the closure of $\mathcal{G}_1 \cup \mathcal{G}_2 \cup \mathcal{G}_3$. The DAE is regular with tractability index 1 on each region $\mathcal{G}_\ell$, $\ell = 1,2,3$.
All linearizations along functions $x_* \in \mathcal{C}^1_*(\mathcal{G}_i)$ are regular linear DAEs with tractability index 1. Through each point $(\bar{x},\bar{t}) \in \mathcal{G}_i$ such that $\bar{x} \in \mathcal{M}_0(\bar{t})$ there passes exactly one solution (cf. Theorem 3.53). This is what we expect. Solutions moving along the borders of the regularity domains or crossing these borders may behave differently as discussed below.

Inspecting solutions of the DAE (3.27) one realizes that different kinds of problems may actually happen if the solution approaches or crosses the critical point set. We take a closer look at special situations.
Set $\gamma(t) = \frac{1}{4} - t^2$, and fix $\bar{t} = 0$ and $\bar{x} = (0, \frac{1}{2}, 0) \in \mathcal{M}_0(0)$ and $\mathcal{M}_0(0) =$

$\{x \in \mathbb{R}^3 : x_2 = \frac{1}{2}, x_1 + x_3 = 0\}$. There are two solutions passing through $(\bar{x}, \bar{t})$. One solution x_* has the second and third components

$$x_{*2}(t) = t + \frac{1}{2}, \quad x_{*3}(t) = \frac{1}{2t-1}((1+2t)x_{*1}(t) - 2t),$$

while the first component is the unique solution $x_{*1} \in \mathcal{C}^1$ of the standard IVP

$$x_1'(t) = \frac{1}{2t-1}((1+2t)x_1(t) - 2t), \quad x_1(0) = 0. \tag{3.28}$$

If t increases and tends to $\frac{1}{2}$, the component $x_{*2}(t)$ approaches the border plane $x_2 = 1$ and the ODE for the first solution component undergoes a singularity. The third component grows unboundedly.
The second solution through $(\bar{x}, \bar{t})$ has the components

$$x_{*2}(t) = -t + \frac{1}{2}, \quad x_{*3}(t) = -\frac{1}{2t+1}((1-2t)x_{*1}(t) - 2t),$$

while the first component is the unique solution $x_{*1} \in \mathcal{C}^1$ of the standard IVP

$$x_1'(t) = -\frac{1}{2t+1}((1-2t)x_1(t) - 2t), \quad x_1(0) = 0.$$

This solution stays, for $t > 0$, within the regularity domain $\mathcal{G}_3$.
The bifurcations observed at the border between $\mathcal{G}_1$ and $\mathcal{G}_2$ and the singularity in the first ODE on the border between $\mathcal{G}_2$ and $\mathcal{G}_3$ indicate that we are in fact confronted with critical points.
Figure 3.3 shows the isocline field of x_1 related to the ODE of (3.28). Figures 3.4, 3.5 and 3.6 show the three components of further solutions of (3.27), which start on the border between $\mathcal{G}_1$ and $\mathcal{G}_2$. The initial values are $\begin{bmatrix} 1 \\ \frac{1}{2} \\ -1 \end{bmatrix}$ and $\begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \\ -\frac{1}{3} \end{bmatrix}$ (solid and dashed lines). We have two solutions in every case. The left-hand side shows the solutions that went to $\mathcal{G}_1$. The other side shows the solutions which enter $\mathcal{G}_2$ and then approach the border between $\mathcal{G}_2$ and $\mathcal{G}_3$, and undergo bifurcations at the border. Note that in numerical computations at this point the continuation of the solution is quite arbitrary.

The linearization along a reference function x_* lying on the border plane $x_2 = 1$,

$$x_*(t) := \begin{bmatrix} \alpha(t) \\ 1 \\ \beta(t) \end{bmatrix},$$

with α, β arbitrary smooth functions, leads to

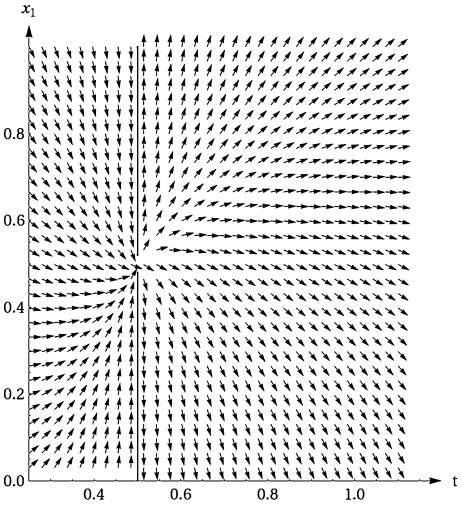


Fig. 3.3 Isocline field of x_1 of (3.28)

$$G_{*0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_{*0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_{*0} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & \alpha - \beta & 0 \end{bmatrix},$$
$$G_{*1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & -1 & 0 \\ 0 & \alpha - \beta & 0 \end{bmatrix}, \quad Q_{*1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad Q_{*1}Q_{*0} = 0.$$

Further, $\Pi_{*1} = \Pi_{*0}P_{*1} = 0$, thus $D\Pi_{*1}D^- = 0$, and

$$G_{*2} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & -1 & 0 \\ 1 & \alpha - \beta & 0 \end{bmatrix}, \quad \det G_{*2} = -1.$$

This means that the DAE linearized along x_* is regular with tractability index 2. It is worth emphasizing that we do not call the original nonlinear DAE regular with index 2 on the set $\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x_2 = 1\}$ because this set is not open in $\mathbb{R}^3 \times \mathbb{R}$. In contrast, the linearization along functions

$$x_*(t) = \begin{bmatrix} \alpha(t) \\ \frac{1}{2} \\ \beta(t) \end{bmatrix}$$

leads to

$$B_{*0} = \begin{bmatrix} 0 & 0 & -1 \\ 0 & 0 & 0 \\ \frac{1}{2} & \alpha - \beta & \frac{1}{2} \end{bmatrix}, \quad G_{*1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & \alpha - \beta & \frac{1}{2} \end{bmatrix}.$$

If $\alpha = \beta$, then $N_{*0} \cap N_{*1} = \{z \in \mathbb{R}^3 : z_1 = 0, z_3 = 0\}$ is nontrivial so that the linearized DAE is no longer regular. If $\alpha \neq \beta$, then we can choose

$$Q_{*1} = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2(\beta-\alpha)} & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad Q_{*1}Q_{*0} = 0.$$

It follows that $\Pi_{*1} = \Pi_{*0}P_{*1} = 0$, $D\Pi_{*1}D^- = 0$, which means, Q_{*0}, Q_{*1} are admissible. From

$$G_{*2} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ \frac{1}{2} & \alpha - \beta & \frac{1}{2} \end{bmatrix}$$

and $N_{*0} + N_{*1} = \ker \Pi_{*1} = \mathbb{R}^3$ we see that $(N_{*0} + N_{*1}) \cap N_{*2} = N_{*2}$, i.e., the necessary regularity condition fails again.

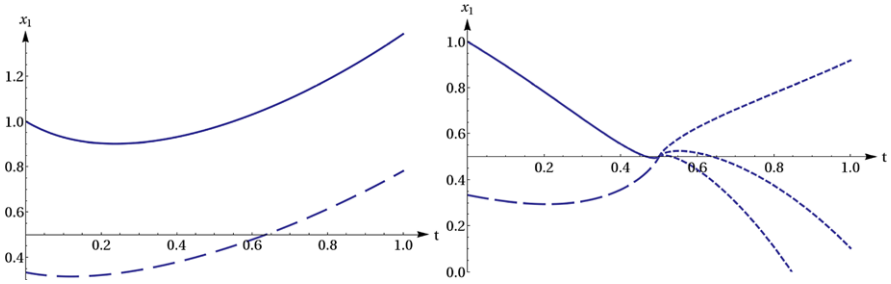


Fig. 3.4 Solution component x_1 of the DAE (3.27)

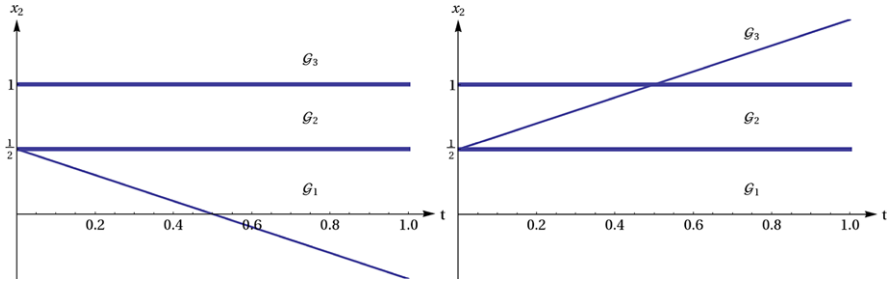


Fig. 3.5 Solution component x_2 of the DAE (3.27)

□

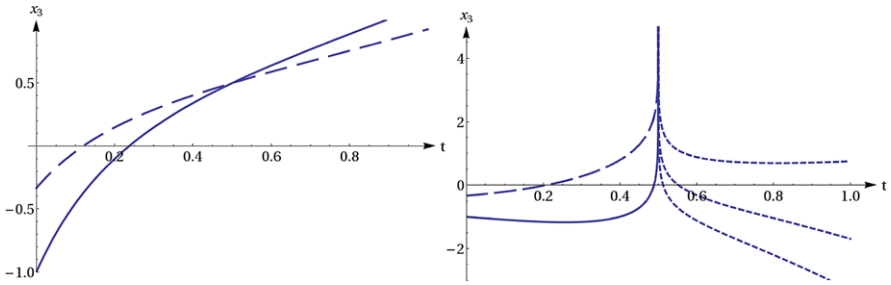


Fig. 3.6 Solution component x_3 of the DAE (3.27)

Remark 3.35. If one puts $\gamma = 0$, the DAE (3.27) simplifies to

$$\begin{aligned} x_1'(t) - x_3(t) &= 0, \\ x_2(t)(1 - x_2(t)) &= 0, \\ x_1(t)x_2(t) + x_3(t)(1 - x_2(t)) - t &= 0. \end{aligned} \tag{3.29}$$

System (3.29) was originally introduced in [5, p. 235–236] to demonstrate that an index notion should be a local one. It has the only solutions

$$x_*(t) = \begin{bmatrix} t \\ 1 \\ 1 \end{bmatrix} \quad \text{and} \quad x_{**}(t) = \begin{bmatrix} \frac{1}{2}t^2 + c \\ 0 \\ t \end{bmatrix}.$$

The solutions x_{**} with arbitrary $c \in \mathbb{R}$, lie in the index-1 regularity region $\mathcal{G}_1$. The other solution x_* proceeds on the border between $\mathcal{G}_2$ and $\mathcal{G}_3$. The linearization along x_* is a regular DAE with tractability index 2. However, there is no neighborhood of the graph of x_* representing a regularity region with tractability index 2.

By differentiating the second equation, and employing the solution property that $x_2 \neq \frac{1}{2}$, one obtains (cf. [5]) from (3.29) the system

$$\begin{aligned} x_1'(t) - x_3(t) &= 0, \\ x_2'(t) &= 0, \\ x_1(t)x_2(t) + x_3(t)(1 - x_2(t)) - t &= 0. \end{aligned}$$

At a first glance, one could conclude that the index depends on the initial condition, which means, $x_2(0) = 0$ yields the index to be 1, and $x_2(0) = 1$ yields the index to equal 2. However, when trying to apply the corresponding notion of the (differentiation) index along a solution to slightly perturbed problems, one comes into trouble. In our opinion, in spite of practical models and numerical computations, a characterization is wanted that is somehow stable with respect to perturbations. This is in full agreement with the demand, e.g., in [25] that the statements concerning the *dif-*

ferentiation index are taken to hold locally on open subsets of the respective spaces (cf. also the discussion in Section 3.10).

In general we do not expect a DAE (3.1) to be regular on its entire definition domain $\mathcal{D}_f \times \mathcal{I}_f$ as it is the case for the class of Hessenberg form DAEs. It seems to be rather natural, as sketched in Figure 3.7, that $\mathcal{D}_f \times \mathcal{I}_f$ decomposes into several maximal regularity regions the borders of which consist of critical points. In contrast to Example 3.34, it may well happen that the characteristic values on different regularity regions are different, as the next example shows. However, in each regularity region there must be uniform characteristic values. A solution can enter a region with new characteristic values only after passing a critical point.

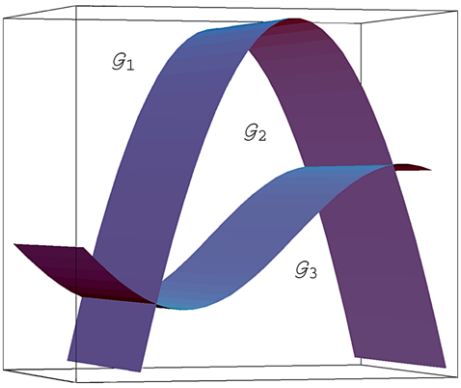


Fig. 3.7 Regularity regions bordered by critical points

Example 3.36 (Regularity regions with different characteristics). Let the function $\alpha \in \mathcal{C}^1((-\infty,\infty),\mathbb{R})$ be given as $\alpha(s) = \begin{cases} s^2 & \text{for } s > 0 \\ 0 & \text{for } s \le 0. \end{cases}$
Consider the DAE

$$\begin{aligned} x_1'(t) - x_2(t) + x_3(t) &= 0, \\ x_2'(t) + x_1(t) &= 0, \\ x_1(t)^3 + \alpha(x_1(t))x_3(t) - (\sin t)^3 &= 0, \end{aligned} \tag{3.30}$$

which has the form (3.1) and satisfies Assumption 3.16 with

$$d(x,t) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad f(y,x,t) = \begin{bmatrix} y_1 - x_2 + x_3 \\ y_2 + x_1 \\ x_1^3 + \alpha(x_1)x_3 - (\sin t)^3 \end{bmatrix},$$
$$y \in \mathbb{R}^2, \quad x \in \mathcal{D}_f = \mathbb{R}^3, \quad t \in \mathcal{J}_f = \mathbb{R}.$$

Compute

$$G_0 = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix}, \qquad B_0 = \begin{bmatrix} & 0 & -1 & 1 \\ & 1 & 0 & 0 \\ 3x_1^2 + \alpha'(x_1)x_3 & 0 & \alpha(x_1) \end{bmatrix},$$

$$Q_0 = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix}, \qquad G_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & \alpha(x_1) \end{bmatrix}.$$

This makes it clear that the DAE is regular with characteristics $r_0 = 2$, $r_1 = 3$, $\mu = 1$ for $x_1 > 0$, i.e., on the regularity region $\mathcal{G}_1 := \{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x_1 > 0\}$.
For $x_1 < 0$, we obtain further

$$G_1 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, Q_1 = \begin{bmatrix} 1 & & \\ & 0 & \\ -1 & & 0 \end{bmatrix}, G_2 = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 3x_1^2 & 0 & 0 \end{bmatrix},$$

and hence the DAE is regular with characteristics $r_0 = r_1 = 2$, $r_2 = 3$, $\mu = 2$ on the regularity region $\mathcal{G}_2 := \{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x_1 < 0\}$.
For every given reference function x_*, the linearization has the form

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} x(t) \right)' + \begin{bmatrix} & 0 & -1 & 1 \\ & 1 & 0 & 0 \\ 3x_{*1}^2 + \alpha'(x_{*1})x_{*3} & 0 & \alpha(x_{*1}) \end{bmatrix} x(t) = q(t). \tag{3.31}$$

This linear DAE is regular with index 1 on intervals where $x_{*1}(t) > 0$, and it is regular with index 2 on intervals where $x_{*1}(t) < 0$. On intervals where $x_{*1}(t) = 0$ the linear DAE (3.31) is no longer regular, because then $G_1 = G_2$ and there does not exist any admissible projector Q_2.
In particular, the reference function $x_*(t) = \begin{bmatrix} \sin t \\ \cos t \\ 0 \end{bmatrix}$ represents a periodic solution of the original nonlinear DAE (3.30). It shuttles between $\mathcal{G}_1$ and $\mathcal{G}_2$. The corresponding linear DAE (3.31) reads

$$\begin{aligned} x_1'(t) - x_2(t) + x_3(t) &= q_1(t), \\ x_2'(t) + x_1(t) &= q_2(t), \\ 3(\sin t)^2 x_1(t) + \alpha(\sin t)x_3(t) - (\sin t)^3 &= q_3(t). \end{aligned}$$

This linear DAE is regular with index 1 on all intervals where $\sin t$ is strictly positive, and regular with index 2 on all intervals where $\sin t$ is strictly negative. □

Theorem 3.37. (Stability with respect to perturbations)
If the DAE (3.1) is, on the open set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$, regular with tractability index μ and characteristics $r_0, \ldots, r_\mu$, then the perturbed DAE

$$f((d(x(t),t))', x(t), t) = q(t), \tag{3.32}$$

with at least continuous perturbation $q : \mathcal{I}_f \to \mathbb{R}^m$, is also regular on $\mathcal{G}$ with the same index and the same characteristics.

Proof. The assertion is evident since the admissible matrix function sequences are constructed just from d, f_y and f_x. □

Theorem 3.33 provides the basis of practical index calculations and index monitoring. The admissible matrix function sequence with the involved partial derivatives of the projector functions is rather intended for theoretical investigations. Even if the partial derivatives were available in practice, the amount seems to be far from being reasonable. Owing to Theorem 3.33, one can choose reference functions and then turn to linearizations involving time derivatives only. In this way, necessary regularity conditions can be checked in practice, see Sections 7.4 and 8.1.

It is favorable if one can benefit from structural properties. For instance, so-called Hessenberg form DAEs of arbitrary size are always regular DAEs (cf. Section 3.5). Also the so-called MNA-DAEs (cf. Section 3.6) show a very useful structure.

To check the regularity of a given DAE or to monitor its index and characteristic values, one can save computations on the last level of the admissible matrix function sequence. Instead of generating the *admissible* projector $Q_{\mu-1}$, the term $B_{\mu-1}$ housing the derivative $(D\Pi_{\mu-1}D^-)'$ and G_μ, one can make do with cheaper expressions due to the next proposition.

Proposition 3.38. (Modified regularity condition) *Let the DAE (3.1) satisfy Assumption 3.16 with $k = m$. Let f and d be sufficiently smooth on the open connected set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$.*
Then the DAE (3.1) is regular on $\mathcal{G}$ with tractability index $\mu \geq 2$, precisely if there are projector functions $Q_0, \ldots, Q_{\mu-2}$ admissible on $\mathcal{G}$, the matrix function $G_{\mu-1}$ has constant rank $r_{\mu-1} < m$ and one of the matrix functions

$$G_{\mu-1} + \mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1} = G_{\mu-1} + \mathcal{W}_{\mu-1}B_{\mu-2}\tilde{Q}_{\mu-1},\ G_{\mu-1} + B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}, \quad (3.33)$$

which are built by an arbitrary projector function $\tilde{Q}_{\mu-1}$ onto $\ker G_{\mu-1}$, is nonsingular.

Proof. Let the DAE be regular with index μ and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions. Let $\tilde{Q}_{\mu-1}$ be an arbitrary projector function onto $\ker G_{\mu-1}$. Then the relations

$$G_\mu = (G_{\mu-1} + B_{\mu-2}P_{\mu-2}Q_{\mu-1})(I - P_{\mu-1}D^-(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-2}Q_{\mu-1})$$
$$G_{\mu-1} + B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1} = (G_{\mu-1} + B_{\mu-2}P_{\mu-2}Q_{\mu-1})(P_{\mu-1} + \tilde{Q}_{\mu-1})$$

show nonsingular factors on the furthermost right-hand side. By Proposition 3.20, it holds that

$$\operatorname{im} G_\mu = \operatorname{im} G_{\mu-1} \oplus \operatorname{im} \mathcal{W}_{\mu-1}BQ_{\mu-1} = \operatorname{im} G_{\mu-1} \oplus \operatorname{im} \mathcal{W}_{\mu-1}B_{\mu-2}Q_{\mu-1}.$$

Regarding

$$G_{\mu-1}+\mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1}=(G_{\mu-1}+\mathcal{W}_{\mu-1}BQ_{\mu-1})(P_{\mu-1}+\tilde{Q}_{\mu-1}),$$
$$G_{\mu-1}+\mathcal{W}_{\mu-1}B_{\mu-2}\tilde{Q}_{\mu-1}=(G_{\mu-1}+\mathcal{W}_{\mu-1}B_{\mu-2}Q_{\mu-1})(P_{\mu-1}+\tilde{Q}_{\mu-1}),$$

altogether it follows that the matrix functions (3.33) are nonsingular simultaneously with G_μ.

Conversely, let $Q_0,\dots,Q_{\mu-2}$ be admissible, and $G_{\mu-1}$ have constant rank $r_{\mu-1}<m$. Introduce the subspace $S_{\mu-1}=\ker\mathcal{W}_{\mu-1}B=\ker\mathcal{W}_{\mu-1}B\Pi_{\mu-2}$. The inclusion $N_0+\cdots+N_{\mu-2}\subseteq S_{\mu-1}$ is evident. If the first matrix function $G_{\mu-1}+\mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1}$ is nonsingular, then $N_{\mu-1}\cap S_{\mu-1}=\{0\}$ must be valid, thus $N_0+\cdots+N_{\mu-2}\subseteq S_{\mu-1}\cap N_{\mu-1}=\{0\}$. Therefore, we can choose a projector function $Q_{\mu-1}$ such that $Q_0,\dots,Q_{\mu-1}$ are admissible. The resulting G_μ is nonsingular.
If the other matrix function $G_{\mu-1}+B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}$ is nonsingular, the $G_{\mu-1}+\mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1}$ is so, too, because of the representation

$$\begin{aligned}G_{\mu-1}+B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}&=G_{\mu-1}+\mathcal{W}_{\mu-1}B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}\\&\quad+(I-\mathcal{W}_{\mu-1})B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}\\&=G_{\mu-1}+\mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1}+G_{\mu-1}\tilde{G}_{\mu-1}^{-}B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}\\&=(G_{\mu-1}+\mathcal{W}_{\mu-1}B\tilde{Q}_{\mu-1})(I+\tilde{G}_{\mu-1}^{-}B_{\mu-2}P_{\mu-2}\tilde{Q}_{\mu-1}),\end{aligned}$$

whereby $\tilde{G}_{\mu-1}^{-}$ denotes the reflexive generalized inverse of $G_{\mu-1}$ fixed by the four properties $G_{\mu-1}\tilde{G}_{\mu-1}^{-}G_{\mu-1}=G_{\mu-1}$, $\tilde{G}_{\mu-1}^{-}G_{\mu-1}\tilde{G}_{\mu-1}^{-}=\tilde{G}_{\mu-1}^{-}$, $\tilde{G}_{\mu-1}^{-}G_{\mu-1}=\tilde{P}_{\mu-1}$, $G_{\mu-1}\tilde{G}_{\mu-1}^{-}=(I-\mathcal{W}_{\mu-1})$. The above arguments apply again. □

3.4 Transformation invariance

What happens with the DAE (3.1) if we transform the unknown function $x(t)=k(\tilde{x}(t),t)$ and turn to the transformed DAE

$$\tilde{f}((\tilde{d}(\tilde{x}(t),t))',\tilde{x}(t),t)=0\,? \tag{3.34}$$

Has the new DAE a properly stated leading term, too? Do the characteristic values change, and is regularity actually maintained? We shall find answers to these questions. It is already known by Theorem 2.18 that, in the case of linear DAEs and linear transformations, the characteristic values do not change.

Let the basic assumptions 3.16 for the DAE (3.1) be satisfied, and let $\mathcal{D}\times\mathcal{I}\subseteq\mathcal{D}_f\times\mathcal{I}_f$ be the open set on which we intend to realize a local transformation. Let $h\in\mathcal{C}^\nu(\mathcal{D}\times\mathcal{I},\mathbb{R}^m)$, $k\in\mathcal{C}^\nu(\tilde{\mathcal{D}}\times\mathcal{I},\mathbb{R}^m)$, with $\nu\in\mathbb{N}$, be such that, for each $t\in\mathcal{I}$, $h(\cdot,t)$ and $k(\cdot,t)$ act bijectively from $\mathcal{D}$ onto $\tilde{\mathcal{D}}$ and from $\tilde{\mathcal{D}}$ onto $\mathcal{D}$, respectively, and $h(\cdot,t)$, $k(\cdot,t)$ are inverse to each other, i.e.,

$$\tilde{x} = h(k(\tilde{x},t),t), \quad \tilde{x} \in \tilde{\mathcal{D}},$$
$$x = k(h(x,t),t), \quad x \in \mathcal{D}.$$

Then, the partial derivatives

$$K(\tilde{x},t) := k_{\tilde{x}}(\tilde{x},t), \quad H(x,t) := h_x(x,t),$$

remain nonsingular on their definition domains, and it holds that

$$K(\tilde{x},t) = H(k(\tilde{x},t),t)^{-1}, \quad H(x,t) = K(h(x,t),t)^{-1}.$$

We speak then of *regular local transformations*. The transformed DAE (3.34) is now given by

$$\tilde{f}(y,\tilde{x},t) := f(y,k(\tilde{x},t),t), \quad y \in \mathbb{R}^n,\ \tilde{x} \in \tilde{\mathcal{D}},\ t \in \mathcal{I},$$
$$\tilde{d}(\tilde{x},t) := d(k(\tilde{x},t),t), \quad \tilde{x} \in \tilde{\mathcal{D}},\ t \in \mathcal{I}.$$

The first partial derivatives to be used for the matrix function sequence are

$$\begin{aligned}
\tilde{f}_y(y,\tilde{x},t) &= f_y(y,k(\tilde{x},t),t),\\
\tilde{f}_{\tilde{x}}(y,\tilde{x},t) &= f_x(y,k(\tilde{x},t),t)K(\tilde{x},t),\\
\tilde{d}_{\tilde{x}}(\tilde{x},t) &= d_x(k(\tilde{x},t),t)K(\tilde{x},t),\\
\tilde{d}_t(\tilde{x},t) &= d_t(k(\tilde{x},t),t) + d_x(k(\tilde{x},t),t)k_t(\tilde{x},t).
\end{aligned}$$

Since k is continuously differentiable, the subspaces $\ker \tilde{f}_y$ and $\operatorname{im} \tilde{d}_x$ are $\mathcal{C}^1$-subspaces on $\mathbb{R}^n \times \tilde{\mathcal{D}} \times \mathcal{I}$. From the transversality of $\ker f_y$ and $\operatorname{im} d_x$ it follows that

$$\mathbb{R}^n = f_y(y,k(\tilde{x},t),t) \oplus \operatorname{im} d_x(k(\tilde{x},t),t) = \tilde{f}_y(y,\tilde{x},t) \oplus \operatorname{im} \tilde{d}_{\tilde{x}}(\tilde{x},t),$$

and hence the DAE (3.34) inherits the properly stated leading term from the original DAE (3.1).

Theorem 3.39. *Let the DAE (3.1) satisfy the basic assumptions 3.16. Let $\mathcal{D} \times \mathcal{I} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open, and let the regular local transformations h, k be ν times continuously differentiable, $\nu \geq 1$. If $\ker f_y(y,x,t)$ depends on y, then $\nu \geq 2$.*

(1) *Then the basic assumption 3.16 holds true for the transformed DAE (3.34), too. In particular, the DAE (3.34) has a properly stated leading term.*

(2) *If there are projector functions $Q_0, \ldots, Q_\kappa$ admissible on $\mathcal{D} \times \mathcal{I}$ for the original DAE (3.1), and $\nu \geq \kappa + 1$, then there are also projector functions $\tilde{Q}_0, \ldots, \tilde{Q}_\kappa$ admissible on $\tilde{\mathcal{D}} \times \mathcal{I}$ accompanying the transformed DAE (3.34). It holds that $\tilde{r}_i = r_i$, $\tilde{u}_i = u_i$, $i = 0, \ldots, \kappa$.*

(3) *If the given DAE (3.1) is regular on $\mathcal{D} \times \mathcal{I}$ with index μ, and if $\nu \geq \mu$, then the transformed DAE (3.34) is regular on $\tilde{\mathcal{D}} \times \mathcal{I}$. It has the same index μ as well as the same characteristic values as the DAE (3.1).*

Proof. The first assertion is already verified. The third assertion is an immediate consequence of the second one. To prove the second assertion we form step-

wise the matrix function sequence for the transformed DAE. Let the admissible matrix function sequence $G_0,\dots,G_\kappa$ be given for the DAE (3.1), $G_0 = G_0(x,t)$, $G_i = G_i(x^i,\dots,x^1,x,t)$ if $i \ge 1$.
The transformations h and k provide at the same time the following one-to-one correspondence between the original and transformed jet variables up to level κ:

$$\begin{aligned} x^1 &= K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t), \qquad k^{[0]}(\tilde{x},t) := k_t(\tilde{x},t), \\ \tilde{x}^1 &= H(x,t)x^1 + h^{[0]}(x,t), \qquad h^{[0]}(x,t) := h_t(x,t), \end{aligned} \tag{3.35}$$

and, for $j = 1,\dots,\kappa-1$,

$$\begin{aligned} x^{j+1} &= K(\tilde{x},t)\tilde{x}^{j+1} + k^{[j]}(\tilde{x}^j,\dots,\tilde{x}^1,\tilde{x},t), \\ \tilde{x}^{j+1} &= H(x,t)x^{j+1} + h^{[j]}(x^j,\dots,x^1,x,t), \end{aligned}$$

with

$$\begin{aligned} k^{[j]}(\tilde{x}^j,\dots,\tilde{x}^1,\tilde{x},t) &:= \big(K(\tilde{x},t)\tilde{x}^j + k^{[j-1]}(\tilde{x}^{j-1},\dots,\tilde{x}^1,\tilde{x},t)\big)_{\tilde{x}}\tilde{x}^1 \\ &\quad + \big(K(\tilde{x},t)\tilde{x}^j + k^{[j-1]}(\tilde{x}^{j-1},\dots,\tilde{x}^1,\tilde{x},t)\big)_t + \sum_{\ell=1}^{j-1} k^{[j-1]}_{\tilde{x}^\ell}(\tilde{x}^{j-1},\dots,\tilde{x}^1,\tilde{x},t)\tilde{x}^{\ell+1}, \end{aligned}$$

and an analogous $h^{[j]}(x^j,\dots,x^1,x,t)$. Notice that $\nu \ge \kappa+1$ ensures that all functions $k^{[0]},\dots,k^{[\kappa-1]}$ are continuously differentiable. Denote

$$\begin{aligned} \tilde{D}(\tilde{x},t) &:= \tilde{d}_{\tilde{x}}(\tilde{x},t) = d_x(k(\tilde{x},t),t)K(\tilde{x},t) = D(k(\tilde{x},t),t)K(\tilde{x},t), \\ \tilde{Q}_0(\tilde{x},t) &:= K(\tilde{x},t)^{-1}Q_0(k(\tilde{x},t),t)K(\tilde{x},t). \end{aligned}$$

$\tilde{Q}_0$ is a continuous projector function onto $\ker \tilde{D}(\tilde{x},t)$, thus $\tilde{Q}_0$ is admissible, and we are done if $\kappa = 0$. Assume $\kappa \ge 1$. Introduce further (cf. (3.14)–(3.15))

$$\begin{aligned} \tilde{A}(\tilde{x}^1,\tilde{x},t) &:= f_y(\tilde{D}(\tilde{x},t)\tilde{x}^1 + \tilde{d}_t(\tilde{x},t),k(\tilde{x},t),t), \\ \tilde{B}(\tilde{x}^1,\tilde{x},t) &:= f_x(\tilde{D}(\tilde{x},t)\tilde{x}^1 + \tilde{d}_t(\tilde{x},t),k(\tilde{x},t),t)K(\tilde{x},t), \\ \tilde{G}_0(\tilde{x}^1,\tilde{x},t) &:= \tilde{A}(\tilde{x}^1,\tilde{x},t)\tilde{D}(\tilde{x},t), \end{aligned}$$

to begin the matrix function sequence with. By means of the correspondence (3.35) we derive

$$\begin{aligned} \tilde{D}(\tilde{x},t)\tilde{x}^1 + \tilde{d}_t(\tilde{x},t) &= D(k(\tilde{x},t),t)K(\tilde{x},t)\tilde{x}^1 + d_t(k(\tilde{x},t),t) + D(k(\tilde{x},t),t)k_t(\tilde{x},t) \\ &= D(k(\tilde{x},t),t)\big\{K(\tilde{x},t)\tilde{x}^1 + k_t(\tilde{x},t)\big\} + d_t(k(\tilde{x},t),t) \\ &= D(x,t)x^1 + d_t(x,t), \end{aligned} \tag{3.36}$$

and this yields

$$\begin{aligned}\tilde{A}(\tilde{x}^1,\tilde{x},t) &= A(x^1,x,t),\\ \tilde{B}(\tilde{x}^1,\tilde{x},t) &= B(x^1,x,t)K(\tilde{x},t),\\ \tilde{G}_0(\tilde{x}^1,\tilde{x},t) &= G_0(x^1,x,t)K(\tilde{x},t),\end{aligned}$$

and, further, for the border projector $\tilde{R}(\tilde{x}^1,\tilde{x},t)$ which accompanies the decomposition (cf. (3.16), and Definition 3.15)

$$\mathbb{R}^n = \ker\tilde{A}(\tilde{x}^1,\tilde{x},t)\oplus \operatorname{im}\tilde{D}(\tilde{x},t)$$

we arrive at

$$\tilde{R}(\tilde{x}^1,\tilde{x},t) = R(x^1,x,t) = R\big(K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t),k(\tilde{x},t),t\big).$$

With (cf. (3.18))

$$\tilde{D}(\tilde{x}^1,\tilde{x},t)^- := K(\tilde{x},t)^{-1}D\big(K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t),k(\tilde{x},t),t\big)^-,$$

a continuous generalized inverse of $\tilde{D}(\tilde{x},t)$ is given, and

$$\tilde{D}\tilde{D}^- = \tilde{R},\quad \tilde{D}^-\tilde{D} = \tilde{P}_0.$$

Compute

$$\begin{aligned}\tilde{G}_1(\tilde{x}^1,\tilde{x},t) &= \tilde{G}_0(\tilde{x}^1,\tilde{x},t) + \tilde{B}(\tilde{x}^1,\tilde{x},t)\tilde{Q}_0(\tilde{x},t)\\ &= G_1(K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t),k(\tilde{x},t),t)K(\tilde{x},t),\\ \tilde{r}_1 &= r_1,\\ \tilde{N}_1(\tilde{x}^1,\tilde{x},t) &= K(\tilde{x},t)^{-1}N_1(x^1,x,t),\\ \tilde{N}_1(\tilde{x}^1,\tilde{x},t)\cap\tilde{N}_0(\tilde{x},t) &= K(\tilde{x},t)^{-1}(N_1(x^1,x,t)\cap N_0(x,t)),\\ \tilde{u}_1 &= u_1.\end{aligned}$$

The choice

$$\tilde{Q}_1(\tilde{x}^1,\tilde{x},t) := K(\tilde{x},t)^{-1}Q_1\big(K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t),k(\tilde{x},t),t\big)K(\tilde{x},t)$$

yields a continuous projector function $\tilde{Q}_1$ onto $\tilde{N}_1$ such that $\tilde{X}_1 \subseteq \ker\tilde{Q}_1$, $\tilde{X}_1 := K^{-1}X_1$ (cf. Definition 3.21), and, moreover,

$$\tilde{D}\tilde{\Pi}_1\tilde{D}^- = DKK^{-1}P_0KK^{-1}P_1KK^{-1}D^- = DP_0P_1D^- = D\Pi_1D^-,$$

i.e.,

$$(\tilde{D}\tilde{\Pi}_1\tilde{D}^-)(\tilde{x}_1,\tilde{x},t) = (D\Pi_1D^-)(x^1,x,t) = (D\Pi_1D^-)\big(K(\tilde{x},t)\tilde{x}^1 + k^{[0]}(\tilde{x},t),k(\tilde{x},t),t\big),$$

hence, $\tilde{D}\tilde{\Pi}_1\tilde{D}^-$ inherits the $\mathcal{C}^1$ property from $D\Pi_1D^-$, and $\tilde{Q}_0$, $\tilde{Q}_1$ are shown to be admissible on $\tilde{\mathcal{D}}\times\mathcal{I}$. Compute further the total derivative

$$
\begin{aligned}
(\tilde{D}\tilde{\Pi}_1\tilde{D}^-)' &= (\tilde{D}\tilde{\Pi}_1\tilde{D}^-)_{\tilde{x}^1}\tilde{x}^2 + (\tilde{D}\tilde{\Pi}_1\tilde{D}^-)_{\tilde{x}}\tilde{x}^1 + (\tilde{D}\tilde{\Pi}_1\tilde{D}^-)_t \\
&= (D\Pi_1 D^-)_{x^1} K\tilde{x}^2 + (D\Pi_1 D^-)_{x^1}\left(K\tilde{x}^1 + k^{[0]}\right)_{\tilde{x}}\tilde{x}^1 + (D\Pi_1 D^-)_x K\tilde{x}^1 \\
&\quad + (D\Pi_1 D^-)_{x^1}\left(K\tilde{x}^1 + k^{[0]}\right)_t + (D\Pi_1 D^-)_x k_t + (D\Pi_1 D^-)_t \\
&= (D\Pi_1 D^-)_{x^1} x^2 + (D\Pi_1 D^-)_x x^1 + (D\Pi_1 D^-)_t = (D\Pi_1 D^-)',
\end{aligned}
$$

as well as $\tilde{B}_1 = B_1 K$. To apply induction we assume $\tilde{Q}_0, \ldots, \tilde{Q}_i$ to be admissible on $\tilde{\mathcal{D}} \times \mathcal{I}$, and

$$
\tilde{G}_j = G_j K, \quad \tilde{Q}_j = K^{-1} Q_j K, \quad \tilde{B}_j = B_j K, \quad j = 0, \ldots, i.
$$

Form $\tilde{G}_{i+1} = \tilde{G}_i + \tilde{B}_i \tilde{Q}_i = G_{i+1} K$ and choose $\tilde{Q}_{i+1} := K^{-1} Q_{i+1} K$. $\tilde{G}_{i+1}$ has constant rank $\tilde{r}_{i+1} = r_{i+1}$. $\tilde{Q}_{i+1}$ is continuous and projects onto $\tilde{N}_{i+1} = K^{-1} N_{i+1}$. Due to

$$
(\tilde{N}_0 + \cdots + \tilde{N}_i) \cap \tilde{N}_{i+1} = K^{-1}\left((N_0 + \cdots + N_i) \cap N_{i+1}\right)
$$

it follows that $\tilde{u}_{i+1} = u_{i+1}$. Further, it holds that

$$
\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^- = \tilde{D}\tilde{\Pi}_i\tilde{P}_{i+1}\tilde{D}^- = D\Pi_i P_{i+1} D^- = D\Pi_{i+1} D^-,
$$

and in more detail,

$$
\begin{aligned}
&(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)(\tilde{x}^{i+1}, \ldots, \tilde{x}^1, \tilde{x}, t) = (D\Pi_{i+1} D^-)(x^{i+1}, \ldots, x^1, x, t) \\
&= (D\Pi_{i+1} D^-)\left(K(\tilde{x}, t)\tilde{x}^{i+1} + k^{[i]}(\tilde{x}^i, \ldots, \tilde{x}^1, \tilde{x}, t), \ldots, K(\tilde{x}, t)\tilde{x}^1 + k^{[0]}(\tilde{x}, t), k(\tilde{x}, t), t\right).
\end{aligned}
$$

Since $D\Pi_{i+1} D^-$ is continuously differentiable so is $\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-$, thus $\tilde{Q}_0, \ldots, \tilde{Q}_{i+1}$ are admissible. Compute the partial derivatives

$$
\begin{aligned}
(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_{\tilde{x}^{i+1}} &= (D\Pi_{i+1} D^-)_{x^{i+1}} K, \\
(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_{\tilde{x}^j} &= \sum_{\ell=j+1}^{i+1} (D\Pi_{i+1} D^-)_{x^\ell} k^{[\ell-1]}_{\tilde{x}^j} + (D\Pi_{i+1} D^-)_{x^j} K, \quad j = 1, \ldots, i, \\
(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_{\tilde{x}} &= \sum_{\ell=1}^{i+1} (D\Pi_{i+1} D^-)_{x^\ell}\left(K\tilde{x}^\ell + k^{[\ell-1]}\right)_{\tilde{x}} + (D\Pi_{i+1} D^-)_x K, \\
(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_t &= \sum_{\ell=1}^{i+1} (D\Pi_{i+1} D^-)_{x^\ell}\left(K\tilde{x}^\ell + k^{[\ell-1]}\right)_t + (D\Pi_{i+1} D^-)_x k_t + (D\Pi_{i+1} D^-)_t
\end{aligned}
$$

and then the total derivative

$$
\begin{aligned}
(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)' &= \sum_{j=1}^{i+1}(\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_{\tilde{x}^j}\tilde{x}^{j+1} + (\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_{\tilde{x}}\tilde{x}^1 + (\tilde{D}\tilde{\Pi}_{i+1}\tilde{D}^-)_t \\
&= \sum_{\ell=1}^{i+1}(D\Pi_{i+1}D^-)_{x^\ell}\Big\{K\tilde{x}^{\ell+1} + \sum_{j=1}^{\ell-1}k_{\tilde{x}^j}^{[\ell-1]}\tilde{x}^{j+1} + \big(K\tilde{x}^\ell + k^{[\ell-1]}\big)_{\tilde{x}}\tilde{x}^1 \\
&\qquad + \big(K\tilde{x}^\ell + k^{[\ell-1]}\big)_t\Big\} + (D\Pi_{i+1}D^-)_x\big(K\tilde{x}^1 + k_t\big) + (D\Pi_{i+1}D^-)_t \\
&= \sum_{\ell=1}^{i+1}(D\Pi_{i+1}D^-)_{x^\ell}x^{\ell+1} + (D\Pi_{i+1}D^-)_x x^1 + (D\Pi_{i+1}D^-)_t \\
&= (D\Pi_{i+1}D^-)'.
\end{aligned}
$$

Finally, $\tilde{B}_{i+1} = B_{i+1}K$ follows, and this completes the proof. □

Theorem 3.39 applies to general DAEs (3.1) comprising k equations but the unknown function has m components.

3.5 Hessenberg form DAEs of arbitrary size

Hessenberg form DAEs are semi-explicit systems with a special structure

$$
\begin{aligned}
x_1'(t) \qquad\qquad\qquad &+ b_1(x_1(t),\dots,x_{r-1}(t),x_r(t),t) = 0,\\
x_2'(t) \qquad\qquad &+ b_2(x_1(t),\dots,x_{r-1}(t),t) = 0,\\
x_3'(t) \qquad &+ b_3(x_2(t),\dots,x_{r-1}(t),t) = 0,\\
&\cdots\\
x_{r-1}'(t) &+ b_{r-1}(x_{r-2}(t),x_{r-1}(t),t) = 0,\\
&\quad b_r(x_{r-1}(t),t) = 0,
\end{aligned}
\tag{3.37}
$$

with $m_1 + \cdots + m_{r-1} + m_r = m$ equations, $m_r > 0$, and a function $b : \mathcal{D}_b \times \mathcal{I}_b \to \mathbb{R}^m$ being at least continuous together with the partial derivative b_x. $\mathcal{D}_b \subseteq \mathbb{R}^m$ is open, $\mathcal{I}_b \subseteq \mathbb{R}$ is an interval, $r \geq 2$ is an integer. The partial derivative

$$
b_x = \begin{bmatrix}
B_{11} & \dots & B_{1,r-1} & B_{1r} \\
B_{21} & \ddots & \vdots & 0 \\
 & \ddots & B_{r-1,r-1} & \\
 & & B_{r,r-1} & 0
\end{bmatrix}
\begin{matrix}
\}m_1\\ \}m_2\\ \\ \}m_{r-1}\\ \}m_r
\end{matrix}
$$

with $B_{ij} := b_{i,x_j}$ shows the *Hessenberg structure* from which the name comes.

Definition 3.40. The system (3.37) is said to be a DAE in *Hessenberg form of size* r, if the matrix function product

$$
B_{r,r-1}\cdots B_{21}B_{1r} \tag{3.38}
$$

remains nonsingular on $\mathcal{D}_b \times \mathcal{I}_b$.

We put (3.37) into the general form (3.1) by means of $n = m_1 + \cdots + m_{r-1}$,

$$f(y,x,t) = Ay + b(x,t), \quad d(x,t) = Dx, \quad x \in \mathcal{D}_b,\ t \in \mathcal{I}_b,\ y \in \mathbb{R},$$

and

$$A := \begin{bmatrix} I & & \\ & \ddots & \\ & & I \\ & & 0 \end{bmatrix}, \quad D := \begin{bmatrix} I & & \\ & \ddots & \\ & & I\ 0 \end{bmatrix}, \quad D^- := A,$$

such that $\ker A = \{0\}$, $\operatorname{im} D = \mathbb{R}^n$. Then the DAE (3.37) has a properly stated leading term, the border projector is simply $R = I$, and Assumption 3.16 applies.
At this point we call attention to the fact that in the present section the integer r indicates the size of the Hessenberg structure. We do not use it here for $\operatorname{rank} D = m - m_r$, but we use only $r_0 = \operatorname{rank} G_0 = \operatorname{rank} D$ so that no confusion can arise.

Hessenberg form size-2 DAEs were already considered in Section 3.3, Example 3.32. It is shown there that any Hessenberg size-2 DAE is regular with tractability index 2 on the definition domain $\mathcal{D}_b \times \mathcal{I}_b$. Example 2.11 in Subsection 2.2.2 provides an admissible matrix function sequence for a linear Hessenberg size-3 DAE and the characteristic values $r_0 = r_1 = r_2 = m - m_3 < m$ and $r_3 = m$, thus $\mu = 3$. This sequence further shows the impact of the time-varying subspaces $\operatorname{im} B_{13}$ and $\operatorname{im} B_{21}B_{13}$, which is responsible for the time derivatives of the projector functions within the admissible matrix functions. In the case of nonlinear DAEs considered now, these subspaces may additionally depend on x, so that the jet variables may come in. The most important class of Hessenberg form DAEs are those of size 3, among them the equations describing the motion of constrained multibody systems.

Example 3.41 (Constrained multibody system). After [63, Chapter 1] the general first-order form of the equation of motion of a constrained multibody system reads

$$p' = Z(p)v, \tag{3.39}$$
$$Mv' = f_a(t,p,v,s) - Z(p)^* G(p)^* \lambda, \tag{3.40}$$
$$s' = f_s(t,p,v,s), \tag{3.41}$$
$$0 = g(p), \tag{3.42}$$

where p and s contain position coordinates, v velocities, and λ is a Lagrange multiplier.
The constraint (3.42) defines a manifold of free motion. $G(p) := g_p(p)$ is the constraint matrix, and the generalized constraint forces $G(p)^*\lambda$ are responsible for the constraint to be satisfied. M denotes the given positive definite mass matrix, and f_a contains the applied forces. Equation (3.40) arises from Newton's law completed by d'Alembert's principle.

$Z(p)$ is a nonsingular transformation matrix. Equation (3.39) is the kinematic differential equation. Equation (3.41) represents the influence of further components (electromagnetic forces, hydraulic components, control devices, etc.).
The positions and velocities are expected to be continuously differentiable, while the Lagrange multipliers are usually less smooth.
Multiply equation (3.40) by M^{-1} and move the top equation (3.39) to the third place so that the semi-explicit system

$$\begin{aligned} v' &= M^{-1} f_a(t,p,v,s) - M^{-1} Z(p)^* G(p)^* \lambda, \\ s' &= f_s(t,p,v,s), \\ p' &= Z(p)v, \\ 0 &= g(p), \end{aligned}$$

results. Set $x_1 := \begin{bmatrix} v \\ s \end{bmatrix}$, $x_2 = p$, $x_3 := \lambda$, which allows us to write the system in Hessenberg form (3.37) with size $r = 3$. The resulting partial Jacobian b_x has the particular entries

$$B_{13} = \begin{bmatrix} M^{-1} Z^* G^* \\ 0 \end{bmatrix}, \quad B_{21} = \begin{bmatrix} Z & 0 \end{bmatrix}, \quad B_{32} = G,$$

yielding the product

$$B_{32} B_{21} B_{13} = G Z M^{-1} Z^* G^*.$$

The common demand for $G = g_p$ to have full row rank, which excludes redundant constraints, ensures the product $B_{32}B_{21}B_{13}$ remains nonsingular (cf. [63], also Lemma 3.44 below). □

Theorem 3.42. *Any Hessenberg form DAE (3.37) with size r and sufficiently smooth $b : \mathcal{D}_b \times \mathcal{I}_b \to \mathbb{R}^m$ is regular on $\mathcal{D}_b \times \mathcal{I}_b$ with tractability index r and characteristic values*

$$r_0 = \cdots = r_{r-1} = m - m_r, \quad r_r = m.$$

Theorem 3.42 attests to the structure of Hessenberg systems to be very special. In particular, the nilpotent matrix $\mathcal{N}$ within the Weierstraß–Kronecker form (cf. Proposition 1.3) of a linear constant coefficient DAE in Hessenberg size-r form consists exclusively of nilpotent Jordan blocks of uniform order r.

Proof. This statement is proved by providing an admissible projector function sequence yielding a nonsingular matrix function G_r. We apply an inductive proof including a certain amount of technical computations.
Since the product (3.38) remains nonsingular, the blocks

$$B_{1r},\ B_{21}B_{1r},\ \ldots,\ B_{r-1,r-2}\cdots B_{21}B_{1r} \tag{3.43}$$

have full column rank m_r. Then, the subspaces $\operatorname{im} B_{1r}$, $\operatorname{im} B_{21}B_{1r}, \ldots,$ $\operatorname{im} B_{r-1,r-2}\cdots B_{21}B_{1r}$ are at least $\mathcal{C}$-subspaces. We suppose that b is smooth enough

to make them $\mathcal{C}^1$-subspaces. Introduce continuously differentiable projectors $\Omega_1, \ldots, \Omega_{r-1}$ onto $\operatorname{im} B_{1r}$, $\operatorname{im} B_{21}B_{1r}, \ldots, \operatorname{im} B_{r-1,r-2}\cdots B_{21}B_{1r}$, respectively. With the use of generalized inverses we may represent

$$\Omega_1 = B_{1r}B_{1r}^-, \quad \Omega_2 = B_{21}B_{1r}(B_{21}B_{1r})^-, \ldots,$$
$$\Omega_{r-1} = B_{r-1,r-2}\cdots B_{21}B_{1r}(B_{r-1,r-2}\cdots B_{21}B_{1r})^-.$$

Since the blocks (3.43) have full column rank, it holds that

$$\begin{aligned} B_{1r}^- B_{1r} = I,\ (B_{21}B_{1r})^- B_{21}B_{1r} = I, \ldots, \\ (B_{r-1,r-2}\cdots B_{21}B_{1r})^- B_{r-1,r-2}\cdots B_{21}B_{1r} = I. \end{aligned} \tag{3.44}$$

Then, for $\ell = 1, \ldots, r-2$, it is easily checked that

$$\operatorname{im} B_{\ell+1,\ell}\Omega_\ell = \operatorname{im} \Omega_{\ell+1}.$$

Now we compose a matrix function sequence (3.19)–(3.21) and admissible projector functions for (3.37). We begin with $G_0 = AD$, $B_0 = B = b_x$,

$$G_0 = \begin{bmatrix} I & & \\ & \ddots & \\ & & I \\ & & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & & B_{1r} \\ & \ddots & \\ & & I \\ & & 0 \end{bmatrix},$$

$\Pi_0 = P_0 = G_0$, and $r_0 = m - m_r$, $r_1 = r_0$. Describe the nullspace of G_1 by

$$\begin{aligned} N_1 &= \{z \in \mathbb{R}^m : z_1 + B_{1r}z_r = 0, z_2 = 0, \ldots, z_{r-1} = 0\} \\ &= \{z \in \mathbb{R}^m : z_1 = \Omega_1 z_1, z_r = -B_{1r}^- z_1, z_2 = 0, \ldots, z_{r-1} = 0\}, \end{aligned}$$

such that we immediately find a projector onto N_1, namely

$$Q_1 = \begin{bmatrix} \Omega_1 & & \\ & 0 & \\ & & \ddots \\ -B_{1r}^- & & 0 \end{bmatrix}.$$

Observe that $Q_1Q_0 = 0$ is true. Form

$$\Pi_1 = \Pi_0 P_1 = \begin{bmatrix} I - \Omega_1 & & & \\ & I & & \\ & & \ddots & \\ & & & I \\ & & & 0 \end{bmatrix}, \quad D\Pi_1 D^- = \begin{bmatrix} I - \Omega_1 & & & \\ & I & & \\ & & \ddots & \\ & & & I \end{bmatrix},$$

and see that Q_0, Q_1 are admissible. Next we compute

$$B_1 = BP_0 - G_1 D^-(D\Pi_1 D^-)'D = \begin{bmatrix} B_{11} & \dots & B_{1,r-1} & 0 \\ B_{21} & \ddots & \vdots & \vdots \\ & \ddots & B_{r-1,r-1} & 0 \\ & & B_{r,r-1} & 0 \end{bmatrix} + \begin{bmatrix} \Omega_1' & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

and

$$G_2 = \begin{bmatrix} I & & & B_{1r} \\ B_{21}\Omega_1 & I & & \\ & \ddots & & \\ & & I & \\ & & & 0 \end{bmatrix} + C_2,$$

where the matrix function C_2 has the single nontrivial entry $C_{2,11} = (B_{11} + \Omega_1')\Omega_1$. All other blocks in C_2 are zero-blocks. G_2 has constant rank $r_2 = r_0$, if $r > 2$, and full rank $r_2 = m$, if $r = 2$. This can be verified by applying Proposition 3.20 (3), and taking the projectors

$$\mathcal{W}_0 = \mathcal{W}_1 = \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & & I \end{bmatrix} \quad \text{such that} \quad \mathcal{W}_0 B = \mathcal{W}_1 B = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & B_{r,r-1} & 0 \end{bmatrix},$$

and

$$\mathcal{W}_1 B Q_1 = 0 \text{ for } r > 2, \quad \mathcal{W}_1 B Q_1 = \begin{bmatrix} 0 & 0 \\ B_{21} & 0 \end{bmatrix} \begin{bmatrix} \Omega_1 & 0 \\ -B_{12}^- & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ B_{21}\Omega_1 & 0 \end{bmatrix} \text{ for } r = 2.$$

From $\operatorname{rank} G_2 = \operatorname{rank} G_1 + \operatorname{rank} \mathcal{W}_1 B Q_1$ we conclude $r_2 = r_1 = r_0 = m - m_r$ for $r > 2$, and $r_2 = r_1 + \operatorname{rank} B_{21}\Omega_1 = m - m_2 + m_2 = m$ for $r = 2$. For $r = 2$ we are done.

Assume that $r > 2$. For an induction proof we assume the following: $k + 1 \le r$.

(1) $Q_0, \dots, Q_k$ are admissible projectors, and Q_j has the block structure

$$Q_j = \begin{bmatrix} 0 & & * & & & \\ & \ddots & \vdots & & & \\ & & 0 & * & & \\ & & & \Omega_j & & \\ & & & 0 & 0 & \\ & & & \vdots & & \ddots \\ & & & 0 & & \ddots \\ & & & * & & & 0 \end{bmatrix}, \quad j = 1, \dots, k, \tag{3.45}$$

whereby the nontrivial entries in column number j have the property $Q_{j,ij} = Q_{j,ij}\Omega_j$, $i = 1, \dots, j-1$, and $i = r$.

(2) The matrix function G_k has the structure

$$G_k = \begin{bmatrix} I & & & & & B_{1r} \\ B_{21}\Omega_1 & I & & & & \\ & \ddots & \ddots & & & \\ & & B_{k,k-1}\Omega_{k-1} & I & & \\ & & & I & & \\ & & & & \ddots & \\ & & & & I & \\ & & & & & 0 \end{bmatrix} + C_k, \tag{3.46}$$

where $C_k = \begin{bmatrix} * & \dots & * & & & \\ & \ddots & \vdots & & & \\ & & * & & & \\ & & 0 & & & \\ & & & \ddots & & \\ & & & & 0 & \end{bmatrix} \begin{matrix} \\ \\ \\ k-1 \\ k \\ \\ \end{matrix}$

is upper block triangular with nontrivial entries in the first $k-1$ columns. These entries satisfy the condition

$$C_{k,ij} = C_{k,ij}\Omega_j, \quad j = 1,\dots,k-1,\ i = j,\dots,k-1, \tag{3.47}$$

and G_k has constant rank $r_k = m - m_r$.

(3) The projector product Π_k has the structure

$$\Pi_k = \begin{bmatrix} I-\Omega_1 & * & \dots & * & & & \\ & \ddots & \ddots & \vdots & & & \\ & & \ddots & * & & & \\ & & & I-\Omega_k & & & \\ & & & & I & & \\ & & & & & \ddots & \\ & & & & & & I \\ & & & & & & 0 \end{bmatrix}, \tag{3.48}$$

where the nontrivial entries indicated by stars have the properties

$$\Pi_{k,ij} = (I-\Omega_i)\Pi_{k,ij}\Omega_j, \quad i = 1,\dots,k-1,\ j = i+1,\dots,k.$$

(4) The matrix function B_k has the structure

$$B_k = \begin{bmatrix} * & \cdots & \cdots & \cdots & \cdots & * & 0 \\ * & \ddots & & & & \vdots & \vdots \\ & \ddots & \ddots & & & \vdots & \vdots \\ & & * & \ddots & & \vdots & \vdots \\ & & & B_{k+1,k} & \ddots & \vdots & \vdots \\ & & & & \ddots & * & \\ & & & & & B_{r,r-1} & 0 \end{bmatrix}, \tag{3.49}$$

such that

$$B_k Q_k = \begin{bmatrix} 0 & & & * & & & \\ & \ddots & & \vdots & & & \\ & & 0 & \vdots & & & \\ & & & * & & & \\ & & & B_{k+1,k}\Omega_k & & & \\ & & & 0 & & 0 & \\ & & & \vdots & & \ddots & \\ & & & 0 & & & 0 \\ & & & k & & & \end{bmatrix} \begin{matrix} \\ \\ \\ \\ k+1 \\ \\ \\ \\ \end{matrix} \tag{3.50}$$

results.

We have to verify that assumptions (3.45)–(3.49) lead to the same properties for k replaced by $k+1$. First we form

$$G_{k+1} = \begin{bmatrix} I & & & & & & B_{1r} \\ B_{21}\Omega_1 & I & & & & & \\ & \ddots & \ddots & & & & \\ & & B_{k+1,k}\Omega_k & I & & & \\ & & & & I & & \\ & & & & & \ddots & \\ & & & & & I & \\ & & & & & & 0 \end{bmatrix} + C_{k+1},$$

where C_{k+1} results to be upper block triangular with the entries from C_k in columns 1 to $k-1$ and entries from $B_k Q_k$ in the kth column, i.e.,

$$
\begin{aligned}
C_{k+1,ij} &= C_{k,ij}, && j = 1,\dots,k-1,\ i = j,\dots,k-1,\\
C_{k+1,ik} &= (B_k Q_k)_{ik}, && i = 1,\dots,k.
\end{aligned}
$$

Since the nontrivial entries of Q_k have the property $Q_{k,ik} = Q_{k,ik}\Omega_k$, $i = 1,\dots,k$, it follows that

$$C_{k+1,ik} = C_{k+1,ik}\Omega_k, \quad i = 1,\dots,k,$$

and G_{k+1} has the right shape.
Next we describe the nullspace of G_{k+1}. $G_{k+1}z = 0$ implies in detail

$$
\begin{aligned}
z_1 + B_{1r}z_r + \sum_{\ell=1}^{k} C_{k+1,1\ell}z_\ell &= 0,\\
B_{21}\Omega_1 z_1 + z_2 + \sum_{\ell=2}^{k} C_{k+1,2\ell}z_\ell &= 0,\\
&\dots\\
B_{k,k-1}\Omega_{k-1}z_{k-1} + z_k + C_{k+1,kk}z_k &= 0,\\
B_{k+1,k}\Omega_k z_k + z_{k+1} &= 0,\\
z_{k+2} &= 0,\\
&\dots\\
z_{r-1} &= 0.
\end{aligned}
\tag{3.51}
$$

Using the properties resulting from the nonsingularity of the product (3.38) which are described at the beginning of this proof we realize that it makes sense to multiply the first equation in (3.51) by $I - \Omega_1$ and B_{1r}^-, the second one by $(I - \Omega_2)$ and $B_{1r}(B_{21}B_{1r})^-$, and so on, to obtain the equivalent system

$$
\begin{aligned}
(I-\Omega_1)z_1 + \sum_{\ell=1}^{k} (I-\Omega_1)C_{k+1,1\ell}\Omega_\ell z_\ell &= 0,\\
B_{1r}^-\Omega_1 z_1 + z_r + \sum_{\ell=1}^{k} B_{1r}^- C_{k+1,1\ell}\Omega_\ell z_\ell &= 0,\\
(I-\Omega_2)z_2 + \sum_{\ell=2}^{k} (I-\Omega_2)C_{k+1,2\ell}\Omega_\ell z_\ell &= 0,\\
\Omega_1 z_1 + B_{1r}(B_{21}B_{1r})^-\Omega_2 z_2 + \sum_{\ell=2}^{k} B_{1r}(B_{21}B_{1r})^- C_{k+1,2\ell}\Omega_\ell z_\ell &= 0,\\
&\dots\\
(I-\Omega_k)z_k + (I-\Omega_k)C_{k+1,kk}\Omega_k z_k &= 0,\\
\Omega_{k-1}z_{k-1} + B_{k-1,k-2}\cdots B_{21}B_{1r}(B_{k,k-1}\cdots B_{21}B_{1r})^-\Omega_k z_k&\\
+B_{k-1,k-2}\cdots B_{21}B_{1r}(B_{k,k-1}\cdots B_{21}B_{1r})^- C_{k+1,kk}\Omega_k z_k &= 0,\\
(I-\Omega_{k+1})z_{k+1} &= 0,\\
\Omega_k z_k + B_{k,k-1}\cdots B_{21}B_{1r}(B_{k+1,k}\cdots B_{21}B_{1r})^-\Omega_{k+1}z_{k+1} &= 0,\\
z_{k+2} &= 0,\\
&\dots\\
z_{r-1} &= 0.
\end{aligned}
$$

From this new system we observe that $z_{k+2} = 0, \ldots, z_{r-1} = 0$, $z_{k+1} = \Omega_{k+1} z_{k+1}$, and all resulting components can be successively expressed as $z_i = E_i \Omega_{k+1} z_{k+1}$ (with certain coefficients E_i) $i = 1, \ldots, k$ and $i = r$. Therefore,

$$Q_{k+1} = \begin{bmatrix} 0 & & * & & & \\ & \ddots & \vdots & & & \\ & & 0 & * & & \\ & & & \Omega_{k+1} & & \\ & & & 0 & 0 & \\ & & & \vdots & & \ddots \\ & & & 0 & & \\ & & & * & & & 0 \end{bmatrix},$$

with the only nontrivial entries

$$\begin{aligned} Q_{k+1,i,k+1} &= E_i \Omega_{k+1}, \quad i = 1, \ldots, k,\ i = r, \\ Q_{k+1,k+1,k+1} &= \Omega_{k+1}, \end{aligned} \tag{3.52}$$

is a continuous projector onto $N_{k+1} = \ker G_{k+1}$, and it satisfies $Q_{k+1} Q_j = 0$, $j = 0, \ldots, k$.

Notice that G_{k+1} has constant rank $r_{k+1} = m - m_r$, if $k \le r - 1$. Now the projector Π_{k+1} has the structure

$$\Pi_{k+1} = \Pi_k P_{k+1} = \begin{bmatrix} I - \Omega_1 & * & \ldots & * & & & \\ & \ddots & \ddots & \vdots & & & \\ & & & * & & & \\ & & & I - \Omega_{k+1} & & & \\ & & & & I & & \\ & & & & & \ddots & \\ & & & & & & I \\ & & & & & & & 0 \end{bmatrix},$$

where the indicated nontrivial off-diagonal entries possess by construction the properties

$$\Pi_{k+1,ij} = (I - \Omega_i) \Pi_{k+1,ij} \Omega_j, \quad i = 1, \ldots, k,\ j = i+1, \ldots, k+1.$$

Π_{k+1} is continuous, and $D\Pi_{k+1} D^-$ is continuously differentiable, if these off-diagonal entries are so. Taking a closer look at these entries we know that they are formed by the coefficients E_i arising in (3.52), and resulting from the linear system (3.51). This makes it clear that supposing the function b to be smooth enough we may consider $D\Pi_{k+1} D^-$ to be continuously differentiable, and hence $Q_0, \ldots, Q_{k+1}$ to be admissible.

The remaining matrix function B_{k+1} has the structure

$$B_{k+1} = B_k P_k - G_{k+1} D^- (D\Pi_{k+1} D^-)' D\Pi_k$$

$$= \begin{bmatrix} * & \cdots & \cdots & \cdots & * & 0 \\ * & & & & \vdots & \vdots \\ & \ddots & & & \vdots & \vdots \\ & & * & & \vdots & \vdots \\ & & & B_{k+2,k+1} & \vdots & \vdots \\ & & & & * & \vdots \\ & & & & B_{r,r-1} & 0 \end{bmatrix} - \begin{bmatrix} * & \cdots & \cdots & * & & \\ * & \ddots & & \vdots & & \\ & \ddots & \ddots & \vdots & & \\ & & * & * & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & 0 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \\ \end{matrix} k+1$$

$$k+1$$

$$= \begin{bmatrix} * & \cdots & \cdots & \cdots & \cdots & * & 0 \\ * & & & & & \vdots & \vdots \\ & \ddots & & & & \vdots & \vdots \\ & & * & & & \vdots & \vdots \\ & & & B_{k+2,k+1} & & \vdots & \vdots \\ & & & & \ddots & * & \vdots \\ & & & & & B_{r,r-1} & 0 \end{bmatrix},$$

which fits into (3.49) and completes the induction. It turns out that we may form these admissible projectors $Q_0, \ldots, Q_k$ as long as we reach $k+1 = r$.
Let $Q_0, \ldots, , Q_{r-1}$ be already given. We have $r_0 = \cdots = r_{r-1} = m - m_r$, and

$$G_r = G_{r-1} + B_{r-1} Q_{r-1} = \begin{bmatrix} I & & & & B_{1r} \\ B_{21}\Omega_1 & I & & & \\ & \ddots & \ddots & & \\ & & \ddots & I & \\ & & & B_{r,r-1}\Omega_{r-1} & 0 \end{bmatrix} + C_r.$$

It remains to show that G_r is nonsingular. Apply again Proposition 3.20 (3) and take into account that $\operatorname{im} G_{r-1} = \operatorname{im} G_{r-2} = \cdots = \operatorname{im} G_0$ such that we can use $\mathcal{W}_{r-1} = \mathcal{W}_0$. This leads to $r_r = \operatorname{rank} G_r = \operatorname{rank} G_{r-1} + \operatorname{rank} \mathcal{W}_{r-1} B Q_{r-1}$, and with

$$\mathcal{W}_{r-1} B Q_{r-1} = \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & B_{r,r-1} \; 0 \end{bmatrix} \begin{bmatrix} 0 & & & * \\ & \ddots & & \vdots \\ & & 0 & * \\ & & & \Omega_{r-1} \\ & & & * \quad 0 \end{bmatrix} = \begin{bmatrix} 0 & & & \\ & \ddots & & \\ & & 0 & \\ & & B_{r,r-1}\Omega_{r-1} & 0 \end{bmatrix}$$

we find

$$r_r = m - m_r + \operatorname{rank} B_{r,r-1}\Omega_{r-1} = m - m_r + m_r = m.$$

This completes the proof. □

3.6 DAEs in circuit simulation

Circuit simulation was one of the driving motivations to study differential-algebraic equations. The behavior of circuits depends, on the one hand, on the kind of network elements involved and, on the other hand, on the connection of the network elements. Kirchhoff's laws describe algebraic relations between branch currents and branch voltages depending on the network structure. Additionally, the characteristics of dynamic elements like capacitances and inductances lead to differential equations. Hence, one is always confronted with a differential-algebraic equation system when modeling electrical circuits.

Due to their high complexity, integrated circuits need an automatic treatment for generating the model equations. One of the most commonly used techniques is Modified Nodal Analysis (MNA). Let us have a more detailed look into this analysis in order to get information on the structure of the resulting equation system.
In this section we use the notation common in circuit theory (cf. [51], [58]) to make things more transparent for readers from this area.
The transient behavior of the circuit is described by its branch voltages $u = u(t)$ and branch currents $j = j(t)$. Due to Kirchhoff's voltage law, all branch voltages u can be written as a linear combination of nodal potentials e,

$$u = A^{\mathrm{T}} e, \tag{3.53}$$

where $A_a \in \mathbb{R}^{n\times b}$ denotes the so-called incidence matrix with the entries

$$a_{ik} = \begin{cases} 1 & \text{if branch } k \text{ leaves node } i \\ -1 & \text{if branch } k \text{ enters node } i \\ 0 & \text{if branch } k \text{ is not incident with node } i. \end{cases}$$

Here, n and b denote the number of nodes and branches of the circuit. Since the number of nodes is usually much smaller than the number of branches, a network description using nodal potentials instead of branch voltages is advantageous. The modified nodal analysis (MNA) uses all node potentials and all branch currents of current controlled elements as the vector of unknowns and describes the electrical network as follows.

1. Fix one node as the datum node and set the potential of the datum node to be zero.
2. Express all branch voltages by nodal potentials using (3.53).

3. Write the node equations by applying Kirchhoff's current law (KCL) to each node except for the datum node:

$$A j = 0. \tag{3.54}$$

   The vector j represents the branch current vector. Here, A is the reduced incidence matrix that coincides with the incidence matrix A_a apart from the fact that the row corresponding to the datum node is neglected.
4. Replace the currents j_k of voltage controlled elements by the voltage–current relation of these elements in equation (3.54).
5. Add the current–voltage relations for all current controlled elements.

We want to demonstrate this with the following simple example circuit.

Example 3.43 (Small RCL circuit). We consider a circuit, consisting of a capacitance, an inductance, a resistor and a voltage source (see Figure 3.8). We denote the

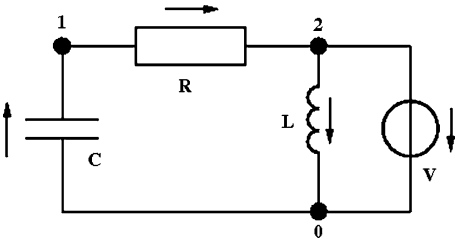


Fig. 3.8 Circuit with one capacitance, resistor, inductance and voltage source

branch currents and branch voltages of the voltage source, the resistance, the capacitance and the inductance by j_V, j_R, j_C, j_L and v_V, v_R, v_C, v_L. First we fix the node 0 as the datum node. Then, we find

$$v_V = e_2, \quad v_R = e_1 - e_2, \quad v_C = -e_1, \quad v_L = e_2.$$

In the third step, we write the KCL for the nodes 1 and 2, i.e.,

$$\begin{aligned} -j_C + j_R &= 0 \quad \text{(node 1)}, \\ -j_R + j_L + j_V &= 0 \quad \text{(node 2)}. \end{aligned}$$

The element characteristics are given by

$$j_R = G v_R = G(e_1 - e_2), \quad j_C = C v_C' = -C e_1', \tag{3.55}$$

and

$$e_2 = v_V = v_{\text{input}}(t), \quad e_2 = v_L = L j_L'. \tag{3.56}$$

Here, G denotes the conductance of the resistance, which means $G = R^{-1}$. The relations (3.55) are inserted into the KCL equations for node 1 and 2 which implies

$$\begin{aligned} Ce_1' + G(e_1 - e_2) &= 0, \\ G(e_2 - e_1) + j_L + j_V &= 0. \end{aligned}$$

It remains to add the characteristic equations (3.56) and we get the differential-algebraic equation system

$$\begin{aligned} Ce_1' + G(e_1 - e_2) &= 0, \\ G(e_2 - e_1) + j_L + j_V &= 0, \\ j_L' - e_2 &= 0, \\ e_2 &= v_{\text{input}}(t) \end{aligned}$$

in the variables e_1, e_2, j_L and j_V. □

In general, the MNA results in quasi-linear DAEs of the form

$$\hat{A}(d(x,t))' + b(x,t) = 0. \tag{3.57}$$

What do the matrix $\hat{A}$ and the functions d and b look like? In order to see this, we split the incidence matrix A into the element-related incidence matrices

$$A = \begin{bmatrix} A_C & A_L & A_R & A_V & A_I \end{bmatrix},$$

where A_C, A_L, A_R, A_V and A_I describe the branch–current relation for capacitive branches, inductive branches, resistive branches, branches of voltage sources and branches of current sources, respectively. Using the element characteristics

$$j_C = \frac{\mathrm{d}}{\mathrm{d}t} q(v_C,t) = \frac{\mathrm{d}}{\mathrm{d}t} q(A_C^{\mathrm{T}} e,t), \quad \frac{\mathrm{d}}{\mathrm{d}t} \phi(j_L,t) = v_L = A_L^{\mathrm{T}} e$$

for capacitances and inductances as well as

$$j_R = g(v_R,t) = g(A_R^{\mathrm{T}} e,t), \quad j_I = i_s(A^{\mathrm{T}} e, j_L, j_V, t), v_V = v_s(A^{\mathrm{T}} e, j_L, j_V, t)$$

for resistances, current and voltages sources, we obtain

$$\begin{aligned} A_C \frac{\mathrm{d}}{\mathrm{d}t} q(A_C^{\mathrm{T}} e,t) + A_R g(A_R^{\mathrm{T}} e,t) + A_L j_L + A_V j_V + A_I i_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} \phi(j_L,t) - A_L^{\mathrm{T}} e &= 0, \\ A_V^{\mathrm{T}} e - v_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0, \end{aligned} \tag{3.58}$$

where i_s and v_s are input functions. Next we aim to uncover beneficial DAE structures of the network system (3.58).

Supposing that the capacitance matrix $C(v,t)$ and the inductance matrix $L(j,t)$ are positive definite, which means passivity of capacitances and inductances, the DAE (3.58) has a properly stated derivative term. In order to see this we formulate a lemma which is also useful in the analysis later.

Lemma 3.44. *If M is a positive definite $m \times m$ matrix and A is a rectangular matrix of dimension $k \times m$, then we have*

$$\ker AMA^{\mathrm{T}} = \ker A^{\mathrm{T}} \quad \textit{and} \quad \operatorname{im} AMA^{\mathrm{T}} = \operatorname{im} A$$

and

$$\ker A \oplus \operatorname{im} MA^{\mathrm{T}} = \mathbb{R}^k.$$

Furthermore, the matrix $AMA^{\mathrm{T}} + Q_A^{\mathrm{T}} Q_A$ is nonsingular for any projector Q_A onto $\ker A^{\mathrm{T}}$.

Proof. The first two equations of Lemma 3.44 follow immediately from the definition of positive definite matrices. For the third equation we assume z to be an element of $\ker A$ and $\operatorname{im} MA^{\mathrm{T}}$. Then, we find an element y such that $z = MA^{\mathrm{T}} y$, thus $AMA^{\mathrm{T}} y = Az = 0$. Since M is positive definite, we get $A^{\mathrm{T}} y = 0$, i.e., $z = 0$. Consequently, the intersection of $\ker A$ and $\operatorname{im} MA^{\mathrm{T}}$ is trivial. Consider now z as an arbitrary element of $\mathbb{R}^k$ and choose a projector $\bar{Q}_A$ onto $\ker A$ with $\operatorname{im} MA^{\mathrm{T}} \subseteq \ker \bar{Q}_A$. Then we have, due to the non-singularity of M,

$$\operatorname{rank}(MA^{\mathrm{T}}) = \operatorname{rank}(A^{\mathrm{T}}) = \operatorname{rank}(A) = \dim \ker \bar{Q}_A,$$

in other words $\operatorname{im} MA^{\mathrm{T}} = \ker \bar{Q}_A$. This implies

$$z = \bar{Q}_A z + (I - \bar{Q}_A) z \in \operatorname{im} \bar{Q}_A \oplus \ker \bar{Q}_A = \ker A \oplus \operatorname{im} MA^{\mathrm{T}}$$

and the third equation of the theorem is satisfied. It remains to show the nonsingularity of the matrix $AMA^{\mathrm{T}} + Q_A^{\mathrm{T}} Q_A$. Assume z to belong to the kernel of this matrix. Then,

$$0 = Q_A^{\mathrm{T}} (AMA^{\mathrm{T}} + Q_A^{\mathrm{T}} Q_A) z = Q_A^{\mathrm{T}} Q_A z,$$

which means $Q_A z = 0$. This implies $AMA^{\mathrm{T}} z = 0$ and, finally, $z \in \ker A^{\mathrm{T}} = \operatorname{im} Q_A$. Since Q_A is a projector, we conclude that $z = Q_A z = 0$. □

We rewrite the system (3.58) as

$$\bar{A}(d(x,t))' + b(x,t) = 0, \tag{3.59}$$

with

$$\bar{A} = \begin{bmatrix} A_C & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix}, \quad d(x,t) = \begin{bmatrix} q(A_C^{\mathrm{T}} e, t) \\ \phi(j_L, t) \end{bmatrix}, \quad x = \begin{bmatrix} e \\ j_L \\ j_V \end{bmatrix},$$

and

$$b(x,t) = \begin{bmatrix} A_R g(A_R^{\mathrm{T}} e, t) + A_L j_L + A_V j_V + A_I i_s(A^{\mathrm{T}} e, j_L, j_V, t), \\ -A_L^{\mathrm{T}} e \\ A_V^{\mathrm{T}} e - v_s(A^{\mathrm{T}} e, j_L, j_V, t) \end{bmatrix}.$$

We use the bar notation $\bar{A}$ in order to distinguish it from the incidence matrix introduced before. By

$$\ker \bar{A} = \ker A_C \times \{0\}, \quad \operatorname{im} d_x(x,t) = \operatorname{im} C(A_C^{\mathrm{T}} e, t) A_C^{\mathrm{T}} \times \operatorname{im} L(j_L, t) \tag{3.60}$$

Lemma 3.44 implies that

$$\ker \bar{A} \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^{n_C} \times \mathbb{R}^{n_L},$$

with n_C, n_L being the numbers of capacitances and inductances, respectively. Remembering that $\bar{A}$ represents a constant matrix, we find our general Assumption 3.16 to be satisfied, if d is continuously differentiable and b_x is continuous.

The solutions of the network system (3.58) are expected to consist of continuous nodal potentials and branch currents such that the charges and fluxes are continuously differentiable. This makes sense from the physical point of view, and it is consistent with the solution notion (Definition 3.2) for the DAE (3.59).

For the study of the regularity and index of (3.58), we introduce $G(u,t) := \partial_u g(u,t)$. Further, we denote projectors onto

$$\ker A_C^{\mathrm{T}},\ \ker A_V^{\mathrm{T}} Q_C,\ \ker A_R^{\mathrm{T}} Q_C Q_{V-C},\ \ker A_V^{\mathrm{T}},\ \ker A_C,\ \ker Q_C^{\mathrm{T}} A_V,$$

by

$$Q_C, \qquad Q_{V-C}, \qquad Q_{R-CV}, \qquad Q_V, \qquad \bar{Q}_C, \qquad \bar{Q}_{V-C},$$

respectively[1]. The complementary projectors shall be denoted by $P := I - Q$, with the corresponding subindex. We observe that

$$\operatorname{im} P_C \subset \ker P_{V-C}, \quad \operatorname{im} P_{V-C} \subset \ker P_{R-CV} \quad \text{and} \quad \operatorname{im} P_C \subset \ker P_{R-CV},$$

and that thus $Q_C Q_{V-C}$ is a projector onto $\ker (A_C A_V)^{\mathrm{T}}$, and $Q_C Q_{V-C} Q_{R-VC}$ is a projector onto $\ker (A_C A_R A_V)^{\mathrm{T}}$. In order to shorten denotations, we use the abbreviation $Q_{CRV} := Q_C Q_{V-C} Q_{R-CV}$. Note that the projector P_{CRV}, in general, does not coincide with the projector P_{R-CV}.

We start our analysis with a lemma that describes certain network topological properties in terms of the introduced incidence matrices and projectors.

Lemma 3.45. *[206, 70] Given a lumped circuit with capacitances, inductances and resistances as well as independent voltage sources and current sources, then, the following relations are satisfied.*

(1) *The matrix* $\begin{bmatrix} A_C & A_L & A_R & A_V \end{bmatrix}$ *has full row rank, because cutsets of current sources are forbidden.*

(2) *The matrix* A_V *has full column rank, since loops of voltage sources are forbidden.*

(3) *The matrix* $\begin{bmatrix} A_C & A_R & A_V \end{bmatrix}$ *has full row rank if and only if the circuit does not contain cutsets consisting of inductances and current sources only.*

(4) *The matrix* $Q_C^{\mathrm{T}} A_V$ *has full column rank if and only if the circuit does not contain loops with at least one voltage source and consisting of capacitances and voltage sources only.*

[1] An explicit description of such projectors is given in [67].

For simplicity, we assume all current and voltage sources to be independent, which means $i_s(A^{\mathrm{T}}e, j_L, j_V, t) = i_s(t)$ and $v_s(A^{\mathrm{T}}e, j_L, j_V, t) = v_s(t)$. For the general case we refer to [70]. In order to describe all possible DAE index cases for electric networks we need the following lemma.

Lemma 3.46. *Consider lumped electric circuits containing resistances, capacitances, inductances, as well as independent voltage and current sources. Let the capacitance, inductance and conductance matrices of all capacitances, inductances, and resistances, respectively, be positive definite.*[2] *Furthermore, assume that the circuit neither contains a loop of voltage sources nor a cutset of current sources.*[3] *Then, the auxiliary matrix functions*

$$\begin{aligned} H_1(v,t) &:= A_C C(v,t) A_C^{\mathrm{T}} + Q_C^{\mathrm{T}} Q_C, \\ H_2(j,t) &:= Q_{CRV}^{\mathrm{T}} A_L L^{-1}(j,t) A_L^{\mathrm{T}} Q_{CRV} + P_{CRV}^{\mathrm{T}} P_{CRV}, \\ H_3(v,t) &:= \bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} H_1^{-1}(v,t) A_V \bar{Q}_{V-C} + \bar{P}_{V-C}^{\mathrm{T}} \bar{P}_{V-C} \end{aligned}$$

are nonsingular.

Proof. Regarding Lemma 3.44, it remains to show the matrices $C(v,t)$, $L^{-1}(j,t)$ and $H_1^{-1}(v,t)$ to be positive definite and the projectors Q_C, P_{CRV} and $\bar{P}_{V-C}$ to be projectors onto the nullspaces $\ker A_C^{\mathrm{T}}$, $\ker A_L^{\mathrm{T}} Q_{CRV}$ and $\ker A_V \bar{Q}_{V-C}$, respectively.
First, the capacitance matrix $C(v,t)$ is positive definite due to the assumption. The relation $\operatorname{im} Q_C = \ker A_C^{\mathrm{T}}$ follows directly from the definition of Q_C; see the page before. Consequently, $H_1(v,t)$ is nonsingular. Furthermore, it is positive definite since

$$x^{\mathrm{T}} H_1(v,t) x = (A_C^{\mathrm{T}} x)^{\mathrm{T}} C(v,t)(A_C^{\mathrm{T}} x) + (Q_c x)^{\mathrm{T}}(Q_c x) \geq 0.$$

Since the inverse of a positive definite matrix is always positive definite, we get $H_1^{-1}(v,t)$ to be positive definite. The assumption that the inductance $L(j,t)$ is positive definite implies $L^{-1}(j,t)$ is also positive definite.
Since cutsets of current sources are forbidden, the incidence matrix $[A_C A_R A_V A_L]$, containing all noncurrent source branches, has full row rank. This implies

$$\ker \begin{bmatrix} A_C^{\mathrm{T}} \\ A_R^{\mathrm{T}} \\ A_V^{\mathrm{T}} \\ A_L^{\mathrm{T}} \end{bmatrix} = \{0\}$$

and, further,

$$\ker A_L^{\mathrm{T}} Q_{CRV} = \ker Q_{CRV} = \operatorname{im} P_{CRV}.$$

The matrix A_V^{T} has full row rank since loops of voltage sources are forbidden. From that we may conclude that

[2] For capacitances and inductances with affine characteristics the positive definiteness implies that they are strictly locally passive (cf. [77]).

[3] Loops of voltage sources and cutsets of current sources would lead to a short-circuit.

$$\ker A_V \bar{Q}_{V-C} = \ker \bar{Q}_{V-C} = \operatorname{im} \bar{P}_{V-C}.$$

□

Now we can formulate the following theorem describing all possible DAE index cases for electric circuits.

Theorem 3.47. *Let the assumptions of Lemma 3.46 be satisfied. Furthermore, let all current and voltage sources be independent.*[4] *Then, the following statements are true.*

(1) *If the network contains neither L-I cutsets nor C-V loops then the network system (3.58) leads to a regular DAE system of index* ≤ 1. *The index is 0 if and only if there is a capacitive path from each node to the datum node and the network does not contain voltage sources.*

(2) *If the network contains L-I cutsets or C-V loops then the network system (3.58) leads to a regular index-2 DAE system.*

(3) *If the network system yields an index-1 or index-2 DAE system, then* $G_0 := \bar{A}d_x$ *has constant rank and*

$$Q_0 = \begin{bmatrix} Q_C & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix} \tag{3.61}$$

is a projector onto the nullspace of G_0. *Further, the matrix function* $G_1 := G_0 + b_x Q_0$ *also has constant rank, and*

$$Q_1 = \begin{bmatrix} H_1^{-1} A_V \bar{Q}_{V-C} H_3^{-1} \bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} P_C & Q_{CRV} H_2^{-1} Q_{CRV}^{\mathrm{T}} A_L & 0 \\ 0 & L^{-1} A_L^{\mathrm{T}} Q_{CRV} H_2^{-1} Q_{CRV}^{\mathrm{T}} A_L & 0 \\ -\bar{Q}_{V-C} H_3^{-1} \bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} P_C & 0 & 0 \end{bmatrix} \tag{3.62}$$

is a projector onto the nullspace of G_1. Q_0 *and* Q_1 *are continuous and satisfy the condition* $Q_1 Q_0 = 0$.

Proof. First, we form

$$G_0 = \begin{bmatrix} A_C & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} \begin{bmatrix} C(\cdot) A_C^{\mathrm{T}} & 0 & 0 \\ 0 & L(\cdot) & 0 \end{bmatrix} = \begin{bmatrix} A_C C(\cdot) A_C^{\mathrm{T}} & 0 \\ 0 & L(\cdot) & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Regarding (3.60), we see that G_0 has full row rank if and only if A_C has full row rank and the equations corresponding to the voltage sources disappear. Simple arguments from graph theory show that A_C has full row rank if and only if there is a capacitive path from each node of the network to the datum node. Consequently, we have shown the index-0 case.

In order to investigate the index-1 case, we use the projector Q_0 given by (3.61) and form

[4] For the general case we refer to [70].

$$G_1 = G_0 + B_0 Q_0 = \begin{bmatrix} A_C C(\cdot) A_C^{\mathrm{T}} & 0 \\ 0 & L(\cdot)\,0 \\ 0 & 0\;\;0 \end{bmatrix} + \begin{bmatrix} A_R G(\cdot) A_R^{\mathrm{T}} & A_L & A_V \\ -A_L^{\mathrm{T}} & 0 & 0 \\ A_V^{\mathrm{T}} & 0 & \end{bmatrix} \begin{bmatrix} Q_C & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}$$
$$= \begin{bmatrix} A_C C(\cdot) A_C^{\mathrm{T}} + A_R G(\cdot) A_R^{\mathrm{T}} Q_C & 0 & A_V \\ -A_L^{\mathrm{T}} Q_C & L(\cdot) & 0 \\ A_V^{\mathrm{T}} Q_C & 0 & 0 \end{bmatrix}.$$

It is not difficult to verify that Q_1 is a projector and $G_1 Q_1 = 0$ holds for Q_1 given by (3.62) if one regards the relations

$$\begin{aligned} P_C^{\mathrm{T}} &= A_C C(\cdot) A_C^{\mathrm{T}} H_1^{-1}(\cdot), \\ Q_{CRV}^{\mathrm{T}} &= Q_{CRV}^{\mathrm{T}} A_L L^{-1}(\cdot) A_L^{\mathrm{T}} Q_{CRV} H_2^{-1}(\cdot), \\ \bar{Q}_{V-C}^{\mathrm{T}} &= \bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} H_1^{-1}(\cdot) A_V \bar{Q}_{V-C} H_3^{-1}(\cdot) \end{aligned}$$

as well as

$$Q_C H_1^{-1}(\cdot) = H_1^{-1}(\cdot) Q_C^{\mathrm{T}}, \quad Q_C^{\mathrm{T}} A_V \bar{Q}_{V-C} = 0, \quad Q_C Q_{CRV} = Q_{CRV}$$

and

$$A_C^{\mathrm{T}} Q_{CRV} = 0, \quad A_R^{\mathrm{T}} Q_{CRV} = 0, \quad A_V^{\mathrm{T}} Q_{CRV} = 0,$$

that follow directly from the definitions of the projectors Q_* and the matrices H_i, $i = 1,2,3$. Consequently, $\operatorname{im} Q_1 \subseteq \ker G_1$. In order to show that $\ker G_1 \subseteq \operatorname{im} Q_1$, we assume $z \in \ker G_1(\cdot)$. Then,

$$A_C C(\cdot) A_C^{\mathrm{T}} z_e + A_R G(\cdot) A_R^{\mathrm{T}} Q_C z_e + A_V z_V = 0, \tag{3.63}$$
$$-A_L^{\mathrm{T}} Q_C z_e + L(\cdot) z_L = 0, \tag{3.64}$$
$$A_V^{\mathrm{T}} Q_C z_e = 0. \tag{3.65}$$

Considering (3.65) we see that

$$z_e = Q_{V-C} z_e. \tag{3.66}$$

Multiplying (3.63) by $z_e^{\mathrm{T}} Q_{V-C}^{\mathrm{T}} Q_C^{\mathrm{T}}$ yields

$$z_e^{\mathrm{T}} Q_{V-C}^{\mathrm{T}} Q_C^{\mathrm{T}} A_R G(\cdot) A_R^{\mathrm{T}} Q_C Q_{V-C} z_e = 0.$$

Since $G(\cdot)$ is positive definite, we find $Q_C Q_{V-C} z_e = 0$. Taking into account (3.66), we get

$$Q_C z_e = Q_{CRV} z_e. \tag{3.67}$$

Relation (3.64) leads to

$$z_L = L^{-1}(\cdot) A_L^{\mathrm{T}} Q_C z_e = L^{-1}(\cdot) A_L^{\mathrm{T}} Q_{CRV} z_e. \tag{3.68}$$

Multiplying (3.63) by Q_C^{T} now yields $Q_C^{\mathrm{T}} A_V z_V = 0$ and, hence,

$$z_V = \bar{Q}_{C-C} z_V. \tag{3.69}$$

Regarding (3.67)–(3.69), we obtain $Q_1 z = z$ which implies $z \in \operatorname{im} Q_1$. Consequently, $\operatorname{im} Q_1 = \ker G_1$.

Obviously, we have $Q_1 Q_0 = 0$ for the projectors Q_0 and Q_1 given by (3.61) and (3.62).

The matrix G_1 is nonsingular if and only if $Q_1 = 0$. The latter relation is satisfied if and only if

$$\bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} = 0 \quad \text{and} \quad Q_{CRV}^{\mathrm{T}} A_L = 0.$$

Since loops of voltage sources only are forbidden, the matrix A_V^{T} has full row rank. Furthermore, the matrix $[A_C A_R A_V A_L]$ has full row rank since cutsets of current sources only are forbidden. Both relations allow the conclusion that G_1 is nonsingular if and only if

$$\bar{Q}_{V-C}^{\mathrm{T}} = 0 \quad \text{and} \quad Q_{CRV}^{\mathrm{T}} = 0.$$

The first condition reflects the case that there is no C-V loop in the network. The second one corresponds to the condition that the network does not contain L-I cutsets. Consequently, the index-1 case has been completely proven.

Finally, applying the modified regularity condition given by Proposition 3.38, and taking into account that, owing to the relation

$$\tilde{G}_2 := G_1 + BP_0Q_1 = (G_1 + \mathcal{W}_0 BQ_1)(I + P_1 G_0^- BP_0 Q_1),$$

the matrix functions $\tilde{G}_2$ and $G_1 + \mathcal{W}_0 BQ_1$ must share their rank, it suffices to show that the matrix function

$$\begin{aligned}\tilde{G}_2 &= G_1 + BP_0Q_1 \\ &= \begin{bmatrix} A_C C(\cdot) A_C^{\mathrm{T}} + A_R G(\cdot) A_R^{\mathrm{T}} Q_C & 0 & A_V \\ -A_L^{\mathrm{T}} Q_C & L(\cdot) & 0 \\ A_V^{\mathrm{T}} Q_C & 0 & 0 \end{bmatrix} + \begin{bmatrix} A_R G(\cdot) A_R^{\mathrm{T}} P_C & A_L & 0 \\ -A_L^{\mathrm{T}} P_C & 0 & 0 \\ A_V^{\mathrm{T}} P_C & 0 & 0 \end{bmatrix} Q_1\end{aligned}$$

remains nonsingular. Let z be an element of $\ker \tilde{G}_2$. Then we have

$$0 = \begin{bmatrix} 0 & 0 & \bar{Q}_{V-C}^{\mathrm{T}} \end{bmatrix} G_2 z = \bar{Q}_{V-C}^{\mathrm{T}} A_V^{\mathrm{T}} P_C z_e$$

and

$$0 = \begin{bmatrix} Q_{CRV}^{\mathrm{T}} & 0 & 0 \end{bmatrix} G_2 z = Q_{CRV}^{\mathrm{T}} A_L z_L.$$

Both conclusions yield $Q_1 z = 0$, and hence $G_1 z = \tilde{G}_2 z = 0$. In other words, z belongs to $\ker Q_1$ and also to $\ker G_1 = \operatorname{im} Q_1$. In consequence, $z = 0$ holds true and $\tilde{G}_2$ is nonsingular. □

We want to finish this section with a summary of structural properties of circuit systems. We have seen by (3.59) that circuit systems are given in quasi-linear form

$$\bar{A}(d(x,t))' + b(x,t) = 0 \tag{3.70}$$

with a constant matrix $\bar{A}$. The subspace $\operatorname{im} Q_0Q_1 = N_0 \cap S_0$ is always independent of the choice of projectors. From the decoupled versions of linear index-2 DAEs we know that exactly the part of the solution belonging to $\operatorname{im} Q_0Q_1$ describes the index-2 components of the system. By an index-2 component we mean a component which involves first derivatives of algebraically given components. Interestingly, these components appear only linearly in circuit systems as the following proposition shows.

Theorem 3.48. *Let the assumptions of Theorem 3.47 be satisfied and let the index-1 or index-2 case be valid.*

(1) *Then, the circuit systems (3.70) have the special structure*

$$\bar{A}(d(P_0x,t))' + \bar{b}(Ux,t) + \bar{B}Tx = 0 \tag{3.71}$$

with constant coefficient matrix $\bar{B}$ and constant projectors

$$P_0 = \begin{bmatrix} P_C & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0. \end{bmatrix}, \quad T = \begin{bmatrix} Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}. \end{bmatrix}, \quad U := I - T.$$

(2) *The projectors Q_0 and Q_1 described in (3.61) and (3.62) satisfy the relations*

$$\operatorname{im} T = \operatorname{im} Q_0Q_1 = \operatorname{im} Q_{CRV} \times \{0\} \times \operatorname{im} \bar{Q}_{V-C}, \quad Q_0 = I - P_0.$$

Remark 3.49. Theorem 3.48 remains valid also for controlled current and voltage sources if they do not belong to C-V loops or L-I cutsets and their controlling voltages and currents do not belong to C-V loops or L-I cutsets. Using the results from [70], the proof given here can also be applied to systems with controlled sources.

Proof. We start by proving (2). Using (3.61) and (3.62) we find

$$Q_0Q_1 = \begin{bmatrix} 0 & Q_{CRV}H_2^{-1}(\cdot)Q_{CRV}^{\mathrm{T}}A_L & 0 \\ 0 & 0 & 0 \\ -\bar{Q}_{V-C}H_3^{-1}(\cdot)\bar{Q}_{V-C}^T A_V^T P_C & 0 & 0 \end{bmatrix}.$$

Obviously, $\operatorname{im} Q_0Q_1 \subseteq \operatorname{im} Q_{CRV} \times \{0\} \times \operatorname{im} \bar{Q}_{V-C}$. On the other hand, we have

$$Q_0Q_1 \begin{bmatrix} -H_1^{-1}(\cdot)A_V\bar{Q}_{V-C}z_V \\ L^{-1}(\cdot)A_L^{\mathrm{T}}Q_{CRV}z_e \\ 0 \end{bmatrix} = \begin{bmatrix} Q_{CRV}z_e \\ 0 \\ \bar{Q}_{V-C}z_V \end{bmatrix}$$

for any $z_e \in \mathbb{R}^{n_e}$ and $z_V \in \mathbb{R}^{n_V}$ which implies also $\operatorname{im} Q_0Q_1 \supseteq \operatorname{im} Q_{CRV} \times \{0\} \times \operatorname{im} \bar{Q}_{V-C}$.

(1) Since we have assumed all voltage and current sources to be independent, the function $b(x,t)$ in (3.71) has the form

$$b(x,t) = \begin{bmatrix} A_R g(A_R^{\mathrm{T}} e,t) + A_L j_L + A_V j_V + A_I i_s(t) \\ -A_L^{\mathrm{T}} e \\ A_V^{\mathrm{T}} e - v_s(t) \end{bmatrix}.$$

Defining

$$\bar{b}(x,t) := \begin{bmatrix} A_R g(A_R^{\mathrm{T}} e,t) + A_L j_L + A_V (I - \bar{Q}_{V-C}) j_V + A_I i_s(t) \\ -A_L^{\mathrm{T}} (I - Q_{CRV}) e \\ A_V^{\mathrm{T}} e - v_s(t) \end{bmatrix},$$

and

$$\bar{B} := \begin{bmatrix} 0 & 0 & A_V \bar{Q}_{V-C} \\ -A_L^{\mathrm{T}} Q_{CRV} e & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

we get $b(x,t) = \bar{b}(x,t) + BTx$ with the projector

$$T = \begin{bmatrix} Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}. \end{bmatrix}.$$

Notice that $\bar{b}(x,t) = \bar{b}(Ux,t)$ for $U = I - T$ since $A_R^{\mathrm{T}} Q_{CRV} e = 0$ and $A_V^{\mathrm{T}} Q_{CRV} e = 0$. Owing to the properly stated leading term, the projector $Q_0 = I - P_0$ is at the same time a projector onto $\ker d_x$. This implies

$$d(x,t) - d(P_0 x,t) = \int_0^1 d_x(sx + (1-s)P_0 x,t) Q_0 x \mathrm{d}s = 0, \quad \text{for all arguments } x \text{ and } t.$$

□

Observe that here the intersection subspace $N_0 \cap S_0 = \operatorname{im} T$ is even a constant one. It is trivial that $\operatorname{im} T = \{0\}$ in the index-1 case, but it has dimension ≥ 1 for index-2 DAEs. We close this section by discussing solvability.

Theorem 3.50. *Let the function d in (3.71) be continuously differentiable, let $\bar{b}$ be continuous together with $\bar{b}_x$ and let the underlying network system (3.58) satisfy the assumptions of Theorem 3.47 which yield a regular index-1 DAE.*
Then, to each point (x_0,t_0) such that $\bar{b}(x_0,t_0) \in \operatorname{im} \bar{A}$, there exists at least one solution x_* of the DAE passing through $x_*(t_0) = x_0$.
If d also has continuous second partial derivatives d_{xx}, d_{tx}, then this solution is unique.

Proof. The DAE satisfies Assumption 3.16 and it is regular with index 1. Additionally, $\ker d_x = \ker P_0$ is constant. The existence of the solution is now an immediate consequence of Theorem 3.55.
The uniqueness follows from Theorem 3.53. □

The index-2 case is less transparent. The local solvability Theorem 3.56 applies only to DAEs having a linear derivative part. For this reason we rewrite the DAE as

$$\bar{A}d_x(P_0x,t)(P_0x)' + \bar{A}d_t(P_0x,t) + \bar{b}(Ux,t) + \bar{B}Tx = 0. \tag{3.72}$$

If d is continuously differentiable and also has continuous second partial derivatives d_{xx}, d_{tx}, and $\bar{b}$ is continuous together with $\bar{b}_x$, then the DAE (3.72) meets Assumption 3.16. Also the structural conditions demanded by Theorem 3.56 are fulfilled by equation (3.72). The consistency condition for the initial point (x_0,t_0) resulting from the obvious constraint reads $\bar{b}(Ux_0,t_0) \in \operatorname{im}\bar{A}$. A second much more subtle consistency condition for (x_0,t_0) results from formula (3.99) in Theorem 3.56. Then, supposing slight additional smoothness, Theorem 3.56 guarantees the existence and uniqueness of a solution x_* with $x_*(t_0) = x_0$.

3.7 Local solvability

Each regular linear DAE with sufficiently smooth coefficients and excitations is solvable. Nonlinear DAEs are much more complex. Regularity does not necessarily imply solvability. For instance, if a nonlinear Hessenberg form DAE has size r on its definition domain $\mathcal{D}_b \times \mathcal{I}_b$, i.e., this DAE is regular with tractability index r (cf. Theorem 3.42), then this does not at all mean that there is a solution passing through a given point $(x_0,t_0) \in \mathcal{D}_b \times \mathcal{I}_b$. For the existence of such a solution, x_0 must be a *consistent* value. The following two examples illustrate the situation.

Example 3.51 (Semi-explicit index-1 DAE). We consider once again the DAE

$$\begin{aligned} x_1'(t) + x_1(t) &= 0, \\ x_1(t)^2 + x_2(t)^2 - 1 &= \gamma(t), \end{aligned} \tag{3.73}$$

given on $\mathcal{D}_f = \{x \in \mathbb{R}^2 : x_2 > 0\}$, $\mathcal{I}_f = \mathbb{R}$ from Example 3.7. As shown in Example 3.18, it is a semi-explicit DAE being regular with index 1 on $\mathcal{D}_f \times \mathcal{I}_f$. Every solution value at time t must lie in the set

$$\mathcal{M}_0(t) := \{x \in \mathcal{D}_f : (x_1)^2 + (x_2)^2 - 1 - \gamma(t) = 0\},$$

and, obviously, through points outside there are no solutions. Through each point $t_0 \in \mathcal{I}_f$, $x_0 \in \mathcal{M}_0(t_0)$ passes through exactly one solution. □

Example 3.52 (Hessenberg size-2 DAE). Reconsider the DAE

$$\begin{aligned} x_1'(t) + x_1(t) &= 0, \\ x_2(t)x_2'(t) - x_3(t) &= 0, \\ x_1(t)^2 + x_2(t)^2 - 1 &= \gamma(t), \end{aligned} \tag{3.74}$$

given on $\mathcal{D}_f = \{x \in \mathbb{R}^3 : x_2 > 0\}$, $\mathcal{I}_f = \mathbb{R}$, which is investigated in Examples 3.8 and 3.19. The DAE is regular with tractability index 2. The solution values must belong to the obvious constraint set

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

and also to the hidden constraint set

$$\mathcal{H}(t) := \{x \in \mathcal{D}_f : -2(x_1)^2 + 2x_3 - \gamma'(t) = 0\}.$$

The obvious constraint set $\mathcal{M}_0(t)$ contains points which are no longer consistent, but the proper subset $\mathcal{M}_1(t) := \mathcal{M}_0(t) \cap \mathcal{H}(t) \subset \mathcal{M}_0(t)$ consists of consistent points, that is, through each point $t_0 \in \mathbb{R}$, $x_0 \in \mathcal{M}_1(t_0)$ passes through a solution. □

In the present section we prove that the obvious constraint set of a general regular DAE (3.1) with tractability index 1 is filled by solutions as it is the case in Example 3.51.
Furthermore, we also prove the local solvability of a class of regular DAEs with tractability index 2, which meets the structure of MNA DAEs and applies to Example 3.52.

By definition, a function $x_* \in \mathcal{C}(\mathcal{I}_*, \mathbb{R}^m)$ is a solution of the DAE (3.1), if $x_*(t) \in \mathcal{D}_f$, $t \in \mathcal{I}_*$, $d(x_*(.),.) \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^n)$, and the DAE (3.1) is satisfied pointwise on $\mathcal{I}_*$. In our basic setting, d is always a $\mathcal{C}^1$ function, and $D(x,t) = d_x(x,t)$ has constant rank. The inclusion

$$u_*'(t) - d_t(x_*(t),t) \in D(x_*(t),t), t \in \mathcal{I}_* \tag{3.75}$$

is valid for all solutions (cf. Proposition C.1). Due to the constant rank of $D(x,t)$ there is a continuous functions w_* such that

$$u_*'(t) - d_t(x_*(t),t) = D(x_*(t),t)w_*(t), t \in \mathcal{I}_*. \tag{3.76}$$

In particular, for $d(x,t) = D(t)x$ it holds that

$$\begin{aligned} u_*'(t) - D'(t)x_*(t) &= (D(t)P_0(t)x_*(t))' - D'(t)x_*(t) \\ &= D'(t)P_0(t)x_*(t) + D(t)(P_0(t)x_*(t))' - D'(t)x_*(t) \\ &= -D'(t)Q_0(t)x_*(t) + D(t)(P_0(t)x_*(t))' \\ &= D(t)Q_0'(t)x_*(t) + D(t)(P_0(t)x_*(t))' \\ &= D(t)\{(P_0(t)x_*(t))' - P_0'(t)x_*(t)\} \end{aligned}$$

with any $\mathcal{C}^1$-projector Q_0 onto $\ker D$, $P_0 = I - Q_0$.

3.7.1 Index-1 DAEs

Let the DAE (3.1) be regular with tractability index 1 on the open set $\mathcal{G} \subset \mathcal{D}_f \times \mathcal{I}_f$. All solution values have to remain within the obvious constraint set

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

We prove that through each $(\bar{x},\bar{t}) \in \mathcal{G}$, $\bar{x} \in \mathcal{M}_0(\bar{t})$ passes through exactly one solution. This means that the obvious constraint is at the same time the set of consistent values.

Theorem 3.53. *Let the DAE (3.1) satisfy Assumption 3.16 and be regular with tractability index 1 on the open set $\mathcal{G} \subset \mathcal{D}_f \times \mathcal{I}_f$. Let d have the additional continuous partial derivatives d_{xx}, d_{xt}.*
Then, for each $(\bar{x},\bar{t}) \in \mathcal{G}$, $\bar{x} \in \mathcal{M}_0(\bar{t})$, there is exactly one solution $x_* \in \mathcal{C}(\mathcal{I}_*,\mathbb{R}^m)$ of the DAE which satisfies $x_*(\bar{t}) = \bar{x}$.

Proof. Suppose all assumptions to be given. Owing to the index-1 property, the matrix function $G_1 = G_0 + B_0Q_0$ remains nonsingular on $\mathbb{R}^m \times \mathcal{G}$ independently of the special choice of the continuous projector function Q_0 onto $N_0 = \ker G_0 = \ker D$. Owing to the properly stated leading term the obvious constraint set is (cf. Proposition 3.10)

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

We take use of the subspaces

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

and

$$S_0(x^1,x,t) := \{z \in \mathbb{R}^m : B_0(x^1,x,t)z \in \operatorname{im} G_0(x^1,x,t)\},$$

defined (cf. Section 3.1) for $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$, $y \in \mathbb{R}^n$, and $x^1 \in \mathbb{R}^m$.

For each $(x,t) \in \mathcal{G}$, the nonsingularity of $G_1(x^1,x,t)$, for all $x^1 \in \mathbb{R}^m$, is equivalent to the transversality condition (cf. Lemma A.9)

$$N_0(x,t) \oplus S_0(x^1,x,t) = \mathbb{R}^m, \quad x^1 \in \mathbb{R}^m. \tag{3.77}$$

We fix an arbitrary $\bar{x} \in \mathcal{M}_0(\bar{t})$, $(\bar{x},\bar{t}) \in \mathcal{G}$. There is exactly one $\bar{y} \in \mathbb{R}^n$ with $\bar{y} - d_t(\bar{x},\bar{t}) \in \operatorname{im} D(\bar{x},\bar{t})$, $f(\bar{y},\bar{x},\bar{t}) = 0$.
Denote by $\bar{x}^1 \in \mathbb{R}^m$ the value that is uniquely determined by the conditions

$$Q_0(\bar{x},\bar{t})\bar{x}^1 = 0, \quad D(\bar{x},\bar{t})\bar{x}^1 = \bar{y} - d_t(\bar{x},\bar{t}).$$

Then it holds that $S(\bar{y},\bar{x},\bar{t}) = S_0(\bar{x}^1,\bar{x},\bar{t})$. The index-1 property (3.77) yields

$$N_0(\bar{x},\bar{t}) \oplus S(\bar{y},\bar{x},\bar{t}) = \mathbb{R}^m,$$

and further, if $\mathcal{N}_{(\bar{x},\bar{t})} \subset \mathcal{G}$ denotes a suitable neighborhood of $(\bar{x},\bar{t})$,

$$N_0(x,t) \oplus S(\bar{y},\bar{x},\bar{t}) = \mathbb{R}^m, \quad (x,t) \in \mathcal{N}_{(\bar{x},\bar{t})}. \tag{3.78}$$

This allows us to choose $Q_0(x,t)$ in accordance with the decomposition (3.78), such that

$$\operatorname{im} Q_0(x,t) = N_0(x,t), \quad \ker Q_0(x,t) = S(\bar{y},\bar{x},\bar{t}), \quad \operatorname{im} P_0(x,t) = S(\bar{y},\bar{x},\bar{t}).$$

Denote

$$\bar{Q}_0(t) := Q_0(\bar{x},t), \quad \bar{P}_0(t) := P_0(\bar{x},t), \quad \bar{D}(t) := D(\bar{x},t),$$

and further,

$$\bar{R}(t) := R(\bar{x}^1,\bar{x},t), \quad \bar{D}(t)^- := D(\bar{x}^1,\bar{x},t)^-,$$

and emphasize that this construction leads to the property

$$\ker \bar{Q}_0(t) = \ker Q_0(x,t) = S(\bar{y},\bar{x},\bar{t}), \quad (x,t) \in \mathcal{N}_{(\bar{x},\bar{t})}.$$

Since the projectors $Q_0(x,t)$ and $\bar{Q}_0(t)$ have the common nullspace $S(\bar{y},\bar{x},\bar{t})$, it follows that

$$Q_0(x,t) = Q_0(x,t)\bar{Q}_0(t), \quad \bar{Q}_0(t) = \bar{Q}_0(t)Q_0(x,t), \quad P_0(x,t) = \bar{P}_0(t)P_0(x,t).$$

Because of

$$D(x^1,x,t)^- = P_0(x,t)D(x^1,x,t)^- = \bar{P}_0(t)P_0(x,t)D(x^1,x,t)^-,$$

we obtain the useful property

$$\bar{Q}_0(t)D(x^1,x,t)^- = 0, \quad (x,t) \in \mathcal{N}_{(\bar{x},\bar{t})},\ x^1 \in \mathbb{R}^m.$$

Introduce the additional values

$$\bar{u} := d(\bar{x},\bar{t}), \quad \bar{\mu} := D(\bar{x},\bar{t})\bar{x} = \bar{D}(\bar{t})\bar{x}, \quad \bar{w} := Q_0(\bar{x},\bar{t})\bar{x} + D(\bar{x}^1,\bar{x},\bar{t})^-(\bar{y} - d_t(\bar{x},\bar{t})).$$

By construction, it holds that $\bar{Q}_0(\bar{t})\bar{w} = \bar{Q}_0(\bar{t})\bar{x}$, and

$$\bar{x} = \bar{P}_0(\bar{t})\bar{x} + \bar{Q}_0(\bar{t})\bar{x} = \bar{D}(\bar{t})^-\bar{D}(\bar{t})\bar{x} + \bar{Q}_0(\bar{t})\bar{x} = \bar{D}(\bar{t})^-\bar{\mu} + \bar{Q}_0(\bar{t})\bar{w}.$$

After these preparations we apply the standard implicit function theorem twice. In the first step we define the function

$$\mathcal{D}(\mu,u,w,t) := \bar{R}(t)\big\{u - d\big(\bar{D}(t)^-\mu + \bar{Q}_0(t)w\ ,\ t\big)\big\} - (I - \bar{R}(t))\mu \tag{3.79}$$

for $(\mu,u,w,t) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}$ from a neighborhood of $(\bar{\mu},\bar{u},\bar{w},\bar{t})$. The function $\mathcal{D}$ is continuous, and has the continuous partial derivatives $\mathcal{D}_\mu$, $\mathcal{D}_u$, $\mathcal{D}_w$.
Due to

$$\mathcal{D}(\bar{\mu},\bar{u},\bar{w},\bar{t}) = \bar{R}(\bar{t})\{\bar{u} - d(\bar{x},\bar{t})\} - (I - \bar{R}(\bar{t}))\bar{D}(\bar{t})\bar{x} = 0,$$

and

$$\mathcal{D}_\mu(\bar{\mu},\bar{u},\bar{w},\bar{t}) = \bar{R}(\bar{t})\{-\bar{R}(\bar{t})\} - (I - \bar{R}(\bar{t})) = -I,$$

the equation $\mathcal{D}(\mu,u,w,t) = 0$ implicitly defines a unique function $\mu = h(u,w,t)$ on a neighborhood $\mathcal{N}_{(\bar{u},\bar{w},\bar{t})}$ of $(\bar{u},\bar{w},\bar{t})$. This function h is continuous and has continuous partial derivatives h_u, h_w. Its values belong to a neighborhood of $\bar{\mu}$. It holds that

$h(\bar{u},\bar{w},\bar{t}\,)=\bar{\mu}$ and, for $(u,w,t)\in\mathcal{N}_{(\bar{u},\bar{w},\bar{t})}$,

$$\mathcal{D}(h(u,w,t),u,w,t)=0,\quad h(u,w,t)=\bar{R}(t)h(u,w,t),\quad h(u,w,t)=h(u,\bar{Q}_0(t)w,t).$$

As the second step we denote

$$\xi(u,w,t):=\bar{D}(t)^{-}h(u,w,t)+\bar{Q}_0(t)w,\quad (u,w,t)\in\mathcal{N}_{(\bar{u},\bar{w},\bar{t})},$$

and define the function $\mathcal{F}$ on $\mathcal{N}_{(\bar{u},\bar{w},\bar{t})}$ by

$$\mathcal{F}(u,w,t):=f(D(\xi(u,w,t),t)w+d_t(\xi(u,w,t),t),\xi(u,w,t),t). \tag{3.80}$$

Observe that $\xi(\bar{u},\bar{w},\bar{t}\,)=\bar{x}$, and

$$\mathcal{F}(\bar{u},\bar{w},\bar{t}\,)=f(D(\bar{x},\bar{t}\,)\bar{w}+d_t(\bar{x},\bar{t}\,),\bar{x},\bar{t}\,)=f(\bar{y},\bar{x},\bar{t}\,)=0.$$

The functions ξ and $\mathcal{F}$ are continuous. ξ, respectively $\mathcal{F}$, have continuous partial derivatives ξ_w, ξ_u, respectively $\mathcal{F}_w$, $\mathcal{F}_u$. We show the nonsingularity of the partial Jacobian $\mathcal{F}_w(\bar{u},\bar{w},\bar{t}\,)$. Consider the equation $\mathcal{F}_w(\bar{u},\bar{w},\bar{t}\,)z=0$, i.e.,

$$\bar{f}_y\{\bar{D}z+\bar{D}_x\bar{\xi}_w z\bar{w}+\bar{d}_{tx}\bar{\xi}_w z\}+\bar{f}_x\bar{\xi}_w z=0 \tag{3.81}$$

(the accent bar indicates that the functions are to be evaluated at the arguments $\bar{u}$, $\bar{w}$, $\bar{t}$, $\bar{x}$, $\bar{y}$). Recall that $h(u,w,t)\equiv h(u,\bar{Q}_0(t)w,t)$, which leads to $\xi(u,w,t)=\xi(u,\bar{Q}_0(t)w,t)$, and hence

$$\xi_w(u,w,t)=\xi_w(u,w,t)\bar{Q}_0(t).$$

That means that we have $\bar{\xi}_w z=\bar{\xi}_w\bar{Q}_0(t)z$ in (3.81). Rearrange (3.81) to

$$(\bar{f}_y\bar{D}+\bar{f}_x\bar{Q}_0)z+f_y\{\bar{D}_x\bar{\xi}_w z\bar{w}+\bar{d}_{tx}\bar{\xi}_w z\}+\bar{f}_x(\bar{\xi}_w-I)\bar{Q}_0 z=0.$$

A closer look at the first matrix in front of z shows that

$$\begin{aligned}\bar{f}_y\bar{D}+\bar{f}_x\bar{Q}_0 &= f_y(\bar{y},\bar{x},\bar{t}\,)D(\bar{x},\bar{t}\,)+f_x(\bar{y},\bar{x},\bar{t}\,)Q_0(\bar{x},\bar{t}\,)\\ &= A(\bar{x}^1,\bar{x},\bar{t}\,)D(\bar{x},\bar{t}\,)+B(\bar{x}^1,\bar{x},\bar{t}\,)Q_0(\bar{x},\bar{t}\,)=G_1(\bar{x}^1,\bar{x},\bar{t}\,)\end{aligned}$$

is nonsingular. It follows that

$$z+\bar{G}_1^{-1}\bar{A}\{\bar{D}_x\bar{\xi}_w z w+\bar{d}_{tx}\bar{\xi}_w z\}+\bar{G}_1^{-1}\bar{B}_0(\bar{\xi}_w-I)\bar{Q}_0 z=0. \tag{3.82}$$

Since $\bar{G}_1^{-1}\bar{A}=\bar{G}_1^{-1}\bar{G}_0\bar{D}^{-}=\bar{P}_0\bar{D}^{-}$, multiplication of (3.82) by $\bar{Q}_0$ cancels out the second term such that

$$\bar{Q}_0 z+\bar{Q}_0\bar{G}_1^{-1}\bar{B}_0(\bar{\xi}_w-I)\bar{Q}_0 z=0. \tag{3.83}$$

Because of $\operatorname{im}\bar{Q}_0=N_0(\bar{x},\bar{t}\,)$, $\ker\bar{Q}_0=S(\bar{y},\bar{x},\bar{t}\,)$, it holds that $\bar{Q}_0=\bar{Q}_0\bar{G}_1^{-1}\bar{B}_0$ (cf. Lemma A.10). Compute

$$\bar{Q}_0(\bar{\xi}_w - I)\bar{Q}_0 = \bar{Q}_0\bar{\xi}_w - \bar{Q}_0 = \bar{Q}_0\{\bar{D}^- \bar{h}_w + \bar{Q}_0\} - \bar{Q}_0 = \bar{Q}_0 - \bar{Q}_0 = 0.$$

We then obtain from (3.83) that $\bar{Q}_0 z = 0$, thus $\bar{\xi}_w z = \bar{\xi}_w \bar{Q}_0 z = 0$. Finally, (3.82) yields $z = 0$. In consequence, the partial Jacobian $\mathcal{F}_w(\bar{u}, \bar{w}, \bar{t})$ is nonsingular. Again, by the implicit function theorem, the equation $\mathcal{F}(u, w, t) = 0$ defines a solution function $w = \omega(u,t)$ on a neighborhood $\mathcal{N}_{(\bar{u},\bar{t})}$ of $(\bar{u},\bar{t})$. The function ω is continuous with a continuous partial derivative ω_u.
Use the shorter expressions

$$\begin{aligned}
\kappa(u,t) &:= \bar{D}(t)^- h(u, \bar{Q}_0(t)\omega(u,t), t) + \bar{Q}_0(t)\omega(u,t) = \xi(u, \omega(u,t), t),\\
\phi(u,t) &:= D(\kappa(u,t),t)\omega(u,t) + d_t(\kappa(u,t),t), \quad (u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}.
\end{aligned}$$

These two functions are continuous with continuous partial derivatives κ_u, ϕ_u.

Now we are ready to construct a solution of the DAE (3.1) which satisfies the condition $x(\bar{t}) = \bar{x}$. First we solve the IVP

$$u'(t) = \phi(u(t),t), \quad u(\bar{t}) = \bar{u}, \tag{3.84}$$

and denote by $u_* \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^n)$ its solution, $\bar{t} \in \mathcal{I}_*$. The interval $\mathcal{I}_*$ is sufficiently small so that all values $(u_*(t),t)$ remain in the definition domain $\mathcal{N}_{(\bar{u},\bar{t})}$ of ϕ. The solution u_* exists and is unique owing to the continuity of ϕ and ϕ_u.
In the final step we set

$$\begin{aligned}
w_*(t) &:= \omega(u_*(t),t),\\
\mu_*(t) &:= h(u_*(t), w_*(t), t),\\
x_*(t) &:= \kappa(u_*(t),t) = \xi(u_*(t), w_*(t), t), \quad t \in \mathcal{I}_*,
\end{aligned}$$

and prove that x_* solves the DAE, possibly on an interval $\mathcal{I}_{**} \subseteq \mathcal{I}_*$ containing $\bar{t}$. By construction, it hold that $x_*(\bar{t}) = \kappa(\bar{u},\bar{t}) = \bar{x}$, i.e., the function x_* passes through the given point.
For $t \in \mathcal{I}_*$ compute

$$\begin{aligned}
f(u_*'(t), x_*(t), t) &= f\Big(\phi(u_*(t),t), \kappa(u_*(t),t), t\Big)\\
&= f\Big(D\big(\xi(u_*(t), w_*(t), t), t\big) w_*(t) + d_t\big(\xi(u_*(t), w_*(t), t), t\big),\\
&\qquad\qquad \xi(u_*(t), w_*(t), t), t\Big) = 0.
\end{aligned}$$

If we succeed in proving the relation $u_*(t) = d(x_*(t),t)$ we obtain, with x_*, the required DAE solution. Remember that x_* and w_* are just continuous functions, and it holds that

$$\begin{aligned}
u_*'(t) &= D(x_*(t),t) w_*(t) + d_t(x_*(t),t), \quad t \in \mathcal{I}_*,\\
u_*(\bar{t}) &= \bar{u} = d(\bar{x},\bar{t}).
\end{aligned} \tag{3.85}$$

The construction further yields $\mathcal{D}(\mu_*(t), u_*(t), w_*(t), t) = 0$, $t \in \mathcal{I}_*$, and hence

$$\bar{R}(t)u_*(t) = \bar{R}(t)d(x_*(t), t), \quad t \in \mathcal{I}_*. \tag{3.86}$$

This implies, in particular, that the function $\bar{R}(.)d(x_*(.),.)$ is continuously differentiable. We derive

$$(\bar{R}(t)d(x_*(t),t))' = \bar{R}'(t)d(x_*(t),t) + \bar{R}(t)d_t(x_*(t),t) + \mathfrak{L}(t), \tag{3.87}$$

with

$$\mathfrak{L}(t) := \lim_{\tau \to 0} \int_0^1 \bar{R}(t) \frac{1}{\tau} D\big(x_*(t) + s(x_*(t+\tau) - x_*(t)), t + s\tau\big)(x_*(t+\tau) - x_*(t))\mathrm{d}s.$$

The limit $\mathfrak{L}(t)$ is well defined and continuous with respect to t since the derivative on the left side of (3.87) exists and is continuous. By means of (3.85) and (3.86) we find the expression

$$\begin{aligned}(\bar{R}(t)d(x_*(t),t))' &= (\bar{R}(t)u_*(t))' = \bar{R}'(t)u_*(t) + \bar{R}(t)u_*'(t) \\ &= \bar{R}'(t)u_*(t) + \bar{R}(t)(D(x_*(t),t)w_*(t) + d_t(x_*(t),t)),\end{aligned}$$

and therefore

$$\mathfrak{L}(t) = \bar{R}(t)D(x_*(t),t)w_*(t) + \bar{R}'(t)(u_*(t) - d(x_*(t),t)).$$

The difference quotient

$$\frac{1}{\tau}(d(x_*(t+\tau),t+\tau) - d(x_*(t),t)) = \mathfrak{K}(t,\tau) + d_t(x_*(t),t),$$

with

$$\mathfrak{K}(t,\tau) := \int_0^1 \frac{1}{\tau} D\big(x_*(t) + s(x_*(t+\tau) - x_*(t)), t + s\tau\big)(x_*(t+\tau) - x_*(t))\mathrm{d}s,$$

possesses a limit for $\tau \to 0$, if $\mathfrak{K}(t,\tau)$ does. To prove the latter, we recall that $\bar{R}(t)$ projects onto $\operatorname{im} D(\bar{x},t)$. Denote $\bar{N}_A(t) := \ker \bar{R}(t)$ such that

$$\operatorname{im} D(\bar{x},t) \oplus \bar{N}_A(t) = \mathbb{R}^n, \quad t \in \mathcal{I}_*.$$

In a sufficiently small neighborhood of $(\bar{x},\bar{t})$, say for $x \in \mathcal{N}_{\bar{x}}$, $t \in \mathcal{I}_{**}$, the decomposition

$$\operatorname{im} D(x,t) \oplus \bar{N}_A(t) = \mathbb{R}^n$$

is valid, too. Denote by $\tilde{R}(x,t)$ the projector onto $\operatorname{im} D(x,t)$ along $\bar{N}_A(t)$. Since $\bar{R}(t)$ and $\tilde{R}(x,t)$ share their nullspace $\bar{N}_A(t)$, it holds that $\tilde{R}(x,t) = \tilde{R}(x,t)\bar{R}(t)$, and hence

$$D(x,t) = \tilde{R}(x,t)D(x,t) = \tilde{R}(x,t)\bar{R}(t)D(x,t), \quad x \in \mathcal{N}_{\bar{x}}, t \in \mathcal{I}_{**}.$$

By this means we obtain the limit $\mathfrak{K}(t) := \lim_{\tau\to 0} \mathfrak{K}(t,\tau) = \tilde{R}(x_*(t),t)\mathfrak{L}(t)$, and consequently, the derivative $d(x_*(t),t)'$ exists. Compute further

$$\begin{aligned} d(x_*(t),t)' =& d_t(x_*(t),t) + \mathfrak{K}(t) \\ =& d_t(x_*(t),t) + \tilde{R}(x_*(t),t)\bar{R}(t)D(x_*(t),t)w_*(t) \\ & + \tilde{R}(x_*(t),t)\bar{R}'(t)(u_*(t) - d(x_*(t),t)) \\ =& d_t(x_*(t),t) + D(x_*(t),t)w_*(t) + R(x_*(t),t)\bar{R}'(t)(u_*(t) - d(x_*(t),t)) \\ =& u_*'(t) + R(x_*(t),t)\bar{R}'(t)(u_*(t) - d(x_*(t),t)). \end{aligned}$$

Now it is evident that the function $\delta_* := u_* - d(x_*(.),.)$ is the solution of the standard homogeneous IVP $\delta' + \tilde{R}\bar{R}\delta = 0$, $\delta(\bar{t}) = 0$, and therefore δ_* vanishes identically. This completes the proof of the relation $u_*(t) = d(x_*(t),t)$, $t \in \mathcal{I}_{**}$, thus, x_* is in fact a solution of the DAE. The uniqueness of the solution x_* is a consequence of the related properties of the implicitly given functions, and of the solution of the IVP (3.84). □

Considering the assumptions in detail we see that they are in fact appropriate, and in general problems they cannot be weakened without losing, e.g., uniqueness.

Having local solutions, one can extend these solutions as long as the solution does not leave the domain of regularity with index 1. Till now, such characterizations of the maximal existence intervals as they are known for explicit regular ODEs are not available. And there is no general answer to the question of whether there are extensions through critical points and what they look like. This highly interesting topic needs future research.

In particular, the general Theorem 3.53 applies to the DAE (3.88) in Example 3.54 which is not covered by the theory in Part II.

Example 3.54 ($\ker f_y$ *varies with* y*).* Put $n = m = m_1 + m_2$, $k = m$, and

$$f(y,x,t) := \begin{bmatrix} y_1 + \varphi(y_2,x,t) \\ x_2 - \psi(x_1,t) \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \in \mathbb{R}^m, t \in \mathbb{R},$$
$$d(x,t) := \begin{bmatrix} x_1 \\ \psi(x_1,t) \end{bmatrix}, \quad f_y = \begin{bmatrix} I & \varphi_{y_2} \\ 0 & 0 \end{bmatrix}, \quad d_x = \begin{bmatrix} I & 0 \\ \psi_{x_1} & 0 \end{bmatrix}, \quad d_t = \begin{bmatrix} 0 \\ \psi_t \end{bmatrix},$$

where φ and ψ are smooth functions such that the matrix $I + \varphi_{y_2}\psi_{x_1}$ remains nonsingular. The leading term of the resulting DAE (3.1) is properly stated: f_y and d_x have constant rank m_1, and $\ker f_y \oplus \operatorname{im} d_x = \mathbb{R}^m$ is valid. This special DAE (3.1) can be written in detail

$$\begin{aligned} x_1'(t) + \varphi((\psi(x_1(t),t))', x_1(t), x_2(t), t) &= 0, \\ x_2(t) - \psi(x_1(t),t) &= 0. \end{aligned} \tag{3.88}$$

The nonsingularity of $I+\varphi_{y_2}\psi_{x_1}$ is responsible for the fact that the implicit equation for $x_1'(t)$, namely,

$$x_1'(t)+\varphi(\psi_{x_1}(x_1(t),t)x_1'(t)+\psi_t(x_1(t),t),x_1(t),\psi(x_1(t),t),t)=0, \tag{3.89}$$

is regular, that is, it contains a uniquely defined expression for $x_1'(t)$, if any.
The matrix function G_0, as well as an admissible Q_0 and the resulting $G_1=G_0+f_xQ_0$, are

$$G_0=Ad_x=\begin{bmatrix} I+\varphi_{y_2}\psi_{x_1} & 0\\ 0 & 0\end{bmatrix},\quad Q_0=\begin{bmatrix}0\\ I\end{bmatrix},\quad G_1=\begin{bmatrix} I+\varphi_{y_2}\psi_{x_1} & \varphi_{x_2}\\ 0 & I\end{bmatrix}.$$

Since the matrix function G_1 remains nonsingular, by definition, the DAE is regular with tractability index 1. Compute the constraint sets

$$\widetilde{\mathcal{M}}_0(t)=\{x\in\mathbb{R}^m : x_2=\psi(x_1,t)\},$$

and

$$\begin{aligned}\mathcal{M}_0(t)&=\{x\in\mathbb{R}^m : x_2=\psi(x_1,t),\\ &\qquad \exists y\in\mathbb{R}^n : y_1+\varphi(y_2,x,t)=0, y_2-\psi_t(x_1,t)=\psi_{x_1}(x_1,t)y_1\}\\ &=\{x\in\mathbb{R}^m : x_2=\psi(x_1,t), \exists y_2\in\mathbb{R}^{m_2} : y_2-\psi_t(x_1,t)=-\psi_{x_1}(x_1,t)\varphi(y_2,x,t)\}\\ &=\{x\in\widetilde{\mathcal{M}}_0(t) : \exists y_2\in\mathbb{R}^{m_2} : y_2+\psi_{x_1}(x_1,t)\varphi(y_2,x,t)=\psi_t(x_1,t)\}.\end{aligned}$$

The set $\mathcal{M}_0(t)$ is seemingly a proper subset of $\widetilde{\mathcal{M}}_0(t)$, however, we are not aware of a concrete case where this actually happens.

For a very large class of DAEs the subspace $\ker D(x,t)$ is a $\mathcal{C}^1$-subspace independent of x or even a constant subspace, and we may denote

$$\ker D(x,t)=:N_0(t),\quad \text{for } x\in\mathcal{D}_f,\ t\in\mathcal{I}_f. \tag{3.90}$$

For instance, the MNA-DAEs discussed in Section 3.6 and the DAEs describing constrained mechanical motion show a constant leading nullspace. Also, this property is given in Example 3.54.
Let Q_0 be any continuously differentiable projector function onto N_0, $P_0=I-Q_0$. The property (3.90) implies

$$d(x,t)-d(P_0(t)x,t)=\int_0^1 d_x(sx+(1-s)P_0(t)x,t)Q_0(t)x\,\mathrm{d}s=0,\quad x\in\mathcal{D}_f, t\in\mathcal{I}_f,$$

supposing that the definition domain $\mathcal{D}_f$ contains $P_0(t)x$ together with x. Now the following identities result:

$$
\begin{aligned}
d(x,t) &\equiv d(P_0(t)x,t),\\
d_x(x,t) &\equiv d_x(P_0(t)x,t),\\
d_t(x,t) &\equiv d_t(P_0(t)x,t) + d_x(P_0(t)x,t)P_0'(t)x.
\end{aligned}
$$

The fact that $d(x,t)$ is independent of the nullspace component $Q_0(t)x$ allows us slightly to reduce the smoothness requirement for d in the local solvability assertion.

Theorem 3.55. *Let the DAE (3.1) satisfy Assumption 3.16 and be regular with tractability index 1 on the open set $\mathcal{G} \subset \mathcal{D}_f \times \mathcal{I}_f$. Additionally, let* $\ker D(x,t)$ *be a $\mathcal{C}^1$-subspace independent of x.*
Then, for each $(\bar{x},\bar{t}) \in \mathcal{G}$, $\bar{x} \in \mathcal{M}_0(\bar{t})$, there exists at least one solution $x_* \in C(\mathcal{I}_*,\mathbb{R}^m)$ of the DAE passing through $x_*(\bar{t}) = \bar{x}$.

Proof. We repeat the arguments of the previous proof. The special property of d leads to

$$
\begin{aligned}
\mathcal{D}(\mu,u,w,t) &\equiv \mathcal{D}(\mu,u,0,t),\\
h(u,w,t) &\equiv h(u,0,t),\\
\xi(u,w,t) &\equiv \bar{D}(t)^- h(u,0,t) + \bar{Q}_0(t)w,\\
\mathcal{F}(u,w,t) &\equiv f(D(\bar{D}(t)^- h(u,0,t),t)(I + P_0'(t)Q_0(t))w\\
&\qquad + d_t(\bar{D}(t)^- h(u,0,t),t), \bar{D}(t)^- h(u,0,t) + Q_0(t)w, t).
\end{aligned}
$$

This makes it clear that the partial derivative $\mathcal{F}_w$ exists without assuming d to have the second partial derivatives. However, now the resulting function $\omega(u,t)$ is just continuous. For the existence of ω_u second derivatives of d would be needed. □

3.7.2 Index-2 DAEs

In higher index DAEs, new difficulties arise from hidden constraints and inherent differentiations. We restrict our interest to regular DAEs with tractability index 2 and provide a solvability result which meets, in particular, the situation in circuit modeling. For transparency we recall and modify some relevant facts from Chapter 2.

3.7.2.1 Advanced decoupling of linear index-2 DAEs

Any regular DAE

$$
A(Dx)' + Bx = q, \tag{3.91}
$$

with tractability index two and fine decoupling projector functions Q_0, Q_1, decomposes into the system

$$
\begin{aligned}
u' - (D\Pi_1 D^-)' u + D\Pi_1 G_2^{-1} B D^- u &= D\Pi_1 G_2^{-1} q,\\
-Q_0 Q_1 D^- (D P_0 Q_1 x)' & \\
+ Q_0 x + (Q_0 P_1 G_2^{-1} B + Q_0 Q_1 D^- (D P_0 Q_1 D^-)' D)\Pi_1 x &= Q_0 P_1 G_2^{-1} q,\\
P_0 Q_1 x &= P_0 Q_1 G_2^{-1} q,
\end{aligned}
$$

where $u = D\Pi_1 x$, and the DAE solution is $x = \Pi_1 D^- u + P_0 Q_1 x + Q_0 x$.
Notice that $Q_1 G_2^{-1} B \Pi_1 = 0$ is true for fine decoupling.
Here the matrix function D is supposed to be continuously differentiable. This allows us to choose a continuously differentiable projector function Q_0. In consequence, all D^-, Π_1, $P_0 Q_1$ are continuously differentiable, too. The subspace $\operatorname{im} Q_0 Q_1 = N_0 \cap S_0$ is a $\mathcal{C}$-subspace of dimension $m - r_1$. The matrix function $Q_0 Q_1$ has constant rank $m - r_1$. We introduce the additional continuous projector functions T and $U = I - T$ such that

$$
\operatorname{im} T = \operatorname{im} Q_0 Q_1, \quad T Q_0 = T = Q_0 T, \quad P_0 U = P_0 = U P_0
$$

is fulfilled. With the help of these projectors we advance the splittings of the DAE solution and the DAE itself to $x = \Pi_1 D^- u + P_0 Q_1 x + U Q_0 x + T x$ and

$$
\begin{aligned}
u' - (D\Pi_1 D^-)' u + D\Pi_1 G_2^{-1} B D^- u &= D\Pi_1 G_2^{-1} q,\\
-Q_0 Q_1 D^- (D P_0 Q_1 x)' & \\
+ T x + (T G_2^{-1} B + Q_0 Q_1 D^- (D P_0 Q_1 D^-)' D)\Pi_1 x &= T Q_0 P_1 G_2^{-1} q,\\
U Q_0 x + U Q_0 P_1 G_2^{-1} B D^- u &= U Q_0 G_2^{-1} q,\\
P_0 Q_1 x &= P_0 Q_1 G_2^{-1} q.
\end{aligned}
$$

Since $Z := P_0 Q_1 + U Q_0$ is also a projector function, it is reasonable to write $x = \Pi_1 D^- u + Z x + T x$ and

$$
\begin{aligned}
u' - (D\Pi_1 D^-)' u + D\Pi_1 G_2^{-1} B D^- u &= D\Pi_1 G_2^{-1} q,\\
-Q_0 Q_1 D^- (D P_0 Q_1 x)' & \\
+ T x + (T G_2^{-1} B + Q_0 Q_1 D^- (D P_0 Q_1 D^-)' D)\Pi_1 x &= (T - Q_0 Q_1) G_2^{-1} q,\\
Z x + U Q_0 G_2^{-1} B \Pi_1 x &= Z G_2^{-1} q.
\end{aligned}
$$

It is worth mentioning that this system could also be obtained by splitting the original DAE (3.91) via $(\Pi_1 + Z + T) G_2^{-1}$.
Our construction is accompanied by the properties

$$
\begin{aligned}
\ker Z G_2^{-1} &= \operatorname{im} A,\\
G_2^{-1} B T &= T,\\
Z G_2^{-1} B &= Z G_2^{-1} B \Pi_1 + P_0 Q_1 + U Q_0 = Z G_2^{-1} B \Pi_1 + Z,\\
D Z G_2^{-1} B \Pi_1 &= D P_0 Q_1 G_2^{-1} B \Pi_1 = 0,
\end{aligned}
$$

which play their role in the nonlinear case later on, too.
The obvious constraint of the linear DAE (Definition 3.9 and Proposition 3.10) is

$$\begin{aligned}\mathcal{M}_0(t) &= \{x \in \mathbb{R}^m : \exists! y \in \mathbb{R}^n : y - D'(t)x \in \operatorname{im} D(t),\, A(t)y + B(t)x = q(t)\}\\ &= \{x \in \mathbb{R}^m : B(t)x - q(t) \in \operatorname{im} A(t)\}\\ &= \{x \in \mathbb{R}^m : Z(t)G_2(t)^{-1}(B(t)x - q(t)) = 0\}\\ &= \{x \in \mathbb{R}^m : Z(t)x = Z(t)G_2(t)^{-1}(q(t) - B(t)\Pi_1(t)x)\},\end{aligned}$$

which shows the component $Z(t)x$ to be fixed in terms of $\Pi_1(t)x$. We also use the description

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^m : \exists! D(t)x^1, x^1 \in \mathbb{R}^m :\ A(t)(D(t)x^1 + D'(t)x) + B(t)x = q(t)\},$$

where the symbol $\exists! D(t)x^1$ indicates that x^1 is not necessarily unique, but the image $D(t)x^1$ is so. The above decoupled system unveils in its second line the hidden constraint which additionally fixes the component $T(t)x$. This allows us to describe the set of consistent values as

$$\begin{aligned}\mathcal{M}_1(t) = \big\{x \in \mathbb{R}^m :\ & Z(t)x = Z(t)G_2(t)^{-1}(q(t) - B(t)\Pi_1(t)x),\\ & -Q_0(t)Q_1(t)D(t)^-(D(t)P_0(t)Q_1(t)G_2(t)^{-1}q(t))' + T(t)x\\ & + \big(T(t)G_2(t)^{-1}B(t) + Q_0(t)Q_1(t)D(t)^-(D(t)P_0(t)Q_1(t)D(t)^-)'D(t)\big)\Pi_1(t)x\\ & = (T(t) - Q_0(t)Q_1(t))G_2(t)^{-1}q(t)\big\}.\end{aligned}$$

It is not difficult to check that the more transparent description

$$\begin{aligned}\mathcal{M}_1(t) = \big\{x \in \mathbb{R}^m : \exists! D(t)x^1, x^1 \in \mathbb{R}^m :\ & A(t)(D(t)x^1 + D'(t)x) + B(t)x = q(t),\\ & D(t)P_0(t)Q_1(t)x^1 = (D(t)P_0(t)Q_1(t)G_2(t)^{-1}q(t))' - (D(t)P_0(t)Q_1(t))'x\big\}\end{aligned}$$

is true by an analogous decoupling of the equation

$$A(t)(D(t)x^1 + D'(t)x) + B(t)x = q(t).$$

This new description does not explicitly determine the hidden constraint as the previous version does, but instead fixes the corresponding part of the jet variable x^1. This understanding appears to be helpful in the nonlinear case.

3.7.2.2 Nonlinear index-2 DAEs

We consider the nonlinear DAE with linear derivative term

$$f((D(t)x(t))', x(t), t) = 0, \tag{3.92}$$

which is supposed to satisfy Assumption 3.16 and to be regular with tractability index 2 on the open set $\mathcal{G} \in \mathcal{D}_f \times \mathcal{I}_f$. The obvious constraint set is

$$\begin{aligned}\mathcal{M}_0(t) &= \{x \in \mathcal{D}_f : \exists! y \in \mathbb{R}^n : y - D'(t)x \in \operatorname{im} D(t),\ f(y,x,t) = 0\} \\ &= \{x \in \mathcal{D}_f : \exists! D(t)x^1, x^1 \in \mathbb{R}^m :\ f((D(t)x^1 + D'(t)x, x, t) = 0\}.\end{aligned}$$

We apply the subspaces

$$N_0(t) = \ker D(t) \quad \text{and} \quad S(y,x,t) = \{z \in \mathbb{R}^m : f_x(y,x,t)z \in \operatorname{im} f_y(y,x,t)\},$$

and notice that the intersection $N_0(t) \cap S(y,x,t)$ has the dimension $m - r_1$ for all $y \in \mathbb{R}^n, (x,t) \in \mathcal{G}, x \in \mathcal{M}_0(t)$.

Consider a fixed point $(\bar{x}, \bar{t}) \in \mathcal{G}$ such that $\bar{x} \in \mathcal{M}_0(\bar{t})$, and denote by $\bar{y} \in \mathbb{R}^n$, $\bar{x}^1 \in \mathbb{R}^m$ associated values with

$$\bar{y} - D'(\bar{t})\bar{x} = D(\bar{t})\bar{x}^1,\ f(\bar{y}, \bar{x}, \bar{t}) = 0.$$

We intend to take use of the above advanced decoupling of linear index-2 DAEs applying it via linearization. For this aim we construct a reference function $\tilde{x} \in \mathcal{C}^2(\mathcal{I}, \mathbb{R}^m)$ such that its values lie in $\mathcal{G}$, i.e., $(\tilde{x}(t), t) \in \mathcal{G}$, for $t \in \mathcal{I}$, and $\tilde{x}(\bar{t}) = \bar{x}$, $(D\tilde{x})'(\bar{t}) = \bar{y}$. We can take, for instance, $\tilde{x}(t) = \bar{x} + (t - \bar{t})\bar{x}^1$. Then we apply linearization along the function $\tilde{x}$ and arrive at the linear DAE

$$\tilde{A}(Dx)' + \tilde{B}x = q \tag{3.93}$$

with coefficients

$$\tilde{A}(t) = f_y\big((D(t)\tilde{x}(t))', \tilde{x}(t), t\big),\ \tilde{B}(t) = f_x\big((D(t)\tilde{x}(t))', \tilde{x}(t), t\big),\ t \in \mathcal{I}.$$

The linear DAE (3.93) inherits regularity with tractability index 2 and the characteristic values r_0, r_1, r_2 from the nonlinear DAE (3.92) as guaranteed by Theorem 3.33. Choose the projector function Q_0 onto N_0 to be continuously differentiable. This is possible owing to the continuous differentiability of D.
Without loss of generality (cf. Proposition 3.79) we suppose the subspace $\ker f_y(y,x,t)$ to be independent of the variables y and x, such that $\ker f_y(y,x,t) = \ker R(t)$ holds true. Since D, R and P_0 are continuously differentiable, so is D^-.

The following condition restricts the possible structure of the DAE (3.92), but it is valid, for instance, in the MNA DAE in the previous section:

$$\operatorname{im} f_y(y,x,t) \text{ and } N_0(t) \cap S(y,x,t) \text{ are independent of } y \text{ and } x. \tag{3.94}$$

Assuming this structure to be given, $T(t) \in L(\mathbb{R}^m)$ denotes an additional projector such that

$$\operatorname{im} T(t) = N_0(t) \cap S(y,x,t), \quad y \in \mathbb{R}^n, (x,t) \in \mathcal{G}. \tag{3.95}$$

We choose T as in the linear case so that the relations

$$TQ_0 = T = Q_0T, \quad U := I - T, \quad UP_0 = P_0 = P_0U$$

become true. Supposing the intersection subspace to be a $\mathcal{C}^1$-subspace we may arrange T to be continuously differentiable.
Next, starting with Q_0, we form an admissible matrix function sequence and admissible projector functions for the linear DAE (3.93). We indicate by a tilde, if they may depend on the function $\tilde{x}$, e.g., $\tilde{Q}_1$, etc. The resulting projector functions $\tilde{\Pi}_1$ and $\tilde{Z} = P_0\tilde{Q}_1 + UQ_0$ are also continuously differentiable.
Denote

$$\begin{aligned}
\bar{u} &= D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x},\\
\bar{z} &= \tilde{Z}(\bar{t})\bar{x},\\
\bar{w} &= T(\bar{t})\bar{x},\\
\bar{u}^1 &= D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}^1 + (D\tilde{\Pi}_1)'(\bar{t})\bar{x},\\
\bar{v}^1 &= D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\bar{x}^1 + (DP_0\tilde{Q}_1)'(\bar{t})\bar{x},
\end{aligned}$$

so that

$$\bar{x} = \tilde{\Pi}_1(\bar{t})D(\bar{t})^-\bar{u} + \bar{z} + \bar{w}, \quad \bar{y} = \bar{u}^1 + \bar{v}^1.$$

Since the subspace $\operatorname{im} f_y(y,x,t)$ does not vary with y and x, it holds that

$$\ker \tilde{Z}(t)\tilde{G}_2(t)^{-1} = \operatorname{im} f_y(y,x,t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}. \tag{3.96}$$

Moreover, the condition (3.94) yields

$$\tilde{Z}(t)\tilde{G}_2(t)^{-1} f_x(y,x,t)T(t) = 0, \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}. \tag{3.97}$$

Namely, for each $\xi \in \mathbb{R}^m$, $T(t)\xi$ belongs to $S(y,x,t)$, thus $f_x(y,x,t)T(t)\xi \in \operatorname{im} f_y(y,x,t)$, and hence $\tilde{Z}(t)\tilde{G}_2(t)^{-1} f_x(y,x,t)T(t)\xi = 0$.

Further, we derive for $y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}$ that

$$\begin{aligned}
&\tilde{Z}(t)\tilde{G}_2(t)^{-1}\big\{f(y,x,t) - f(0,(\tilde{\Pi}_1(t) + \tilde{Z}(t))x,t)\big\}\\
&\quad = \int_0^1 \tilde{Z}(t)\tilde{G}_2(t)^{-1}\big\{f_y(sy, sx + (1-s)(\tilde{\Pi}_1(t) + \tilde{Z}(t))x,t)y\\
&\qquad\qquad + f_x(sy, sx + (1-s)(\tilde{\Pi}_1(t) + \tilde{Z}(t))x,t)T(t)x\big\}\mathrm{d}s = 0.
\end{aligned}$$

This yields the identities

$$\begin{aligned}
\tilde{Z}(t)\tilde{G}_2(t)^{-1} f(y,x,t) &= \tilde{Z}(t)\tilde{G}_2(t)^{-1} f(0,(\tilde{\Pi}_1(t) + \tilde{Z}(t))x,t),\\
\tilde{Z}(t)\tilde{G}_2(t)^{-1} f_x(y,x,t) &= \tilde{Z}(t)\tilde{G}_2(t)^{-1} f_x(0,(\tilde{\Pi}_1(t) + \tilde{Z}(t))x,t)(\tilde{\Pi}_1(t) + \tilde{Z}(t)),\\
&\qquad y \in \mathbb{R}^n,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}.
\end{aligned}$$

Define the auxiliary function g on a neighborhood $\mathcal{N}_{(\bar{u},\bar{z},\bar{t})}$ of $(\bar{u},\bar{z},\bar{t})$ in $\mathbb{R}^{n+m+1}$ by

$$g(u,z,t) := \tilde{Z}(t)\tilde{G}_2(t)^{-1} f(0, \tilde{\Pi}_1(t)D(t)^{-}u + \tilde{Z}(t)z, t), \quad (u,z,t) \in \mathcal{N}_{(\bar{u},\bar{z},\bar{t})}. \quad (3.98)$$

This gives $g(\bar{u},\bar{z},\bar{t}) = \tilde{Z}(\bar{t})\tilde{G}_2(\bar{t})^{-1} f(\bar{y},\bar{x},\bar{t}) = 0$. The function g plays its role in our solvability assertion, which takes up the idea of applying the structural condition (3.94) from [205] and [211]. A priori, g is continuous together with its partial derivatives g_u and g_z.

Now we are in a position to formulate the solvability assertion.

Theorem 3.56. *Let the DAE (3.92) satisfy Assumption 3.16 and be regular with tractability index 2 on the open set $\mathcal{G}$. Let the structural restriction (3.94) be given, and the intersection $N_0 \cap S$ be a $\mathcal{C}^1$-subspace.*
Let $(\bar{x},\bar{t}) \in \mathcal{G}$ be fixed, $\bar{x} \in \mathcal{M}_0(\bar{t})$, and let $\bar{y} \in \mathbb{R}^n$, $\bar{x}^1 \in \mathbb{R}^m$ denote associated values with

$$\bar{y} - D'(\bar{t})\bar{x} = D(\bar{t})\bar{x}^1, \quad f(\bar{y},\bar{x},\bar{t}) = 0.$$

Let the linearized DAE (3.93) have the fine decoupling projector functions $Q_0, \tilde{Q}_1$. Let the function g (see (3.98)) continuously differentiable.
Let the consistency condition

$$D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\bar{x}^1 + (DP_0\tilde{Q}_1)'(\bar{t})(I - \tilde{Z}(\bar{t}))\bar{x} + (Dg)_t(D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}, \tilde{Z}(\bar{t})\bar{x}, \bar{t}) = 0 \quad (3.99)$$

be satisfied.

(1) *Then there exists at least one solution $x_* \in \mathcal{C}_D^1(\mathcal{I}_*, \mathbb{R}^m)$ of the DAE passing through $x_*(\bar{t}) = \bar{x}$.*

(2) *The IVP*

$$f((D(t)x(t))', x(t), t) = 0, \quad D(\bar{t})\tilde{\Pi}_1(\bar{t})x(\bar{t}) = D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x} + \delta,$$

is solvable for all sufficiently small $\delta \in \operatorname{im} D(\bar{t})\tilde{\Pi}_1(\bar{t})$.

(3) *If, additionally, g has second partial derivatives $g_{uu}, g_{zz}, g_{zu}, g_{tu}, g_{tz}$, then the DAE solutions in items (1) and (2) are unique.*

The idea behind the proof is to benefit from the structural restrictions and to decompose the original DAE in the following way.
Because of $\tilde{\Pi}_1 + \tilde{Z} + T = I$ and the nonsingularity of $\tilde{G}_2$ we may turn from the given DAE (3.92) to

$$(\tilde{\Pi}_1(t) + \tilde{Z}(t) + T(t))\tilde{G}_2(t)^{-1} f((D(t)x(t))', x(t), t) = 0.$$

Owing to the projector properties, the latter equation decomposes into the three parts

$$\begin{aligned}
\tilde{\Pi}_1(t)\tilde{G}_2(t)^{-1} f((D(t)x(t))', x(t), t) &= 0,\\
\tilde{Z}(t)\tilde{G}_2(t)^{-1} f((D(t)x(t))', x(t), t) &= 0,\\
T(t)\tilde{G}_2(t)^{-1} f((D(t)x(t))', x(t), t) &= 0,
\end{aligned}$$

which is the basic idea of the following investigation.

Lemma 3.57. *Let g have the additional continuous partial derivative g_t.*

(1) *Then there is a unique continuously differentiable function $h : \mathcal{N}_{(\bar{u},\bar{t})} \to \mathbb{R}^m$ such that $h(\bar{u},\bar{t}) = \bar{z}$, $h_u(\bar{u},\bar{t}) = \tilde{Z}(\bar{t})\tilde{G}_2(\bar{t})^{-1}\tilde{B}(\bar{t})\tilde{\Pi}_1(\bar{t})$, and $g(u,h(u,t),t) = 0$,*

$$h(u,t) = \tilde{Z}(t)h(u,t) = h(D(t)\tilde{\Pi}_1(t)D(t)^- u,t),\ (u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}.$$

(2) *If $Q_0, \tilde{Q}_1$ provide a fine decoupling of the linearized DAE (3.93), then it follows that $D(\bar{t})h_u(\bar{u},\bar{t}) = 0$.*

(3) *If g has the additional continuous partial derivatives $g_{uu}, g_{zz}, g_{uz}, g_{tz}, g_{tu}$, then the function h has the continuous second partial derivatives h_{uu}, h_{tu}.*

Proof. Introduce the continuously differentiable function

$$\mathcal{H}(u,z,t) := g(u,z,t) + (I - \tilde{Z}(t))z, \quad (u,z,t) \in \mathcal{N}_{(\bar{u},\bar{z},\bar{t})}.$$

Compute $\mathcal{H}(\bar{u},\bar{z},\bar{t}) = 0$ and

$$\mathcal{H}_z(\bar{u},\bar{z},\bar{t}) = g_z(\bar{u},\bar{z},\bar{t}) + (I - \tilde{Z}(t)) = \tilde{Z}(t)\tilde{G}_2(t)^{-1}\tilde{B}(t)\tilde{Z}(t) + (I - \tilde{Z}(t)) = I.$$

The first and third assertions are now direct consequences of the standard implicit function theorem, and the fact that $\mathcal{H}(u,h(u,t),t) = 0$ implies $g(u,h(u,t),t) = 0$. The second assertion follows from

$$D(\bar{t})\tilde{Z}(\bar{t})\tilde{G}_2(\bar{t})^{-1}\tilde{B}(\bar{t})\tilde{\Pi}_1(\bar{t}) = D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\tilde{G}_2(\bar{t})^{-1}\tilde{B}(\bar{t})\tilde{\Pi}_1(\bar{t}) = 0.$$

□

Define the further auxiliary function

$$v^1(u,t,u^1) := D'(t)h(u,t) + D(t)\{h_u(u,t)u^1 + h_t(u,t)\}, \quad u^1 \in \mathbb{R}^n, (u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}.$$

If g is continuously differentiable, then the function v^1 is continuous together with its partial derivative $v^1_{u^1}(u,t,u^1) = D(t)h_u(u,t)$. If g has additional second partial derivatives, then v^1 also has the continuous partial derivative v^1_u. Compute

$$v^1(\bar{u},\bar{t},\bar{u}^1) = D'(\bar{t})\bar{z} + D(\bar{t})h_t(\bar{u},\bar{t})$$

and

$$h_t(\bar{u},\bar{t}) = -\mathcal{H}_t(\bar{u},\bar{z},\bar{t}) = -g_t(\bar{u},\bar{z},\bar{t}) + \tilde{Z}'(\bar{t})\bar{z},$$

$$D(\bar{t})h_t(\bar{u},\bar{t}) = -D(\bar{t})g_t(\bar{u},\bar{z},\bar{t}) + D(\bar{t})\tilde{Z}'(\bar{t})\bar{z} = -D(\bar{t})g_t(\bar{u},\bar{z},\bar{t}) + (D\tilde{Z})'(\bar{t})\bar{z} - D'(\bar{t})\bar{z}.$$

It follows that

$$v^1(\bar{u},\bar{t},\bar{u}^1) = (DP_0\tilde{Q}_1)'(\bar{t})\bar{z} - D(\bar{t})g_t(\bar{u},\bar{z},\bar{t}). \tag{3.100}$$

Assume for the moment that there is already a solution $x_* \in \mathcal{C}^1_D(\mathcal{I}_*, \mathbb{R}^n)$ passing through $x_*(\bar{t}) = \bar{x}$, with $\mathcal{I}_*$ a small neighborhood of $\bar{t}$. Then

$$\begin{aligned}
\tilde{Z}(t)x_*(t) &= h(D(t)\tilde{\Pi}_1(t)x_*(t),t),\\
D(t)P_0(t)\tilde{Q}_1(t)x_*(t) &= D(t)h(D(t)\tilde{\Pi}_1(t)x_*(t),t),\\
(D(t)P_0(t)\tilde{Q}_1(t)x_*(t))' &= v^1\big(D(t)\tilde{\Pi}_1(t)x_*(t),t,(D(t)\tilde{\Pi}_1(t)x_*(t))'\big) \quad \text{for } t\in\mathcal{I}_*,
\end{aligned}$$

must necessarily be valid, and in particular also

$$\bar{v}^1 = v^1(\bar{u},\bar{t},\bar{u}^1). \tag{3.101}$$

The last condition (3.101) reads in detail (cf. (3.100))

$$D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\bar{x}^1 + (DP_0\tilde{Q}_1)'(\bar{t})\bar{x} = (DP_0\tilde{Q}_1)'(\bar{t})\bar{z} - D(\bar{t})g_t(\bar{u},\bar{z},\bar{t}).$$

Taking into account that

$$D(\bar{t})g_t(\bar{u},\bar{z},\bar{t}) = (Dg)_t(\bar{u},\bar{z},\bar{t}) - D'(\bar{t})g(\bar{u},\bar{z},\bar{t}) = (Dg)_t(\bar{u},\bar{z},\bar{t})$$

we obtain the equivalent condition (3.99), namely

$$D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\bar{x}^1 + (DP_0\tilde{Q}_1)'(\bar{t})(I-\tilde{Z}(\bar{t}))\bar{x} + (Dg)_t(D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x},\tilde{Z}(\bar{t})\bar{x},\bar{t}) = 0.$$

For a linear DAE (3.91), the condition (3.99) simplifies to the demand

$$D(\bar{t})P_0(\bar{t})Q_1(\bar{t})\bar{x}^1 + (DP_0Q_1)'(\bar{t})\bar{x} = (DP_0Q_1G_2^{-1}q)'(\bar{t})$$

already known in this context and used to describe the set of consistent values $\mathcal{M}_1(t)$ in the linear case.

Proof (of Theorem 3.56). We verify assertion (2) first. We introduce the function $\mathcal{K}$ on a neighborhood $\mathcal{N}_{(\bar{u},\bar{w},\bar{u}^1,\bar{t})}$ of $(\bar{u},\bar{w},\bar{u}^1,\bar{t})$ by

$$\begin{aligned}
\mathcal{K}(u,w,u^1,t) &:= (I-D(t)\tilde{\Pi}_1(t)D(t)^-)(u^1-(D(t)\tilde{\Pi}_1(t)D(t)^-)'u)\\
&\quad + D(t)\tilde{\Pi}_1(t)\tilde{G}_2(t)^{-1}f\big(u^1+v^1(u,t,u^1),\ \tilde{\Pi}_1(t)D(t)^-u+h(u,t)+T(t)w,\ t\big).
\end{aligned}$$

The function $\mathcal{K}$ is continuous and has continuous partial derivatives $\mathcal{K}_{u^1}$ and $\mathcal{K}_w$. We obtain $\mathcal{K}(\bar{u},\bar{w},\bar{u}^1,\bar{t}) = 0$ and $\mathcal{K}_{u^1}(\bar{u},\bar{w},\bar{u}^1,\bar{t}) = I$. Then the implicit function theorem provides a unique continuous function $k : \mathcal{N}_{(\bar{u},\bar{w},\bar{t})} \to \mathbb{R}^n$ such that $k(\bar{u},\bar{w},\bar{t}) = \bar{u}^1$; further (cf. (3.97)) $k_w(\bar{u},\bar{w},\bar{t}) = 0$ and

$$\begin{aligned}
&\mathcal{K}(u,w,k(u,w,t),t) = 0,\\
&k(u,w,t) - (D(t)\tilde{\Pi}_1(t)D^-(t))'u\\
&\qquad = D(t)\tilde{\Pi}_1(t)D^-(t)(k(u,w,t)-(D(t)\tilde{\Pi}_1(t)D^-(t))'u),\\
&D(t)\tilde{\Pi}_1(t)\tilde{G}_2(t)^{-1}f\big(k(u,w,t)\\
&\qquad + v^1(u,t,k(u,w,t)),\tilde{\Pi}_1(t)D(t)^-u+h(u,t)+T(t)w,\ t\big) = 0,\\
&k(u,w,t) = k(D(t)\tilde{\Pi}_1(t)D^-(t)u,w,t) = k(u,T(t)w,t), \quad (u,w,t)\in\mathcal{N}_{(\bar{u},\bar{w},\bar{t})}.
\end{aligned}$$

The composed function $\gamma(u,w,t) := k(u,w,t) + v^1(u,t,k(u,w,t))$ is continuous together with its partial derivative γ_w. We have $\gamma(\bar{u},\bar{w},\bar{t}) = \bar{u}^1 + \bar{v}^1 = \bar{y}$ and $\gamma_w(\bar{u},\bar{w},\bar{t}) = 0$.
We build the last auxiliary function $\mathcal{L}$ on a neighborhood $\mathcal{N}_{(\bar{u},\bar{w},\bar{t})}$ of $(\bar{u},\bar{w},\bar{t})$ by

$$\begin{aligned}\mathcal{L}(u,w,t) :=& (I - T(t))w \\ &+ T(t)\tilde{G}_2(t)^{-1} f\big(\gamma(u,w,t), \tilde{\Pi}_1(t)D(t)^- u + h(u,t) + T(t)w, t\big).\end{aligned}$$

$\mathcal{L}$ is continuous and has a continuous partial derivative $\mathcal{L}_w$. It holds that $\mathcal{L}(\bar{u},\bar{w},\bar{t}) = 0$ and $\mathcal{L}_w = (I-T) + T\tilde{G}_2^{-1}(f_y\gamma_w + f_x T)$, and hence
$\mathcal{L}_w(\bar{u},\bar{w},\bar{t}) = I - T(\bar{t}) + T(\bar{t})\tilde{G}_2(\bar{t})^{-1}\tilde{B}(\bar{t})T(\bar{t}) = I$. The implicit function theorem yields a continuous function $l : \mathcal{N}_{(\bar{u},\bar{t})} \to \mathbb{R}^m$ such that $l(\bar{u},\bar{t}) = \bar{w}$ and, for $(u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}$,

$$\begin{aligned}&\mathcal{L}(u,l(u,t),t) = 0, \quad l(u,t) = T(t)l(u,t) = l(D(t)\tilde{\Pi}_1(t)D(t)^- u, t), \\ &T(t)\tilde{G}_2(t)^{-1} f(\gamma(u,l(u,t),t), \tilde{\Pi}_1(t)D(t)^- u + h(u,t) + l(u,t), t) = 0.\end{aligned}$$

Now we are prepared to construct a solution of the IVP in assertion (2). Owing to Peano's theorem, the standard IVP

$$u'(t) = k(u(t), l(u(t),t),t) =: \varphi(u(t),t), \quad u(\bar{t}) = \bar{u} + \delta \tag{3.102}$$

possesses at least one solution $u_* \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^n)$.
Multiplying the identity

$$u_*'(t) - k(u_*(t), l(u_*(t),t),t) \equiv 0$$

by $I - D(t)\tilde{\Pi}_1(t)D(t)^-$ we prove that the function $\alpha_* := (I - D\tilde{\Pi}_1 D^-)u_*$ is the solution of the IVP $\alpha' = (I - D\tilde{\Pi}_1 D^-)'\alpha$, $\alpha(\bar{t}) = 0$. Therefore, α_* vanishes identically, which means that $u_* = D\tilde{\Pi}_1 D^- u_*$ is true.
We compose the continuous function

$$x_*(t) := \tilde{\Pi}_1(t)D(t)^- u_*(t) + h(u_*(t),t) + l(u_*(t),t), \quad t \in \mathcal{I}_*$$

such that

$$T(t)x_*(t) = l(u_*(t),t), \quad \tilde{Z}(t)x_*(t) = h(u_*(t),t), \quad D(t)\tilde{\Pi}_1(t)x_*(t) = u_*(t).$$

The part $Dx_* = u_* + Dh(u_*(.),.)$ is continuously differentiable.
The initial condition $D(\bar{t})\tilde{\Pi}_1(\bar{t})x_*(\bar{t}) = \bar{u} + \delta$ is fulfilled. Finally, due to the construction of the auxiliary functions h, l and k, we have the three identities on $\mathcal{I}_*$:

$$\begin{aligned}
\tilde{Z}(t)\tilde{G}_2(t)^{-1} f((D(t)x_*(t))',x_*(t),t) &= \tilde{Z}(t)\tilde{G}_2(t)^{-1} f(0,\tilde{\Pi}_1(t)D(t)^- u_*(t) \\
&\qquad + \tilde{Z}(t)x_*(t),t) \\
&= g(u_*(t),h(u_*(t),t),t) = 0, \\
T(t)\tilde{G}_2(t)^{-1} f((D(t)x_*(t))',x_*(t),t) &= \mathcal{L}(u_*(t),l(u_*(t),t),t) = 0, \\
D(t)\tilde{\Pi}_1(t)\tilde{G}_2(t)^{-1} f((D(t)x_*(t))',x_*(t),t) &= \mathcal{K}(u_*(t),w_*(t),k(u_*(t),w_*(t),t),t) = 0.
\end{aligned}$$

Altogether this yields for $t \in \mathcal{I}_*$

$$\begin{aligned}
&f((D(t)x_*(t))',x_*(t),t) \\
&= \tilde{G}_2(t)(\tilde{Z}(t) + T(t) + D(t)^- D(t)\tilde{\Pi}_1(t))\tilde{G}_2(t)^{-1} f((D(t)x_*(t))',x_*(t),t) = 0.
\end{aligned}$$

Assertion (1) is a consequence of (2). Namely, we put $\delta = 0$ in the initial condition. Then, for the solution provided by (2) it follows that

$$x_*(\bar{t}) := \tilde{\Pi}_1(\bar{t})D(\bar{t})^- \bar{u} + h(\bar{u},t) + l(\bar{u},t) = \bar{x}.$$

Assertion (3) follows immediately since now the function φ in (3.102) possesses the continuous partial derivative φ_u. □

3.7.2.3 Index reduction step

The smoothness demands for g (see (3.98)) in the previous part require only parts of the function f to be smoother than is supposed in the basic Assumption 3.16. A generous overall assumption would be that f is twice continuously differentiable.

Theorem 3.56 provides the following local description of the constraint sets: for $(x,t) \in \mathcal{N}_{(\bar{x},\bar{t})}$ it holds that

$$\begin{aligned}
x \in \mathcal{M}_0(t) &\iff \tilde{Z}(t)x = h(D(t)\tilde{\Pi}_1(t)x,t), \\
x \in \mathcal{M}_1(t) &\iff (\tilde{Z}(t) + T(t))x = h(D(t)\tilde{\Pi}_1(t)x,t) + l(D(t)\tilde{\Pi}_1(t)x,t).
\end{aligned}$$

Agreeing upon the meaning of the projectors $\tilde{Q}_1$ and $\tilde{\Pi}_1$ as given via linearization and fine decoupling we can describe the set of consistent initial values as

$$\mathcal{M}_1(\bar{t}) = \{\bar{x} \in \mathcal{M}_0(\bar{t}) : D(\bar{t})P_0(\bar{t})\tilde{Q}_1(\bar{t})\bar{x}^1 = (DP_0\tilde{Q}_1)'(\bar{t})\bar{z} - (Dg)_t(\bar{u},\bar{z},\bar{t})\}.$$

With the background of Theorem 3.56, we can write

$$\begin{aligned}
f((D(t)x(t))',x(t),t) &= f((D\tilde{\Pi}_1 x)'(t) + (DP_0Q_1x)'(t),x(t),t) \\
&= f\big((D\tilde{\Pi}_1 D^-)(t)(D\tilde{\Pi}_1 x)'(t) \\
&\qquad + (D\tilde{\Pi}_1 D^-)'(t)(D\tilde{\Pi}_1 x)(t) + (DP_0\tilde{Q}_1 x)'(t),x(t),t\big),
\end{aligned}$$

for all functions $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ with values in $\mathcal{D}_f$. In the above analysis we learned that the term $D\tilde{\Pi}_1 x$ corresponds to the inherent ODE while the term $DP_0\tilde{Q}_1 x$ is to

be differentiated. The function v^1 defined via Lemma 3.57 as

$$v^1(u,t,u^1) := D'(t)h(u,t) + D(t)\{h_u(u,t)u^1 + h_t(u,t)\}, \quad u^1 \in \mathbb{R}^n, (u,t) \in \mathcal{N}_{(\bar{u},\bar{t})},$$

represents the tool to carry out the inherent differentiation in the given index-2 DAE, that is, for each solution x_* it holds that

$$(D(t)P_0(t)\tilde{Q}_1(t)x_*(t))' = v^1(D(t)\tilde{\Pi}_1(t)x_*(t), t, (D(t)\tilde{\Pi}_1(t)x_*(t))').$$

Inserting such an expression into the original DAE, we arrive at the new DAE

$$\check{f}((\check{D}(t)x(t))', x(t), t) = 0, \tag{3.103}$$

with functions

$$\begin{aligned}\check{D}(t) :=& D(t)\tilde{\Pi}_1(t),\\ \check{f}(\check{y},x,t) :=& f\big((D\tilde{\Pi}_1 D^-)(t)\check{y}\\ &+ (D\tilde{\Pi}_1 D^-)'(t)D(t)\tilde{\Pi}_1(t)x + v^1(D(t)\tilde{\Pi}_1(t)x, t, \check{y}), x, t\big),\end{aligned}$$

defined for $\check{y} \in \mathbb{R}^n$, $x \in \mathcal{N}_{\bar{x}}$, $t \in \mathcal{N}_{\bar{t}}$.
We expect this procedure to provide a local index reduction by one. In fact, the new DAE is regular with tractability index 1 as we show by Theorem 3.58.
Recall the function g playing its role in Lemma 3.57, and in the definition of the function h, in turn used to obtain the function v^1:

$$g(u,z,t) := \tilde{Z}(t)\tilde{G}_2(t)^{-1} f(0, \tilde{\Pi}_1(t)D(t)^- u + \tilde{Z}(t)z, t), \ (u,z,t) \in \mathcal{N}_{(D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}, \tilde{Z}(\bar{t})\bar{x}, \bar{t})}.$$

We stress once more that the function g inherits its smoothness from the function f.

Theorem 3.58. *Let the DAE (3.92) satisfy Assumption 3.16 and be regular with tractability index 2 on the open set $\mathcal{G}$. Let the structural restriction (3.94) be given, and the intersection $N_0 \cap S$ be a $\mathcal{C}^1$-subspace.*
Let $(\bar{x},\bar{t}) \in \mathcal{G}$ be fixed, $\bar{x} \in \mathcal{M}_0(\bar{t})$, and let $\bar{y} \in \mathbb{R}^n$, $\bar{x}^1 \in \mathbb{R}^m$ denote associated values such that

$$\bar{y} - D'(\bar{t})\bar{x} = D(\bar{t})\bar{x}^1, \quad f(\bar{y},\bar{x},\bar{t}) = 0.$$

Let the linearized DAE (3.93) have the fine decoupling projector functions $Q_0, \tilde{Q}_1$. Let the function g be continuously differentiable and have continuous second partial derivatives $g_{uu}, g_{zz}, g_{zu}, g_{tu}, g_{tz}$.
Let the consistency condition (3.99) be satisfied.

(1) *Then the DAE (3.103) is regular with tractability index 1 on a neighborhood of $(\bar{x}^1,\bar{x},\bar{t})$, and $\bar{x}$ belongs to the obvious constraint set $\check{\mathcal{M}}_0(\bar{t})$ of this DAE.*
(2) *The DAEs (3.92) and (3.103) are locally equivalent in the following sense:*

If $x_* \in \mathcal{C}^1_D(\mathcal{I}_*, \mathbb{R}^m)$ is a solution of the index-2 DAE (3.92) passing through the reference point $(\bar{x}, \bar{t})$, then x_* belongs to the function space $\mathcal{C}^1_{D\tilde{\Pi}_1}(\mathcal{I}_*, \mathbb{R}^m)$ and solves the index-1 DAE (3.103). If $x_* \in \mathcal{C}^1_{D\tilde{\Pi}_1}(\mathcal{I}_*, \mathbb{R}^m)$ is a solution of the index-1 DAE (3.103) passing through the reference point $(\bar{x}, \bar{t})$, then x_* belongs also to the function space $\mathcal{C}^1_D(\mathcal{I}_*, \mathbb{R}^m)$ and solves the index-2 DAE (3.92).

To be more precise, regularity with tractability index 1 on a neighborhood of $(\bar{x}^1, \bar{x}, \bar{t})$ is meant in the sense of Definition 3.62 while the previous Definition 3.28 explains regularity with index 1 only if there is no restriction on the jet variable x^1. Definition 3.62 allows a localization also of the jet variables.
In the simpler particular case, if $f(y,x,t) = A(t)y + b(x,t)$, the resulting DAE (3.103) is regular with index 1 on a neighborhood $\mathcal{N}_{(\bar{x},\bar{t})}$ of $(\bar{x}, \bar{t})$ in the sense of Definition 3.28. In the context of the more general Definition 3.62 this is regularity with tractability index 1 on $\mathbb{R}^m \times \mathcal{N}_{(\bar{x},\bar{t})}$.

Proof. Owing to Lemma 3.57, the function h is continuously differentiable and has the continuous second partial derivatives h_{uu} and h_{tu}. In consequence, the function v^1 is continuous with a continuous partial derivative v^1_u, and it depends linearly on u^1.
Moreover, the property $h_u = h_u D\tilde{\Pi}_1 D^-$ is given, and hence $v^1_{u^1} = Dh_u = v^1_{u^1} D\tilde{\Pi}_1 D^-$. The matrix function $\check{D}$ is continuously differentiable, the function $\check{f}$ is continuous and has continuous partial derivatives with respect to $\check{y}$ and x,

$$\check{f}_{\check{y}} = f_y\{D\tilde{\Pi}_1 D^- + v^1_{u^1}\} = f_y(I + Dh_u)D\tilde{\Pi}_1 D^-,$$
$$\check{f}_x = f_y\{(D\tilde{\Pi}_1 D^-)' D\tilde{\Pi}_1 + v^1_u D\tilde{\Pi}_1\} + f_x.$$

We show that $\ker \check{f}_{\check{y}} = \operatorname{im}(I - D\tilde{\Pi}_1 D^-)$ is valid, but then the derivative term of (3.103) is properly involved and the DAE satisfies the basic Assumption 3.16. Consider a $w \in \ker \check{f}_{\check{y}}$, which means

$$(I + Dh_u)D\tilde{\Pi}_1 D^- w \in \ker f_y \cap \operatorname{im} D = \{0\}.$$

Due to $Dh_u = DP_0\tilde{Q}_1 h_u \tilde{\Pi}_1$ (cf. Lemma 3.57), the matrix function

$$I + Dh_u = I + DP_0\tilde{Q}_1 h_u \tilde{\Pi}_1$$

is nonsingular, but this implies $D\tilde{\Pi}_1 D^- w = 0$.
Next we verify the index-1 property. Choose $\check{Q}_0 = I - \tilde{\Pi}_1$ and indicate in the same way the members of the matrix function sequence associated with the DAE (3.103). Compute

$$\begin{aligned}
\breve{A}(\bar{x}^1,\bar{x},\bar{t}) &= \breve{f}_{\dot{y}}(\breve{D}(\bar{t})\bar{x}^1+\breve{D}'(\bar{t})\bar{x},\bar{x},\bar{t}) = f_y(\bar{u}^1+\bar{v}^1,\bar{x},\bar{t}) = \tilde{A}(\bar{t}),\\
\breve{B}(\bar{x}^1,\bar{x},\bar{t})\breve{Q}_0(\bar{t}) &= \breve{f}_x(\breve{D}(\bar{t})\bar{x}^1+\breve{D}'(\bar{t})\bar{x},\bar{x},\bar{t})(I-\tilde{\Pi}_1(\bar{t}))\\
&= f_x(\bar{u}^1+\bar{v}^1,\bar{x},\bar{t})(I-\tilde{\Pi}_1(\bar{t})) = \tilde{B}(\bar{t})(I-\tilde{\Pi}_1(\bar{t})),\\
\breve{G}_1(\bar{x}^1,\bar{x},\bar{t}) &= \breve{A}(\bar{x}^1,\bar{x},\bar{t})\breve{D}(\bar{t})+\breve{B}(\bar{x}^1,\bar{x},\bar{t})\breve{Q}_0(\bar{t})\\
&= \tilde{A}(\bar{t})D(\bar{t})\tilde{\Pi}_1(\bar{t})+\tilde{B}(\bar{t})(I-\tilde{\Pi}_1(\bar{t}))\\
&= \tilde{A}(\bar{t})D(\bar{t})\tilde{\Pi}_1(\bar{t})+\tilde{B}(\bar{t})(\tilde{Z}(\bar{t})+T(\bar{t})).
\end{aligned}$$

It follows that

$$\tilde{G}_2(\bar{t})^{-1}\breve{G}_1(\bar{x}^1,\bar{x},\bar{t}) = \tilde{\Pi}_1(\bar{t})+\tilde{G}_2(\bar{t})^{-1}\tilde{B}(\bar{t})\tilde{Z}(\bar{t})+T(\bar{t}).$$

Let ζ belong to the nullspace of $\breve{G}_1(\bar{x}^1,\bar{x},\bar{t})$, that is

$$\{\tilde{A}(\bar{t})D(\bar{t})\tilde{\Pi}_1(\bar{t})+\tilde{B}(\bar{t})((\tilde{Z}(\bar{t})+T(\bar{t}))\}\zeta = 0.$$

Multiplication by $\tilde{Z}(\bar{t})\tilde{G}_2(\bar{t})^{-1}$ yields $\tilde{Z}(\bar{t})\zeta = 0$. Then, set $\tilde{Z}(\bar{t})\zeta = 0$ and multiply by $\tilde{G}_2(\bar{t})^{-1}$ so that

$$\tilde{\Pi}_1(\bar{t})\zeta + T(\bar{t})\zeta = 0$$

results. Finally, this implies $\zeta = 0$, and the matrix $\breve{G}_1(\bar{x}^1,\bar{x},\bar{t})$ is nonsingular. It is clear that the matrix function $\breve{G}_1$ preserves nonsingularity on a neighborhood of our reference point $(\bar{x}^1,\bar{x},\bar{t})$, and therefore the DAE is regular with tractability index 1 on a neighborhood of this point.

Notice that, for the special case $f(y,x,t) = A(t)y+b(x,t)$, the matrix function $\breve{G}_1$ is independent of the variable x^1, and therefore it remains nonsingular uniformly for all $x^1 \in \mathbb{R}^m$.

Next we show that the reference point $\bar{x}$ belongs to the obvious constraint set associated with the DAE (3.103) at time $\bar{t}$, that is

$$\breve{\mathcal{M}}_0(\bar{t}) := \{x \in \mathcal{N}_{\bar{x}} : \exists \breve{x}^1 \in \mathbb{R}^m : \breve{f}(\breve{D}(\bar{t})\breve{x}^1+\breve{D}'(\bar{t})x,x,\bar{t}) = 0\}.$$

We set $\bar{\breve{x}}^1 := \bar{x}^1$ and show that, in fact, $\breve{f}(\breve{D}(\bar{t})\bar{x}^1+\breve{D}'(\bar{t})\bar{x},\bar{x},\bar{t}) = 0$ is valid. We have

$$\breve{D}(\bar{t})\bar{x}^1+\breve{D}'(\bar{t})\bar{x} = D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}^1+(D\tilde{\Pi}_1)'(\bar{t})\bar{x} = \bar{u}^1,$$

and, taking the condition (3.99) into account, which means $v^1(D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x},\bar{t},\bar{u}^1) = \bar{v}^1$, it follows that

$$\begin{aligned}
\breve{f}(\breve{D}(\bar{t})\bar{x}^1+\breve{D}'(\bar{t})\bar{x},\bar{x},\bar{t}) =& f\big((D\tilde{\Pi}_1D^-)(\bar{t})\{D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}^1\\
&+(D\tilde{\Pi}_1)'(\bar{t})\bar{x}\}+(D\tilde{\Pi}_1D^-)'(\bar{t})D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}+\bar{v}^1,\bar{x},\bar{t}\big)\\
=& f(D(\bar{t})\tilde{\Pi}_1(\bar{t})\bar{x}^1+(D\tilde{\Pi}_1)'(\bar{t})\bar{x}+\bar{v}^1,\bar{x},\bar{t})\\
=& f(\bar{u}^1+\bar{v}^1,\bar{x},\bar{t}) = f(\bar{y},\bar{x},\bar{t}) = 0.
\end{aligned}$$

Now the proof of our first assertion is complete. We turn to the second assertion. Let $x_* \in \mathcal{C}^1_D(\mathcal{I}_*, \mathbb{R}^m)$ be a solution of the DAE (3.92) with values in the definition domain of the functions h and v^1. Then it necessarily holds that

$$\begin{aligned}
\tilde{Z}(t)x_*(t) &= h(D(t)\tilde{\Pi}_1(t)x_*(t),t),\\
D(t)P_0(t)\tilde{Q}_1(t)x_*(t) &= D(t)h(D(t)\tilde{\Pi}_1(t)x_*(t),t),\\
(D(t)P_0(t)\tilde{Q}_1(t)x_*(t))' &= v^1(D(t)\tilde{\Pi}_1(t)x_*(t),t,(D(t)\tilde{\Pi}_1(t)x_*(t))') \quad \text{for } t \in \mathcal{I}_*.
\end{aligned}$$

Because of $D\tilde{\Pi}_1 x_* = D\tilde{\Pi}_1 D^- Dx_*$, the component $D\tilde{\Pi}_1 x_*$ inherits its smoothness from that of $D\tilde{\Pi}_1 D^-$ and Dx_*. Inserting the above expression into the identity

$$\begin{aligned}
0 &= f((D(t)x_*(t))',x_*(t),t) = f((D\tilde{\Pi}_1 x_*)'(t) + DP_0\tilde{Q}_1 x_*)'(t),\ x_*(t),t)\\
&= f\big((D\tilde{\Pi}_1 D^-)(t)(D\tilde{\Pi}_1 x_*)'(t) + (D\tilde{\Pi}_1 D^-)'(t)D(t)\tilde{\Pi}_1(t)x_*(t)\\
&\qquad + (DP_0\tilde{Q}_1 x_*)'(t),\ x_*(t),t\big)
\end{aligned}$$

we see that the new DAE (3.103) is satisfied.
Conversely, let $x_* \in \mathcal{C}^1_{D\tilde{\Pi}_1}(\mathcal{I}_*, \mathbb{R}^m)$ be a solution of the DAE (3.103), i.e.,

$$\begin{aligned}
0 &= \check{f}((D\tilde{\Pi}_1 x_*)'(t),x_*(t),t)\\
&= f((D\tilde{\Pi}_1 x_*)'(t) + v^1(D(t)\tilde{\Pi}_1(t)x_*(t),t,(D\tilde{\Pi}_1 x_*)'(t)),x_*(t),t).
\end{aligned}$$

The structural restrictions (3.96), (3.97) lead to

$$0 = \tilde{Z}(t)\tilde{G}_2(t)^{-1} f(0, \tilde{\Pi}_1(t)x_*(t) + \tilde{Z}(t)x_*(t),t) = g(D(t)\tilde{\Pi}_1(t)x_*(t), \tilde{Z}(t)x_*(t),t),$$

and, with regard to Lemma 3.57, we find the relation

$$D(t)P_0(t)\tilde{Q}_1(t)x_*(t) = D(t)h(D(t)\tilde{\Pi}_1(t)x_*(t),t).$$

Since h and $D\tilde{\Pi}_1 x_*$ are continuously differentiable, the component $DP_0\tilde{Q}_1 x_*$ on the left side is so, too. We derive

$$(DP_0\tilde{Q}_1 x_*)'(t) = v^1(D(t)\tilde{\Pi}_1(t)x_*(t),t,(D\tilde{\Pi}_1 x_*)'(t))$$

and insert this expression into the above identity. This makes it clear that x_* solves the DAE (3.92). □

3.8 Advanced localization of regularity: including jet variables

The class of nonlinear DAEs that are regular on their entire definition domain, which means that there is just a single maximal regularity region, comprises, for instance, the MNA-DAE and the Hessenberg form DAEs of arbitrary size. A different situation is given in Example 3.34, where the definition domain $\mathcal{D}_f \times \mathcal{I}_f$ is split into three

maximal regularity regions $\mathcal{G}_i$, $i = 1,2,3$, whose borders consist of critical points. The special DAE (3.27) is regular with tractability index 1 locally on each region $\mathcal{G}_i$; however, neither of the regions $\mathcal{G}_i$ covers the obvious constraint set, and there are solutions crossing the borders. The class of DAEs which show several maximal regularity regions represents a straightforward generalization of the DAEs having just a single maximal regularity region. However, also this notion needs a further generalization.

The regularity notion given in Section 3.3 is local with respect to the basic variables $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$. Admissible projectors $Q_0, \ldots, Q_{\mu-1}$ may depend not only on $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ but also on jet variables $x^1, \ldots, x^{\mu-1} \in \mathbb{R}^m$. The demands yielding regularity on a region $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ are meant to hold true for all $(x,t) \in \mathcal{G}$ and globally for all $x^1, \ldots, x^{\mu-1} \in \mathbb{R}^m$, as well.

There are quite simple nonlinear problems where the regularity Definition 3.28 does not apply. It is natural to advance the localization to apply to jet variables, too.

*Example 3.59 (*rank G_1 *depends on* x^1*).* Set $k = m = 2$, $n = 1$, $\alpha \in \mathbb{R}$ is a parameter, and $\beta : \mathbb{R} \to \mathbb{R}$ is a continuous function. The DAE

$$\begin{aligned} (x_1(t) + x_2(t))' - t^2 - \alpha &= 0, \\ x_1(t)(x_1(t) + x_2(t))' - \beta(t) &= 0, \end{aligned} \tag{3.104}$$

has the solutions

$$x_{*1}(t) = \frac{\beta(t)}{t^2 + \alpha}, \quad x_{*1}(t) + x_{*2}(t) = z_0 + \alpha(t - t_0) + \frac{1}{3}(t^3 - t_0^3), \quad t \in \mathcal{I}_*,$$

which satisfy the initial condition $x_{*1}(t_0) + x_{*2}(t_0) = z_0$, $z_0 \in \mathbb{R}$. The initial time point t_0 as well as the existence interval $\mathcal{I}_*$ are bound with the requirement for the expression $t^2 + \alpha$ not to have zeros.

If $\alpha > 0$, then $t_0 \in \mathbb{R}$ can be chosen arbitrarily, and $\mathcal{I}_* = \mathbb{R}$. If $\alpha < 0$, then $t_0 \neq \pm\sqrt{-\alpha}$ is allowed, but the interval $\mathcal{I}_*$ is restricted. If $\alpha = 0$, then $t_0 \neq 0$ is allowed, and it results that $\mathcal{I}_* = (0, \infty)$ for $t_0 > 0$, and $\mathcal{I}_* = (-\infty, 0)$ for $t_0 < 0$.

We put the DAE (3.104) into the general form (3.1) by

$$f(y,x,t) := \begin{bmatrix} 1 \\ x_1 \end{bmatrix} y - \begin{bmatrix} t^2 + \alpha \\ \beta(t) \end{bmatrix}, \quad d(x,t) := x_1 + x_2, \quad x \in \mathbb{R}^2,\ y \in \mathbb{R},\ t \in \mathbb{R}.$$

Assumption 3.16 is valid with $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = \mathbb{R}$, and $\ker f_y = 0$. The DAE has a properly stated leading term. The obvious constraint

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^2 : (\alpha + t^2)x_1 = \beta(t)\}$$

is well-defined and nonempty for all $t \in \mathbb{R}$ with $t^2 + \alpha \neq 0$. For $\bar{t} \in \mathbb{R}$ with $\alpha + \bar{t}^2 = 0$ and $\beta(\bar{t}) = 0$ it follows that $\mathcal{M}_0(\bar{t}) = \mathbb{R}^2$. To each fixed $t_0 \in \mathbb{R}$, $x_0 \in \mathcal{M}_0(t_0)$, with $t_0^2 + \alpha \neq 0$, there is exactly one solution x_* (given on its individual interval $\mathcal{I}_*$) passing through it. That is, the DAE (3.104) behaves as a regular index-1 DAE. Derive

$$A = \begin{bmatrix} 1 \\ x_1 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 1 \end{bmatrix}, \quad D^- = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad R = 1,$$
$$Q_0 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \quad P_0 = \Pi_0 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ x_1^1 + x_2^1 & 0 \end{bmatrix},$$
$$G_0 = \begin{bmatrix} 1 & 1 \\ x_1 & x_1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & 1 \\ x_1 + x_1^1 + x_2^1 & x_1 \end{bmatrix}, \quad \det G_1 = -(x_1^1 + x_2^1).$$

Inspecting the matrix function G_1 and their determinant we see that Definition 3.28 does not apply here, since G_1 becomes singular for $x_1^1 = -x_2^1$. Therefore we are led to extend the regularity notion for suitable sets concerning the jet variables, too. It makes sense to say that the DAE (3.104) is regular with tractability index 1 on the two open sets $\mathcal{G}_-^{[1]} := \mathcal{D}_-^{(1)} \times \mathcal{D}_f \times \mathcal{I}_f$ and $\mathcal{G}_+^{[1]} := \mathcal{D}_+^{(1)} \times \mathcal{D}_f \times \mathcal{I}_f$, with $\mathcal{D}_+^{(1)} := \{x^1 \in \mathbb{R}^2 : x_1^1 + x_2^1 > 0\}$, $\mathcal{D}_-^{(1)} := \{x^1 \in \mathbb{R}^2 : x_1^1 + x_2^1 < 0\}$.
Points belonging to the border set $\{(x^1, x, t) \in \mathbb{R}^2 \times \mathcal{D}_f \times \mathcal{I}_f : x_1^1 + x_2^1 = 0\}$ are considered to be critical ones. These points correspond to zeros of the expression $t^2 + \alpha$ in the solutions, that is, they are in fact critical.
If one linearizes the DAE (3.104) along a smooth reference function with values only in $\mathcal{G}_-^{[1]}$ or only in $\mathcal{G}_+^{[1]}$, then the resulting linear DAE (3.11) is regular with tractability index 1. In contrast, if one linearizes along a reference function x_* with $x'_{*,1}(t) + x'_{*,2}(t) = 0$, $t \in \mathcal{I}_*$, that is, along a function with values on the border between the sets $\mathcal{G}_-^{[1]}$ and $\mathcal{G}_+^{[1]}$, then the resulting linear DAE (3.11) fails to be regular on $\mathcal{I}_*$. □

Example 3.60 ([125], rank G_3 depends on x and x^1). Set $n = 2$, $k = m = 3$, $\eta \in \mathbb{R}$ is a parameter, $\mathcal{D}_f = \mathbb{R}^3$, $\mathcal{I}_f = \mathbb{R}$, $q \in C^1(\mathbb{R}, \mathbb{R}^3)$, $q_3 \in C^2(\mathbb{R}, \mathbb{R})$. The DAE

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}}_{A} \Big(\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D} x(t) \Big)' + \begin{bmatrix} x_1(t) - t \\ x_2(t)(x_1(t) + \eta) - 1 \\ x_2(t)(1 - \frac{1}{2}x_2(t)) + x_3(t) \end{bmatrix} - q(t) = 0 \tag{3.105}$$

has a properly stated leading term and satisfies Assumption 3.16. The obvious constraint is

$$\mathcal{M}_0(t) = \Big\{ x \in \mathbb{R}^3 : x_2\big(1 - \frac{1}{2}x_2\big) + x_3 = q_3(t) \Big\}.$$

The DAE has the only solutions

$$\begin{aligned} x_{*1}(t) &= -x'_{*2}(t) + t + q_1(t), \\ x_{*2}(t) &= \frac{1 + q_2(t) - q_3'(t)}{t + \eta + q_1(t)}, \\ x_{*3}(t) &= -x_{*2}(t)\Big(1 - \frac{1}{2}x_{*2}(t)\Big) + q_3(t), \quad t \in \mathcal{I}_*, \end{aligned}$$

given on intervals $\mathcal{I}_*$ such that $t+\eta+q_1(t)\neq 0$, $t\in\mathcal{I}_*$. The solutions satisfy the relations

$$x_{*1}(t)+\eta+x'_{*2}(t)=t+\eta+q_1(t),\quad x_{*2}(t)(t+\eta+q_1(t))=1+q_2(t)-q'_3(t),\quad t\in\mathcal{I}_*.$$

We construct an admissible matrix function sequence starting with

$$G_0=\begin{bmatrix}0&1&0\\0&1&1\\0&0&0\end{bmatrix},\quad Q_0=\begin{bmatrix}1&0&0\\0&0&0\\0&0&0\end{bmatrix},\quad B_0=\begin{bmatrix}1&0&0\\x_2&x_1+\eta&0\\0&1-x_2&1\end{bmatrix},$$
$$D^-=\begin{bmatrix}0&0\\1&0\\0&1\end{bmatrix},\quad P_0=\Pi_0=\begin{bmatrix}0&0&0\\0&1&0\\0&0&1\end{bmatrix},\quad G_1=\begin{bmatrix}1&1&0\\x_2&1&1\\0&0&0\end{bmatrix}.$$

Choose further

$$Q_1:=\begin{bmatrix}0&\alpha&\beta\\0&-\alpha&-\beta\\0&(1-x_2)\alpha&(1-x_2)\beta\end{bmatrix},\quad \alpha:=(1-x_2)\beta-1,$$

$\beta\in\mathcal{C}^1(\mathbb{R}^3\times\mathbb{R},\mathbb{R})$ an arbitrary function, such that Q_0,Q_1 are admissible projectors. It holds that $Q_1Q_0=0$, and

$$D\Pi_1D^-=\begin{bmatrix}1+\alpha&\beta\\-(1-x_2)\alpha&1-(1-x_2)\beta\end{bmatrix}.$$

It follows that (cf. [125])

$$G_2=\begin{bmatrix}1&1-\alpha\beta x_2^1&-\beta^2x_2^1\\x_2&1-\alpha(x_1+\eta+x_2^1)-\alpha\beta x_2x_2^1&1-\beta(x_1+\eta+x_2^1)-\beta^2x_2x_2^1\\0&0&0\end{bmatrix}.$$

Since, with $\mathcal{W}_1:=\begin{bmatrix}0&0&0\\0&0&0\\0&0&1\end{bmatrix}$, it holds that $\mathcal{W}_1B_0Q_1=0$, we conclude that $\operatorname{im}G_2=\operatorname{im}G_1$, and hence $\mathcal{W}_2=\mathcal{W}_1$,

$$S_2=\ker\mathcal{W}_2B_0=\{z\in\mathbb{R}^3:(1-x_2)z_2+z_3=0\}.$$

Consider $z\in S_2\cap N_2$, that is

$$\begin{aligned}(1-x_2)z_2+z_3&=0,\\ z_1+(1-\alpha\beta x_2^1)z_2-\beta^2x_2^1z_3&=0,\\ x_2z_1+(1-\alpha(x_1+\eta+x_2^1)-\alpha\beta x_2x_2^1)z_2+(1-\beta(x_1+\eta+x_2^1)-\beta^2x_2x_2^1)z_3&=0,\end{aligned}$$

or, equivalently, the system

$$
\begin{aligned}
z_3 &= -(1-x_2)z_2,\\
z_1 + (1+\beta x_2^1)z_2 &= 0,\\
x_2 z_1 + (1+(x_1+\eta+x_2^1)-(1-x_2)+\beta x_2 x_2^1)z_2 &= 0,
\end{aligned}
$$

that is

$$
\begin{aligned}
z_3 &= -(1-x_2)z_2,\\
z_1 &= -(1+\beta x_2^1)z_2,\\
(x_1+\eta+x_2^1)z_2 &= 0.
\end{aligned}
$$

Because of $N_0+N_1 \subseteq S_2$, the relation $S_2 \cap N_2 = \{0\}$ implies $(N_0+N_1)\cap N_2 = \{0\}$. At the same time, without computing a particular projector Q_2, we know the matrix function G_3 (cf. Lemma A.9) remains nonsingular, supposing $x_1+\eta+x_2^1 \neq 0$. In consequence, the DAE (3.105) is regular with tractability index 3 on the open sets

$$\mathcal{G}_+^{[2]} = \{(x^2,x^1,x,t) \in \mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R} : x_1+\eta+x_2^1 > 0\}$$

and

$$\mathcal{G}_-^{[2]} = \{(x^2,x^1,x,t) \in \mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R}^3\times\mathbb{R} : x_1+\eta+x_2^1 < 0\}.$$

If one takes a smooth reference function x_* with values in just one of these sets, then the linearized DAE (3.11) is regular with index 3. In contrast, for a reference function x_* that satisfies $x_{*,1}(t)+\eta+x'_{*,2}(t) = 0$, $t \in \mathcal{I}_*$, the resulting linear DAE (3.11) fails to be regular. All corresponding matrix functions G_{*i} are singular. Furthermore, letting

$$q_1(t) = -t^2,\ q_2(t) = 0,\ q_3(t) = 0, \quad \text{for } t \in \mathcal{I}_* = [0,2), \qquad \text{and } \eta = 2,$$

the function x_* given by

$$
\begin{aligned}
x_{*1}(t) &= -x'_{*2}(t) + t - t^2,\\
x_{*2}(t) &= \frac{1}{t+2-t^2},\\
x_{*3}(t) &= -x_{*2}(t) + \frac{1}{2}x_{*2}(t)^2,\ t \in \mathcal{I}_*,
\end{aligned}
$$

is a solution of the original DAE and has values $(x_*''(t),x_*'(t),x_*(t),t) \in \mathcal{G}_+^{[2]}$. However, if t approaches 2, $x_*(t)$ grows unboundedly, which indicates the singularity at the border between $\mathcal{G}_+^{[2]}$ and $\mathcal{G}_-^{[2]}$. □

We call those open connected sets $\mathcal{G}_-^{[1]}$ and $\mathcal{G}_+^{[1]}$ *regularity regions*, too. In the previous two examples, linearizations along reference functions with values belonging to the border of such a maximal regularity region fail to be a regular DAE. In different cases it may also happen that regularity is maintained, but the index changes. We refer to Example 3.34 which shows three regularity regions; on one border the lin-

earized DAEs are regular (with changed index), while they fail to be regular on the other border. As before, to allow for small perturbations without losing regularity we tie regularity to open sets of the corresponding spaces. The following two definitions generalize the Definition 3.21 of admissible matrix function sequences and the Definition 3.28 of regular DAEs.

By construction (cf. Section 3.2), for $i \geq 1$, the matrix function G_i depends on the variables x, t and $x^1, \ldots, x^i$, and so does the nullspace projector function Q_i. On the next level, the additional variable x^{i+1} comes in, and G_{i+1} depends on the variables $x, t, x^1, \ldots, x^{i+1}$, and so on. Of course, it may happen in more special cases, e.g., the last examples, that these matrix functions do not actually vary with all these variables.

For an easier description, if we deal with the level κ, we now suppose that all matrix functions of the lower levels depend on all jet variables up to x^κ. The lower level matrix functions are constant functions with respect to the jet variables coming in later.

We first extend the previous Definition 3.21 of admissible projector functions.

Definition 3.61. Let the DAE (3.1) satisfy the basic Assumption 3.16. Let $\kappa \in \mathbb{N}$ be the given level and let the sets $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ and $\mathcal{G}^{[\kappa]} \subseteq \mathbb{R}^{m\kappa} \times \mathcal{D}_f \times \mathcal{I}_f$ be open connected.

Let the projector function Q_0 onto $\ker D$ be continuous on $\mathcal{G}^{[\kappa]}$, $P_0 = I - Q_0$, and D^- be determined there by (3.18).

We call the sequence $G_0, \ldots, G_\kappa$ an *admissible matrix function sequence* associated with the DAE on the set $\mathcal{G}^{[\kappa]}$, if it is built by the rule

Set $G_0 := AD$, $B_0 := B$, $N_0 := \ker G_0$.
For $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1}Q_{i-1},$$
$$B_i := B_{i-1}P_{i-1} - G_i D^-(D\Pi_i D^-)' D\Pi_{i-1}$$
$$N_i := \ker G_i, \quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i,$$

fix a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,
choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,
set $P_i := I - Q_i$, $\Pi_i := \Pi_{i-1}P_i$

and, additionally,

(a) the matrix function G_i has constant rank r_i on $\mathcal{G}^{[\kappa]}$, $i = 0, \ldots, \kappa$,

(b) the intersection $\widehat{N}_i$ has constant dimension $u_i := \dim \widehat{N}_i$ there,

(c) the product function Π_i is continuous and $D\Pi_i D^-$ is continuously differentiable on $\mathcal{G}^{[\kappa]}$, $i = 0, \ldots, \kappa$.

The projector functions $Q_0, \ldots, Q_\kappa$ in an admissible matrix function sequence are said to be *admissible* themselves.

The numbers $r_0 := \operatorname{rank} G_0, \ldots, r_\kappa := \operatorname{rank} G_\kappa$ and $u_1, \ldots, u_\kappa$ are named *characteristic values* of the DAE on $\mathcal{G}^\kappa$.
The matrix functions $G_0, \ldots, G_\kappa$ are said to be *admissible on* $\mathcal{G}$, if they are admissible on $\mathcal{G}^{[\kappa]} = \mathbb{R}^{m\kappa} \times \mathcal{G}$.

Having this more general notion of admissible matrix function sequences, which maintains the algebraic properties figured out in Section 3.2, we are ready to extend also the regularity notion correspondingly.

Definition 3.62. Let the DAE (3.1) satisfy Assumption 3.16, $k = m$, and let the sets $\mathcal{G}^{[\mu]} \subseteq \mathbb{R}^{m\mu} \times \mathcal{D}_f \times \mathcal{I}_f$ and $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open connected.

(1) If $r_0 = m$, then equation (3.1) is said to be regular with index 0.
(2) The DAE is said to be *regular with tractability index* $\mu \in \mathbb{N}$ *on* $\mathcal{G}^{[\mu]}$, if there is a matrix function sequence admissible on $\mathcal{G}^{[\mu]}$ such that $r_{\mu-1} < r_\mu = m$. Then $\mathcal{G}^{[\mu]}$ is named a *regularity region* of the DAE, with characteristic values $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$ and tractability index μ.
(3) A jet $(x^\mu, \ldots, x^1, x, t) \in \mathbb{R}^{m\mu} \times \mathcal{D}_f \times \mathcal{I}_f$ is named a *regular index* μ *jet*, if there is a neighborhood in $\mathbb{R}^{m\mu} \times \mathbb{R}^m \times \mathbb{R}$ which is a regularity region with tractability index μ.
(4) If the DAE is regular with tractability index μ on $\mathcal{G}^{[\mu]} = \mathbb{R}^{m\mu} \times \mathcal{G}$, then we say simply the *DAE is regular on* $\mathcal{G}$, and $\mathcal{G}$ is called a *regularity region*.
(5) The point $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ is called a *regular point*, if it has a neighborhood $\mathfrak{N}_{(x,t)} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ such that $\mathbb{R}^{m\mu} \times \mathfrak{N}_{(x,t)}$ is a regularity region.

Definition 3.62 is consistent with the previous Definition 3.28. Examples 3.59 and 3.60 provide actual regularity regions with index 1 and index 3, respectively.
By construction, if a nonlinear DAE (3.1) is regular with tractability index μ on $\mathcal{G}^{[\mu]}$, then all linearizations along smooth reference functions x_* with values in $\mathcal{G}^{[\mu]}$, i.e.,

$$(x_*^{(\mu)}(t), \ldots, x_*'(t), x_*(t), t) \in \mathcal{G}^{[\mu]}, \quad t \in \mathcal{I}_*,$$

are regular with uniform tractability index μ, and uniform characteristics $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$. The linearizations may behave differently, if the reference function crosses critical points. A systematic study of possible coherences needs future research.

Next we reconsider the local solvability assertion in Subsection 3.7.1 and adapt Theorem 3.53 to the advanced localization.

Theorem 3.63. *Let the DAE (3.1) satisfy Assumption 3.16 and be regular with tractability index 1 on the open set* $\mathcal{G}^{[1]} \subset \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f$. *Let* d *have the additional continuous partial derivatives* d_{xx}, d_{xt}.
Then, for each $(\bar{x}^1, \bar{x}, \bar{t}) \in \mathcal{G}^{[1]}$, $\bar{x} \in \mathcal{M}_0(\bar{t})$, *there exists exactly one solution* $x_* \in \mathcal{C}(\mathcal{I}_*, \mathbb{R}^m)$ *such that* $x_*(\bar{t}) = \bar{x}$.

Proof. In contrast to Theorem 3.53, now a value $\bar{x}^1$ is already given, and $D(\bar{x},\bar{t})\bar{x}^1 = \bar{y} - d_t(\bar{x},\bar{t})$.
The matrix functions have the following special property:

For each arbitrary $\bar{\bar{x}}^1 \in \mathbb{R}^m$ with $D(\bar{x},\bar{t})\bar{x}^1 = D(\bar{x},\bar{t})\bar{\bar{x}}^1$, it follows that $G_1(\bar{x}^1,\bar{x},\bar{t}) = G_1(\bar{\bar{x}}^1,\bar{x},\bar{t})$. This allows us to select the $\bar{\bar{x}}^1$ in such a way that $Q_0(\bar{x},\bar{t})\bar{\bar{x}}^1 = 0$. Then we apply the former proof starting with $(\bar{\bar{x}}^1,\bar{x},\bar{t})$ instead of $(\bar{x}^1,\bar{x},\bar{t})$. □

We emphasize that here the solutions are not necessarily continuously differentiable. Moreover, even if they are, the relation $x_*'(\bar{t}) = \bar{x}^1$ cannot be expected to be valid, as the following example shows.

Example 3.64 (Inconsistency of $\bar{x}^1$). Set $k = m = 3, n = 2$, and turn to the special DAE of the form (3.1) given by

$$f(y,x,t) = \begin{bmatrix} y_1 - x_3 \\ y_2 - x_1 - t \\ x_3 y_2 - x_1 x_3 + x_2 \end{bmatrix}, \quad d(x,t) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad y \in \mathbb{R}^2, x \in \mathbb{R}^3, t \in \mathbb{R}.$$

This DAE has a properly stated leading term and satisfies Assumption 3.16. In more detail, it reads

$$\begin{aligned} x_1'(t) - x_3(t) &= 0, \\ x_2'(t) - x_1(t) - t &= 0, \\ x_3(t)x_2'(t) - x_1(t)x_3(t) + x_2(t) &= 0, \end{aligned} \tag{3.106}$$

and in compact formulation,

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & x_3(t) & 1 \end{bmatrix} \begin{bmatrix} x_1'(t) - x_3(t) \\ x_2'(t) - x_1(t) \\ x_2(t) \end{bmatrix} - \begin{bmatrix} 0 \\ t \\ 0 \end{bmatrix} = 0.$$

The last version suggests that this would be a regular index-3 DAE. However, this is wrong. Actually the DAE (3.106) is regular with tractability index 1 on the two open sets

$$\mathcal{G}_+^{[1]} = \{(x^1,x,t) \in \mathbb{R}^{m+m+11} : x_2^1 - x_1 > 0\}, \quad \mathcal{G}_-^{[1]} = \{(x^1,x,t) \in \mathbb{R}^{m+m+1} : x_2^1 - x_1 < 0\}.$$

To show this we provide the matrix function sequence

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & x_3 & 0 \end{bmatrix}, B_0 = \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ -x_3 & 1 & x_2^1 - x_1 \end{bmatrix}, Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, G_1 = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & x_3 & x_2^1 - x_1 \end{bmatrix}.$$

The matrix $G_1(x^1,x,t)$ is nonsingular, exactly if $x_2^1 \neq x_1$, which proves regularity with index 1 on the open sets $\mathcal{G}_-^{[1]}$ and $\mathcal{G}_+^{[1]}$.
The obvious constraint set of the DAE (3.106) is given as

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

For each fixed $\bar{t} \in \mathbb{R}$, $\bar{x} \in \mathcal{M}_0(\bar{t})$ such that $\bar{t} \neq 0$, there is a unique solution passing through it. In fact, the first two components of this solution are given by means of

the explicit IVP

$$\begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = \frac{1}{t} \begin{bmatrix} 0 & -1 \\ t & 0 \end{bmatrix} \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ t \end{bmatrix}, \quad x_1(\bar{t}) = \bar{x}_1,\ x_2(\bar{t}) = \bar{x}_2, \tag{3.107}$$

while the third component is $x_3(t) = -\frac{1}{t} x_2(t)$.
This is exactly what we expect for a regular index-1 DAE. The condition $\bar{t} \neq 0$ ensures that this solution proceeds in $\mathcal{G}_-^{[1]}$ or $\mathcal{G}_+^{[1]}$. If $\bar{t} > 0$, the condition $f(D\bar{x}^1, \bar{x}, \bar{t}) = 0$ defines the first two components of $\bar{x}^1$ uniquely, and $\bar{x}_2^1 - \bar{x}_1 = \bar{t} > 0$ holds true. We get $(\bar{x}^1, \bar{x}, \bar{t}) \in \mathcal{G}_+^{[1]}$, for all $\bar{x}_3^1 \in \mathbb{R}$.

Each linearization (3.11) along a smooth reference function with values in just one of the regularity regions is regular with tractability index 1, however, linearization along a function lying on the border of these regions yields a regular DAE with tractability index 3. Namely, let x_* denote an arbitrary smooth reference function with values on the border set, i.e., $x_{*2}'(t) - x_{*1}(t) = 0$, $t \in \mathcal{I}_*$. The DAE (3.11) linearized along this function has the coefficients

$$A_* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & x_{*3} \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B_* = \begin{bmatrix} 0 & 0 & -1 \\ -1 & 0 & 0 \\ -x_{*3} & 1 & 0 \end{bmatrix}.$$

The following matrix function sequence for this DAE shows singular matrix functions G_{*0}, G_{*1}, and G_{*2}, but ends up with a nonsingular G_{*3}:

$$G_{*0} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & x_{*3} & 0 \end{bmatrix}, \quad Q_{*0} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad G_{*1} = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 1 & 0 \\ 0 & x_{*3} & 0 \end{bmatrix}, \quad Q_{*1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix},$$

$$G_{*2} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ -x_{*3} & x_{*3} & 0 \end{bmatrix}, \ Q_{*2} = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 0 \end{bmatrix}, \ G_{*3} = \begin{bmatrix} 1 & 0 & 1 \\ -1 & 1 & 0 \\ -x_{*3} & 1+x_{*3} & 0 \end{bmatrix}, \ \det G_{*3} = -1,$$

which proves in fact that the linearized DAE is regular with tractability index 3. For instance, the special reference functions

$$x_*(t) = \begin{bmatrix} a \\ at + b \\ -\frac{1}{t}(at + b) \end{bmatrix},$$

with certain parameters $a, b \in \mathbb{R}$, have values $x_*(t) \in \mathcal{M}_0(t)$, but, because of $x_{*2}'(t) - x_{*1}(t) = 0$, they are located on the border between the regularity regions.

Now we put $\bar{t} = 1$, $\bar{x}_1 = 1$, $\bar{x}_2 = 1$, $\bar{x}_3 = -1$, $\bar{x}_1^1 = -1$, $\bar{x}_2^1 = 2$, $\bar{x}_3^1 = 7$.
The condition $\bar{x} \in \mathcal{M}_0(\bar{t})$ is fulfilled, since $\bar{x}_2 + \bar{x}_3 = 0$. Further, $(\bar{x}^1, \bar{x}, \bar{t})$ belongs to $\mathcal{G}_+^{[1]}$. Taking a look at the corresponding solution of the explicit IVP (3.107), one recognizes that $x_{*,1}'(1) = \bar{x}_1^1$, $x_{*,2}'(1) = \bar{x}_2^1$. The third solution component is

$$x_{*,3}(t) = -\frac{1}{t}x_{*,2}(t), \quad \text{thus} \quad x'_{*,3}(t) = +\frac{1}{t^2}x_{*,2}(t) - \frac{1}{t}x'_{*,2}(t).$$

Finally it follows that $x'_{*,3}(1) = -1 \neq \bar{x}^1_3$, which proves the inconsistency of $\bar{x}^1_3$. Notice further that we could choose an arbitrary $x'_{*,3}(1) \neq \bar{x}^1_3$, and we would come to the same conclusion. □

Why does it happen that Theorem 3.63 works, though $\bar{x}^1$ may fail to be consistent? In general, the matrix function sequence is determined from the coefficients

$$\begin{aligned} A(x^1,x,t) &:= f_y(D(x,t)x^1 + d_t(x,t),x,t) \\ B(x^1,x,t) &:= f_x(D(x,t)x^1 + d_t(x,t),x,t) \\ D(x,t) &:= d_x(x,t), \end{aligned}$$

and x^1 is involved exclusively via the term $D(x,t)x^1$. Therefore, it follows that

$$G_1(x^1,x,t) = G_1(x^1+z,x,t) \quad \text{for all} \quad z \in \ker D(x,t),$$

and $(\bar{x}^1,\bar{x},\bar{t}) \in \mathcal{G}^{[1]}$ implies $(\bar{x}^1+z,\bar{x},\bar{t}) \in \mathcal{G}^{[1]}$ for all $z \in \ker D(x,t)$. This explains why the consistency of x^1 and $x'_*(\bar{t})$ cannot be expected even if the solution is smooth.

3.9 Operator settings

In the present section we restrict our interest to IVPs in DAEs comprising an equal number of equations and unknowns, $k = m$. The DAE is now specified as

$$f((D(t)x(t))',x(t),t) = 0. \tag{3.108}$$

It is assumed to satisfy Assumption 3.16. In particular, the derivative is properly involved and D is continuously differentiable.
Let $\mathcal{I} \subseteq \mathcal{I}_f$ be a *compact* interval. Let $\mathcal{D}_F$ denote an open set in the function space $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ such that $x \in \mathcal{D}_F$ implies $x(t) \in \mathcal{D}_f, t \in \mathcal{I}$. Define the operator F acting on $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ by

$$(Fx)(t) := f((D(t)x(t))',x(t),t), \quad t \in \mathcal{I}, \quad x \in \mathcal{D}_F, \tag{3.109}$$

so that the range $\operatorname{im} F$ resides in the continuous function space,

$$F : \mathcal{D}_F \subseteq \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m).$$

Since D is continuously differentiable, the inclusion

$$\mathcal{C}^\nu(\mathcal{I},\mathbb{R}^m) \subseteq \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$$

is valid for each $\nu \in \mathbb{N}$. Equipped with the natural norm

$$\|x\|_{\mathcal{C}_D^1} := \|x\|_\infty + \|(Dx)'\|_\infty, \quad x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m),$$

the linear function space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ becomes a Banach space and the DAE (3.108) is represented as the operator equation

$$Fx = 0 \tag{3.110}$$

in a Banach space setting. At this point it is worth emphasizing that the operator equation (3.110) reflects the classical view on DAEs: the solutions belong to $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ and satisfy the DAE pointwise for all $t \in \mathcal{I}$.
For each arbitrary $x_* \in \mathcal{D}_F$ we continue to denote

$$A_*(t) := f_y((D(t)x_*(t))', x_*(t), t), \quad B_*(t) := f_x((D(t)x_*(t))', x_*(t), t), \quad t \in \mathcal{I}.$$

Next, for arbitrarily fixed $x_* \in \mathcal{D}_F$ and any $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ the directional derivative

$$F_x(x_*)x := \lim_{\tau \to 0} \frac{1}{\tau}(F(x_* + \tau x) - F(x_*)) = A_*(Dx)' + B_* x$$

is well defined. In fact, the resulting map $F_x(x_*) : \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m)$ is linear and bounded. Moreover, $F_x(\bar{x})$ varies continuously with respect to $\bar{x}$. This means that the linear bounded map

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

is the Fréchet derivative of F at x_*. The linear operator equation $F_x(x_*)x = q$ stands now for the *linearization* of the original DAE, that is, for the linear DAE

$$A_*(Dx)' + B_* x = q. \tag{3.111}$$

While in the context of differential equations, one usually speaks about linearization *along the function* x_*, in the context of operator equations one rather applies the wording *linearization at* x_*.
The linearizations (3.111) inherit from the nonlinear DAE (3.108) the properly stated leading term. The function space $\mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ accommodates also the solutions of the linearized DAE.
Based on Theorem 3.33, we state that the DAE operator F is regular with characteristics $r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$, exactly if all linearizations $F_x(x_*)x = q$, with $x_* \in \mathcal{D}_F \cap \mathcal{C}^m(\mathcal{I}, \mathbb{R}^m)$ are so.

We complete the DAE (3.108) by the initial condition

$$Cx(t_0) = z_0, \tag{3.112}$$

with fixed $t_0 \in \mathcal{I}$ and a matrix $C \in L(\mathbb{R}^m, \mathbb{R}^d)$ to be specified later. The composed operator associated with the IVP (3.108), (3.112),

$$\mathcal{F} : \mathcal{D}_F \subseteq \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d, \quad \mathcal{F}x := (Fx, Cx(t_0) - z_0), \quad x \in \mathcal{D}_F,$$

is as smooth as F. The equation $\mathcal{F}x = 0$ represents the IVP (3.108), (3.112), whereas the equation $\mathcal{F}x = (q,\delta)$ is the operator setting of the perturbed IVP

$$f((D(t)x(t))',x(t),t) = q(t), \quad t \in \mathcal{I}, \quad Cx(t_0) - z_0 = \delta. \tag{3.113}$$

3.9.1 Linear case

The linear case, if $f(y,x,t) = A(t)y + B(t)x$, is of particular interest, and we introduce the extra symbol L for the linear mapping given by

$$Lx := A(Dx)' + Bx, \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m). \tag{3.114}$$

The linear operator equation $Lx = q$ now stands for the linear DAE

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \quad t \in \mathcal{I}. \tag{3.115}$$

The linear operator L which maps $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ into $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$ is bounded, since the inequality

$$\|Lx\|_\infty \le \|A\|_\infty \|(Dx)'\|_\infty + \|B\|_\infty \|x\|_\infty \le \max\{\|A\|_\infty, \|B\|_\infty\} \|x\|_{\mathcal{C}^1_D}$$

holds true for all $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$.
We complete the DAE (3.115) with the initial condition (3.112) and compose the map

$$\mathcal{L} : \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d, \quad \mathcal{L}x := (Lx,\, Cx(t_0)), \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m),$$

so that the operator equation $\mathcal{L}x = (q, z_0)$ represents the IVP (3.115), (3.112). The operator $\mathcal{L}$ is bounded simultaneously with L.

For operators acting in Banach spaces, the closure or nonclosure of the range is a very important feature. To apply, e.g., Fredholm theory and generalized inverses one needs to have closed range operators. Therefore, to know the precise range of the linear DAE operator L would be helpful. We take a look at a simple special case and figure out the range.

Example 3.65 (Nonclosed range). The operator L given by

$$Lx = \underbrace{\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}}_{A} \Big(\underbrace{\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}}_{D} x \Big)' + x = \begin{bmatrix} x_2' + x_1 \\ x_2 \end{bmatrix},$$

$x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^2) = \{x \in \mathcal{C}(\mathcal{I},\mathbb{R}^2) : x_2 \in \mathcal{C}^1(\mathcal{I},\mathbb{R})\}$ has the range $\operatorname{im} L = \mathcal{C}(\mathcal{I},\mathbb{R}) \times \mathcal{C}^1(\mathcal{I},\mathbb{R})$, which is a nonclosed subset in $\mathcal{C}(\mathcal{I},\mathbb{R}^2)$. Note that the equation $Lx = q$ represents a regular index-2 DAE.

The basic assumption on the DAE coefficients to be just continuous is mild. If necessary, certain additional smoothness of the coefficients is required to obtain regularity and solvability results. One needs the technical machinery of Chapter 2 to describe the requirements in detail. In this section, we do not give a rigorous description of the smoothness demands concerning the coefficients of the DAE, but instead we use the vague formulation *sufficiently smooth*. However we are precise in view of the right-hand sides. The following theorem is a consequence of Proposition 2.58 and Theorem 2.59.

Theorem 3.66. *Let the DAE (3.115) be regular with tractability index $\mu \in \mathbb{N}$ and characteristic values $0 \le r_0 \le \cdots \le r_{\mu-1} < r_\mu = m$ on the compact interval $\mathcal{I}$, $d := m - \sum_{j=1}^{\mu}(m - r_{j-1})$, and let the data of the DAE be sufficiently smooth.*
Let the matrix C which forms the initial condition (3.112) satisfy $\ker C = \ker \Pi_{\mu-1}(t_0)$. Then the following assertions are true:

(1) *The map L has a d-dimensional nullspace, and the map $\mathcal{L}$ is injective.*
(2) *If $\mu = 1$, then L is surjective and $\mathcal{L}$ is a bijection.*
(3) *If $\mu \ge 2$, then the ranges $\operatorname{im} L$ and $\operatorname{im} \mathcal{L}$ are nonclosed proper subsets in $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$, respectively $\mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$.*
(4) *The inverse map $\mathcal{L}^{-1}$ is bounded for $\mu = 1$, but unbounded for $\mu > 1$.*
(5) *For every $z_0 \in \mathbb{R}^d$ and $q \in \mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m)$, the IVP (3.115), (3.112) is uniquely solvable, and there is a constant K such that the inequality*

$$\|x\|_{\mathcal{C}^1_D} \le K\Big(|z_0| + \|q\|_\infty + \sum_{j=1}^{\mu-1} \|q^{(j)}\|_\infty\Big)$$

is valid for all these solutions.

Remember that, for a linear operator $\mathfrak{L} : X \to Y$ acting in the Banach spaces X, Y, the equation $\mathfrak{L}x = y$ is said to be a *well-posed problem in the sense of Hadamard*, if $\mathfrak{L}$ is bijective and there is a continuous inverse $\mathfrak{L}^{-1}$. Otherwise this linear equation is called an *ill-posed problem*. If the range of the operator $\mathfrak{L}$ is a nonclosed subset in Y, then the linear equation is said to be *essentially ill-posed in Tikhonov's sense*.

Owing to Theorem 3.66, the IVP (3.115), (3.112) is a well-posed problem solely for $\mu \le 1$, but otherwise this IVP is ill-posed. The typical solution behavior of ill-posed problems can be observed in higher index DAEs: small perturbations of the right-hand side yield large changes of the solution. Already the simple constant coefficient DAE in Example 1.5 gives an impression of this ill-posedness. We take a further look to this DAE.

Example 3.67 (A simple index-4 DAE). The operator L associated to the DAE

$$\begin{bmatrix} 1&0&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 0&0&0&0&0 \end{bmatrix} \left(\begin{bmatrix} 1&0&0&0&0\\ 0&0&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix} x(t) \right)' + \begin{bmatrix} -\alpha&-1&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix} x(t) = q(t),$$

is given on the function space

$$\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^5) = \{x \in \mathcal{C}(\mathcal{I},\mathbb{R}^5) : x_1, x_3, x_4, x_5 \in \mathcal{C}^1(\mathcal{I},\mathbb{R})\}$$

and its image is

$$\begin{aligned} \operatorname{im} L = \big\{ q \in \mathcal{C}(\mathcal{I},\mathbb{R}^5) : q_5 \in \mathcal{C}^1(\mathcal{I},\mathbb{R}), q_4 - q_5' \in \mathcal{C}^1(\mathcal{I},\mathbb{R}),\\ q_3 - (q_4 - q_5')' \in \mathcal{C}^1(\mathcal{I},\mathbb{R}) \big\} =: \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^5). \end{aligned}$$

Namely, it is easily checked that for each $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^5)$ it follows that $Lx \in \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^m)$. Conversely, $q \in \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^m)$ appears to be an admissible excitation such that the equation $Lx = q$ is solvable.
Obviously the inclusion

$$\mathcal{C}^3(\mathcal{I},\mathbb{R}^5) \subset \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^5) \subset \mathcal{C}(\mathcal{I},\mathbb{R}^5)$$

is valid. Introducing the norm

$$\|q\|_{ind\,4} := \|q\|_\infty + \|q_5'\|_\infty + \|(q_4 - q_5')'\|_\infty + \|(q_3 - (q_4 - q_5')')'\|_\infty,$$

on the linear function space $C^{ind\,4}(\mathcal{I},\mathbb{R}^5)$ we obtain a further Banach space. Moreover, owing to the inequality

$$\|Lx\|_{ind\,4} = \|Lx\|_\infty + \|x_5'\|_\infty + \|x_4'\|_\infty + \|x_3'\|_\infty \le K_L \|x\|_{\mathcal{C}^1_D}, \quad x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^5),$$

the operator L is bounded and surjective in the new setting

$$L : \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^5) \to \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^5),$$

which implies that the respective operator corresponding to the IVP,

$$\mathcal{L} = \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^5) \to \mathcal{C}^{ind\,4}(\mathcal{I},\mathbb{R}^5) \times \operatorname{im} C,$$

is bounded and bijective. In turn, in this setting, the inverse $\mathcal{L}^{-1}$ is continuous, and hence the IVP is well-posed. However, we keep in mind the actual enormous error amplification shown by the figures of Example 1.5, a fact that is completely independent of the mathematical setting.

□

Confronted with the nonclosed range of the operator L in the original setting, we are led to consider the map L in the new advanced setting, namely in the spaces $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ and $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$, where $\mathcal{C}^{ind\,\mu}(\mathcal{I},\mathbb{R}^m)$ is defined to be the linear space

$\operatorname{im} L$ equipped with a natural stronger norm $\|.\|_{ind\,\mu}$ such that one creates a Banach space. L is surjective in this advanced setting. Provided the inequality

$$\|Lx\|_{ind\,\mu} \le K_L \|x\|_{\mathcal{C}_D^1}, \quad x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m), \tag{3.116}$$

is valid with a constant K_L, the operator L and the accompanying IVP operator $\mathcal{L}$ are bounded in the advanced setting. Then, as a bounded map acting bijectively on Banach spaces, $\mathcal{L}$ has a bounded inverse.
In Example 3.65, it holds that $\|Lx\|_{ind\,2} = \|x\|_{\mathcal{C}_D^1}$, and hence the operator L is continuous in the advanced setting.
In Example 1.5 the inequality (3.116) is also given, but in the general case, the procedure to provide a suitable stronger norm as well as to check whether L becomes bounded is somewhat cumbersome. The advanced setting, both the function space and its norm, depends strongly on the special DAE. Nevertheless it seems to work. However, we do not advance far in this direction and restrict our further interest to the index-2 case. We think that although there is a couple of interesting perturbation results, this road has rather a dead end.

Proposition 3.68. *Let the DAE (3.115) be fine with tractability index 2, and let Q_0, Q_1 denote completely decoupling projector functions. Then, the operator L has the range*

$$\operatorname{im} L = \{q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^m) : D\Pi_0 Q_1 G_2^{-1} q \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)\}.$$

The linear space $\operatorname{im} L$ *equipped with the norm*

$$\|q\|_{ind\,2} := \|q\|_\infty + \|(D\Pi_0 Q_1 G_2^{-1} q)'\|_\infty, \quad q \in \operatorname{im} L,$$

yields the Banach space $\mathcal{C}^{ind\,2}(\mathcal{I}, \mathbb{R}^m)$, and $L : \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}^{ind\,2}(\mathcal{I}, \mathbb{R}^m)$ is bounded.

Proof. The inclusion $\mathcal{C}^1_{D\Pi_0 Q_1 G_2^{-1}}(\mathcal{I}, \mathbb{R}^m) \subseteq \operatorname{im} L$ follows from the solvability statements in Section 2.6. We verify the reverse implication. Consider an arbitrary $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ and the resulting continuous $q := Lx = A(Dx)' + Bx$. Compute

$$D\Pi_0 Q_1 G_2^{-1} q = D\Pi_0 Q_1 G_2^{-1} Bx = D\Pi_0 Q_1 G_2^{-1} B_1 Q_1 x = D\Pi_0 Q_1 x = D\Pi_0 Q_1 D^- Dx$$

which shows $D\Pi_0 Q_1 G_2^{-1} q$ to be continuously differentiable together with Dx and $D\Pi_0 Q_1 D^-$, and hence the assertion concerning $\operatorname{im} L$ is valid.
By standard arguments, one proves the function space $\mathcal{C}^{ind\,2}(\mathcal{I}, \mathbb{R}^m)$ to be complete. Furthermore, because of

$$\begin{aligned} \left\|(D\Pi_0 Q_1 G_2^{-1} Lx)'\right\|_\infty &= \left\|(D\Pi_0 Q_1 D^- Dx)'\right\|_\infty \\ &\le \max\{\|(D\Pi_0 Q_1 D^-)' D\|_\infty, \|D\Pi_0 Q_1 D^-\|_\infty\} \|x\|_{\mathcal{C}_D^1}, \end{aligned}$$

the operator L is bounded in the new setting. □

3.9.2 Nonlinear case

We turn back to the nonlinear DAE (3.108) and the accompanying nonlinear map F (3.109) acting from the space $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ into the continuous function space $\mathcal{C}(\mathcal{I},\mathbb{R}^m)$.

Denote now by $x_* \in C^1_D(\mathcal{I},\mathbb{R}^m)$ a solution of the DAE (3.108), and put $z_0 := Cx_*(t_0)$ so that $Fx_* = 0$, $\mathcal{F}x_* = 0$. Is then the perturbed IVP (3.113), respectively the operator equation $\mathcal{F}x_* = (q,\delta)$ solvable, and how does the solution depend on the perturbations? Does the *implicit function theorem* answer these questions? We have a continuously differentiable map $\mathcal{F}$, and $\mathcal{F}x_* = 0$ is satisfied. If the derivative $\mathcal{F}_x(x_*)$ were a homeomorphism, then we would obtain the good answers by means of the Banach fixed point theorem. From Theorem 3.66 it is known that $\mathcal{F}_x(x_*)$ is bijective provided the index is 1 and the initial condition is such that $d = m - r_0$, $\ker C = \ker \Pi_0(t_0)$. Regarding the relation $\ker \Pi_0(t_0) = \ker D(t_0)$ the following theorem results immediately.

Theorem 3.69. *Let the DAE (3.108) satisfy Assumption 3.16. Let x_* be a solution of the DAE (3.108), $z_0 := Cx_*(t_0)$, and let the linearization (3.111) at x_* be regular with tractability index 1.*
Let the matrix C satisfy $\ker C = \ker D(t_0)$, $\operatorname{im} C = \mathbb{R}^d$, $d = \operatorname{rank} D(t_0)$.
Then, for every sufficiently small perturbation $(q,\delta) \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$, the perturbed IVP (3.113) possesses exactly one solution $x(q,\delta)$ in the neighborhood of x_*, and the inequality

$$\|x(q,\delta) - x_*\|_{\mathcal{C}^1_D} \le K_1(|\delta| + \|q\|_\infty) \tag{3.117}$$

is valid for all these solutions, whereby K_1 is a constant.
$x(q,\delta)$ is defined on a neighborhood of the origin in $\mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$ such that $x(0,0) = x_*$. Furthermore, $x(q,\delta)$ is continuously differentiable with respect to (q,δ).

In particular, in the index-1 case, the function value $(x(0,\delta))(t) =: x(t;\delta)$ depends continuously differentiably on the initial data δ for $t \in \mathcal{I}$. The IVP

$$f((D(t)x(t))',x(t),t) = 0, \quad t \in \mathcal{I}, \quad C(x(t_0) - x_*(t_0)) = \delta,$$

is uniquely solvable for all sufficiently small $\delta \in \mathbb{R}^d$, the solution $x(t;\delta)$ is continuously differentiable with respect to δ, and the sensitivity matrix $X(t;\delta) := x_\delta(t;\delta) \in L(\mathbb{R}^d,\mathbb{R}^m)$ satisfies the variational system

$$\begin{aligned} f_y\big((D(t)x(t;\delta))',x(t;\delta),t\big)\big(D(t)X(t;\delta)\big)' \\ + f_x\big((D(t)x(t;\delta))',x(t;\delta),t\big)X(t;\delta) = 0, \quad t \in \mathcal{I}, \\ CX(t_0;\delta) = I_d. \end{aligned}$$

The columns of the matrix function $X(.;\delta)$ belong to the function space $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$.

Similarly, IVPs and BVPs in regular index-1 DAEs, whose data depend smoothly on additional parameters, have solutions that are continuously differentiable with respect to these parameters, and the sensitivity matrix satisfies the corresponding variational system.
Evidently, with these properties, regular index-1 DAEs are very close to regular ODEs.

Higher index DAEs are essentially different. Let x_* again denote a solution of the nonlinear DAE (3.108). Let the linearized equation (3.111) be regular with tractability index $\mu > 1$, and let the matrix C in the initial condition be such that unique solvability of the linear IVP $\mathcal{F}_x(x_*)x = (q, \delta)$ is ensured for every q from $\operatorname{im} F_x(x_*)$ (cf. Theorem 3.66). Then the linear mapping

$$\mathcal{F}_x(x_*) : \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}(\mathcal{I}, \mathbb{R}^m) \times \mathbb{R}^d$$

is injective but has an unbounded inverse. In the advanced setting

$$\mathcal{F}_x(x_*) : \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \to \mathcal{C}_*^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m) \times \mathbb{R}^d,$$

a bounded inverse exists, where $\mathcal{C}_*^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m)$ denotes the function space that arises from $\operatorname{im} \mathcal{F}_x(x_*)$ by introducing a suitable norm to reach a Banach space. If the nonlinear operator $\mathcal{F}$ also maps into this space, i.e.,

$$\mathcal{F}x \in \mathcal{C}_*^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m), \quad x \in \mathcal{D}_F, \tag{3.118}$$

and if the Fréchet differentiability is not lost in the new setting, then the *implicit function theorem* yields a perturbation result analogous to Theorem 3.69, but now instead of the inequality (3.117) it follows that

$$\|x(q, \delta) - x_*\|_{\mathcal{C}_D^1} \leq K_\mu(|\delta| + \|q\|_{ind\,\mu}).$$

In particular, for q being $\mu - 1$ times continuously differentiable, and for sufficiently smooth DAE coefficients, the inequality

$$\|x(q, \delta) - x_*\|_\infty \leq \|x(q, \delta) - x_*\|_{\mathcal{C}_D^1} \leq \tilde{K}_\mu(|\delta| + \|q\|_\infty + \sum_{j=1}^{\mu-1} \|q^{(j)}\|_\infty) \tag{3.119}$$

follows.
The above only seemingly solves the perturbation problem, since there are serious difficulties concerning condition (3.118). This condition can only be forced by means of strong structural restrictions, for instance

$$\mathcal{W}_{*0}(t)\{f(y, x, t) - f(0, P_0(t)x, t)\} \in \operatorname{im} \mathcal{W}_{*0}(t)B_*(t)Q_0(t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}, \tag{3.120}$$

where $\mathcal{W}_{*0}(t) := I - A_*(t)A_*^-(t)$ denotes a projector along $\operatorname{im} A_*(t)$. At least Hessenberg form size-2 DAEs meet this condition.
If the DAE (3.111) linearized at x_* is regular with tractability index 2, then the actual

Banach space is given by (cf. Proposition 3.68)

$$\mathcal{C}_*^{ind\,2}(\mathcal{I},\mathbb{R}^m) := \{q \in \mathcal{C}(\mathcal{I},\mathbb{R}^m) : D\Pi_{*0}Q_{*1}G_{*2}^{-1}q \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^m)\},$$
$$\|q\|_{*ind\,2} := \|q\|_\infty + \|(D\Pi_{*0}Q_{*1}G_{*2}^{-1}q)'\|_\infty, \quad q \in \mathcal{C}_*^{ind\,2}(\mathcal{I},\mathbb{R}^m).$$

Proposition 3.70. *Let the DAE (3.108) satisfy Assumption 3.16. Let x_* be a solution of (3.108), $z_0 := Cx_*(t_0)$, and let the linearization (3.111) at x_* be fine with tractability index 2 and characteristic values $r_0 \le r_1 < r_2 = m$, $d = r_0 - (m - r_1)$. Let Q_{*0}, Q_{*1} be completely decoupling projector functions to the linearized DAE (3.111).*
Let the matrix C satisfy the conditions $\ker C = \ker \Pi_{*1}(t_0)$, $\operatorname{im} C = \mathbb{R}^d$.
Additionally, let all data be sufficiently smooth and let the function f satisfy the structural condition (3.120) at least in a neighborhood of the extended graph of x_*. Then, for all sufficiently small perturbation $(q,\delta) \in \mathcal{C}_*^{ind\,2}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$, the perturbed IVP (3.113) possesses exactly one solution $x(q,\delta)$ in the neighborhood of x_*, and the inequality

$$\|x(q,\delta) - x_*\|_{\mathcal{C}_D^1} \le K_2(|\delta| + \|q\|_{*ind\,2}) \tag{3.121}$$

is valid for all these solutions, where K_2 is a constant.
$x(q,\delta)$ is defined on a neighborhood of the origin in $\mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$ such that $x(0,0) = x_*$. Furthermore, $x(q,\delta)$ is continuously differentiably with respect to (q,δ).

Proof. The assertion is proved in [163] for perturbed index-2 DAEs in modified standard form $\mathfrak{f}(P(t)x(t))' - P'(t)x(t), x(t), t) = q(t)$. The same arguments apply to the DAE (3.108). □

One should pay attention to the fact that d and C in Proposition 3.70 differ from those in Theorem 3.69.

In consequence of Proposition 3.70, similar to those of Theorem 3.69, the function value $(x(0,\delta))(t) =: x(t;\delta)$ again depends continuously differentiable on the initial data δ for $t \in \mathcal{I}$. The IVP

$$f((D(t)x(t))', x(t), t) = 0, \quad t \in \mathcal{I}, \quad C(x(t_0) - x_*(t_0)) = \delta,$$

is uniquely solvable for all sufficiently small $\delta \in \mathbb{R}^d$, the solution $x(t;\delta)$ is continuously differentiable with respect to δ, and the sensitivity matrix $X(t;\delta) := x_\delta(t;\delta) \in L(\mathbb{R}^d,\mathbb{R}^m)$ satisfies the variational system

$$f_y((D(t)x(t;\delta))', x(t;\delta), t)(D(t)X(t;\delta))' + f_x((D(t)x(t;\delta))', x(t;\delta), t)X(t;\delta) = 0, \quad t \in \mathcal{I},$$
$$CX(t_0;\delta) = I_d.$$

The columns of the matrix function $X(.;\delta)$ belong to the function space $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$.

Theorem 3.70 confirms once again the ambivalent character of higher index DAEs, and the conflicting nature of their solutions: On the one hand they behave as we would expect coming from regular ODE theory. On the other hand they behave as solutions of ill-posed problems.

As adumbrated in the previous sections of this chapter, we hope to reach new transparent solvability assertions without somewhat artificial structural restrictions as in Proposition 3.70 by applying the theory of regularity regions and linearizations. We emphasize the uniform structural characteristics of all linearizations within a regularity region. It is hoped to verify the following conjecture which would pertinently generalize Theorem 3.69.

Notice at this place only that the term $\tilde{\Pi}_{*,can}$ in the conjecture below stands for the canonical projector function associated with the linearization of the DAE at a close smooth approximation $\tilde{x}_*$ of x_*. If x_* itself is smooth enough, we put $\tilde{x}_* = x_*$.

Conjecture 3.71. Let the DAE

$$f((D(t)x(t))',x(t),t) = 0, \quad t \in \mathcal{I},$$

satisfy Assumption 3.16 and have the regularity region $\mathcal{G}$ with tractability index μ. Let the data be sufficiently smooth. Let $x_* \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ be a solution with values in $\mathcal{G}$, $\mathcal{I}$ be compact and $z_0 := Cx_*(t_0)$. Let the matrix C satisfy the condition $\ker C = \ker \tilde{\Pi}_{*,can}(t_0)$, $\operatorname{im} C = \mathbb{R}^d$, $d = \operatorname{rank} \tilde{\Pi}_{*,can}(t_0)$.

Then, for every sufficiently small perturbation $(q,\delta) \in \mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$, the IVP

$$f((D(t)x(t))',x(t),t) = q(t), \quad t \in \mathcal{I}, \quad C(x(t_0)) = z_0 + \delta, \tag{3.122}$$

possesses exactly one solution $x(q,\delta) \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ in the neighborhood of x_*, and the inequality

$$\|x(q,\delta) - x_*\|_{\mathcal{C}_D^1} \le K_\mu(|\delta| + \|q\|_\infty + \sum_{j=1}^{\mu-1} \|q^{(j)}\|_\infty) \tag{3.123}$$

is valid for all these solutions, where K_μ is a constant.

$x(q,\delta)$ is defined on a neighborhood of the origin in $\mathcal{C}^{\mu-1}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^d$ such that $x(0,0) = x_*$. Furthermore, for fixed q, the function $x(q,\delta)$ is continuously differentiable with respect to δ.

3.10 A glance at the standard approach via the derivative array and differentiation index

The derivative array approach aiming at a so-called *completion ODE* associated with the standard form DAE

$$\mathfrak{f}(x'(t),x(t),t) = 0 \tag{3.124}$$

works for smooth nonlinear systems in a similar way as for linear ones (cf. [44, 50]). In this way, the rank of the partial Jacobian $\mathfrak{f}_{x'}(x',x,t)$ is allowed to vary except for the index-1 case (cf. Note (3) in Section 2.12).

To begin, one provides, for a certain index $\kappa \in \mathbb{N}$, the *prolongated system* or *derivative array system*

$$\mathcal{E}_\kappa(x^{\kappa+1},\dots,x^2,x^1,x,t) = 0, \tag{3.125}$$

where the *derivative array function*

$$\mathcal{E}_\kappa(x^{\kappa+1},\dots,x^2,x^1,x,t) := \begin{bmatrix} \mathfrak{f}(x^1,x,t) \\ \mathfrak{f}_{x^1}(x^1,x,t)x^2 + \mathfrak{f}_x(x^1,x,t)x^1 + \mathfrak{f}_t(x^1,x,t) \\ \vdots \\ \mathfrak{f}_{x^1}(x^1,x,t)x^{\kappa+1} + \dots + \mathfrak{f}_{\underbrace{t\cdots t}_{\kappa}}(x^1,x,t) \end{bmatrix},$$

is defined for $t \in \mathcal{I}_\mathfrak{f}$, $x \in \mathcal{D}_\mathfrak{f}$ and $x^1,\dots,x^{\kappa+1} \in \mathbb{R}^m$. It results from the given function $\mathfrak{f}$ by taking the total derivatives in jet variables up to order κ and collecting all these expressions row-wise into the array. Then it is asked whether the prolongated system (3.125) determines on $\mathcal{D}_\mathfrak{f} \times \mathcal{I}_\mathfrak{f}$ (or on an open subset) a continuous function $\mathcal{S}$ such that

$$x^1 = \mathcal{S}(x,t), \quad (x,t) \in \mathcal{D}_\mathfrak{f} \times \mathcal{I}_\mathfrak{f}$$

holds true. Then one solves the resulting explicit ODE

$$x'(t) = \mathcal{S}(x(t),t).$$

The basic tool for deciding whether a vector field description $\mathcal{S}$ can be extracted from equation (3.125) consists in the *fullness property* (e.g. [25]). In particular, one has explicitly to prepare at each level κ the Jacobian

$$\mathcal{J}_\kappa = [\mathcal{E}_{\kappa,x^1}\ \mathcal{E}_{\kappa,w}], \quad w := [x^2,\dots,x^{\kappa+1}]$$

and to check whether it has constant rank and is smoothly 1-full. If there is no such function $\mathcal{S}$, one tries again on the next level $\kappa+1$. This procedure needs highly smooth data. The amount increases enormously with κ and the dimension.
The commonly used index characterization of general standard form DAEs (3.124) is the *differentiation index*, at the beginning called merely the index without an epithet, and sometimes named the *differential index* (e.g. [25], [44], [45], [105]). The differentiation index supposes derivative array systems (3.125).

Definition 3.72. Let $\mathfrak{f}$ be sufficiently smooth.
If $\mathfrak{f}_{x^1}(x^1,x,t)$ remains nonsingular, equation (3.124) is called a DAE with differentiation index 0, as well as a regular ODE.
Otherwise, if there is an index $\mu \in \mathbb{N}$ such that the prolongated system $\mathcal{E}_\mu(x^{\mu+1},\dots,x^2,x^1,x,t) = 0$ determines the variable x^1 as a continuous function $\mathcal{S}$ in terms of x and t on (an open set in) $\mathcal{D}_\mathfrak{f} \times \mathcal{I}_\mathfrak{f}$, and if μ is the smallest such index,

then μ is named the *differentiation index* of the DAE (3.124). The resulting explicit ODE

$$x'(t) = \mathcal{S}(x(t),t)$$

is called a *completion ODE* (e.g. [45]) as well as an *underlying ODE* (e.g. [105]).

Example 3.73 (Consistency with regularity regions). The equation $\mathcal{E}_1(x^1,x,t) = 0$ for the DAE in Example 3.34 reads

$$\begin{aligned} x_1^1 - x_3 &= 0,\\ x_2(1-x_2) - \gamma(t) &= 0,\\ x_1x_2 + x_3(1-x_2) - t &= 0, \end{aligned}$$

$$\begin{aligned} x_1^2 - x_3^1 &= 0,\\ x_2^1(1-x_2) - x_2x_2^1 - \gamma'(t) &= 0,\\ x_1^1x_2 + x_1x_2^1 + x_3^1(1-x_2) - x_3x_2^1 - 1 &= 0. \end{aligned}$$

Looking for a function $x^1 = \mathcal{S}(x,t)$ one is led to the system

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

which provides the required functions $x^1 = \mathcal{S}(x,t)$ precisely on each of the regularity regions

$$\begin{aligned} \mathcal{G}_1 &:= \left\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x_2 < \frac{1}{2}\right\},\\ \mathcal{G}_2 &:= \left\{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : \frac{1}{2} < x_2 < 1\right\},\\ \mathcal{G}_3 &:= \{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : 1 < x_2\}, \end{aligned}$$

given in Example 3.34. It follows that the DAE has differentiation index 1 on the (tractability-) index-1 regularity regions. □

For a large class of DAEs the constant-rank conditions supporting the tractability index and regularity regions are exactly the same as needed to determine the differentiation index and completion ODE. This indicates a certain consistency. We highlight the essential differences later on in this section.

Index-1 DAEs are those whose solutions fill up the obvious constraint

$$\mathcal{M}_0(t) = \{x \in \mathcal{D}_\mathfrak{f} : \exists x^1 : \mathfrak{f}(x^1,x,t) = 0\}$$

as in the particular case of Example 3.7. In general, one has to expect further constraints that are not so obvious, but hidden in the DAE formulation as in Exam-

ple 3.8. The general expectation is a sequence of constraint sets

$$\mathcal{M}_0(t) \supset \mathcal{M}_1(t) \supset \cdots \supset \mathcal{M}_{\mu-1}(t) = \mathcal{M}_\mu(t), \quad t \in \mathcal{I}_\dagger,$$

becoming stationary at level $\mu - 1$, and just the set $\mathcal{M}_{\mu-1}(t)$ is filled up by solutions. This idea is incorporated, e.g., in the notion of *geometrical solvability* (cf. [44, 50]) of standard form DAEs (3.124) saying that there is a well-behaved manifold of solutions and a given solution is uniquely determined by an appropriate initial condition. The DAE solutions are embedded into the flow of the completion ODE. More precisely, the DAE solutions are those solutions of the completion ODE which are located at the final constraint set $\mathcal{M}_{\mu-1}(t)$, that is,

$$x'(t) = \mathcal{S}(x(t),t), \quad x(t) \in \mathcal{M}_{\mu-1}(t). \tag{3.126}$$

The framework of the completion ODE is taken to hold on open sets so that, with wise foresight in view of a numerical treatment, perturbations by excitations can be incorporated. The solution of the IVP

$$x'(t) = \mathcal{S}(x(t),t), \quad x(t_0) = x_0 \in \mathcal{M}_{\mu-1}(t_0),$$

proceeds in $\mathcal{M}_{\mu-1}(t)$, however, in numerical computations it drifts away from this constraint set. This phenomenon is caused by the stability behavior of the completed flow in the neighborhood of the constraint set. It is well known that, in general, completion ODEs are not uniquely determined by their original DAEs.

Example 3.74 (Different completion ODEs). Supposing the real functions α, β and γ are sufficiently smooth, the autonomous Hessenberg size-2 DAE

$$\begin{aligned} x_1' &= (\alpha(x_1) - x_2)\beta(x_1) + \gamma(x_1), \\ x_2' &= x_3, \\ 0 &= x_2 - \alpha(x_1), \end{aligned} \tag{3.127}$$

leads to the following two specific autonomous completion ODEs

$$\begin{aligned} x_1' &= (\alpha(x_1) - x_2)\beta(x_1) + \gamma(x_1), \\ x_2' &= x_3, \\ x_3' &= \alpha''(x_1)\big((\alpha(x_1) - x_2)\beta(x_1) + \gamma(x_1)\big)^2 - \alpha'(x_1)\beta(x_1)x_3 + \alpha'(x_1)\big(\alpha'(x_1)\beta(x_1) \\ &\qquad + (\alpha(x_1) - x_2)\beta'(x_1) + \gamma'(x_1)\big)\big(\alpha(x_1)\beta(x_1) - x_2\beta(x_1) + \gamma(x_1)\big), \end{aligned} \tag{3.128}$$

and

$$
\begin{aligned}
x_1' &= (\alpha(x_1)-x_2)\beta(x_1)+\gamma(x_1),\\
x_2' &= \alpha'(x_1)\big(\alpha(x_1)\beta(x_1)-x_2\beta(x_1)+\gamma(x_1)\big),\\
x_3' &= \alpha''(x_1)\big((\alpha(x_1)-x_2)\beta(x_1)+\gamma(x_1)\big)^2-\alpha'(x_1)\beta(x_1)x_3+\alpha'(x_1)\big(\alpha'(x_1)\beta(x_1)\\
&\quad +(\alpha(x_1)-x_2)\beta'(x_1)+\gamma'(x_1)\big)\big(\alpha(x_1)\beta(x_1)-x_2\beta(x_1)+\gamma(x_1)\big).
\end{aligned}
\tag{3.129}
$$

The constraint sets also being independent of t are

$$\mathcal{M}_0=\{x\in\mathbb{R}^3 : x_2=\alpha(x_1)\}\supset\mathcal{M}_1=\{x\in\mathbb{R}^3 : x_2=\alpha(x_1), x_3=\gamma(x_1)\}.$$

Assume $\gamma(c)=0,\gamma'(c)\neq 0$, for a certain fixed $c\in\mathbb{R}$, and consider the stationary solution x_* of the DAE given by $x_{*,1}=c$, $x_{*,2}=\alpha(c)$, $x_{*,3}=0$. Owing to Lyapunov's theorem the eigenstructure of the corresponding Jacobian matrix $\mathcal{S}_x(x_*)$ is responsible for the stability behavior of the reference solution x_*. In the first case, these eigenvalues are

$$\lambda_1=\gamma'(c),\ \lambda_2=0,\ \lambda_3=0,$$

and the two zero eigenvalues belong to a second-order Jordan chain. In the second case, the eigenvalues are

$$\lambda_1=\gamma'(c),\ \lambda_2=0,\ \lambda_3=-\alpha'(c)\beta(c).$$

This explains why numerical solutions often drift away from the constraint set they should remain on. Even if $\gamma'(c)<0$, the stationary solution x_* fails to be asymptotically stable as a solution of the completion ODE.
Setting $\alpha(\xi)=-\xi$, $\beta(\xi)=5$ and $\gamma(\xi)=1-\xi^2$ then we obtain $c=1$. The resulting stationary solution is $x_*=(1,-1,0)$. The DAE (3.127) is now

$$
\begin{aligned}
x_1' &= -5(x_1+x_2)+1-x_1^2,\\
x_2' &= x_3,\\
0 &= x_1+x_2,
\end{aligned}
$$

with the obvious constraint set $\mathcal{M}_0=\{x\in\mathbb{R}^3 : x_1+x_2=0\}$ and the set of consistent values $\mathcal{M}_1=\{x\in\mathbb{R}^3 : x_1+x_2=0,\ x_3=x_1^2-1\}$.
The completion ODEs (3.128) and (3.129) simplify to

$$
\begin{aligned}
x_1' &= -5(x_1+x_2)+1-x_1^2,\\
x_2' &= x_3,\\
x_3' &= 5x_3+(5+2x_1)(-5(x_1+x_2)+1-x_1^2),
\end{aligned}
\tag{3.130}
$$

respectively to

$$
\begin{aligned}
x_1' &= -5(x_1+x_2)+1-x_1^2,\\
x_2' &= 5(x_1+x_2)-1+x_1^2,\\
x_3' &= 5x_3+(5+2x_1)(-5(x_1+x_2)+1-x_1^2).
\end{aligned}
\tag{3.131}
$$

The solution x_* is asymptotically stable as a solution of the DAE, that is, IVPs with slightly perturbed but consistent initial values have solutions on the entire interval $[0,\infty)$ tending to x_*.

Notice that solutions of the completion ODEs which start from points in $\mathcal{M}_1$ behave in exactly the same way. However, if the initial value of a solution of the completion ODEs does not exactly belong to $\mathcal{M}_1$, then the solution fails to approach x_*, but drifts away.

Figure 3.9 shows the solution (solid line) of the DAE starting at $t = 0$ in $(1.1,-1.1,0.21) \in \mathcal{M}_1$ which solves at the same time the completion ODEs, and the solutions of the ODEs (3.130) (dashed line) and (3.131) (dot-dashed line) starting from the initial value $(1.1,-1.101,0.21)$ which is close to the previous one but does not belong to $\mathcal{M}_1$. While the solution of (3.131) (dot-dashed line) moves away quickly the solution of (3.130) (dashed line) drifts slowly. □

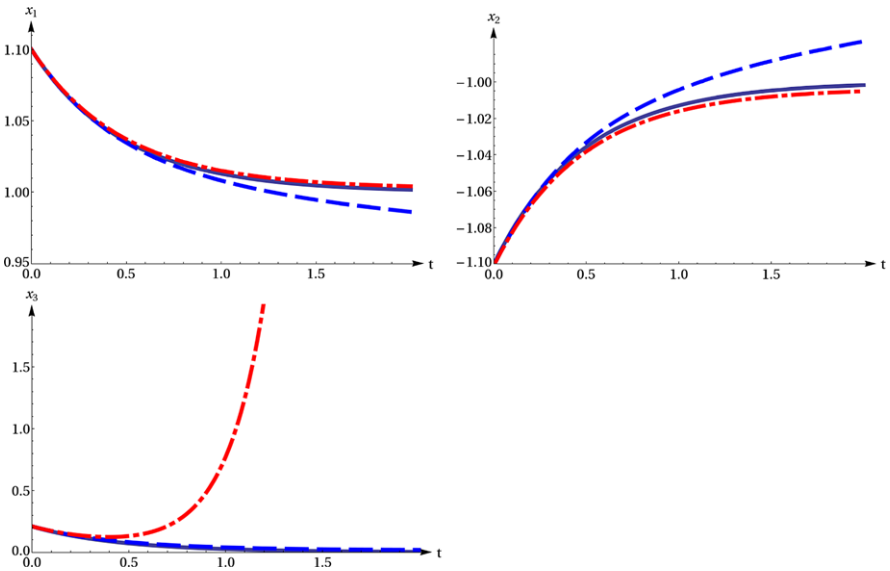


Fig. 3.9 Solution components x_1, x_2, x_3 related to Example 3.74

It is not at all a simple task to extract the description of the completion ODE and the constraint manifold from equation (3.125). Even if the full information about the completion ODE is available, i.e., the vector field $\mathcal{S}$ is given, and the constraint set is described by an equation, say $\mathcal{M}_{\mu-1}(t) = \{x \in \mathcal{D}_f : h(x,t) = 0\}$, and $h_x(x,t)$ has full row rank, then in view of the numerical treatment, it is proposed ([87], cf. [25]) to change to the Hessenberg size-2 DAE

$$\begin{aligned} x'(t) &= \mathcal{S}(x(t),t) + h_x(x(t),t)^* \lambda(t), \\ 0 &= h(x(t),t), \end{aligned}$$

whereby the new variable λ has the size $\operatorname{rank} h_x$.

There are various special approaches to DAEs, which are based in a similar way on derivative array functions, such as reduction techniques (e.g., [95, 189]) and transformations into special forms (e.g., [130]). In any case one has to provide derivative array functions with all their included derivatives. We refer to the monograph [130] which is devoted to derivative array approaches for a further discussion. To avoid the shortcomings of the completion ODE, e.g., in [130], one sets stronger priority in regarding the constraints and tries to extract from equation (3.125) an index-1 DAE of the special form

$$\mathrm{x}_1'(t) = \mathcal{L}(\mathrm{x}_1(t), t), \quad \mathrm{x}_2(t) = \mathcal{R}(\mathrm{x}_1(t), t), \tag{3.132}$$

whereby the given components of the unknown x are suitably partitioned into x_1 and x_2. It should be pointed out that the ODE in (3.132) is not the same as an IERODE. To see the difference we turn to the simple constant coefficient DAE in Example 1.5 (cf. also Example 3.67).

Example 3.75 (Different resulting ODEs). For the regular index-4 DAE

$$\underbrace{\begin{bmatrix} 1&0&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1\\ 0&0&0&0&0 \end{bmatrix}}_{E} x'(t) + \underbrace{\begin{bmatrix} -\alpha&-1&0&0&0\\ 0&1&0&0&0\\ 0&0&1&0&0\\ 0&0&0&1&0\\ 0&0&0&0&1 \end{bmatrix}}_{F} x(t) = q(t)$$

the corresponding equation $\mathcal{E}_4(x^5, x^4, x^3, x^2, x^1, x, t) = 0$ comprises 25 equations, from which a completion ODE as well as an index-1 DAE (3.132) can be extracted, namely

$$\begin{aligned} x_1' &= \alpha x_1 + x_2 + q_1,\\ x_2' &= q_2' - q_3'' + q_4^{(3)} - q_5^{(4)},\\ x_3' &= -x_2 + q_2,\\ x_4' &= -x_3 + q_3,\\ x_5' &= -x_4 + q_4, \end{aligned}$$

and

$$\begin{aligned} x_1' &= \alpha x_1 + q_1 - q_2 + q_3' - q_4^{(2)} + q_5^{(3)},\\ x_2 &= q_2 - q_3' + q_4^{(2)} - q_5^{(3)},\\ x_3 &= q_3 - q_4' + q_5^{(2)},\\ x_4 &= q_4 - q_5',\\ x_5 &= q_5. \end{aligned} \tag{3.133}$$

In contrast, the projector based decoupling given in Example 3.67 leads to

$$(x_1+x_3-\alpha x_4+\alpha^2 x_5)' = \alpha(x_1+x_3-\alpha x_4+\alpha^2 x_5)+q_1+q_2-\alpha q_3+\alpha^2 q_4-\alpha^3 q_5, \tag{3.134}$$

$$\begin{aligned} x_2 &= q_2-(q_3-(q_4-q_5')')', \\ x_3 &= q_3-(q_4-q_5')', \\ x_4 &= q_4-q_5', \\ x_5 &= q_5. \end{aligned}$$

The ODE (3.133) and the IERODE (3.134) have the same dimension. We recognize that the IERODE 3.134 precisely reflects the smoothest part of the unknown x, being independent of the derivatives of the excitation. This part is captured by means of the projector Π_3 which is the spectral projector associated with the matrix pencil $\lambda E+F$. In general, one cannot expect standard basis vectors like $(1,0,\ldots,0)^T$ to belong to the finite eigenspace of a given regular matrix pencil. □

We turn to another example to highlight further differences.

Example 3.76 (Campbell's DAE). Consider the system of $m = m_3+m_1+m_2$, $m_2 > m_1$, equations

$$\begin{aligned} \mathcal{A}(x_1(t),t)x_1'(t)+\varphi(x_1(t),t)+x_3(t) &= 0, \\ x_1(t)-\gamma(t) &= 0, \\ x_2'(t)+\psi(x_1(t),x_2(t),x_3(t),t) &= 0, \end{aligned}$$

where $\mathcal{A},\varphi,\gamma,\psi$ are sufficiently smooth on the domain $\mathbb{R}^m\times\mathbb{R}$. The matrix function $\mathcal{A}(x_1,t)\in L(\mathbb{R}^{m_1},\mathbb{R}^{m_2})$ is assumed to have different rank on different subdomains. We assume in detail

$$\mathcal{A}(x_1,t)=0, \text{ if } x_1\in\mathcal{D}_1^{[1]}, \text{ and } \operatorname{rank}\mathcal{A}(x_1,t)=m_1, \text{ if } x_1\in\mathcal{D}_1^{[2]},$$

with open connected sets $\mathcal{D}_1^{[1]},\mathcal{D}_1^{[2]}$ in $\mathbb{R}^{m_1}$. This DAE can be easily solved. It serves as a special case to emphasize the advantages of the derivative array approach (e.g., [46]). To apply this approach we form the array functions

$$\mathcal{E}_1(x^2,x^1,x,t) := \begin{bmatrix} \mathcal{A}(x_1,t)x_1^1+\varphi(x_1,t)+x_3 \\ x_1-\gamma(t) \\ x_2^1+\psi(x,t) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{A}(x_1,t)x_1^2+\mathcal{A}_{x_1}(x_1,t)x_1^1x_1^1+\mathcal{A}_t(x_1,t)x_1^1 \\ \qquad +\varphi_{x_1}(x_1,t)x_1^1+\varphi_t(x_1,t)+x_3^1 \\ x_1^1-\gamma'(t) \\ x_2^2+\psi_{x_1}(x,t)x_1^1+\psi_{x_2}(x,t)x_2^1+\psi_{x_3}(x,t)x_3^1+\psi_t(x,t) \end{bmatrix}$$

and

$$\mathcal{E}_2(x^3,x^2,x^1,x,t) := \begin{bmatrix} \mathcal{A}(x_1,t)x_1^1 + \varphi(x_1,t) + x_3 \\ x_1 - \gamma(t) \\ x_2^1 + \psi(x,t) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{A}(x_1,t)x_1^2 + \mathcal{A}_{x_1}(x_1,t)x_1^1x_1^1 + \mathcal{A}_t(x_1,t)x_1^1 \\ \qquad + \varphi_{x_1}(x_1,t)x_1^1 + \varphi_t(x_1,t) + x_3^1 \\ x_1^1 - \gamma'(t) \\ x_2^2 + \psi_{x_1}(x,t)x_1^1 + \psi_{x_2}(x,t)x_2^1 + \psi_{x_3}(x,t)x_3^1 + \psi_t(x,t) \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \mathcal{A}(x_1,t)x_1^3 + \cdots + x_3^2 \\ x_1^2 - \gamma''(t) \\ x_2^3 + \psi_{x_1}(x,t)x_1^2 + \cdots + \psi_{tt}(x,t) \end{bmatrix}.$$

First we check whether the equation

$$\mathcal{E}_1(x^2,x^1,x,t) = 0$$

contains a relation $x^1 = \mathcal{S}(x,t)$ with an at least continuous function $\mathcal{S}$ defined on an open set in $\mathbb{R}^m \times \mathbb{R}$. This happens in fact if $\mathcal{A}(x_1,t)$ vanishes identically, for instance, if $x_1 \in \mathcal{D}_1^{[1]}$. Therefore, the DAE has differentiation index 1 on the corresponding region

$$\mathcal{G}_1 := \mathcal{D}_1^{[1]} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} \times \mathbb{R}.$$

A look at the DAE system itself shows that then the solution does not depend on the derivative of the function γ.

If $\mathcal{A}(x_1,t)$ does not vanish identically, but if it disappears just on a lower-dimensional subset $\Omega \subset \mathbb{R}^m \times \mathbb{R}$, then the prolongated system $\mathcal{E}_1(x^2,x^1,x,t) = 0$ determines a vector field $\mathcal{S}_\Omega$ just on this subset Ω, that is, $\mathcal{S}_\Omega$ is no longer given on an open set in $\mathbb{R}^m \times \mathbb{R}$, and the definition of the completion ODE does not apply.

Therefore, if $\mathcal{A}(x_1,t)$ does not vanish identically, we must turn to the equation

$$\mathcal{E}_2(x^3,x^2,x^1,x,t) = 0$$

and ask again for a relation $x^1 = \mathcal{S}(x,t)$. Now we actually attain such a relation globally and independently of the behavior of $\mathcal{A}(x_1,t)$. We obtain a completion ODE $x'(t) = \mathcal{S}(x(t),t)$ and the DAE has differentiation index 2 on its definition domain $\mathbb{R}^m \times \mathbb{R}$.

On the other hand, since the rank of $\mathcal{A}$ varies, this DAE cannot be rewritten as a DAE with a properly leading term. However writing the above system as

$$\underbrace{\begin{bmatrix} \mathcal{A}(x_1(t),t) & 0 \\ 0 & 0 \\ 0 & I \end{bmatrix}}_{A(x_1(t),t)} \Big(\underbrace{\begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \end{bmatrix}}_{D} x(t) \Big)' + \begin{bmatrix} \varphi(x_1(t),t) + x_3(t) \\ x_1(t) - \gamma(t) \\ \psi(x_1(t),x_2(t),x_3(t),t) \end{bmatrix} = 0, \tag{3.135}$$

we arrive at a DAE with quasi-proper leading term (cf. Section 3.12 and Chapter 9). Therefore it does not matter if the matrix $\mathcal{A}(x_1,t)$ changes its rank. Observe that the leading term in (3.135) becomes even properly stated if the matrix function $\mathcal{A}$ has full column rank, for instance, if $x_1 \in \mathcal{D}_2^{[2]}$.
We form a matrix function sequence from only the first partial derivatives of the coefficients $\mathcal{A}, \varphi, \gamma, \psi$ starting with

$$G_0 := AD = \begin{bmatrix} \mathcal{A} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \end{bmatrix}, \quad B_0 := \begin{bmatrix} \varphi_{x^1} + (\mathcal{A}y)_{x^1} & 0 & I \\ I & 0 & 0 \\ \psi_{x^1} & \psi_{x^2} & \psi_{x^3} \end{bmatrix}.$$

We first compute a projector Q_0 onto $\ker D$, and then $G_1 := G_0 + B_0 Q_0$, that is

$$Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} \mathcal{A} & 0 & I \\ 0 & 0 & 0 \\ 0 & I & \psi_{x^3} \end{bmatrix}.$$

Next we ask whether G_1 is nonsingular. This is not the case, but G_1 has constant rank. We compute the continuous projector function Q_1 onto $\ker G_1$,

$$Q_1 = \begin{bmatrix} I & 0 & 0 \\ \psi_{x^3}\mathcal{A} & 0 & 0 \\ -\mathcal{A} & 0 & 0 \end{bmatrix}.$$

Next we set $P_0 := I - Q_0$, $P_1 := I - Q_1$ and compute

$$\tilde{G}_2 := G_1 + B_0 P_0 Q_1 = \begin{bmatrix} \mathcal{A} + \varphi_{x^1} & 0 & I \\ I & 0 & 0 \\ \psi_{x^1} + \psi_{x^2}\psi_{x^3}\mathcal{A} & I & \psi_{x^3} \end{bmatrix}.$$

The matrix function $\tilde{G}_2$ is everywhere nonsingular, which means that the DAE is quasi-regular on the definition domain $\mathbb{R}^m \times \mathbb{R}$ (cf. Chapter 9).
At the same time, on open sets where $\mathcal{A}(x_1,t)$ has full column rank, a regular index-2 DAE results. In particular,

$$\mathcal{G}_2 := \mathcal{D}_1^{[2]} \times \mathbb{R}^{m_2} \times \mathbb{R}^{m_3} \times \mathbb{R}$$

is a regularity region with index 2.
On open sets where $\mathcal{A}$ identically vanishes, by replacing D in (3.135) with $D_{new} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \end{bmatrix}$ a proper reformulation results and there is a regularity region with index-1. For instance, $\mathcal{G}_1$ is such an index-1 regularity region. □

The notion of quasi-regularity (see Chapter 9) is somewhat close to the differential index. It allows rank changes but it is weak in the sense that the restrictions of the given DAE to subdomains do not inherit the global characteristics.

3.11 Using structural peculiarities to ease models

A more comfortable version of a DAE with proper leading term is the equation

$$f((D(t)x(t))',x(t),t)=0, \tag{3.136}$$

in which the derivative term is linear. This special form arises in the general DAE (3.1) for $d(x,t)=D(t)x$. It might be easier to handle than the fully nonlinear DAE. In particular, there is a linear function space, $\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^n)$, which accommodates the solutions. Therefore, sometimes it is reasonable to turn from equation (3.1) to an equivalent auxiliary enlarged system which possesses such a simpler structure. The following proposition ensures the change.
Let the DAE (3.1) satisfy Assumption 3.16 and let $\ker f_y(y,x,t)$ be independent of y and x. Then a projector valued function $R_A\in\mathcal{C}^1(\mathcal{I}_f,L(\mathbb{R}^n))$ is available such that

$$\ker R_A(t)=\ker f_y(y,x,t),\quad y\in\mathbb{R}^n,x\in\mathcal{D}_f,t\in\mathcal{I}_f.$$

For instance, the orthoprojector function along $\ker f_y(y,x,t)$ can be chosen. Because of the identity

$$f(y,x,t)-f(R_A(t)y,x,t)=\int_0^1 f_y(sy+(1-s)R_A(t)y,x,t)(I-R_A(t))y\mathrm{d}s=0 \tag{3.137}$$

we can rewrite the DAE (3.1) as

$$f(R_A(t)(d(x(t),t))',x(t),t)=0,$$

and hence as

$$f((R_A(t)d(x(t),t))'-R_A'(t)d(x(t),t),x(t),t)=0. \tag{3.138}$$

The latter equation suggests we turn to a slightly weaker solution notion.

Definition 3.77. Let the DAE (3.1) satisfy Assumption 3.16 and show the nullspace $\ker f_y(y,x,t)$ to be independent of y and x. Each function $x_*\in\mathcal{C}(\mathcal{I}_*,\mathbb{R}^m)$ with values in $\mathcal{D}_f$ and a continuously differentiable term $R_A(.)d(x_*(.),.)$, which satisfies the DAE pointwise on the interval $\mathcal{I}_*$, is said to be a *solution* of this DAE.

One can check immediately that this solution notion is invariant with respect to the special choice of the projector function R_A. Of course, if $(I-R_A(.))d(x_*(.),.)$ is also continuously differentiable, then we attain a solution in the previous sense.

The enlarged system

$$f\big((R_A(t)u(t))'-R_A'(t)d(x(t),t),x(t),t\big)=0, \tag{3.139}$$
$$u(t)-d(x(t),t)=0, \tag{3.140}$$

actually has the required form (3.136). We have

$$\hat{f}((\hat{D}(t)\hat{x}(t))',\hat{x}(t),t) = 0, \tag{3.141}$$

with

$$\hat{x} = \begin{bmatrix} u \\ x \end{bmatrix}, \quad \hat{f}(y,\hat{x},t) = \begin{bmatrix} f(y - R_A'(t)d(x,t),x,t) \\ u - d(x,t) \end{bmatrix}, \quad y \in \mathbb{R}^n, u \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}_f,$$
$$\hat{D}(t) = \begin{bmatrix} R_A(t) & 0 \end{bmatrix}.$$

Since the original DAE (3.1) satisfies the Assumption 3.16 so does the enlarged version (3.141). In particular, it holds that

$$\ker \hat{f}_y = \ker f_y = \ker R_A \quad \text{and} \quad \operatorname{im} \hat{D} = \operatorname{im} R_A,$$

so that $\ker \hat{f}_y$ and $\operatorname{im} \hat{D}$ are actually transversal $\mathcal{C}^1$-subspaces.
If even $\ker f_y = \{0\}$, then $R_A = I$ and the enlarged system (3.141) simplifies to

$$f(u'(t),x(t),t) = 0,$$
$$u(t) - d(x(t),t) = 0.$$

Proposition 3.78. *Let equation (3.1) satisfy Assumption 3.16 and* $\ker f_y$ *be independent of y and x.*

(1) *Then the enlarged system (3.141) is a DAE of the form (3.136) and satisfies Assumption 3.16, too.*
(2) *If x_* is a solution of the DAE (3.1) in the sense of Definition 3.77, then $\hat{x}_* := (u_*, x_*)$, $u_* := d(x_*(.),.)$, is a solution of the enlarged DAE (3.141), and vice versa.*
(3) *If f_y has full column rank, then $R_A = I$, and the enlarged DAE comprises also a full-column-rank partial derivative $\hat{f}_y$. If x_* is a solution of (3.1), then the pair $u_* := d(x_*(.),.)$, x_* is a solution of the enlarged system, and vice versa.*

Proof. (1) and the first part of (3) are evident and are shown before Proposition 3.78. It remains to show (2) and the second part of (3).
If x_* is a solution of (3.1) in the sense of Definition 3.77, $u_* := d(x_*(.),.)$, then the second row (3.140) of the enlarged DAE is satisfied. Furthermore, the component $R_A u_*$ is continuously differentiable and

$$\begin{aligned} f\big((R_A(t)u_*(t))' - R_A'(t)d(x_*(t),t),x_*(t),t\big) &= f(R_A(t)u_*'(t),x_*(t),t) \\ &= f(u_*'(t),x_*(t),t) = 0. \end{aligned}$$

Conversely, if $\hat{x}_* := (u_*, x_*)$ is a solution of the enlarged DAE, then $R_A u_*$ is continuously differentiable and

$$\begin{aligned}0 &= u_*(t) - d(x_*(t),t),\\ 0 &= f((R_A(t)u_*(t))' - R_A'(t)d(x_*(t),t),x_*(t),t)\\ &= f(R_A(t)(d(x_*(t),t))' - R_A'(t)d(x_*(t),t),x_*(t),t).\end{aligned}$$

This proves the assertion. □

The analysis simplifies if one has subspaces which do not at all depend on y and x. In the standard applications—circuit simulation and mechanical motion simulation—the partial Jacobian f_y is a constant matrix such that $\ker f_y$ is constant. We are not aware of any applications resulting in subspaces $\ker f_y$, $\operatorname{im} d_x$ that actually depend on y and x. Of course, theoretically, such a dependence is imaginable.
If just one of the two relevant subspaces has the desired property, then the DAE can be slightly modified to acquire the property for the other subspace, too. This fact is worth considering in the modeling process at the very beginning.

More precisely, let DAE (3.1) satisfy Assumption 3.1, let $\operatorname{im} d_x$ be a $\mathcal{C}^1$-subspace. Assuming $\operatorname{im} d_x(x,t)$ to be independent of x, we find a projector function $R_D \in \mathcal{C}^1(\mathcal{I}_f, L(\mathbb{R}^n))$ such that $\operatorname{im} R_D(t) = \operatorname{im} d_x(x,t)$, for all $x \in \mathcal{D}_f, t \in \mathcal{I}_f$. In particular, the orthoprojector onto $\operatorname{im} d_x(x,t)$ can serve as $R_D(t)$. Then, we turn from (3.1) to the modified DAE

$$\tilde{f}((d(x(t),t))',x(t),t) = 0, \tag{3.142}$$

where

$$\begin{aligned}\tilde{f}(y,x,t) &:= f(R_D(t)y + (I - R_D(t))d_t(x,t),\ x,t),\\ \tilde{f}_y(y,x,t) &= f_y(R_D(t)y + (I - R_D(t))d_t(x,t),x,t)R_D(t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}_f.\end{aligned}$$

In contrast, in the opposite case, if $\operatorname{im} d_x(x,t)$ depends on x, but $\ker f_y(y,x,t)$ is independent of (y,x), supposing that $\ker f_y$ is a $\mathcal{C}^1$-subspace, we take a projector function $R_A \in \mathcal{C}^1(\mathcal{I}_f, L(\mathbb{R}^n))$ such that $\ker R_A(t) = \ker f_y(y,x,t)$, and modify the DAE as

$$\tilde{f}((\tilde{d}(x(t),t))',x(t),t) = 0, \tag{3.143}$$

with

$$\begin{aligned}\tilde{d}(x,t) &:= R_A(t)d(x,t),\\ \tilde{f}(y,x,t) &:= f(R_A(t)y - R_A'(t)d(x,t),\ x,t),\\ \tilde{f}_y(y,x,t) &= f_y(R_A(t)y - R_A'(t)d(x,t),\ x,t)R_A(t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f, t \in \mathcal{I}_f.\end{aligned}$$

Proposition 3.79. *Let the DAE (3.1) satisfy Assumption 3.16.*

(1) *If $\operatorname{im} d_x(x,t)$ is independent of x, then the DAE (3.142) satisfies Assumption 3.16, too, and it holds that*

$$\ker \tilde{f}_y(y,x,t) = \ker R_D(t), \quad \operatorname{im} d_x(x,t) = \operatorname{im} R_D(t).$$

Moreover, the DAEs (3.1) and (3.142) are equivalent.

(2) *Let* $\ker f_y(y,x,t)$ *be independent of* y *and* x*. Then the DAE* (3.143)*, satisfies Assumption 3.16, too, and it holds that*

$$\ker \tilde{f}_y(y,x,t) = \ker R_A(t), \quad \operatorname{im} \tilde{d}_x(x,t) = \operatorname{im} R_A(t).$$

The solutions of the DAE (3.1) are at the same time solutions of the modified DAE (3.143)*, whereas the solutions of* (3.143) *are solutions of (3.1) in the sense of Definition 3.77.*

Proof. (1) For each arbitrary function $x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$, with values in $\mathcal{D}_f$, such that $d(x(.),.) \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)$, Proposition C.1 provides the expression

$$(d(x(t),t))' = d_x(x(t),t)w(t) + d_t(x(t),t),\ t \in \mathcal{I},$$

with a certain continuous function w. This yields

$$(I - R_D(t))(d(x(t),t))' = (I - R_D(t))d_t(x(t),t),$$

and hence

$$\begin{aligned} f((d(x(t),t))',x(t),t) &= f\big(R_D(t)(d(x(t),t))' + (I - R_D(t))d_t(x(t),t),\ x(t),t\big) \\ &= \tilde{f}\big((d(x(t),t))',x(t),t\big). \end{aligned}$$

Consequently, each solution of (3.1) also solves (3.142), and vice versa.
Since the DAE (3.1) has a properly stated leading term, its transversality condition implies $\ker f_y(y,x,t) \cap \operatorname{im} R_D(t) = \{0\}$, thus $\ker \tilde{f}_y(y,x,t) = \ker R_D(t)$, and hence $\ker \tilde{f}_y(y,x,t) \oplus \operatorname{im} d_x(x,t) = \ker R_D(t) \oplus \operatorname{im} R_D(t) = \mathbb{R}^n$.
(2) Choosing a projector function $R_A \in \mathcal{C}^1(\mathcal{I}_f, L(\mathbb{R}^n))$, $\ker R_A(t) \subseteq \ker f_y(y,x,t)$ we apply relation (3.137). For each arbitrary $x \in \mathcal{C}(\mathcal{I},\mathbb{R}^m)$, with values in $\mathcal{D}_f$, $d(x(.),.) \in \mathcal{C}^1(\mathcal{I},\mathbb{R}^n)$, we derive

$$\begin{aligned} f\big((d(x(t),t))',x(t),t\big) &= f\big(R_A(t)(d(x(t),t))',x(t),t\big) \\ &= f\big((R_A(t)d(x(t),t))' - R_A'(t)d(x(t),t),x(t),t\big) \\ &= \tilde{f}((\tilde{d}(x(t),t))',x(t),t). \end{aligned}$$

This shows that each solution of (3.1) also solves the modified DAE (3.143). If x_* is a solution of the modified DAE, then just $R_A d(x_*(.),.)$ is continuously differentiable, so that Definition 3.77 applies.
Since (3.1) has a properly stated leading term, it holds that $\ker R_A(t) \cap \operatorname{im} d_x(x,t) = \{0\}$. This yields $\operatorname{im} \tilde{d}_x(x,t) = \operatorname{im} R_A(t) d_x(x,t) = R_A(t) \operatorname{im} d_x(x,t)$; further $\operatorname{im} \tilde{d}_x(x,t) = \operatorname{im} R_A(t)$, and $\ker \tilde{f}_y(y,x,t) \oplus \operatorname{im} \tilde{d}_x(x,t) = \ker R_A(t) \oplus \operatorname{im} R_A(t) = \mathbb{R}^n$. □

3.12 Regularity regions of DAEs with quasi-proper leading terms

It may well happen that a given DAE (3.1) satisfies the Assumption 3.1, and it actually shows a regular behavior, but fails to have a properly involved derivative. To capture this situation we relax the constant-rank condition for f_y and apply local reformulations.

Definition 3.80. Equation (3.1) which satisfies Assumption 3.1 is said to be a DAE with *quasi-proper leading term*, if $\operatorname{im} d_x$ is a $\mathcal{C}^1$-subspace, $\ker d_x$ is nontrivial, and there exists a further $\mathcal{C}^1$-subspace N_A, possibly depending on y,x,t, such that the inclusion

$$N_A(y,x,t) \subseteq \ker f_y(y,x,t), \quad y \in \mathbb{R}^n, x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{3.144}$$

and the transversality condition

$$N_A(y,x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^n, \quad x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{3.145}$$

are valid.

For what concerns the solution notion, we continue to apply Definition 3.2.

There is a simple idea to create a DAE with quasi-proper leading term: One arranges things in such a way that d_x is rectangular and has full row rank $r = n \leq m-1$. In this case, the trivial subspace $N_A = \{0\}$ satisfies both conditions (3.144) and (3.145).

Often the substitution of a singular square matrix D_{inc} into a standard form DAE

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

such that $\ker D_{inc} \cap \operatorname{im} D_{inc} = \{0\}$ and $\mathfrak{f}(x^1,x,t) = \mathfrak{f}(D_{inc}x^1,x,t)$ holds for all arguments, will do. Having such an *incidence matrix*, its entries are mostly zeros and ones, and the standard form DAE can be rewritten as

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

One attains a quasi-proper leading term by letting $N_A := \ker D_{inc}$.

Example 3.81 (Quasi-proper leading term by an incidence matrix). Consider the nonlinear system

$$\begin{aligned} \alpha(x_2(t),x_3(t),t)\, x_2'(t) + x_1(t) - q_1(t) &= 0,\\ \beta(x_3(t),t)\, x_3'(t) + x_2(t) - q_2(t) &= 0,\\ x_3(t) - q_3(t) &= 0, \end{aligned}$$

with smooth functions α and β. Assume the function $\alpha(x_2,x_3,t)$ vanishes identically for $x_2 \leq 1$ and remains positive elsewhere. The function β has no zeros at all. This system cannot be written globally as a DAE with proper leading term. However, choosing $D_{inc} = \operatorname{diag}(0,1,1)$ we obtain the DAE with quasi-proper leading term

$$\begin{bmatrix} 0 & \alpha(x_2(t),x_3(t),t) & 0 \\ 0 & 0 & \beta(x_3(t),t) \\ 0 & 0 & 0 \end{bmatrix} \Big(\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D_{inc}} x(t) \Big)' + x(t) - q(t) = 0, \qquad (3.146)$$

with

$$f(y,x,t) = \begin{bmatrix} 0 & \alpha(x_2,x_3,t) & 0 \\ 0 & 0 & \beta(x_3,t) \\ 0 & 0 & 0 \end{bmatrix} y + x - q(t), \quad d(x,t) = \begin{bmatrix} 0 \\ x_2 \\ x_3 \end{bmatrix} = D_{inc}x,$$

and $N_A = \ker D_{inc}$. We introduce the open connected sets

$$\mathcal{G}_+ = \{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x_2 > 1\}, \quad \mathcal{G}_- = \{(x,t) \in \mathbb{R}^3 \times \mathbb{R} : x_2 < 1\},$$

and consider the DAE on these sets separately. On $\mathcal{G}_+$, this is a DAE with properly stated leading term. Further, computing an admissible matrix function sequence, one knows the DAE to be regular with index 3 and characteristics $r_0 = r_1 = r_2 = 2$, $r_3 = 3$.

In contrast, on $\mathcal{G}_-$, the leading term of the DAE is no longer properly stated and N_A is a proper subspace of $\ker f_y$. Observe that f_y has constant rank on $\mathcal{G}_-$, and $R_{new} := \operatorname{diag}(0,0,1)$ is the orthoprojector along $\ker f_y$. Replacing in the DAE (3.146) the function d by $d_{new} = R_{new}d = R_{new}D_{inc}x$ we arrive at the DAE

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & \beta(x_3(t),t) \\ 0 & 0 & 0 \end{bmatrix} \Big(\underbrace{\begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D_{inc}} x(t) \Big)' + x(t) - q(t) = 0, \qquad (3.147)$$

with properly stated leading term, and which is regular with index 2 and characteristic values $r_0 = 1, r_1 = 2, r_2 = 3$. That means that the reformulated DAE (3.147) is regular on $\mathcal{G}_-$ in the sense of Definition 3.28, whereas this definition does not apply to the original quasi-proper DAE (3.146). □

Similarly as in this example, the leading term in a quasi-proper DAE is often locally somewhat too generously stated and could be reformulated locally in a proper version. Nevertheless we agree to speak of regularity regions also in those cases.

Definition 3.82. Let the DAE (3.1) satisfy Assumption 3.1 and have a quasi-proper leading term. The open connected set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ is said to be a *regularity region* of this DAE, if f_y has constant rank on $\mathcal{G}$ and the DAE can be reformulated on $\mathcal{G}$ such that the resulting DAE has a properly stated leading term and is regular on $\mathcal{G}$ in the sense of Definition 3.28.

We emphasize at this point that, as in the above example, a proper reformulation comes along with a lower level smoothness demand concerning the solution. The following proposition provides sufficient conditions for proper reformulations.

Proposition 3.83. *Let the DAE (3.1) satisfy Assumption 3.1 and have a quasi-proper leading term. Let* $\ker f_y$ *be a* $\mathcal{C}^1$*-subspace on the open set* $\mathbb{R}^n \times \mathcal{G}$, $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$. *Let* $\ker f_y$ *be independent of* y *and* x *there,*

$$\ker f_y(y,x,t) =: N_f(t), \quad (y,x,t) \in \mathbb{R}^n \times \mathcal{G}.$$

Let $R_{new}(t) \in L(\mathbb{R}^n)$ *denote the orthoprojector along* $N_f(t)$, *and further*

$$d_{new}(x,t) := R_{new}(t)d(x,t), \quad (x,t) \in \mathcal{G}.$$

Then the DAE

$$f((d_{new}(x(t),t))' - R'_{new}(t)d(x(t),t),x(t),t) = 0 \tag{3.148}$$

is a reformulation of equation (3.1), which has a properly involved derivative on $\mathbb{R}^n \times \mathcal{G}$.
Each solution of the DAE (3.1) that resides in $\mathcal{G}$ *is also a solution of the new DAE (3.148). Conversely, if* x_* *is a solution of (3.148) with values in* $\mathcal{G}$ *and* $u_* := d(x_*(.),.)$, *then* x_* *is a solution of the DAE (3.1), supposing the part* $(I - R_{new})u_*$ *is also continuously differentiable.*

Proof. Owing to the quasi-proper leading term it holds that $N_A \oplus \operatorname{im} d_x = \mathbb{R}^n$. Denote by R the projector function onto $\operatorname{im} d_x$ along N_A. On $\mathbb{R}^n \times \mathcal{G}$ the relations

$$\operatorname{im}(I-R) = N_A \subseteq N_f = \ker f_y = \ker R_{new}$$

are valid, and therefore $R_{new}(I-R) = 0$, $R_{new} = R_{new}R$. It follows that

$$\operatorname{im} d_{new\,x} = \operatorname{im} R_{new} d_x = \operatorname{im} R_{new} R = \operatorname{im} R_{new},$$

and hence $\mathbb{R}^n = \ker R_{new} \oplus \operatorname{im} R_{new} = \ker f_y \oplus \operatorname{im} d_{new\,x}$.
Let x_* be a solution of the DAE (3.1) with path in $\mathcal{G}$. Then $u_* := d(x_*(.),.)$ is continuously differentiable. R_{new} is also continuously differentiable since $\ker f_y$ is a $\mathcal{C}^1$-subspace. Then, $u_{new\,*} := R_{new}d(x_*(.),.) = R_{new}u_*$ is continuously differentiable, too. We derive

$$\begin{aligned} f\big((d_{new}(x_*(t),t))' - R'_{new}(t)d(x_*(t),t),x_*(t),t\big) \\ = f\big((R_{new}(t)u_*(t))' - R'_{new}(t)u_*(t),x_*(t),t\big) \\ = f\big(R_{new}(t)u'_*(t),x_*(t),t\big) = f(u'_*(t),x_*(t),t) = 0, \end{aligned}$$

so that x_* solves the reformulated DAE (3.148).
Solutions x_* of (3.148) are continuous with $R_{new}u_*$ being continuously differentiable. The extra smoothness demand ensures the continuous differentiability of u_*. □

3.13 Notes and references

(1) The material of this chapter is to a large extent new. In constructing the admissible matrix function sequences it follows the lines of [168, 171], which consider DAEs of the form

$$A(d(x(t),t))' + b(x(t),t) = 0,$$

and

$$A(x(t),t)(D(t)x(t))' + b(x(t),t) = 0,$$

respectively. Now we consider fully implicit equations including both previous versions.
The tractability index concept is fully consistent with the knowledge of Hessenberg DAEs. Section 3.5 generalizes the special index-3 results obtained in [200] to Hessenberg form DAEs with arbitrary size.
The achievements concerning DAEs in circuit simulation in Section 3.6 reflect ideas from [70] and [207].
The local solvability assertions in Section 3.7 take up the decoupling ideas of [96], [205] and [211] and put them in a more general context.

(2) There are various interrelations between standard form DAEs and DAEs with proper or quasi-proper leading terms. We believe in the general possibility of formulating DAE models in applications at the very beginning with properly stated derivative terms, as is the case in circuit simulation. That is, one creates more precise models than standard form DAEs can be.
It seems that till now, owing to the well-developed theory on standard form DAEs (including numerical integration methods), one often transforms models that are originally in a properly stated version into standard form (e.g., [118]). This means that supposing continuously differentiable solutions and taking the total derivative in (3.1) one turns from (3.1) to the standard form DAE

$$f(d_x(x(t),t)x'(t) + d_t(x(t),t), x(t),t) = 0. \tag{3.149}$$

However this form again hides the precise information on how the derivative is packet in. We do not recommend turning from the precise model (3.1) to (3.149) for numerical integration, etc. In circuit simulation, it is a well-known experience that numerical integration performs better when using the precise model. Furthermore, often the dimensions are very large and the functions f, d, respectively $\mathfrak{f}$, satisfy low smoothness requirements only. From these points of view, it is rather worse to turn from equation (3.1) to the standard form version (3.149) in practice.

The opposite question of whether a given standard form DAE

$$\mathfrak{f}(x'(t),x(t),t) = 0, \tag{3.150}$$

where $\mathfrak{f} : \mathbb{R}^m \times \mathcal{D}_f \times \mathcal{I}_f \longrightarrow \mathbb{R}^k$ is continuous with continuous partial derivatives $\mathfrak{f}_{x'}, \mathfrak{f}_x$, can be reformulated as a DAE with a properly stated leading term or at least with a quasi-proper leading term is less simple.

If there is a nontrivial, possibly time-varying $\mathcal{C}^1$-subspace N in $\mathbb{R}^m$ such that

$$N(t) \subseteq \ker \mathfrak{f}_{x'}(x',x,t), \quad (x',x,t) \in \mathbb{R}^m \times \mathcal{D}_{\mathfrak{f}} \times \mathcal{I}_{\mathfrak{f}}, \tag{3.151}$$

we find (cf. Appendix A.4) a continuously differentiable projector valued function P such that $\ker P = N = \operatorname{im}(I-P)$. It holds that

$$\mathfrak{f}_{x'}(x',x,t)(I-P(t)) = 0, \quad (x',x,t) \in \mathbb{R}^m \times \mathcal{D}_{\mathfrak{f}} \times \mathcal{I}_{\mathfrak{f}},$$

and hence

$$\mathfrak{f}(x',x,t) - \mathfrak{f}(P(t)x',x,t) = \int_0^1 \mathfrak{f}_{x'}(sx' + (1-s)P(t)x',x,t)(I-P(t))x' \mathrm{d}s = 0,$$

thus $\mathfrak{f}(x',x,t) \equiv \mathfrak{f}(P(t)x',x,t)$, and equation (3.150) is the same as

$$\mathfrak{f}(P(t)x'(t),x(t),t) = 0. \tag{3.152}$$

In the next step we turn to

$$\mathfrak{f}((P(t)x(t))' - P'(t)x(t),x(t),t) = 0, \tag{3.153}$$

and this latter form suggests that solutions should be in $\mathcal{C}^1_P(\mathcal{I},\mathbb{R}^n)$ instead of $\mathcal{C}^1(\mathcal{I},\mathbb{R}^n)$. The DAE (3.153) has at least a quasi-proper leading term.
The DAE (3.153) has a proper leading term, if N and $\ker \mathfrak{f}_{x'}$ coincide. We emphasize that the latter requires $\ker \mathfrak{f}_{x'}(y,x,t)$ to be a subspace independent of the variables x' and x, as it is supposed, e.g., in [96], [160].
If $\mathfrak{f}_{x'}$ has a constant nullspace—as it is often the case in applications—also a constant projector P can be chosen, and equation (3.153) simplifies to

$$\mathfrak{f}((Px(t))',x(t),t) = 0. \tag{3.154}$$

Often a standard form DAE can be changed into a DAE with at least quasi-proper leading term by substituting an incidence matrix (see Section 3.12).
The question of whether a general standard form DAE (3.150), whose leading nullspace $\ker \mathfrak{f}_{x^1}$ depends on (x',x), can be reformulated to a DAE with properly stated leading term is unsolved. No general rules are in sight for this task. However, if it works, one can expect advantages concerning solvability and numerical treatment as in Example 3.3.

(3) Geometric methods, treating systems of smooth differential equations—among them DAEs—such as jet varieties, avoid the difficulties concerning drift and perturbation by consequently working just on related manifolds. The particular *geometric reduction* procedure in [189, Chapter IV] (also [145]), uses local parametrization and the subimmersion theorem for providing a sequence of (sub)manifolds

$$\mathcal{M}_0 \supset \mathcal{M}_1 \supset \cdots \supset \mathcal{M}_{\mu-1} = \mathcal{M}_\mu$$

as well as a vector field defined just on the final manifold. Thereby, certain constant-rank requirements are to support the manifold structure. Then the flow given on the final manifold is studied, in particular singularities of this flow are addressed. A local version of this geometric reduction procedure is developed in [194], and it is pointed out how additional singularities may occur in every step on the way to a final manifold.

In contrast, we aim for an analysis of DAEs which persists in view of arbitrarily small perturbations, similarly as it is done with the completion ODEs; however, we proceed without any derivative array functions. We emphasize several aspects concerning perturbations.

Perturbations may force the solutions to leave the constraint set of the unperturbed DAE. In particular, for linear constant coefficient systems $Ex'(t) + Fx(t) = 0$ the flow is restricted to the finite eigenspace of the matrix pencil, that is, to the range of the spectral projector $\operatorname{im} \Pi_{\mu-1}$. The subspace $\operatorname{im} \Pi_{\mu-1}$ is at the same time the set of consistent initial values for the homogeneous DAE. A nontrivial excitation q may force the flow of the DAE $Ex'(t) + Fx(t) = q(t)$ to spread out over all in $\mathbb{R}^m$. The sets of consistent values strongly depend on q.

Following the idea of characterizing linear DAEs by characterizing just the coefficient pair $\{E, F\}$ *independently of the particular right-hand side* q, our goal is a perturbation invariant characterization of general DAEs. In this context we are not interested in working out the particular constraint sets. In our view, generating the obvious and hidden constraint of a DAE is then an essential part of the particular solution procedure.

The DAEs arising from applications are nothing else than models describing physical phenomena just approximately. They are partly derived from physical laws, but other parts are created by means of quite voluntary ansatz functions and parameter calibrations. Having this in mind we aim for a structural characterization that is invariant with respect to perturbations rather than for an explicit description of the solution varieties of special DAEs.

(4) In essence, in the present chapter we preclude rank changes of the matrix function f_y. We want to emphasize again that rank changes in $f_y(y,x,t)$ lead to somewhat critical problems. As pointed out in various case studies (e.g., in Section 2.9), the resulting critical points may have very different natures. There are quite harmless critical points which could be healed by means of smoother data, but there are also serious critical points, yielding singularities in the flow. We do not go here into further detail in this direction. We refer to Chapter 9 for a discussion of quasi-regular problems including harmless critical points. Our goal in the present chapter is just to discover the basic regularity conditions. As explained to a large extent already in Section 2.9 on linear DAEs, with the object of an analysis which meets rigorous low smoothness requirements, we have to put up with constant-rank requirements and critical points which are no longer visible in smoother systems. These arguments keep their value also for nonlinear DAEs. In general we see the constant-rank condition as a useful tool to detect critical points on early stages of the investigation.

(5) Properly stated leading terms were applied first in [10] (report, revised as [11]), in the context of a unified approach to linear DAEs and their adjoints, but not yet marked by this special name. Slightly later, in [113] (report, revised as [114]), the quasi-linear DAE

$$A(x(t),t)(d(x(t),t))' + b(x(t),t) = 0,$$

which has a separate leading term housing the derivative, was defined to have a properly formulated leading term on $\mathcal{D}_f \times \mathcal{I}_f$, if

$$\ker A(x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^m, \quad \text{for all } (x,t) \in \mathcal{D}_f \times \mathcal{I}_f,$$

and there is a projector function $R \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^n))$ such that

$$\ker A(x,t) = \ker R(t),\ \operatorname{im} d_x(x,t) = \operatorname{im} R(t),\ d(x,t) = R(t)d(x,t),\ (x,t) \in \mathcal{D}_f \times \mathcal{I}_f$$

(cf. [114, Definition 5.1]). A comparison makes clear that Definition 3.4 generalizes this former notion considerably. Equation (3.1) is not necessarily quasi-linear and, moreover, the conditions concerning the projector function R are now exchanged for the demand that the two subspaces $\ker f_y$, $\operatorname{im} d_x$ have to be transversal $\mathcal{C}^1$-subspaces. This is much less restrictive. In particular, now these subspaces may depend also on y and x. We dispense with the condition $d(x,t) = R(t)d(x,t)$.
Although the wording *properly stated leading term* sounds somewhat strange for fully implicit equations which do not show a separate leading term housing the derivative, we keep this traditional notion also for fully nonlinear DAEs. At the same time we also speak of *DAEs with properly involved derivatives*.

(6) The question of whether the set $\mathcal{M}_0(t)$ might actually be a proper subset of $\widetilde{\mathcal{M}}_0(t)$ remains unsolved in the fully implicit case, if $\ker f_y(y,x,t)$ depends on y. In Example 3.54 we have illustrated this situation.

(7) There are open questions concerning the extension of solutions. Having a local solutions of an index-1 DAE, one can extend these solutions as long as the solution does not leave the regularity region. Till now we do not see results on the maximal existence intervals as they are known for explicit regular ODEs. And there is no general answer to the question of whether there are extensions through critical points and what they look like. This highly interesting topic needs future research. We refer just to Examples 3.34, 3.36, 3.59, 3.64, and 3.60 for some typical situations.
Moreover, also maximal regularity regions and their borders need further investigation.

(8) As shown in Section 3.8, the regularity regions depend also on the jet coordinates. One could ask whether this is a technical deficit of the tractability index concept. This is not the case. For instance also the quest for a completion ODE is accompanied by the same problem. Revisit Example 3.59. The derivative array system of size 1 for the special DAE (3.104) is

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

The search for a completion ODE leads to the system

$$\begin{bmatrix} 1 & 1 \\ t^2 + \alpha & 0 \end{bmatrix} \begin{bmatrix} x_1^1 \\ x_2^1 \end{bmatrix} = \begin{bmatrix} t^2 + \alpha \\ \beta'(t) - 2tx_1 \end{bmatrix},$$

as well as to the equivalence of the conditions $t^2 + \alpha = 0$ and $x_1^1 + x_2^1 = 0$. Evidently, one is confronted with the same necessity for an advanced localization including the jet variables as it is the case for the regularity regions. We realize *differentiation index 1 on the open sets* $\mathcal{G}_-^{[1]}$ and $\mathcal{G}_+^{[1]}$.

(9) The basic ill-posedness due to the nonclosed range of the operator representing the linear IVP in a higher index DAE was pointed out in [155], and much work was done to apply methods known for ill-posed problems (cf. [110], [106], [107]). Most of the resulting *regularization methods* for DAEs consist of a singular perturbation of the original problem. Although deep results could be proved, except for a few cases having a nice physical background, these regularization methods did not earn much resonance in practice because of the numerical difficulties in solving the singularly perturbed problems.

(10) For a long time (e.g., [163], [205]) it was hoped that appropriate structural restrictions can be found for f to guarantee the structural condition (3.118), that is

$$\mathcal{F}x \in \mathcal{C}_*^{ind\,\mu}(\mathcal{I}, \mathbb{R}^m), \quad x \in \mathcal{D}_F,$$

for the operator setting. Certain conditions were in fact posed. An improved version of Proposition 3.70 is obtained in [205] by means of an advanced decoupling. Although a quite interesting class of index-2 DAEs satisfies the structural condition (3.120), this condition remains somewhat synthesized. It is not satisfied, e.g., in MNA equations arising in circuit simulation. Moreover, the search for further structural conditions, in particular those for index-3 DAEs, did not yield sufficient success. The proposals have been too cumbersome and partly dependent on computational procedures (cf. [200]), and hence, this way seems to have no future.
The background of the difficulties in this context is the fact that, if a certain linearization to a given nonlinear DAE has index $\mu > 1$, then this does not say anything about the neighboring linearizations (see, for instance, Example 3.34). The structural condition (3.118) was also used to ensure the same characteristics of the neighboring linearizations (e.g., [161]). In the present chapter, nonlinear DAEs are approached anew via the concept of regularity regions and linearizations. This time we dispense with structural restrictions.

(11) The most natural formulation of a network system in circuit simulation is the physically reasonable system (3.58), that is

$$\begin{aligned} A_C \frac{d}{dt} q(A_C^{\mathrm{T}} e, t) + A_R g(A_R^{\mathrm{T}} e, t) + A_L j_L + A_V j_V + A_I i_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0, \\ \frac{d}{dt} \phi(j_L, t) - A_L^{\mathrm{T}} e &= 0, \\ A_V^{\mathrm{T}} e - v_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0. \end{aligned} \tag{3.155}$$

Supposing continuously differentiable solutions and applying the chain rule to the network system (3.58), i.e., expressing

$$\frac{d}{dt} q(A_C^{\mathrm{T}} e, t) = C(A_C^{\mathrm{T}} e, t) A_C^{\mathrm{T}} e' + q_t(A_C^{\mathrm{T}} e, t), \quad \frac{d}{dt} \phi(j_L, t) = L(j_L, t) j_L' + \phi_t(j_L, t),$$

with

$$C(v,t) := \frac{\partial}{\partial v} q(v,t), \quad L(j,t) := \frac{\partial}{\partial j} \phi(j,t),$$

one obtains a DAE in standard formulation, namely

$$\begin{aligned} A_C C(A_C^{\mathrm{T}} e, t) A_C^{\mathrm{T}} e' + A_C q_t(A_C^{\mathrm{T}} e, t) + A_R g(A_R^{\mathrm{T}} e, t) + A_L j_L \\ + A_V j_V + A_I i_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0, \\ L(j_L, t) j_L' + \phi_t(j_L, t) - A_L^{\mathrm{T}} e &= 0, \\ A_V^{\mathrm{T}} e - v_s(A^{\mathrm{T}} e, j_L, j_V, t) &= 0. \end{aligned}$$

This resulting DAE is considered as the *conventional MNA formulation*. It is commonly used to apply results and software given for standard form DAEs.
On the other hand, introducing the charges and fluxes

$$\mathrm{q} := q(A_C^{\mathrm{T}} e, t) \quad \text{and} \quad \varphi := \phi(j_L, t)$$

as additional variables, we obtain the equation system

$$\begin{bmatrix} A_C \mathrm{q}' \\ \varphi' \\ 0 \\ 0 \\ 0 \end{bmatrix} + \underbrace{\begin{bmatrix} A_R g(A_R^{\mathrm{T}} e, t) + A_L j_L + A_V j_V + A_I i_s(A^{\mathrm{T}} e, j_L, j_V, t) \\ -A_L^{\mathrm{T}} e \\ A_V^{\mathrm{T}} e - v_s(A^{\mathrm{T}} e, j_L, j_V, t) \\ \mathrm{q} - q(A_C^{\mathrm{T}} e, t) \\ \varphi - \phi(j_L, t) \end{bmatrix}}_{b(\mathrm{q}, \varphi, e, j_L, j_V, t)} = 0$$

which is regarded as the *charge/flux oriented MNA formulation*. It also represents a DAE in standard form and at the same time a DAE with quasi-proper leading term and linear derivative term,

$$\begin{bmatrix} A_C & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \underbrace{\left(\begin{bmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathrm{q} \\ \varphi \\ e \\ j_L \\ j_V \end{bmatrix} \right)'}_{=\left(\begin{bmatrix} \mathrm{q} \\ \varphi \end{bmatrix} \right)'} + b(\mathrm{q}, \varphi, e, j_L, j_V, t) = 0.$$

The solution understanding of the last equation adopts that for (3.155) as it stands. The charge/flux-oriented MNA formulation is well established as intermediate for contracting and analyzing numerical integration methods to solve the original DAE (3.155). The last system has the useful proper reformulation

$$\begin{bmatrix} A_C & 0 \\ 0 & I \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \underbrace{\left(\begin{bmatrix} \bar{P}_C & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \mathrm{q} \\ \varphi \\ e \\ j_L \\ j_V \end{bmatrix} \right)'}_{=\left(\begin{bmatrix} \bar{P}_C \mathrm{q} \\ \varphi \end{bmatrix} \right)'} + b(\mathrm{q}, \varphi, e, j_L, j_V, t) = 0,$$

which is responsible for the fact that integration methods applied to this formulation behave quite well. Furthermore, this represents one of the origins of the idea of turning to DAEs whose derivative term is housed by an extra function (e.g., [114, 168]).

(12) We know from [25] that the *index* μ of a DAE is the smallest integer μ such that the derivative array system $\mathcal{E}_{\mu+1} = 0$ (cf. Section 3.10) determines the variable x^1 as a continuous function of x, t. To this end it is emphasized that the statement is taken to hold locally on an open subset in the basic space.
In contrast, at times the differentiation index is introduced without the explicit demand for open sets. From [105] we learn that the DAE has *differential index* μ, if μ is the minimal number of analytical differentiations such that the prolongated system $\mathcal{E}_{\mu+1} = 0$ can be transformed by algebraic manipulations into an explicit ODE system. No further comments on the nature of the set on which the vector field $\mathcal{S}$ should be given are added, but this may lead to diverging interpretations. In particular, in Example 3.76, one could think of accepting the equation

$$x'(t) = \mathcal{S}_\Omega(x(t), t), \ (x(t), t) \in \Omega,$$

as an underlying ODE and to say that the DAE has differential index 1 on the lower-dimensional subset Ω.
We adopt the original intention of [25], [45] to apply open sets in the basic spaces. In contrast, a different view comes from geometrical approaches which, supposing manifold structures, consequently work on the (sub)manifolds.

(13) Concerning the practical use of the underlying ODEs, the drift phenomenon needs careful compensation when applying numerical integration methods. This is a particular feature in the simulation of constrained mechanical motion. Several constraint stabilization methods and projection techniques have been developed, forcing the numerical solution to stay close to the constraint. We refer to [63] for a comprehensive survey.

(14) The derivative array approaches including the various reduction procedures are approved definite solution procedures rather than a characterization of the given DAE. They apply to smooth problems in standard form.
Our goal is a different one. We look for criteria characterizing the given DAE without solving this DAE in advance or supposing solvability. Since we do not at all use derivative array functions, we can do with low smoothness requirements. We use linearizations and the projector based structural decomposition of the originally given DAE. To our knowledge, this is the only such treatment. In this framework, not only the constant-rank condition concerning the proper statement of the derivative term, but also the additional constant-rank requirements on further levels of the admissible matrix function sequences are welcome tools to figure out regular problems as well as different kinds of critical points.

(15) We emphasize the great benefit of working with projectors against the use of basis functions. Given is an at least continuous matrix function $M : \mathcal{D}_M \subseteq \mathbb{R}^s \to L(\mathbb{R}^m, \mathbb{R}^k)$ which has constant rank on the open set $\mathcal{D}_M$. We would like to describe its nullspace. With $I - M^+M$, a continuous projector function globally defined on $\mathcal{D}_M$ is available. Thereby the size of s does not matter at all. If $\ker M$ is a $\mathcal{C}^1$-subspace, then this projector function is continuously differentiable.
In contrast, we can expect the existence of basis functions globally defined on $\mathcal{D}_M$ which span the $\mathcal{C}^1$-subspace $\ker M$ if $s = 1$ only. Otherwise there are merely local basis functions. We refer to Remark A.16 in the appendix for an illustrative example.

Part II
Index-1 DAEs: Analysis and numerical treatment

Part II constitutes a self-contained script on regular index-1 DAEs. It constitutes in essence an up-to-date improved and completed version of the early book [96]. While the latter is devoted to standard form DAEs, we now address DAEs of the form

$$f((D(t)x(t))',x(t),t) = 0,$$

with properly involved derivative.

This part starts with a chapter on the structural analysis of index-1 DAEs. It is shown that each solution of a regular index-1 DAE is actually a somewhat wrapped solution of an inherent explicit ODE. A certain decoupling function ω, resembling that in [96], plays its role. This inherent ODE is only implicitly given, but it is uniquely determined by the problem data. With this background, local solvability and perturbation results are proved.

In the chapter on numerical integration, backward differentiation formulas and certain classes of Runge–Kutta methods and general linear methods that are suitable for DAEs are discussed. Then we concentrate on the question of whether a given integration method passes the wrapping unchanged and is handed over to the inherent explicit ODE. The answer appears not to be a feature of the method, but a property of the DAE formulation. If the subspace $\operatorname{im} D(t)$ is actually time-invariant, then the integration method reaches the inherent explicit ODE unchanged. This makes the integration smooth to the extent to which it may be smooth for explicit ODEs. Otherwise one has to expect additional serious stepsize restrictions.

The third chapter addresses stability topics. Contractivity and dissipativity of DAEs are introduced, and it is discussed how integration methods reflect the respective flow properties. Again, one can benefit from a time-invariant subspace $\operatorname{im} D(t)$. Finally, stability in the sense of Lyapunov is addressed and the related solvability assertions on infinite intervals are allocated.

Chapter 4
Analysis

This chapter is devoted to the analysis of nonlinear regular index-1 DAEs of the form

$$f((D(t)x(t))',x(t),t) = 0,$$

which contains m equations and m unknown functions. We want to introduce the analysis of such DAEs explaining their inner structure. In particular, in the following chapter, this serves as helpful background for understanding how numerical integration methods work.

The present chapter is self-contained. Neither the general analysis in Chapter 3 devoted to fully nonlinear arbitrarily high index DAEs

$$f((d(x(t),t))',x(t),t) = 0,$$

nor the general linear theory given in Chapter 2 are supposed to be known. Of course, the presentations are consistent.

The chapter is organized as follows. The basic assumptions and notions are collected in Section 4.1. Section 4.2 provides solvability and perturbation results by means of a structural decoupling of the DAE into the inherent explicit ODE and a certain part wrapping up the ODE solutions to become DAE solutions. Then we describe in Section 4.3 how one can compute consistent initial values.

4.1 Basic assumptions and notions

Looking at the formulation of the DAE

$$f((D(t)x(t))',x(t),t) = 0, \tag{4.1}$$

it is natural to search for continuous solutions x with a continuously differentiable part Dx. Therefore, we introduce

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_4,

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

as the solution space of the DAE (4.1). The same function space was already used for linear DAEs in Chapter 2 (cf. Section 2.6.1, also Definition 3.2).
Throughout the whole chapter, we assume the following assumption to be satisfied.

Assumption 4.1. *Let f be a continuous function mapping $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$ to $\mathbb{R}^m$ and having continuous partial derivatives $f_y(y,x,t)$ and $f_x(y,x,t)$. $\mathcal{D}_f \subseteq \mathbb{R}^m$ is assumed to be an open domain and $\mathcal{I}_f \subseteq \mathbb{R}$ an interval. Let D be a continuous matrix function with constant rank that maps $\mathcal{I}_f$ to $L(\mathbb{R}^m,\mathbb{R}^n)$. Let the subspaces* $\ker f_y$ *and* $\operatorname{im} D$ *form $\mathcal{C}^1$-subspaces (see Definition A.19).*
Let the transversality condition

$$\ker f_y(y,x,t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad \forall y \in \mathbb{R}^n,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{4.2}$$

be valid, and finally, let $\ker f_y(y,x,t)$ *be independent of y and x.*

By Definition 3.4, the DAE (4.1) now has a *properly involved derivative* on $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$, except for the fact that Definition 3.4 requires a continuously differentiable matrix function D, while here as in Chapter 2 we accept also functions D being just continuous.
It is useful to operate with the *border projector* $R(t) \in L(\mathbb{R}^n)$ realizing the decomposition of $\mathbb{R}^n$ given by the transversality condition (4.2), such that

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

The function R is continuously differentiable as a projector function acting on $\mathcal{C}^1$-subspaces.

Lemma 4.2. *Assumption 4.1 implies the identities*

$$f(y,x,t) \equiv f(R(t)y,x,t), \quad f_y(y,x,t) \equiv f_y(R(t)y,x,t) \equiv f_y(y,x,t)R(t).$$

Proof. For $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, $y \in \mathbb{R}^n$, $\eta := (I - R(t))y$, we get

$$f(y,x,t) - f(R(t)y,x,t) = \int_0^1 f_y(sy + (1-s)R(t)y,x,t)\eta \, \mathrm{d}s = 0,$$

since $\eta \in \operatorname{im}(I - R(t)) = \ker f_y(sy + (1-s)R(t)y,x,t)$ independently of s. □

For obvious reasons, if $x_* \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ is a solution of (4.1), then the function values $x_*(t)$ must belong to the set

$$\widetilde{\mathcal{M}}_0(t) := \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : f(y,x,t) = 0\}, \tag{4.3}$$

and hence, in contrast to regular ODEs, the solution values of a DAE are restricted to a certain subset of $\mathbb{R}^m$. Supposing Assumption 4.1, for DAEs (4.1) with continuously differentiable D, the *obvious restriction set* or *obvious constraint* is given (cf. Definition 3.9, Lemma 3.11) as

$$\mathcal{M}_0(t) := \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : y - D'(t)x \in \operatorname{im} D(t),\ f(y,x,t) = 0\} = \widetilde{\mathcal{M}}_0(t),$$

for all $t \in \mathcal{I}_f$. Regarding Lemma 4.2, we can check the representation

$$\mathcal{M}_0(t) = \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : y \in \operatorname{im} D(t),\ f(y,x,t) = 0\} = \widetilde{\mathcal{M}}_0(t), \tag{4.4}$$

and this makes sense also in the case of D being just continuous. In consequence, speaking of the obvious restriction set or obvious constraint of a DAE (4.1) under Assumption 4.1 we have in mind the formula (4.4). Following the lines of Proposition 3.10, one can show that, to each $x \in \mathcal{M}_0(t)$ there is always exactly one corresponding y, which means

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

Below we see, for regular index-1 DAEs, and just for those, the obvious constraint exclusively consists of solution values, that is, through each $\bar{t} \in \mathcal{I}_f, \bar{x} \in \mathcal{M}_0(\bar{t})$, there is a solution $x_*(\cdot)$ such that $x_*(\bar{t}) = \bar{x}$. For the moment, we refer to Example 3.7 which shows this property.

We introduce the subspace

$$S(y,x,t) := \{z \in \mathbb{R}^m :\ f_x(y,x,t)z \in \operatorname{im} f_y(y,x,t)\}$$

which plays its role in the following characterization of regular DAEs of index 1.

Definition 4.3. We call a nonlinear DAE (4.1) which satisfies Assumption 4.1 a *regular DAE with tractability index 1* on the open set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$, or more briefly, a *regular index-1 DAE* on $\mathcal{G}$, if

$$\ker D(t) \cap S(y,x,t) = \{0\} \quad \text{for all} \quad y \in \mathbb{R}^n, (x,t) \in \mathcal{G}.$$

If $\mathcal{G}$ is open and connected, and the DAE is regular with index 1 on $\mathcal{G}$, then $\mathcal{G}$ is said to be a regularity region of the DAE, also an index-1 regularity region. If $\mathcal{G} = \mathcal{D}_f \times \mathcal{I}_f$, we speak of a regular DAE with (tractability) index 1.

This definition is consistent with the previous ones concerning regular index-1 DAEs (see Definitions 2.25, 3.28, regarding the nonsingularity of the matrix function (4.8) below). A more subtle concept arises, if one is content with intersections $\ker D(t) \cap S(y,x,t)$ being trivial on an open set in $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$, only. We refer to Section 3.8 which addresses those questions.
In this chapter, we exclusively deal with regular index-1 DAEs, but often we omit the epithet *regular* as is common in the literature.

By Lemma A.9, the index-1 condition $\ker D(t) \cap S(y,x,t) = \{0\}$ is equivalent to

$$\ker D(t) \oplus S(y,x,t) = \mathbb{R}^m, \tag{4.5}$$

and, in turn, the decomposition (4.5) holds true, exactly if the matrix pencil (cf. Definition 1.4)

$$\lambda f_y(y,x,t)D(t) + f_x(y,x,t) \quad \text{is regular with } \textit{Kronecker index } 1. \tag{4.6}$$

Corresponding to Chapter 2, we denote the uniquely given projector realizing decomposition (4.5) by $\Pi_{can}(y,x,t)$, which means

$$\operatorname{im}\Pi_{can}(y,x,t) = S(y,x,t) \quad \text{and} \quad \ker\Pi_{can}(y,x,t) = \ker D(t). \tag{4.7}$$

Introducing a projector $Q_0(t) \in L(\mathbb{R}^m)$ onto $N(t) := \ker D(t)$ and again applying Lemma A.9, we know that the condition (4.5) is equivalent to the regularity of the matrix

$$G(y,x,t) := f_y(y,x,t)D(t) + f_x(y,x,t)Q_0(t), \tag{4.8}$$

independently of the special choice of projector $Q_0(t)$.

Further, for all $t \in \mathcal{I}_f$, we introduce $P_0(t) := I - Q_0(t)$ and, additionally, $D(t)^-$ to be the reflexive generalized inverse of $D(t)$ with the properties

$$D(t)D(t)^- = R(t), \qquad D(t)^-D(t) = P_0(t).$$

Since the matrix function $D(\cdot)$ is supposed to be continuous and to have constant rank we are allowed to assume, in the following, that $Q_0(\cdot)$, $P_0(\cdot)$ and $D(\cdot)^-$ are continuous as well (see Proposition A.17).

4.2 Structure and solvability of index-1 DAEs

In this section we analyze the inner structure of regular index-1 DAEs and provide results about the existence of solutions. First, we extract the inherent ordinary differential equation from the DAE (4.1). In contrast to the linear case, we do not expect to get it globally. However, a smart separation of components allows an elegant extraction locally as follows. For any vector x, we can write

$$x = P_0(t)x + Q_0(t)x = D(t)^-D(t)x + Q_0(t)x.$$

If we regard Lemma 4.2 then equation (4.1) can be expressed as

$$f(R(t)(D(t)x(t))', D(t)^-D(t)x(t) + Q_0(t)x(t), t) = 0. \tag{4.9}$$

Assuming, for a moment, that there is a solution $x_* \in \mathcal{C}_D(\mathcal{I},\mathbb{R}^m)$, we introduce two new functions by $u_*(t) := D(t)x_*(t)$, and $w_*(t) := D(t)^-(D(t)x_*(t))' + Q_0(t)x_*(t)$, for all $t \in \mathcal{I}$, such that $x_*(t) = D(t)^-u_*(t) + Q_0(t)w_*(t)$, and

$$D(t)w_*(t) = R(t)(D(t)x_*(t))', \quad Q_0(t)w_*(t) = Q_0(t)x_*(t),\ t \in \mathcal{I},$$

and hence the identity coming from (4.9) can be rewritten as

$$f((D(t)w_*(t))', D(t)^-u_*(t) + Q_0(t)w_*(t), t) = 0, \quad t \in \mathcal{I}. \tag{4.10}$$

The last expression suggests that we ask whether the equation in $\mathbb{R}^m$, with unknowns $w \in \mathbb{R}^m$, $u \in \mathbb{R}^n$, given by

$$f(D(t)w, D(t)^- u + Q_0(t)w, t) = 0 \tag{4.11}$$

implicitly determines a continuous solution function $w = \omega(u,t)$, such that $w_*(t) = \omega(u_*(t),t)$. For linear regular index-1 DAEs, i.e., in the case of

$$f(y,x,t) := A(t)y + B(t)x - q(t), \quad G(t) = A(t)D(t) + B(t)Q_0(t),$$

equation (4.11) simplifies to

$$(A(t)D(t) + B(t)Q_0(t))w + B(t)D(t)^- u - q(t) = 0,$$

which uniquely determines the function $w = -G(t)^{-1}(B(t)D(t)^- u - q(t)) =: \omega(u,t)$.

The next lemma provides such a desired function $\omega(u,t)$ yielding the local equivalence of (4.10) with $w_*(t) = \omega(u_*(t),t)$. As one can see later by Theorem 4.5, the function $\omega(u,t)$ enables us to decouple the entire dynamic part of the DAE (4.1) from the constraint part.

Lemma 4.4. *Let equation (4.1) be regular of index 1. For given $\bar{t} \in \mathcal{I}_f$, $\bar{x} \in \mathcal{M}_0(\bar{t})$, $\bar{y} \in \operatorname{im} D(\bar{t})$ such that $f(\bar{y},\bar{x},\bar{t}) = 0$, we introduce*

$$\bar{u} := D(\bar{t})\bar{x}, \quad \bar{w} := D(\bar{t})^- \bar{y} + Q_0(\bar{t})\bar{x}$$

and define

$$\mathcal{F}(w,u,t) := f(D(t)w, D(t)^- u + Q_0(t)w, t)$$

for (w,u,t) within a neighborhood $\mathcal{N}_{(\bar{w},\bar{u},\bar{t})} \subseteq \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ of $(\bar{w},\bar{u},\bar{t})$. Then, we find a neighborhood $\mathcal{N}_{(\bar{u},\bar{t})} \subseteq \mathbb{R}^n \times \mathbb{R}$ of $(\bar{u},\bar{t})$ and a continuous function

$$\omega : \mathcal{N}_{(\bar{u},\bar{t})} \to \mathbb{R}^m$$

satisfying $\omega(\bar{u},\bar{t}) = \bar{w}$ and

$$\mathcal{F}(\omega(u,t),u,t) = 0, \quad \text{for all } (u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}.$$

Furthermore, $\omega(u,t) = \omega(R(t)u,t)$, ω has the continuous partial derivative

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

for $(u,t) \in \mathcal{N}_{(\bar{u},\bar{t})}$ and, in particular,

$$\omega_u(\bar{u},\bar{t}) = -(G^{-1}f_x)(\bar{y},\bar{x},\bar{t})\, D(\bar{t})^-$$

with G defined in (4.8).

Proof. First, we have

$$\mathcal{F}(\bar{w},\bar{u},\bar{t}) = f(D(\bar{t})\bar{w}, D(\bar{t})^{-}\bar{u} + Q_0(\bar{t})\bar{w}, \bar{t}) = f(\bar{y},\bar{x},\bar{t}) = 0.$$

Additionally,

$$\begin{aligned}\mathcal{F}_w(\bar{w},\bar{u},\bar{t}) &= f_y\big(D(\bar{t})\bar{w}, D(\bar{t})^{-}\bar{u} + Q_0(\bar{t})\bar{w}, \bar{t}\big)D(\bar{t}) + f_x\big(D(\bar{t})\bar{w}, D(\bar{t})^{-}\bar{u} + Q_0(\bar{t})\bar{w}, \bar{t}\big)Q_0(\bar{t}) \\ &= f_y(\bar{y},\bar{x},\bar{t})D(\bar{t}) + f_x(\bar{y},\bar{x},\bar{t})D(\bar{t})Q_0(\bar{t}) = G(\bar{y},\bar{x},\bar{t})\end{aligned}$$

is nonsingular since the DAE was assumed to be regular of index 1. Now, the assumption follows from the implicit function theorem. □

Knowledge about the implicitly given function ω allows us to state the following theorem describing the inner structure of regular index-1 DAEs (4.1).

Theorem 4.5. *Each solution $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ of a regular index-1 DAE (4.1) can be represented as*

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

with the continuously differentiable function $u_*(\cdot) := D(\cdot)x_*(\cdot)$ satisfying the inherent ODE

$$u'(t) = R'(t)u(t) + D(t)\omega(u(t),t), \tag{4.12}$$

whereby the continuous function ω mapping from a neighborhood $\mathcal{D}_\omega$ of the set $\{(D(t)x_*(t),t) : t \in \mathcal{I}\}$ into $\mathbb{R}^m$ is implicitly given by

$$f\big(D(t)\omega(u,t), D(t)^{-}u + Q_0(t)\omega(u,t),t\big) = 0, \quad (u,t) \in \mathcal{D}_\omega.$$

The function ω has the continuous partial derivative

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

Proof. For any solution $x_* \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ of the regular index-1 DAE (4.1) we know all solution values $x_*(t)$ to be elements of $\mathcal{M}_0(t)$. Therefore, Lemma 4.4 can be applied to all points $(\bar{x},\bar{t})$, $\bar{x} := x_*(\bar{t})$, $\bar{t} \in \mathcal{I}$, and $\bar{y} = R(\bar{t})u_*'(\bar{t})$. By uniqueness and continuity arguments, we find a continuous function ω mapping from a neighborhood $\mathcal{D}_\omega$ of $\{(D(\bar{t})x_*(\bar{t}),\bar{t}) : \bar{t} \in \mathcal{I}\}$ to $\mathbb{R}^m$ with the properties

$$f(D(t)\omega(u,t), D(t)^{-}u + Q_0(t)\omega(u,t),t) = 0, \quad (u,t) \in \mathcal{D}_\omega,$$

$$\omega(u_*(t),t) = w_*(t) := D(t)^{-}u_*'(t) + Q_0(t)x_*(t), \quad u_*(t) := D(t)x_*(t) \tag{4.13}$$

and there is the continuous partial derivative

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

Consequently,

$$D(t)\omega(u_*(t),t) = R(t)u_*'(t) = (Ru_*)'(t) - R'(t)u_*(t) = u_*'(t) - R'(t)u_*(t)$$

since $(Ru_*)(t) = (RDx_*)(t) = (Dx_*)(t) = u_*(t)$. In this way, we know u_* satisfies the ODE

$$u'(t) = R'(t)u(t) + D(t)\omega(u(t),t).$$

Furthermore, expression (4.13) implies

$$Q_0(t)\omega(u_*(t),t) = Q_0(t)x_*(t) \quad \text{and} \quad D(t)^- u_*(t) = P_0(t)x_*(t),$$

and hence, the solution representation

$$x_*(t) = P_0(t)x_*(t) + Q_0(t)x_*(t) = D(t)^- u_*(t) + Q_0(t)\omega(u_*(t),t).$$

□

The solution representation given by Theorem 4.5 explains the inner structure of an index-1 DAE (4.1): The inherent ODE (4.12) describes the flow, the dynamic part, of the DAE in terms of the component $u_*(t) = D(t)x_*(t)$, while the remaining component $Q_0(t)x_*(t)$ is determined by the implicitly given function ω as

$$Q_0(t)x_*(t) = Q_0(t)\omega(u_*(t),t),$$

which reflects the constraint.

For a given index-1 DAE (4.1), the function ω, and so the ODE (4.12), is locally provided by Lemma 4.4, without supposing any solution. We emphasize the importance of this structure by the following definition.

Definition 4.6. For the regular index-1 DAE (4.1), we call the ordinary differential equation (4.12) *the inherent explicit regular ODE*, and we use the abbreviations *inherent ODE* and IERODE.

Proposition 4.7 below justifies this definition saying that the IERODE is uniquely determined by the index-1 DAE itself. One might think that the function $D\omega$ depends on the choice of the projector function Q_0, but it does not.
For linear regular index-1 DAEs (2.1), the ODE (4.12) is nothing else than the IERODE introduced in Definition 2.26, and we already know (see Proposition 2.33 or Theorem 2.39 in the more general context of fine decouplings) that the IERODE coefficients are independent of the projector choice. This is now confirmed once more.

Proposition 4.7. *Let the DAE (4.1) be regular of index 1.*

(1) *Then, the inherent ODE (4.12) is uniquely determined by the problem data functions f and D, which means it is independent of the choice of the projector Q_0.*
(2) *Furthermore, the time-varying subspace $\operatorname{im} D(t)$ of $\mathbb{R}^n$ is an invariant subspace of the inherent ODE (4.12), such that, if a solution $u(\cdot)$ exists on the interval $\mathcal{I} \subset \mathcal{I}_f$, and starts in $u(t_0) \in \operatorname{im} D(t_0)$ for some $t_0 \in \mathcal{I}$, then the solution value $u(t)$ belongs to $\operatorname{im} D(t)$ for all $t \in \mathcal{I}$.*

(3) *If the subspace* $\operatorname{im} D(t)$ *is actually time-invariant, then the solutions* $u(\cdot)$ *of the ODE (4.12), having a certain value* $u(t_0) \in \operatorname{im} D(t_0)$*, satisfy the simpler ODE*

$$u'(t) = D(t)\omega(u(t),t). \tag{4.14}$$

Proof. (1) We show that the vector field of the ODE (4.12) does not depend on the choice of the projector function Q_0. We assume $Q_0(t)$ and $\hat{Q}_0(t)$ to be two projectors onto $\ker D(t)$. Correspondingly, we get generalized inverses $D(t)^-$ and $\hat{D}(t)^-$. According to Lemma 4.4, we find two functions ω and $\hat{\omega}$ satisfying

$$f\big(D(t)\omega(u,t), D(t)^- u + Q_0(t)\omega(u,t), t\big) = 0$$

and

$$f\big(D(t)\hat{\omega}(u,t), \hat{D}(t)^- u + \hat{Q}_0(t)\hat{\omega}(u,t), t\big) = 0.$$

Let $\mathcal{D}_{\omega,\hat{\omega}}$ be their common definition domain. Regarding (4.12), we have to show that the corresponding vector fields coincide, which means

$$R'(t)u + D(t)\omega(u,t) = R'(t)u + D(t)\hat{\omega}(u,t) \tag{4.15}$$

for all $(u,t) \in \mathcal{D}_{\omega,\hat{\omega}}$. Since $\operatorname{im} \hat{Q}_0 = \ker D(t) = \operatorname{im} Q_0$, we know that (cf. Lemma A.3)

$$\hat{Q}_0(t) = Q_0(t)\hat{Q}_0(t)$$

as well as

$$\begin{aligned} \hat{D}(t)^- &= P_0(t)\hat{D}(t)^- + Q_0(t)\hat{D}(t)^- = D(t)^- R(t) + Q_0(t)\hat{D}(t)^- \\ &= D(t)^- + Q_0(t)\hat{D}(t)^- \end{aligned}$$

and we may conclude

$$f\big(D(t)\hat{\omega}(u,t), D(t)^- u + Q_0(t)(\hat{D}(t)^- u + \hat{Q}_0(t)\hat{\omega}(u,t)), t\big) = 0.$$

Introducing $\tilde{\omega}(u,t) := P_0(t)\hat{\omega}(u,t) + Q_0(t)(\hat{D}(t)^- u + \hat{Q}_0(t)\hat{\omega}(u,t))$, we see that

$$f\big(D(t)\tilde{\omega}(u,t), D(t)^- u + Q_0(t)\tilde{\omega}(u,t), t\big) = 0$$

is satisfied. Since ω is the locally uniquely defined function satisfying

$$f\big(D(t)\omega(u,t), D(t)^- u + Q_0(t)\omega(u,t), t\big) = 0,$$

we obtain

$$\omega(u,t) = \tilde{\omega}(u,t) = P_0(t)\hat{\omega}(u,t) + Q_0(t)(\hat{D}(t)^- u + \hat{Q}_0(t)\hat{\omega}(u,t)).$$

This implies

$$D(t)\omega(u,t) = D(t)\hat{\omega}(u,t)$$

and, consequently, (4.15) is true. In other words, the inherent ODE (4.12) is independent of the choice of the projector Q_0. It is uniquely defined by f and D.

(2) We show that $\operatorname{im} D(t)$ is an invariant subspace for the ODE (4.12). If $u(\cdot)$ is a solution existing on the interval $\mathcal{I}$, then the identity

$$(I-R(t))u'(t) = (I-R(t))(R'(t)u(t)+D(t)\omega(u,t)) = (I-R(t))R'(t)u(t), \quad t\in\mathcal{I},$$

is true. Using $v(t) := (I-R(t))u(t)$, we see that

$$\begin{aligned} v'(t) &= (I-R(t))u'(t) - R'(t)u(t) = (I-R(t))R'(t)u(t) - R'(t)u(t) \\ &= -R(t)R'(t)u(t) = -R'(t)u(t) + R'(t)R(t)u(t) = -R'(t)v(t). \end{aligned}$$

For $u(t_0)\in\operatorname{im} D(t_0)$, which means $v(t_0)=0$, the function $v(\cdot)$ vanishes identically and it holds that

$$u(t) = R(t)u(t) \in \operatorname{im} D(t), \text{ for all } t\in\mathcal{I}.$$

(3) Since the subspace $\operatorname{im} D(t)$ does not vary with t, the orthoprojector R_c onto $\operatorname{im} D(t)$ is also independent of t, and $R(t)R_c = R_c$. For the solutions under consideration, due to assertion (2), it holds that $u(t) = R_c u(t)$, thus $R'(t)u(t) = R'(t)R_c u(t) = (R(t)R_c)'u(t) = (R_c)'u(t) = 0$, and hence the respective term in (4.12) disappears. □

Example 4.8 (Decoupling function and regularity regions). Consider the semi-explicit DAE (cf. Example 3.7)

$$\begin{aligned} x_1'(t) + \beta x_1(t) &= 0, \\ x_1(t)^2 + x_2(t)^2 - 1 &= \gamma(t), \end{aligned}$$

on $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = [0,\infty)$. β is a real parameter. The real function γ is continuous on $\mathcal{I}_f$, and $1+\gamma(t)\geq 0$. We write this DAE in the form (4.1) with $n=1$, $m=2$,

$$f(y,x,t) = \begin{bmatrix} y+\beta x_1 \\ x_1^2+x_2^2-\gamma(t)-1 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad D(t) = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

as a DAE with properly stated leading term. Derive further

$$\mathcal{M}_0(t) = \{x\in\mathcal{D}_f : x_1^2+x_2^2-1-\gamma(t)=0\},$$

$$S(y,x,t) = \{z\in\mathbb{R}^2 : 2x_1z_1+2x_2z_2=0\}, \quad \ker D(t) = \{z\in\mathbb{R}^2 : z_1=0\},$$

and

$$Q_0(t) = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G(y,x,t) = \begin{bmatrix} 1 & 0 \\ 0 & 2x_2 \end{bmatrix}.$$

It becomes evident that $G(y,x,t)$ is nonsingular exactly if the intersection $S(y,x,t)\cap\ker D(t)$ is trivial, and this happens if $x_2\neq 0$. In consequence, the DAE has index 1 on the open connected sets

$$\mathcal{G}_+ := \{x \in \mathbb{R}^2 : x_2 > 0\} \times \mathcal{I}_f, \quad \mathcal{G}_- := \{x \in \mathbb{R}^2 : x_2 < 0\} \times \mathcal{I}_f,$$

being maximal index-1 regularity regions. The subspace $x_2 = 0$ constitutes the border between the regularity regions $\mathcal{G}_+$ and $\mathcal{G}_-$.
The canonical projector function

$$\Pi_{can}(y,x,t) = \begin{bmatrix} 1 & 0 \\ -\frac{x_1}{x_2} & 0 \end{bmatrix}$$

is defined for all $y \in \mathbb{R}$, $(x,t) \in \mathcal{G} := \mathcal{G}_+ \cup \mathcal{G}_-$. It grows unboundedly, if x_2 tends to zero. The decoupling function related to $\mathcal{G}_\pm$ now reads

$$\omega(u,t) = \begin{bmatrix} -\beta u \\ \pm(1+\gamma(t)-u^2)^{\frac{1}{2}} \end{bmatrix}, \quad (u,t) \in \mathrm{dom}_\omega$$

with $\mathrm{dom}_\omega := \{(u,t) \in \mathbb{R}^2 : \pm u < (1+\gamma(t))^{\frac{1}{2}}, t \in \mathcal{I}_f\}$, The IERODE is linear, $u'(t) = -\beta u(t)$. The function $D\omega$ has a smooth extension onto $\mathbb{R} \times \mathcal{I}_f$, and we put $D(t)\omega(u,t) = -\beta u$, $(u,t) \in \mathrm{dom}_{D\omega} := \mathbb{R} \times \mathcal{I}_f$.
To each arbitrary $(x_0,t_0) \in \mathcal{G}$, $x_0 \in \mathcal{M}_0(t_0)$, the IERODE has a unique solution such that $u(t_0) = x_{0,1}$, and a unique DAE solution results such that $x(t_0) = x_0$. The living interval of this solution may be finite or infinite, depending on the parameter β and the function γ.
For instance, if $\beta > 0$ and γ vanishes identically, the solution exists on the infinite interval, and $x(t)$ tends to $(0,1)^T$, if $t \to \infty$. Notice that in this case, from each border point between $\mathcal{G}_+$ and $\mathcal{G}_-$, two solutions emerge, one turns to $\mathcal{G}_+$, and the other to $\mathcal{G}_-$.
If $\beta < 0$, the solutions go the other way round, and there are no solutions emerging at the border points. In contrast, a solution starting in $\mathcal{G}$ ends up in finite time at a border point. Because of this critical flow behavior at the border, we call those border points *critical*. □

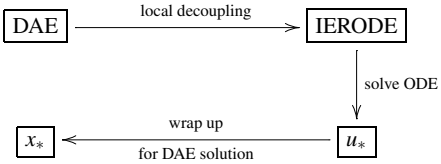


Fig. 4.1 Transfer of solvability results for ODEs to DAEs via local decoupling

The knowledge concerning the inner structure of nonlinear regular index-1 DAEs provided by Lemma 4.4 and Theorem 4.5 allows us to derive solvability results as well as error estimations, and also perturbation results. Mainly, we apply standard results for ODEs to the inherent ODE (4.12) and extend them to the DAE solution regarding the properties of the implicitly given function ω (see Figure 4.1).

Before presenting the results we formulate a useful lemma that provides us with a locally uniquely determined function ω^{pert} which plays the role of ω for slightly perturbed DAEs

$$f((D(t)x(t))',x(t),t) = q(t), \tag{4.16}$$

with a perturbation $q(t) \in \mathbb{R}^m$. Applying Definition 4.3 we see that the original equation (4.1) is a regular index-1 DAE if and only if its perturbed version (4.16) is regular of index-1.

Lemma 4.9. *Let the DAE (4.1) be regular of index 1.*

(1) *For given $\bar{t} \in \mathcal{I} \subset \mathcal{I}_f$, $\bar{x} \in \mathcal{M}_0(\bar{t})$, $\bar{y} \in \operatorname{im} D(\bar{t})$ such that $f(\bar{y},\bar{x},\bar{t}) = 0$, we introduce*

$$\bar{u} := D(\bar{t})\bar{x}, \quad \bar{w} := D(\bar{t})^- \bar{y} + Q_0(\bar{t})\bar{x}$$

and define

$$\mathcal{F}^{\mathrm{pert}}(w,u,t,q) := f(D(t)w, D(t)^- u + Q_0(t)w, t) - q \tag{4.17}$$

for (w,u,t,q) within a neighborhood $\mathcal{N}_{(\bar{w},\bar{u},\bar{t},0)} \subseteq \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ of $(\bar{w},\bar{u},\bar{t},0)$. Then, we find a neighborhood $\mathcal{N}_{(\bar{u},\bar{t},0)} \subseteq \mathbb{R}^n \times \mathbb{R}$ of $(\bar{u},\bar{t},0)$ and a unique continuous function

$$\omega^{\mathrm{pert}} : \mathcal{N}_{(\bar{u},\bar{t},0)} \to \mathbb{R}^m$$

satisfying $\omega^{\mathrm{pert}}(\bar{u},\bar{t},0) = \bar{w}$ and

$$\mathcal{F}^{\mathrm{pert}}(\omega^{\mathrm{pert}}(u,t,q),u,t,q) = 0 \quad \text{for all } (u,t,q) \in \mathcal{N}_{(\bar{u},\bar{t},0)}.$$

Furthermore, $\omega^{\mathrm{pert}}(u,t,q) = \omega^{\mathrm{pert}}(R(t)u,t,q)$, ω^{pert} has the continuous partial derivatives

$$\omega_u^{\mathrm{pert}}(u,t,q) = -(G_1^{-1} f_x)(y,x,t) D(t)^-, \quad \omega_q^{\mathrm{pert}}(u,t,q) = G_1^{-1}(y,x,t)$$

with

$$y := D(t)\omega^{\mathrm{pert}}(u,t,q), \quad x := D(t)^- u + Q_0(t)\omega^{\mathrm{pert}}(u,t,q)$$

for $(u,t,q) \in \mathcal{N}_{(\bar{u},\bar{t},0)}$ and, in particular,

$$\omega_u^{\mathrm{pert}}(\bar{u},\bar{t},0) = -(G_1^{-1} f_x)(\bar{y},\bar{x},\bar{t})\, D(\bar{t})^-, \quad \omega_q^{\mathrm{pert}}(\bar{u},\bar{t},0) = G_1^{-1}(\bar{y},\bar{x},\bar{t}).$$

(2) *Suppose $\mathcal{I}_c \subseteq \mathcal{I}_f$ is a compact interval and $\bar{x} : \mathcal{I}_c \to \mathbb{R}^m$, $\bar{y} : \mathcal{I}_c \to \mathcal{D}_f \subseteq \mathbb{R}^n$ are continuous functions satisfying*

$$f(\bar{y}(t),\bar{x}(t),t) = 0 \qquad \forall t \in \mathcal{I}_c.$$

Define

$$\bar{u}(t) := D(\bar{t})\bar{x}(t), \quad \bar{w}(t) := D(\bar{t})^- \bar{y}(t) + Q_0(\bar{t})\bar{x}(t) \qquad \forall t \in \mathcal{I}_c$$

and

$$\mathcal{F}^{[t]}(w,u,q) := f(D(t)w, D(t)^- u + Q_0(t)w, t) - q \tag{4.18}$$

for all (w,u,q) belonging to a neighborhood $\mathcal{N}_{(\bar{w}(t),\bar{u}(t),0)} \subseteq \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ of $(\bar{w}(t),\bar{u}(t),0)$. Then, we find a radius ρ independent of t and a continuous function

$$\omega^{[t]} : B_\rho(\bar{u}(t),0) \to \mathbb{R}^m$$

satisfying $\omega^{[t]}(\bar{u}(t),0) = \bar{w}(t)$ for all $t \in \mathcal{I}_c$ and

$$\mathcal{F}^{[t]}(\omega^{[t]}(u,q),u,q) = 0 \quad \text{for all } (u,q) \in B_\rho(\bar{u}(t),0) \text{ and } t \in \mathcal{I}_c.$$

The function $\omega^{\mathrm{pert}} : \{(u,t,q) \,|\, t \in \mathcal{I}_c, (u,q) \in B_\rho(\bar{u}(t),0)\} \to \mathbb{R}^m$ defined by $\omega^{\mathrm{pert}}(u,t,q) := \omega^{[t]}(u,q)$ is continuous with respect to t. Furthermore, $\omega^{\mathrm{pert}}(u,t,q) = \omega^{\mathrm{pert}}(R(t)u,t,q)$ for all $t \in \mathcal{I}_c$ and $(u,q) \in B_\rho(\bar{u}(t),0)$. Additionally, ω^{pert} has the continuous partial derivatives

$$\begin{aligned}
\omega_u^{\mathrm{pert}}(u,t,q) &= -(G_1^{-1} f_x)(y,x,t) D(t)^-, \\
\omega_q^{\mathrm{pert}}(u,t,q) &= G_1^{-1}(y,x,t)
\end{aligned}$$

with

$$y := D(t)\omega^{\mathrm{pert}}(u,t,q), \quad x := D(t)^- u + Q_0(t)\omega^{\mathrm{pert}}(u,t,q)$$

for $(u,q) \in B_\rho(\bar{u}(t),0)$ and $t \in \mathcal{I}_c$. In particular,

$$\begin{aligned}
\omega_u^{[t]}(\bar{u}(t),0) &= -(G_1^{-1} f_x)(\bar{y}(t),\bar{x}(t),\bar{t}) D(\bar{t})^-, \\
\omega_q^{[t]}(\bar{u}(t),0) &= G_1^{-1}(\bar{y}(t),\bar{x}(t),\bar{t}).
\end{aligned}$$

Notice that the function $\mathcal{F}^{\mathrm{pert}}$ extends the function $\mathcal{F}$ by the additional perturbation term q. More precisely,

$$\mathcal{F}^{\mathrm{pert}}(w,u,t,q) = \mathcal{F}(w,u,t) - q.$$

Proof. (1) This follows from the implicit function theorem analogously to the proof of Lemma 4.4. The function ω^{pert} extends the previous function ω in the sense

$$\omega^{\mathrm{pert}}(u,t,0) = \omega(u,t).$$

(2) The main work here is to show the existence of such a radius ρ that is independent of $t \in \mathcal{I}_c$. We show the existence of $\omega^{[t]}$ by constructing a fixed point map as follows. Let $\delta > 0$ be so small that

$$\begin{aligned}
\mathcal{N}_{(\bar{y},\bar{x})} &:= \left\{ \left(\bar{y}(t) + D(t)w_\delta, \bar{x}(t) + D(t)^- u_\delta, t\right) | \, t \in \mathcal{I}_c, |w_\delta| \le \delta, |u_\delta| \le \delta \right\} \\
&\subseteq \mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f.
\end{aligned}$$

Since $\bar{y}(\cdot)$, $\bar{x}(\cdot)$, $D(\cdot)$ and $D(\cdot)^-$ are continuous on $\mathcal{I}_c$ we get $\mathcal{N}_{(\bar{y},\bar{x})}$ to be a compact set and, thus, $f_y(\cdot)$ as well as $f_x(\cdot)$ is uniformly continuous on $\mathcal{N}_{(\bar{y},\bar{x})}$. Therefore, we

find an α such that $0 < \alpha \le \delta$ and for all $t \in \mathcal{I}_c$, $w_\delta \in B_\alpha(0)$ and $u_\delta \in B_\alpha(0)$ one has

$$\begin{aligned}&\Big|\big(f_y(\bar y(t)+D(t)w_\delta,\bar x(t)+D(t)^- u_\delta,t)-f_y(\bar y(t),\bar x(t),t)\big)D(t)\Big| \\ &\quad+\Big|\big(f_x(\bar y(t)+D(t)w_\delta,\bar x(t)+D(t)^- u_\delta,t)-f_x(\bar y(t),\bar x(t),t)\big)Q_0(t)\Big| \le \frac{1}{2c_1}\end{aligned} \tag{4.19}$$

with

$$c_1 := \max_{t\in\mathcal{I}_c} |(G_1(\bar y(t),\bar x(t),t))^{-1}|.$$

Define the fixed point map

$$H^{[t]}(w,z) := w - (F_w^{[t]}(\bar w(t),\bar z(t)))^{-1} F^{[t]}(w,z)$$

for $w \in B_\alpha(\bar w(t))$ and $z \in B_\rho(\bar z(t))$ with

$$z := (u,q), \quad \bar z(t) := (\bar u(t),0), \quad \rho := \alpha \cdot \min\Big\{1, \frac{1}{2c_1c_2}\Big\}$$

and

$$c_2 := \max_{t\in\mathcal{I}_c, |w_\delta|\le\alpha, |u_\delta|\le\alpha} (|f_x(\bar y(t)+D(t)w_\delta,\bar x(t)+D(t)^- u_\delta,t)D(t)^-|+1).$$

Next, we show that $H^{[t]}(\cdot,z)$ is a contractive mapping from $B_\alpha(\bar w(t))$ into $B_\alpha(\bar w(t))$. By definition of $F^{[t]}$ and (4.19) we have

$$F^{[t]}(\bar w(t),\bar z(t)) = f(\bar y(t),\bar x(t),t) = 0, \tag{4.20}$$

$$\big|(F_w^{[t]}(\bar w(t),\bar z(t)))^{-1}\big| \le c_1, \tag{4.21}$$

$$\big|F_w^{[t]}(\bar w(t),\bar z(t)) - F_w^{[t]}(w,z)\big| \le \frac{1}{c_2}, \tag{4.22}$$

$$\big|F_z^{[t]}(\bar w(t),\bar z(t)) - F_z^{[t]}(w,z)\big| \le c_2 \tag{4.23}$$

for all $t \in \mathcal{I}_c$, $w \in B_\alpha(\bar w(t))$ and $z \in B_\rho(\bar z(t))$. The contractivity of $H^{[t]}(\cdot,z)$ can be concluded from (4.21), (4.22) and

$$\begin{aligned}&|H^{[t]}(w_1,z) - H^{[t]}(w_1,z)| \\ &\quad= |w_1 - w_2 - (F_w^{[t]}(\bar w(t),\bar z(t)))^{-1} \int_0^1 F_w^{[t]}(sw_1+(1-s)w_2,z)\,\mathrm{d}s\,(w_1-w_2)| \\ &\quad\le c_1 \int_0^1 |F_w^{[t]}(\bar w(t),\bar z(t)) - F_w^{[t]}(sw_1+(1-s)w_2,z)|\,\mathrm{d}s\,|w_1-w_2| \\ &\quad\le c_1 \frac{1}{2c_1}|w_1-w_2| \quad \le \frac{1}{2}|w_1-w_2|\end{aligned}$$

for all $w_1, w_2 \in B_\alpha(\bar{w}(t))$, $z \in B_\rho(\bar{z}(t))$ and $t \in \mathcal{I}_c$. We see that $H^{[t]}(\cdot, z)$ is a self-mapping on $B_\alpha(\bar{w}(t))$ by (4.20)–(4.23) and

$$\begin{aligned}
&|H^{[t]}(w,z) - \bar{w}(t)| \\
&\quad = |w - \bar{w}(t) - (F_w^{[t]}(\bar{w}(t),\bar{z}(t)))^{-1}[F^{[t]}(w,z) - F^{[t]}(\bar{w}(t),\bar{z}(t))]| \\
&\quad \le c_1 |F_w^{[t]}(\bar{w}(t),\bar{z}(t)) - \int_0^1 F_w^{[t]}\big(sw+(1-s)\bar{w}(t), sz+(1-s)\bar{z}(t)\big)\,\mathrm{d}s|\,|w - \bar{w}(t)| \\
&\qquad\qquad + c_1 |\int_0^1 F_z^{[t]}(sw+(1-s)\bar{w}(t), sz+(1-s)\bar{z}(t))\,\mathrm{d}s|\,|z - \bar{z}(t)| \\
&\quad \le c_1 \int_0^1 |F_w^{[t]}(\bar{w}(t),\bar{z}(t)) - F_w^{[t]}\big(sw+(1-s)\bar{w}(t), sz+(1-s)\bar{z}(t)\big)|\,\mathrm{d}s\,|w - \bar{w}(t)| \\
&\qquad\qquad + c_1 c_2 |z - \bar{z}(t)| \\
&\quad \le \frac{1}{2}|w - \bar{w}(t)| + c_1 c_2 |z - \bar{z}(t)| \le \frac{1}{2}\alpha + c_1 c_2 \rho \le \alpha
\end{aligned}$$

for all $t \in \mathcal{I}_c$, $w \in B_\alpha(\bar{w}(t))$ and $z \in B_\rho(\bar{z}(t))$. The Banach fixed point theorem provides a fixed point w of $H^{[t]}(\cdot, z)$. This means that there is a unique $w = w^{[t]}(z) \in B_\alpha(\bar{w}(t))$ such that $H^{[t]}(w^{[t]}(z), z) = w^{[t]}(z)$ for all $t \in \mathcal{I}_c$ and $z \in B_\rho(\bar{z}(t))$. Consequently,

$$F^{[t]}(w^{[t]}(z), z) = 0, \qquad \forall t \in \mathcal{I}_c\ \forall z \in B_\rho(\bar{z}(t)).$$

By standard arguments one obtains $w^{[t]}(z)$ to be continuously differentiable having the derivative

$$w_z^{[t]}(z) = -(F_w^{[t]}(w^{[t]}(z), z))^{-1} F_z^{[t]}(w^{[t]}(z), z), \qquad \forall t \in \mathcal{I}_c\ \forall z \in B_\rho(\bar{z}(t)).$$

Defining

$$\omega^{[t]}(u,q) := w^{[t]}(z) = w^{[t]}(u,q)$$

it remains to show that the function $\tilde{\omega}^{\mathrm{pert}}(u,t,q) := \omega^{[t]}(u,q)$ is continuous with respect to t. Since the locally defined function $\omega^{\mathrm{pert}}(u,t,q)$ from part (1) of this lemma is unique, we can conclude that the function $\tilde{\omega}^{\mathrm{pert}}(u,t,q)$ equals $\omega^{\mathrm{pert}}(u,t,q)$ on a neighborhood $\mathcal{N}_{(\bar{u}(\bar{t}),\bar{t},0)}$ for all $\bar{t} \in \mathcal{I}_c$. Since $\omega^{\mathrm{pert}}(u,t,q)$ is continuous with respect to t, also $\tilde{\omega}^{\mathrm{pert}}(u,t,q)$ is continuous with respect to t. Removing the tilde notation in $\tilde{\omega}^{\mathrm{pert}}(u,t,q)$, the assertion (2) is proven. □

Corollary 4.10 below is an extension of Theorem 4.5, and it can be proven analogously to the proof of Theorem 4.5.

Corollary 4.10. *Let the DAE (4.1) be regular with index 1 and $\mathcal{I}_c \subseteq \mathcal{I}_f$ be a compact interval. Then, each solution $x \in \mathcal{C}_D^1(\mathcal{I}_c, \mathbb{R}^m)$ of the perturbed DAE*

$$f((Dx)'(t), x(t), t) = q(t)$$

with $|(Dx)(t_0) - D(t_0)x^0|$ and $\|q\|_\infty$ being sufficiently small, can be represented as

$$x(t) = D(t)^- u(t) + Q_0(t)\omega^{\mathrm{pert}}(u(t), t, q(t)), \quad t \in \mathcal{I}_c$$

with the continuously differentiable function $u := Dx$ satisfying the (perturbed) IERODE

$$u'(t) = R'(t)u(t) + D(t)\omega^{\text{pert}}(u(t),t,q(t)), \quad u(t_0) = D(t_0)x^0 \tag{4.24}$$

and $\mathcal{F}^{\text{pert}}(\omega^{\text{pert}}(u,t,q),u,t,q) = 0$ for $\mathcal{F}^{\text{pert}}$ defined in (4.17).

Now we are prepared to state the main solvability and perturbation results.

Theorem 4.11. *(Solvability) Let the DAE (4.1) be regular of index 1.*

(1) *Through each $x_0 \in \mathcal{M}_0(t_0)$ there passes exactly one solution of the DAE (4.1). More precisely, we find an open interval $\mathcal{I} \subset \mathcal{I}_f$ and a solution $x \in \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$ satisfying $x(t_0) = x_0 \in \mathcal{M}_0(t_0)$.*

(2) *Let $\mathcal{I}_c \subseteq \mathcal{I}_f$ be a compact interval, $t_0 \in \mathcal{I}_c$. If $x_* \in \mathcal{C}^1_D(\mathcal{I}_c,\mathbb{R}^m)$ is a solution of the DAE (4.1), then all perturbed IVPs*

$$f((Dx)'(t),x(t),t) = q(t), \quad D(t_0)(x(t_0)-x^0) = 0, \quad x^0 \in \mathbb{R}^m, \quad q \in \mathcal{C}(\mathcal{I}_c,\mathbb{R}^m)$$

are uniquely solvable on $\mathcal{C}^1_D(\mathcal{I}_c,\mathbb{R}^m)$ supposing $\|q\|_\infty$ and the deviation $|D(t_0)(x^0 - x_*(t_0))|$ of the initial value x^0 are sufficiently small. The solution x of the perturbed system satisfies

$$\|x - x_*\|_\infty \le C(\,|D(t_0)x(t_0) - D(t_0)x_*(t_0)| + \|q\|_\infty),$$

while its differential component Dx satisfies

$$\max_{t_0\le s\le t} |D(s)x(s) - D(s)x_*(s)| \le e^{c_1(t-t_0)}\left(|D(t)x(t_0) - D(t)x_*(t_0)| + \frac{c_2}{c_1}\max_{t_0\le s\le t}|q(s)|\right),$$

with certain constants $C, c_1, c_2 > 0$.

Proof. (1) Since $x_0 \in \mathcal{M}_0(t_0)$, we may apply Lemma 4.4 (1) for $\bar{t} = t_0$ and $\bar{x} := x_0$ in order to obtain a function ω satisfying

$$f(D(t)\omega(u,t), D(t)^- u + Q_0(t)\omega(u,t),t) = 0 \tag{4.25}$$

and $Q_0(t_0)x_0 = Q_0\omega(D(t_0)x_0,t_0)$. Consider the inherent regular ODE (4.12)

$$u'(t) = R'(t)u(t) + D(t)\omega(u(t),t)$$

and notice that $(D(t_0)x_0,t_0) \in \mathcal{D}_\omega$. Since ω is continuously differentiable with respect to u, the Picard–Lindelöf theorem provides a unique continuously differentiable solution function $u(\cdot)$ existing on a certain neighborhood $\mathcal{I}_u \subseteq \mathcal{I}_f$ of t_0 such that $u(t_0) = D(t_0)x_0$, and $(u(t),t) \in \mathcal{D}_\omega$, for all $t \in \mathcal{I}_u$. Since R is a projector function we find $RR'R = 0$ and

$$\begin{aligned}((I-R(t))u(t))' &= u'(t) - R'(t)u(t) - R(t)u'(t)\\ &= D(t)\omega(u(t),t) - R(t)R'(t)u(t) - D(t)\omega(u(t),t)\\ &= -R(t)R'(t)(I-R(t))u(t).\end{aligned}$$

Regarding $(I-R(t))u(t_0)=0$ we may conclude $(I-R(t))u(t)=0$ for all $t\in\mathcal{I}_u$ and

$$R(t)u'(t) = D(t)\omega(u(t),t), \quad t\in\mathcal{I}_u. \tag{4.26}$$

Next, the function

$$x(t) := D(t)^- u(t) + Q_0(t)\omega(u(t),t), \quad t\in\mathcal{I}_u,$$

is continuous and

$$D(t)x(t) = D(t)D(t)^- u(t) = R(t)u(t) = u(t), \quad t\in\mathcal{I}_u,$$

which means $x\in\mathcal{C}^1_D(\mathcal{I}_u,\mathbb{R}^m)$. Furthermore,

$$x(t_0) = D(t_0)^- D(t_0)x_0 + Q_0(t_0)\omega(D(t_0)x_0,t_0) = P_0(t_0)x_0 + Q_0(t_0)x_0 = x_0,$$

and hence $x(\cdot)$ passes trough x_0. It remains to verify that $x(\cdot)$ satisfies the DAE (4.1). Inserting the definition of $x(\cdot)$ into (4.1) and using (4.26) we find

$$\begin{aligned}f((D(t)x(t))',x(t),t) &= f(u'(t),x(t),t) = f(R(t)u'(t),x(t),t)\\ &= f(D(t)\omega(u(t),t), D(t)^- u(t) + Q_0(t)\omega(u(t),t),t)\\ &= 0, \quad t\in\mathcal{I}_u.\end{aligned}$$

(2) We define $u_*(t) := D(t)x_*(t)$. Since $x_*(\cdot)$ solves the unperturbed DAE (4.1) we get a continuous function $y_*(t) := (D(t)x_*(t))'$ satisfying $f(y_*(t),x_*(t),t)=0$. Then, Lemma 4.9 (2) provides a radius $\rho>0$ and a function $\omega^{\text{pert}}(u,t,q)$ defined on $\{(u,t,q)\,|\,t\in\mathcal{I}_c,\ (u,q)\in B_\rho(u_*(t),0)\}$ such that

$$\mathcal{F}^{\text{pert}}(\omega^{\text{pert}}(u,t,q),u,t,q)=0 \quad \forall\,(u,q)\in B_\rho(u_*(t),0)\ \forall\, t\in\mathcal{I}_c.$$

This implies

$$f\big(D(t)\omega^{\text{pert}}(u,t,q), D(t)^- u + Q_0(t)\omega^{\text{pert}}(u,t,q),t\big) - q = 0$$

for all $(u,q)\in B_\rho(u_*(t),0)$ and $t\in\mathcal{I}_c$. We consider the IVP

$$u'(t) = R'(t)u(t) + D(t)\omega^{\text{pert}}(u(t),t,q(t)), \quad u(t_0) = D(t_0)x^0.$$

Using Peano's theorem we obtain a continuously differentiable solution $u(\cdot)$ on $\mathcal{I}_c$ for sufficiently small perturbations q. With the same arguments as in the proof of part (1), we see that

$$u(t) = R(t)u(t), \quad \forall\, t\in\mathcal{I}_c$$

and the function

$$x(t) := D(t)^{-}u(t) + Q_0(t)\omega^{\text{pert}}(u(t),t,q(t)), \quad t \in \mathcal{I}_c,$$

satisfies the perturbed IVP

$$f((Dx)'(t),x(t),t) = q(t), \quad D(t_0)x(t_0) = R(t_0)u(t_0) = u(t_0) = D(t_0)x^0.$$

It remains to prove the perturbation estimations. We know that

$$\begin{aligned} u'(t) - u'_*(t) = R'(t)(u(t) - u_*(t)) \\ + D(t)(\omega^{\text{pert}}(u(t),t,q(t)) - \omega^{\text{pert}}(u(t),t,0)) \end{aligned}$$

and

$$u(t_0) - u_*(t_0) = D(t_0)(x(t_0) - x_*(t_0)).$$

Taking the mean value we conclude

$$\begin{aligned} u'(t) - u'_*(t) = R'(t)(u(t) - u_*(t)) \\ + D(t)\int_0^1 \omega_u^{\text{pert}}(su(t) + (1-s)u_*(t),t,sq(t))\,\mathrm{d}s\,(u(t) - u_*(t)) \\ + D(t)\int_0^1 \omega_q^{\text{pert}}(su(t) + (1-s)u_*(t),t,sq(t))\,\mathrm{d}s\,q(t). \end{aligned}$$

Since $\mathcal{I}_c$ is assumed to be compact, also the set

$$\left\{(su(t) + (1-s)u_*(t),t,sq(t)) \,|\, t \in \mathcal{I}_c,\, s \in [0,1]\right\}$$

is compact and we obtain uniform bounds for the continuous functions R', $D\omega_u^{\text{pert}} = -DG_1^{-1}f_xD$ and $D\omega_q^{\text{pert}} = -DG_1^{-1}$. Hence, we find constants $c_1 > 0$ and $c_2 > 0$ such that

$$|u'(t) - u'_*(t)| \le c_1|u(t) - u_*(t)| + c_2|q(t)|, \quad \forall t \in \mathcal{I}_c,$$

and Gronwall's lemma implies

$$\max_{t_0\le\tau\le t} |u(\tau) - u_*(\tau)| \le e^{c_1(t-t_0)}\left(|u(t_0) - u_*(t_0)| + \tfrac{c_2}{c_1}\max_{t_0\le\tau\le t}|q(\tau)|\right).$$

Regarding $u(\tau) = D(\tau)x(\tau)$ for all $\tau \in [t_0,t]$, the assertion for the differential components is proven. Additionally, we find a constant $c_3 > 0$ such that

$$\|u - u_*\|_\infty = \max_{t\in\mathcal{I}_c}|u(t) - u_*(t)| \le c_3(|D(t_0)(x(t_0) - x_*(t_0))| + \|q\|_\infty). \tag{4.27}$$

Taking into consideration the solution representation, we derive

$$\begin{aligned} x(t) - x_*(t) &= D(t)^-(u(t) - u_*(t)) \\ &\quad + Q_0(t) \int_0^1 \omega_u^{\text{pert}}(su(t) + (1-s)u_*(t), t, sq(t))\,\mathrm{d}s\,(u(t) - u_*(t)) \\ &\quad + Q_0(t) \int_0^1 \omega_q^{\text{pert}}(su(t) + (1-s)u_*(t), t, sq(t))\,\mathrm{d}s\,q(t). \end{aligned}$$

Again, we find uniform bounds on $\mathcal{I}_c$ for the continuous functions D^-, $Q_0\omega_u^{\text{pert}} = -Q_0 G_1^{-1} f_x D^-$ and $Q_0\omega_q^{\text{pert}} = Q_0 G_1^{-1}$, thus

$$|x(t) - x_*(t)| \le c_4|u(t) - u_*(t)| + c_5|q(t)| \quad \forall t \in \mathcal{I}_c.$$

Together with (4.27), this leads to the perturbation estimation of the theorem. □

4.3 Consistent initial values

This section describes a way to compute consistent initial values y_0, x_0 for a fixed t_0 such that

$$f(y_0, x_0, t_0) = 0. \tag{4.28}$$

This task is relevant for starting integration methods (see Chapter 5). The value y_0 reflects the expression $R(t)(Dx)'(t_0)$. By definition, an initial value x_0 is consistent for

$$f((Dx)'(t), x(t), t) = 0 \tag{4.29}$$

if there is a solution of (4.29) through x_0. As seen in the section before, all values $x_0 \in \mathcal{M}_0(t_0)$ are consistent initial values for index-1 DAEs of the form (4.29). A pair (y_0, x_0) is called a consistent initialization if x_0 is a consistent value and y_0 satisfies (4.28).

We recall Assumption 4.1 and the property $f(y, x, t) = f(R(t)y, x, t)$ resulting from Lemma 4.2. The system (4.28) is underdetermined with m equations and the $n+m$ unknowns (y_0, x_0). Therefore, we aim to complete the system to a regular one. Before doing so, we notice that one sometimes seeks a so-called operation point x_0 satisfying

$$f(0, x_0, t_0) = 0.$$

This is possible by means of Newton-like methods supposing f_x to be nonsingular. An index-1 IVP is described by the DAE (4.29) and the initial condition

$$P_0(t_0)(x(t_0) - x^0) = 0, \quad \text{or equivalently,} \quad D(t_0)(x(t_0) - x^0) = 0.$$

We consider the system of equations

$$\begin{aligned} f(y_0, x_0, t_0) &= 0, \\ P_0(t_0)(x_0 - x^0) &= 0, \\ (I - R(t_0))y_0 &= 0, \end{aligned}$$

which can be condensed to the square system

$$f(y_0, x_0, t_0) = 0, \tag{4.30}$$

$$(I - R(t_0))y_0 + D(t_0)(x_0 - x^0) = 0. \tag{4.31}$$

Lemma 4.12. *The left-hand side of the system (4.30), (4.31) has a nonsingular Jacobian*

$$\mathfrak{J} = \begin{bmatrix} f_y & f_x \\ I - R(t_0) & D(t_0) \end{bmatrix}$$

with respect to y_0, x_0 if (4.29) has tractability index 1. The inverse of $\mathfrak{J}$ is given by

$$\mathfrak{J}^{-1} = \begin{bmatrix} D(t_0)G^{-1} & I - R(t_0) - D(t_0)G^{-1}f_x D^-(t_0) \\ Q_0(t_0)G^{-1} & D^-(t_0) - Q_0(t_0)G^{-1}f_x D^-(t_0) \end{bmatrix}.$$

We omit the arguments of $f_y = f_y(y_0, x_0, t_0)$, $f_x = f_x(y_0, x_0, t_0)$ and $G = G(y_0, x_0, t_0)$.

Proof. The Jacobian of the left-hand side of (4.30), (4.31) with respect to y_0, x_0 is given by

$$\mathfrak{J} = \begin{bmatrix} f_y & f_x \\ I - R(t_0) & D(t_0) \end{bmatrix}.$$

The nonsingularity of $\mathfrak{J}$ is investigated by looking for nontrivial solutions of

$$\begin{bmatrix} f_y & f_x \\ I - R(t_0) & D(t_0) \end{bmatrix} \begin{bmatrix} z_y \\ z_x \end{bmatrix} = 0.$$

Multiplying the second equation by $R(t_0)$, it leads to $D(t_0)z_x = 0$ and, consequently, also $(I - R(t_0))z_y = 0$. Using this, the first equation reads

$$f_y R(t_0)z_y + f_x Q_0(t_0)z_x = 0 \quad \text{or} \quad \underbrace{(f_y D(t_0) + f_x Q_0(t_0))}_{=G_1}(D^-(t_0)z_y + Q_0(t_0)z_x) = 0.$$

From the nonsingularity of G_1 one can conclude $D^-(t_0)z_y = 0$ and $Q_0(t_0)z_x = 0$. Altogether, $z_y = 0$ and $z_x = 0$. This means $\mathfrak{J}$ is nonsingular. The form of the inverse of $\mathfrak{J}$ can be confirmed by direct multiplication. □

The regularity of the nonlinear system (4.30), (4.31) to determine (y_0, x_0) makes it possible to apply Newton-like methods to solve the system. System (4.30), (4.31) has dimension $m + n$, which might be large. The introduction of a new variable $\eta := D^-(t_0)y_0 + Q_0(t_0)x_0$ reduces the dimension of the nonlinear system. We consider the system

$$f(D(t_0)\eta, P_0(t_0)x^0 + Q_0(t_0)\eta, t_0) = 0 \tag{4.32}$$

of dimension m with respect to η. It has a nonsingular Jacobian matrix for index-1 DAEs (4.29). As one can easily verify, consistent initializations (y_0, x_0) can be computed by calculating an η satisfying (4.32) and assembling y_0 and x_0 by

$$y_0 = D(t_0)\eta, \qquad x_0 = P_0(t_0)x^0 + Q_0(t_0)\eta.$$

4.4 Notes and references

(1) To a large extent, our presentation follows the lines of [96, 114]. In [96], similar decoupling functions ω are applied to investigate DAEs in standard form, whereas [114] comes up with a modification of this approach to quasi-linear DAEs of the form $A(x(t),t)(D(t)x(t))' + b(x(t),t) = 0$ with properly stated leading term.

(2) The demand for the nullspace $\ker f_y(y,x,t)$ to be independent of the variables y and x is not really a restriction in the given context. Each DAE that meets all other conditions in Assumption 4.1 can be easily modified to satisfy this requirement, too. Namely, choosing a $\mathcal{C}^1$-projector function onto $\operatorname{im} D$, $\tilde{R} : \mathcal{I}_f \to \mathbb{R}^n$, the equivalent, modified DAE

$$\tilde{f}((D(t)x(t))', x(t), t) = 0,$$

with $\tilde{f}(y,x,t) := f(\tilde{R}(t)y + \tilde{R}'(t)D(t)x, x, t)$, $\ker \tilde{f}_y(y,x,t) = \ker \tilde{R}(t)$, satisfies Assumption 4.1 in all detail.

(3) Definition 4.3 generalizes the corresponding index-1 definitions given for linear DAEs in Chapters 1 and 2.
In the present chapter, dealing exclusively with index-1 DAEs, we apply the notation $G(y,x,t)$, while in Chapters 1 and 2, where also higher level matrix functions come in, the corresponding special cases are G_1, respectively $G_1(t)$. We mention, that in Chapter 3, to handle fully nonlinear DAEs (4.1) of arbitrary index, we use the further generalization $G_1(x^1,x,t)$ of $G_1(t)$ which is slightly different from $G(y,x,t)$.

(4) As addressed in detail in Chapter 3, almost all DAE literature is devoted to standard form DAEs (3.150), i.e.

$$\mathfrak{f}(x'(t), x(t), t) = 0, \tag{4.33}$$

given by a smooth function $\mathfrak{f}$, and one usually applies the *differentiation index* (cf. Remark 3.72). As pointed out in Chapter 3, there are good reasons for supposing a constant-rank partial Jacobian $\mathfrak{f}_{x^1}(x^1,x,t)$, such that $\ker \mathfrak{f}_{x^1}$ becomes a $\mathcal{C}^1$-subspace. By [25, Proposition 2.5.2], the standard form DAE (4.33) has differentiation index 1, exactly if the so-called local pencil

$$\lambda \mathfrak{f}_{x^1}(x^1,x,t) + \mathfrak{f}_x(x^1,x,t) \quad \text{has Kronecker index 1} \tag{4.34}$$

uniformly for all arguments x^1, x, t.
As discovered by Ch. Lubich (cf. Example 9.3), if the subspace $\ker \mathfrak{f}_{x^1}(x^1, x, t)$ actually varies with (x^1, x), then it may happen that the DAE (4.33) has differentiation index 1, but its *perturbation index* is higher. We avoid this situation by supposing the nullspace $\ker \mathfrak{f}_{x^1}(x^1, x, t)$ to be independent of (x^1, x). Note that in applications one usually has such a constant nullspace. Then we put

$$N(t) := \ker \mathfrak{f}_{x^1}(x^1, x, t)$$

choose a $\mathcal{C}^1$-projector function P along N, and turn, as described in Section 3.13, from (4.33) to the DAE (3.153), that is to

$$f((P(t)x(t))', x(t), t) := \mathfrak{f}((P(t)x(t))' - P'(t)x(t), x(t), t) = 0, \tag{4.35}$$

which has a properly involved derivative. The matrix pencil (cf. (4.6))

$$\lambda f_y P + f_x = \lambda \mathfrak{f}_{x^1} P + (\mathfrak{f}_x - \mathfrak{f}_{x^1} P') = \lambda \mathfrak{f}_{x^1} + (\mathfrak{f}_x - \mathfrak{f}_{x^1} P')$$

is regular with Kronecker index 1, exactly if the matrix pencil (4.34) is so (see Lemma A.9). This shows that in this context the tractability index 1 coincides with the differentiation index 1, and hence the results given in this chapter for DAEs of the form (4.1) apply at the same time via (4.35) to the standard form DAEs (4.33).

(5) As an immediate consequence of Theorem 4.11, each DAE (4.1) being regular with index 1 has perturbation index 1.

Chapter 5
Numerical integration

Index-1 DAEs with properly involved derivative have the advantage that there exists a uniquely determined (by the problem data) inherent explicit ODE, which is not the case for standard form DAEs. Investigating numerical integration methods applied to DAEs with properly stated leading term, the central question is how the given method performs on this inherent ODE. We discuss backward differentiation formulas (BDFs), Runge–Kutta methods, and general linear methods (GLMs). In each case, it turns out to be reasonable, to seek a *numerically qualified DAE formulation*, which means a DAE with $\operatorname{im} D(t)$ being independent of t, since then the integration method is passed unchanged to the inherent ODE. In this way, additional restrictions on the integration stepsize, which arise when using DAEs in standard form, are avoided.

Section 5.1 communicates the basic idea by means of an example. Section 5.2 collects material on the methods applied to ODEs and standard form DAEs. Then Section 5.3 describes how these methods can be applied to DAEs with properly leading term. We provide in Section 5.4 a condition which ensures that the given integration method arrives unchanged at the inherent ODE. Then, Section 5.5 provides error estimations and convergence results. We mention that the next chapter on stability issues adds respective results concerning infinite intervals.

In the present chapter, the number n is used for two different quantities. On the one hand, as throughout this monograph, n denotes the dimension of the space where $D(t)$ is mapping into ($D(t)x \in \mathbb{R}^n$ for $x \in \mathbb{R}^m$). On the other hand, n describes the current numerical discretization (cf. x_n, t_n). We keep this common notation for discretizations. It should always be clear from the context which meaning of n is supposed.

In the present chapter we use the acronym ODE for explicit ODEs, as is quite common in numerical analysis.

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_5,

5.1 Basic idea

We apply several well-known ODE methods to DAEs. We start by means of special cases and examples to point out distinctive features.

First, consider the explicit Euler method as a prototype for explicit step by step integration methods. Applied to the explicit linear ODE

$$x'(t) = C(t)x(t) + q(t), \tag{5.1}$$

it reads

$$x_n = x_{n-1} + h(C(t_{n-1})x_{n-1} + q(t_{n-1})),$$

with the stepsize h. Clearly, for given x_{n-1}, the Euler formula uniquely determines the current approximation x_n. Rewriting the Euler method as

$$\frac{1}{h}(x_n - x_{n-1}) = C(t_{n-1})x_{n-1} + q(t_{n-1})$$

we see that the derivative is approximated by the backward difference quotient whereas the right-hand side is evaluated at the previous time point t_{n-1}. Following this idea also in the case of linear DAEs in standard form

$$E(t)x'(t) + F(t)x(t) = q(t) \tag{5.2}$$

we get the method

$$E(t_{n-1})\frac{1}{h}(x_n - x_{n-1}) + F(t_{n-1})x_{n-1} = q(t_{n-1}),$$

and for DAEs with a proper leading term

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \tag{5.3}$$

it follows that

$$A(t_{n-1})\frac{1}{h}(D(t_n)x_n - D(t_{n-1})x_{n-1}) + B(t_{n-1})x_{n-1} = q(t_{n-1}).$$

In the case of DAEs, the matrices $E(t_{n-1})$ and $A(t_{n-1})D(t_{n-1})$ are singular and, obviously, in both cases the current value x_n is no longer uniquely determined. In consequence, the explicit Euler method does not work for DAEs. The same difficulty arises when applying any other explicit method directly. So, if no special structure can be exploited, one is obliged to use implicit integration methods. As a prototype we consider the implicit Euler method. For linear explicit ODEs (5.1), the numerical solution x_n is given by

$$\frac{1}{h}(x_n - x_{n-1}) = C(t_n)x_n + q(t_n).$$

The derivative is again approximated by the backward difference quotient, but the right-hand side is evaluated at the new time point t_n. For linear DAEs (5.2) and (5.3), this idea results in the formulas

$$E(t_n)\frac{1}{h}(x_n - x_{n-1}) + F(t_n)x_n = q(t_n), \tag{5.4}$$

and

$$A(t_n)\frac{1}{h}(D(t_n)x_n - D(t_{n-1})x_{n-1}) + B(t_n)x_n = q(t_n). \tag{5.5}$$

This time, in each case, we obtain a unique solution x_n if the matrices $\frac{1}{h}E(t_n)+F(t_n)$ and $\frac{1}{h}A(t_n)D(t_n)+B(t_n)$, respectively, are nonsingular. It is not difficult to verify that both matrices are nonsingular if the local matrix pencils $\lambda E(t_n)+F(t_n)$ and $\lambda A(t_n)D(t_n)+B(t_n)$, respectively, are regular (see Definition 1.2) and the stepsize h is sufficiently small. As pointed out in Section 4.4, Note (4), index-1 DAEs, both standard form DAEs and DAEs with properly stated leading term, have regular matrix pencils. In the present chapter, we show that large classes of implicit integration methods work well for index-1 DAEs. For the moment we would like to emphasize, that also higher-index DAEs often exhibit regular local matrix pencils. Consequently, implicit numerical integration methods are often formally feasible, but they may generate values far away from the exact solution (cf. Chapter 8).

Before turning to general index-1 DAEs, we consider an example showing a surprising behavior of the implicit Euler method for DAEs. It causes unexpected extra stepsize restrictions compared to its behavior for explicit ODEs. Recall that A-stability is an important feature of numerical ODE methods which allows us to avoid stepsize restrictions caused by stability reasons. The explicit and implicit Euler method applied to the scalar ODE $x'(t) = \lambda x(t)$, with $\lambda < 0$, generate the recursions

$$x_n = (1+h\lambda)x_{n-1}, \quad \text{and} \quad x_n = \frac{1}{1-h\lambda}x_{n-1}.$$

In order to reflect the solution property $|x_n| < |x_{n-1}|$ appropriately, the stepsize restriction $|1+h\lambda| < 1$ or, equivalently, $h < \frac{2}{-\lambda}$, is required for the explicit Euler method. For the implicit Euler method, the corresponding condition $0 < \frac{1}{1-h\lambda} < 1$ is satisfied for any stepsize, which is much more comfortable.

Example 5.1 (The impact of the DAE formulation). Let λ be any real parameter $\lambda < 0$, $\lambda \neq 1$. Consider the DAE

$$(\lambda - 1)x_1' + \lambda t x_2' = 0, \tag{5.6}$$
$$(\lambda - 1)x_1 + (\lambda t - 1)x_2 = 0, \tag{5.7}$$

which has the smooth solutions

$$x_1(t) = -\frac{\lambda t - 1}{\lambda - 1}e^{\lambda(t-t_0)}x_2(t_0), \quad x_2(t) = e^{\lambda(t-t_0)}x_2(t_0). \tag{5.8}$$

The DAE can be written in standard form (5.2) as

$$\begin{bmatrix} \lambda-1 & \lambda t \\ 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} 0 & 0 \\ \lambda-1 & \lambda t-1 \end{bmatrix} x(t) = 0.$$

Turning to the slightly reformulated equivalent version of the given DAE

$$(\lambda-1)x_1' + (\lambda t x_2)' - \lambda x_2 = 0, \tag{5.9}$$
$$(\lambda-1)x_1 + (\lambda t - 1)x_2 = 0 \tag{5.10}$$

one has a DAE with properly stated leading term (5.3),

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

Choosing

$$D(t)^- = \begin{bmatrix} \frac{1}{\lambda-1} \\ 0 \end{bmatrix}, \quad Q_0(t) = \begin{bmatrix} 0 & -\frac{\lambda t}{\lambda-1} \\ 0 & 1 \end{bmatrix},$$

we obtain the decoupling function

$$\omega(u,t) = -\begin{bmatrix} \lambda \frac{t-1}{\lambda-1} \\ -1 \end{bmatrix} u, \quad D(t)\omega(u,t) = \lambda u.$$

Then, the IERODE associated to the DAE version with properly stated leading term applies to the variable $u = Dx = (\lambda-1)x_1 + \lambda t x_2$, and it reads

$$u' = \lambda u. \tag{5.11}$$

Further, observe that, turning from the variables x_1, x_2 to u and v with

$$u := (\lambda-1)x_1 + \lambda t x_2, \qquad v := x_2,$$

the system (5.6)–(5.7) is equivalent to

$$u' = \lambda v, \tag{5.12}$$
$$u = v. \tag{5.13}$$

Next, the implicit Euler method applied to the DAE (5.6)–(5.7) in standard form reads

$$(\lambda-1)\frac{1}{h}(x_{1,n} - x_{1,n-1}) + \lambda t_n \frac{1}{h}(x_{2,n} - x_{2,n-1}) = 0, \tag{5.14}$$
$$(\lambda-1)x_{1,n} + (\lambda t_n - 1)x_{2,n} = 0. \tag{5.15}$$

In terms of the transformed variables

$$u_n = (\lambda-1)x_{1,n} + \lambda t_n x_{2,n}, \qquad v_n = x_{2,n},$$

this leads to

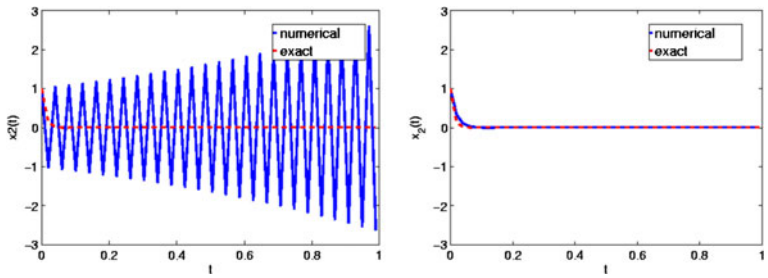
$$\frac{1}{h}(u_n - u_{n-1}) = \lambda v_{n-1}, \tag{5.16}$$

$$u_n = v_n, \tag{5.17}$$

and the recursion

$$\frac{1}{h}(u_n - u_{n-1}) = \lambda u_{n-1}.$$

Surprisingly, this represents the *explicit* Euler method for the ODE $u' = \lambda u$ involved in (5.12)–(5.13) although we have applied the *implicit* Euler method to the DAE (5.6)–(5.7). Consequently, the implicit Euler method applied to the DAE (5.6)–(5.7) provides stability preserving solution approximations only if $h < -\frac{2}{\lambda}$. Such extra stepsize restrictions have already been observed in [3]. Figure 5.1(a) shows the numerical solution x_2 for $\lambda = -100$ and $h = 0.0202$ for the implicit Euler method applied to (5.6)–(5.7). This is not what one expects from an A-stable method. In this sense, the A-stability gets lost when the method is applied to a DAE in standard form. In light of this observation one could think that the implicit Euler method is



(a) DAE formulation (5.6)–(5.7) (b) DAE formulation (5.9)–(5.10)

Fig. 5.1 Solution x_2 of the implicit Euler method with the stepsize $h = 0.0202$ and $\lambda = -100$

not well suited for integrating problems like (5.6)–(5.7). However, the situation becomes much nicer if one applies the same implicit Euler method to the same DAE written with properly stated leading term. Then we get

$$(\lambda - 1)\frac{1}{h}(x_{1,n} - x_{1,n-1}) + \frac{1}{h}(\lambda t_n x_{2,n} - \lambda t_{n-1} x_{2,n-1}) - \lambda x_{2,n} = 0,$$

$$(\lambda - 1)x_{1,n} + (\lambda t_n - 1)x_{2,n} = 0.$$

Using the transformed variables

$$u_n = (\lambda - 1)x_{1,n} + \lambda t_n x_{2,n}, \qquad v_n = x_{2,n},$$

the implicit Euler method for (5.6)–(5.7) implies

$$\frac{1}{h}(u_n - u_{n-1}) = \lambda v_n, \tag{5.18}$$

$$u_n = v_n, \tag{5.19}$$

and the resulting recursion has the form

$$\frac{1}{h}(u_n - u_{n-1}) = \lambda u_n,$$

which is exactly the implicit Euler method for the inherent ODE. In this case, the inherent dynamical system of the DAE with a properly stated leading term is solved by the same numerical method as the one applied to the original DAE. Figure 5.1(b) shows the solution x_2 for the same values as chosen previously. No stepsize restriction for stability reasons occur. □

The preceding example makes it clear that the way we formulate the DAE may have a significant influence on the numerical solution behavior. DAEs with a properly involved derivative seem to have an advantage in contrast to standard form DAEs. The following pages are devoted to a detailed analysis of numerical methods applied to DAEs with a properly involved derivative. Our special interest is directed to the question of how numerical methods applied to DAEs (4.1) act on the inherent ODE (4.12).

The theoretically ideal way of deriving a suitable numerical method for DAEs would be to formulate the method for the inherent ODE and then to compose a numerical solution x_n of the DAE from this ODE solution u_n (see Figure 5.2). However,

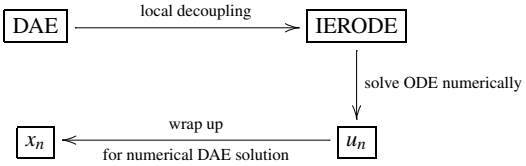


Fig. 5.2 Ideal construction of numerical methods for DAEs via local decoupling

this is not a realistic way to solve DAEs since the local decoupling function ω is usually not known. Even, if a decoupling is known, it is often very costly to compute. What we can do—and we should do it—is to investigate to what extent the numerical methods, being applied to the DAE, generate a correct integration of the IERODE, and whether the constraints are correctly reflected. In general, both concerns might be missed. On the other hand, as we point out in this chapter, there are numerical methods that fulfill the desired properties if they are applied to DAEs in a suitable way. To be precise, we formulate sufficient conditions guaranteeing that the numerical method applied to the DAE (4.1) generates exactly the same method for the IERODE (4.12). Thereby, knowledge of the inner structure of the nonlinear regular index-1 DAE described in Theorem 4.5 plays its role.

As prototypes of linear multistep methods and one-step methods, we discuss the BDF and Runge–Kutta methods. Afterwards we consider general linear methods. First we recall how these methods look for explicit ODEs, and then we consider modification for DAEs in standard form and for DAEs (4.1) with a properly involved derivative.

5.2 Methods applied to ODEs and DAEs in standard form

5.2.1 Backward differentiation formula

The backward differentiation formula (BDF) is an implicit linear multistep method that generalizes the implicit or backward Euler method by approximating the derivative by an eventually more accurate k-step backward difference quotient. This formula has been introduced in [55]. It is widely used, in particular in circuit simulation, on the basis of [86]. For explicit ODEs

$$x'(t) = g(x(t), t),$$

the BDF method is formed by

$$\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{ni} x_{n-i} = g(x_n, t_n).$$

Here, x_n denotes the numerical solution at the time point t_n. The stepsize $t_n - t_{n-1}$ is denoted by h_n. The coefficients α_{ni} are derived from a polynomial interpolation of the ODE solution through the interpolation points $t_n, \ldots, t_{n-k}$. As is well-known, in the case of constant stepsizes, the BDF satisfies the root criterion of Dahlquist for $k \leq 6$, but it does not for $k > 6$ (see, e.g., [86, 105]). For this reason, it is strongly recommended to apply the BDF just with $k \leq 6$, since otherwise dangerous error accumulations may appear. The BDF is feasible in the sense that the nonlinear equation system to be solved, in order to generate the actual value x_n, has the nonsingular Jacobian $I - \frac{h_n}{\alpha_{n0}} g_x$ if the stepsize h_n is sufficiently small. To ensure a smooth numerical integration, one has to consider quite nontrivial aspects concerning the stepsize arrangement (e.g., [99, 96, 100, 34]).

The BDF has a natural extension to DAEs in standard form (4.33) (e.g., [86, 90, 96, 25])

$$\mathfrak{f}\Big(\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{ni} x_{n-i}, x_n, t_n\Big) = 0.$$

Now the system to be solved for x_n has the Jacobian

$$\frac{\alpha_{n0}}{h_n} \mathfrak{f}_{x^1} + \mathfrak{f}_x.$$

As pointed out in Section 4.4 Note (4), for index-1 DAEs, the local matrix pencil (4.34) is regular with Kronecker index 1, and therefore the Jacobian is nonsingular, supposing the stepsize is sufficiently small.

The BDF is implemented in several well-known DAE solver packages such as DASSL [182] and DASPK [181, 148]. The BDF is applied very successfully in many applications, which use in essence index-1 DAEs. In contrast, for higher index DAEs, one has to expect the failure of the method ([90], see also Chapter 8).

5.2.2 Runge–Kutta method

More than a hundred years ago, Runge and Kutta ([196, 136]) introduced their one-step methods, called explicit Runge–Kutta methods today. Implicit Runge–Kutta methods for ODEs

$$x'(t) = g(x(t), t)$$

were promoted in the 1960s (cf. [32, 33, 104, 29]). An s-stage Runge–Kutta method is a one-step method of the form

$$x_n = x_{n-1} + h_n \sum_{i=1}^{s} b_i X'_{ni},$$

where one has to compute the stage derivatives $X'_{n1}, \ldots, X'_{ns}$ as a solution of the equation system

$$X'_{ni} = g\Big(x_{n-1} + h_n \sum_{j=1}^{s} a_{ij} X'_{nj}, t_{ni}\Big), \qquad i = 1, \ldots, s,$$

with stages $t_{ni} := t_{n-1} + c_i h_n$, and coefficients a_{ij}, b_i, c_i which are usually collected in a Butcher tableau

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s. \end{array} = \begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}.$$

If the Runge–Kutta matrix $\mathcal{A}$ is strictly lower triangular, one speaks of an explicit Runge–Kutta method. Otherwise, it is called an implicit one. In case the Runge–Kutta matrix is nonsingular, we denote the entries of its inverse by α_{ij}, that is

$$\mathcal{A}^{-1} =: (\alpha_{ij})_{i,j=1,\ldots,s}\,.$$

As before, x_n denotes the numerical solution at the current time point t_n, and h_n is the stepsize $t_n - t_{n-1}$. The stage derivatives X'_{ni} are thought to approximate the derivative values $x'_*(t_{ni})$ of the solution, and the resulting expressions

$$X_{ni} = x_{n-1} + h_n \sum_{j=1}^{s} a_{ij} X'_{nj},\ i = 1, \ldots, s, \tag{5.20}$$

are called stage approximations for the solution values $x_*(t_{ni})$ at the stages t_{ni}.

Extensions of Runge–Kutta methods to DAEs in standard form

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

are not as evident as in the case of BDF methods. At first glance, one can formulate them as ([183, 96, 25, 103])

$$x_n = x_{n-1} + h_n \sum_{i=1}^{s} b_i X'_{ni}, \tag{5.21}$$

with stage derivatives X'_{ni} to be determined from the system

$$\mathfrak{f}(X'_{ni}, x_{n-1} + h_n \sum_{j=1}^{s} a_{ij} X'_{nj}, t_{ni}) = 0,\ \ i = 1, \ldots, s. \tag{5.22}$$

One has to be aware of the fact that the existence of stage derivatives X'_{ni} is not assured, if one uses an explicit method. This becomes evident by a look at the following simple DAE, with $x = (u, w)$:

$$u' = w, \quad 1 = u^2 + w^2,$$

for which, for instance, any two-stage explicit Runge–Kutta method fails to determine unique stage derivatives $X'_{n1} = (U'_{n1}, W'_{n1})$ and $X'_{n2} = (U'_{n2}, W'_{n2})$ from the respective system (5.22), that is, from

$$\begin{aligned} U'_{n1} &= w_{n-1}, & 1 &= u_{n-1}^2 + w_{n-1}^2 \\ U'_{n2} &= w_{n-1} + h_n a_{21} W'_{n1}, & 1 &= (u_{n-1} + h_n a_{21} U'_{n1})^2 + (w_{n-1} + h_n a_{21} W'_{n1})^2. \end{aligned}$$

On the one hand, W'_{n2} is not defined at all. On the other hand, the system is solvable only if the previous solution (u_{n-1}, w_{n-1}) satisfies the constraint

$$u_{n-1}^2 + w_{n-1}^2 = 1.$$

The situation described above is typical of all explicit Runge–Kutta methods, and we simply decide not to use them. Analogous problems arise for Runge–Kutta methods having a singular Runge–Kutta matrix $\mathcal{A}$, which also rules out these methods. To explain the circumstance we turn, for a moment, to the linear constant coefficient DAE

$$Ex'(t) + Fx(t) = q(t), \tag{5.23}$$

and the respective system of equations (5.22)

$$EX'_{ni} + F\Big(x_{n-1} + h_n \sum_{j=1}^{s} a_{ij} X'_{nj}\Big) = q(t_{ni}), \quad i = 1, \ldots, s. \tag{5.24}$$

We are led to inspect the linear system

$$\mathfrak{A}_n \begin{bmatrix} X'_{n1} \\ \vdots \\ X'_{ns} \end{bmatrix} = \begin{bmatrix} -Fx_{n-1} + q(t_{n1}) \\ \vdots \\ -Fx_{n-1} + q(t_{ns}) \end{bmatrix}, \quad \mathfrak{A}_n := I_s \otimes E + h_n \mathcal{A} \otimes F,$$

whereby the symbol $\otimes$ means the Kronecker product. Clearly, for a feasible method, the coefficient matrix $\mathfrak{A}_n$ should be nonsingular for all sufficiently small stepsizes. For the simplest index-1 DAE, if E is the zero matrix and F is nonsingular, we arrive at $\mathfrak{A}_n = h\mathcal{A} \otimes F$. Therefore, both matrices $\mathfrak{A}_n$ and $\mathcal{A}$ are nonsingular at the same time. In consequence, in order to ensure the feasibility of the method, we have to assume the Runge–Kutta methods to have a nonsingular coefficient matrix $\mathcal{A}$.

From now on, suppose the matrix $\mathcal{A}$ to be nonsingular. Since the DAE (5.23) has index 1, the matrix pair $\{E, F\}$ is regular with Kronecker index 1, and we find nonsingular real valued matrices (cf. Proposition 1.3) L, K such that

$$LEK = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \quad LFK = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}.$$

Multiplying system (5.24) by L, and forming $X'_{ni} = K \begin{bmatrix} U'_{ni} \\ V'_{ni} \end{bmatrix}$, $x_{n-1} = K \begin{bmatrix} u_{n-1} \\ v_{n-1} \end{bmatrix}$, implies that the system (5.24) is decoupled into two parts, the first one

$$U'_{ni} + h_n \sum_{j=1}^{s} a_{ij} U'_{nj} = p(t_{ni}) - W u_{n-1}, \quad i = 1, \ldots, s,$$

being uniquely solvable with respect to $U'_{n1}, \ldots, U'_{ns}$, if the stepsize $h_n > 0$ is small, and the second one

$$h_n \sum_{j=1}^{s} a_{ij} V'_{nj} = r(t_{ni}) - v_{n-1}, \quad i = 1, \ldots, s,$$

which is uniquely solvable, since $\mathcal{A}$ is nonsingular. Thus, the coefficient matrix $\mathfrak{A}_n$ is also nonsingular.

Runge–Kutta methods having a nonsingular coefficient matrix $\mathcal{A}$ exhibit a remarkable property: they provide a one-to-one correspondence between the stage derivatives and the stage approximations via the relation (5.20). Clearly, formula (5.20) fixes the stage approximations $X_{n1}, \ldots, X_{ns}$ to be uniquely determined by the stage derivatives $X'_{n1}, \ldots, X'_{ns}$. Conversely, if the stage approximations $X_{n1}, \ldots, X_{ns}$ are already known, then (5.20) determines the stage derivatives as the solution of the linear system

$$\sum_{j=1}^{s} a_{ij} X'_{nj} = \frac{1}{h_n}(X_{ni} - x_{n-1}), \quad i = 1, \ldots, s.$$

The coefficient matrix of this linear system is $\mathcal{A} \otimes I_m$, and is thus nonsingular, and $(\mathcal{A} \otimes I_m)^{-1} = \mathcal{A}^{-1} \otimes I_m$. The system solution is given as

$$X'_{ni} = \frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij}(X_{nj} - x_{n-1}), \quad i = 1, \ldots, s. \tag{5.25}$$

For all Runge–Kutta methods, we can write

$$\mathfrak{f}(X'_{ni}, X_{ni}, t_{ni}) = 0, \quad i = 1, \ldots, s, \tag{5.26}$$

instead of (5.22). This shows that the stage approximations $X_{n1}, \ldots, X_{ns}$ always satisfy the obvious constraint.

The assumption of a nonsingular Runge–Kutta matrix $\mathcal{A}$ allows us to replace the stage derivatives in the description (5.21), (5.22) by the stage approximations, and to use exclusively the stage approximations. In this way we arrive at another description of the given Runge–Kutta method, namely

$$x_n = x_{n-1} + h_n \sum_{i=1}^{s} b_i \frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij}(X_{nj} - x_{n-1}) = \underbrace{(1 - \sum_{i,j=1}^{s} b_i \alpha_{ij})}_{=:\rho} x_{n-1} + \sum_{i,j=1}^{s} b_i \alpha_{ij} X_{nj}, \tag{5.27}$$

$$\mathfrak{f}\left(\frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij}(X_{nj} - x_{n-1}), X_{ni}, t_{ni}\right) = 0, \qquad i = 1, \ldots, s. \tag{5.28}$$

There is an extensive literature providing stability, convergence and order results for Runge–Kutta methods applied to standard form index-1 DAEs (e.g., [183, 96, 25, 103]). Here we mention the condition $|\rho| \leq 1$ for stability reasons. Additionally, schemes with $|\rho| = 1$ are not recommended for the numerical integration of fully implicit DAEs ([25, p. 100]). Consequently, the Gauss methods are excluded. It is worth mentioning that, in the context of the collocation solution of boundary value problems for index-1 DAEs, also Gauss methods prove their value. Concerning the orders, one has to be aware of certain order reductions when comparing it with the explicit ODE case.

We stress once again that all stage approximations satisfy equation (5.26) which reflects the solution property

$$\mathfrak{f}(x'_*(t_{ni}), x_*(t_{ni}), t_{ni}) = 0, \quad i = 1, \ldots, s.$$

In contrast, the value x_n generated by the Runge–Kutta method via (5.21), or (5.27), does not necessarily satisfy the constraint. There might be no counterpart X'_n such that $\mathfrak{f}(X'_n, x_n, t_n) = 0$ although the true solution satisfies the DAE $\mathfrak{f}(x'_*(t_n), x_*(t_n), t_n) = 0$. No doubt, it would be to the best advantage, if x_n were to

satisfy the DAE, too. This is a question concerning the choice of the coefficients b_i. Luckily, we are able to solve the problem by assuming $c_s = 1$ and $b_i = a_{si}$ for all $i = 1, \ldots, s$. Then, we have $\rho = 0$, $t_n = t_{ns}$ and $x_n = X_{ns}$. Hence x_n satisfies the DAE and, in particular, the constraint.

Another idea is to apply formula (5.21) or (5.27) afterwards to provide a projection onto the constraint set, and then to use the resulting value as the new x_n. The idea of projections back to a constraint has been proposed in [4] for Hessenberg index-2 DAEs. This approach might work for DAEs with special structure, but for index-1 DAEs in general the benefit does not compensate the extra costs.

In the following, an IRK(DAE) is an implicit Runge–Kutta method that is particularly suitable for DAEs because of the nonsingular $\mathcal{A}$, and the conditions $c_s = 1$ and $b_i = a_{si}$, $i = 1, \ldots, s$.

The Radau IIA methods serve as examples for IRK(DAE) methods [105]. The Radau IIA method with stage number $s = 3$ is implemented in the well-known DAE solver package RADAU5 [102]. It has order 5 for ODEs and semi-implicit index-1 DAEs.

An IRK(DAE) method simplifies (5.21), (5.22) to

$$x_n := X_{ns}, \tag{5.29}$$

$$\mathfrak{f}\Big(\frac{1}{h_n}\sum_{j=1}^{s} \alpha_{ij}(X_{nj} - x_{n-1}), X_{ni}, t_{ni}\Big) = 0, \quad i = 1, \ldots, s. \tag{5.30}$$

5.2.3 General linear method

Both integration schemes, linear multistep and one-step methods, are well-established numerical methods with their particular advantages and disadvantages. Linear multistep methods can be implemented efficiently for large problems, but the stability properties are not always satisfactory. One-step methods have superior stability properties but they do suffer from high computational costs.

Several attempts have been made in order to overcome difficulties associated with each class of methods while keeping its advantages. We quote some of them, only. Hybrid methods allow more than one function evaluation in a linear multistep scheme [85]. Using cyclic compositions of multistep methods it became possible to break Dahlquist's barriers [19]. On the other hand, Rosenbrock methods aim at reducing the costs for a Runge–Kutta scheme by linearizing the nonlinear system and incorporating the Jacobian into the numerical scheme [105, 122].

In order to cover both linear multistep and Runge–Kutta methods in one unifying framework, Butcher [28] introduced *general linear methods* (GLMs) for the solution of ordinary differential equations

$$x'(t) = g(x(t), t).$$

As for linear multistep methods, a general linear method uses r input quantities $x_1^{[n-1]},\ldots,x_r^{[n-1]}$ from the past when proceeding from t_{n-1} to $t_n = t_{n-1}+h$ with a stepsize h. For simplicity, we restrict the discussion to constant stepsizes here. Similarly to Runge–Kutta methods, s internal stages $t_{nj} = t_{n-1}+c_j h$, $j = 1,\ldots,s$, are introduced, and the quantities $X_{n1},\ldots,X_{ns}$ have to be calculated from the system

$$X_{ni} = h\sum_{j=1}^{s} a_{ij} g(X_{nj},t_{nj}) + \sum_{j=1}^{r} u_{ij} x_j^{[n-1]}, \qquad i = 1,\ldots,s.$$

Then, the new solution vector $x^{[n]}$ is given by

$$x_i^{[n]} = h\sum_{j=1}^{s} b_{ij} g(X_{nj},t_{nj}) + \sum_{j=1}^{r} v_{ij} x_j^{[n-1]}, \qquad i = 1,\ldots,r.$$

Using the more compact notation

$$X_n = \begin{bmatrix} X_{n1} \\ \vdots \\ X_{ns} \end{bmatrix}, \qquad G(X_n) = \begin{bmatrix} g(X_{n1},t_{n1}) \\ \vdots \\ g(X_{ns},t_{ns}) \end{bmatrix}, \qquad x^{[n-1]} = \begin{bmatrix} x_1^{[n-1]} \\ \vdots \\ x_r^{[n-1]} \end{bmatrix}$$

a general linear method can be written as

$$\begin{aligned} X_n &= (\mathcal{A}\otimes I_m) h G(X_n) + (\mathcal{U}\otimes I_m) x^{[n-1]}, \\ x^{[n]} &= (\mathcal{B}\otimes I_m) h G(X_n) + (\mathcal{V}\otimes I_m) x^{[n-1]}. \end{aligned}$$

The integer m denotes the problem size and $\otimes$ represents the Kronecker product for matrices (cf. [117]). It is only a slight abuse of notation when the Kronecker product is often omitted, i.e.,

$$\begin{bmatrix} X_n \\ x^{[n]} \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{U} \\ \mathcal{B} & \mathcal{V} \end{bmatrix} \begin{bmatrix} hG(X_n) \\ x^{[n-1]} \end{bmatrix}.$$

This formulation of a GLM with s internal and r external stages from the past is due to Burrage and Butcher [27]. The matrices $\mathcal{A}, \mathcal{U}, \mathcal{B}$ and $\mathcal{V}$ contain the method's coefficients.

The internal stages X_{ni} estimate the exact solution values $x_*(t_{ni})$, but the external stages $x_i^{[n-1]}$ are fairly general. Commonly adopted choices are approximations of the form

$$x^{[n-1]} \approx \begin{bmatrix} x_*(t_{n-1}) \\ x_*(t_{n-1}-h) \\ \vdots \\ x_*(t_{n-1}-(r-1)h) \end{bmatrix} \qquad \text{or} \qquad x^{[n-1]} \approx \begin{bmatrix} x_*(t_{n-1}) \\ h x_*'(t_{n-1}) \\ \vdots \\ h^{r-1} x_*^{(r-1)}(t_{n-1}) \end{bmatrix},$$

the former representing a method of multistep type while the latter is a Nordsieck vector. Notice that, compared to Nordsieck's original formulation [179], factorials have been omitted for convenience. General linear methods with Nordsieck type external stages are considered, among other references, in [120, 31, 214, 211]. Different choices of the vector $x^{[n-1]}$ are often related by linear transformations. In this sense the representation of a method using the matrices $\mathcal{A}, \mathcal{U}, \mathcal{B}$ and $\mathcal{V}$ is not unique as two different methods may be equivalent owing to such a transformation [29].

A modification of GLMs to apply to index-1 DAEs in standard form

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

is given by [199]:

$$\begin{aligned}
&\mathfrak{f}(X'_{ni}, X_{ni}, t_{ni}) = 0, \quad i = 1, \ldots, s,\\
&X_n = (\mathcal{A} \otimes I_m) h X'_n + (\mathcal{U} \otimes I_m) x^{[n-1]},\\
&x^{[n]} = (\mathcal{B} \otimes I_m) h X'_n + (\mathcal{V} \otimes I_m) x^{[n-1]}.
\end{aligned}$$

For the same reasons as in the case of Runge–Kutta methods, the matrix $\mathcal{A}$ is assumed to be nonsingular.

5.3 Methods applied to DAEs with a properly involved derivative

Again, as prototypes of linear multistep methods and one-step methods, we discuss the BDF and Runge–Kutta methods. Afterwards we turn to general linear methods.

5.3.1 Backward differentiation formula

A self-evident extension of the BDF to DAEs with a properly involved derivative (4.1)

$$f((Dx)'(t), x(t), t) = 0$$

is given by

$$f\Big(\frac{1}{h_n} \sum_{i=0}^{k} \alpha_{ni} D(t_{n-i}) x_{n-i}, x_n, t_n\Big) = 0. \tag{5.31}$$

Here, just the derivative term $(Dx)'$ is approximated by a backward difference quotient. By construction, the resulting estimate x_n belongs to the constraint set $\mathcal{M}_0(t_n)$ which reflects the true solution property $x_*(t_n) \in \mathcal{M}_0(t_n)$. The Jacobian of the system to be solved for x_n, which means $\frac{1}{h_n}\alpha_{n0} f_y D + f_x$, is nonsingular for all suffi-

ciently small stepsizes, since the related matrix pencil (4.6) is regular with Kronecker index 1.

5.3.2 Runge–Kutta method

IRK(DAE) methods, given by a Butcher tableau

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}, \quad \mathcal{A} \text{ nonsingular}, \quad \mathcal{A}^{-1} =: (\alpha_{ij})_{i,j=1,\ldots,s},$$

are applicable to DAEs in standard formulation (cf. Section 5.2.2). How can we extend these methods to nonlinear DAEs

$$f((Dx)'(t), x(t), t) = 0,$$

with a properly involved derivative? The equation itself suggests that we adapt the method in such a way that only approximations for $(Dx)'(t_{ni})$ are needed. Therefore, we compose the method by means of stage approximations $X_{n1}, \ldots, X_{ns}$ for estimating the solution values $x_*(t_{n1}), \ldots, x_*(t_{ns})$ and stage derivatives $[DX]'_{ni}$ for approximating $(Dx)'(t_{ni})$, $i = 1, \ldots, s$ (cf. (5.26)). Naturally,

$$f([DX]'_{ni}, X_{ni}, t_{ni}) = 0, \qquad i = 1, \ldots, s$$

is satisfied. In order to obtain a reasonable formula containing stage approximations only, we follow the idea behind the Runge–Kutta formula in terms of stage approximations (5.28). We introduce

$$[DX]'_{ni} := \frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij} (D(t_{nj}) X_{nj} - D(t_{n-1}) x_{n-1}), \qquad i = 1, \ldots, s,$$

which is equivalent to

$$D(t_{ni}) X_{ni} = D(t_{n-1}) x_{n-1} + h_n \sum_{j=1}^{s} a_{ij} [DX]'_{nj}, \qquad i = 1, \ldots, s. \tag{5.32}$$

This yields

$$f\left(\frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij} (D(t_{nj}) X_{nj} - D(t_{n-1}) x_{n-1}), X_{ni}, t_{ni}\right) = 0, \qquad i = 1, \ldots, s, \tag{5.33}$$

for computing the stage approximations $X_{n1}, \ldots, X_{ns}$.

In order to answer the basic question of whether the Jacobian of the system remains nonsingular, we turn for a moment, for simplicity, to linear constant coefficient DAEs

$$A(Dx(t))' + Bx(t) = q(t),$$

and the corresponding system

$$A\frac{1}{h_n}\sum_{j=1}^{s} \alpha_{ij}(DX_{nj} - Dx_{n-1}) + BX_{ni} = q(t_{ni}), \qquad i = 1,\ldots,s. \tag{5.34}$$

The coefficient matrix $\mathfrak{A}_n$ of this linear $ms \times ms$ system with respect to the unknowns $X_{n1},\ldots,X_{ns}$ reads

$$\mathfrak{A}_n = \mathcal{A} \otimes (AD) + h_n I_s \otimes B.$$

Due to the index-1 property of the matrix pencil $\{AD, B\}$, the matrix $G := AD + BQ_0$ is nonsingular with Q_0 being the projector onto $\ker AD$ along $\{z \in \mathbb{R}^m : Bz \in \operatorname{im} AD\}$. Letting $P_0 = I - Q_0$ it holds that $G^{-1}AD = P_0$, $G^{-1}BQ_0 = Q_0$, and, additionally, $Q_0 = Q_0 G^{-1}B$. Derive

$$(I_s \otimes G^{-1})\mathfrak{A}_n = \mathcal{A} \otimes P_0 + h_n I_s \otimes (G^{-1}B),$$

and consider the homogeneous equation

$$(\mathcal{A} \otimes P_0 + h_n I_s \otimes (G^{-1}B))Z = 0.$$

Multiplying the homogeneous system by $I_s \otimes Q_0$, and taking into account the relation $(I_s \otimes Q_0)(\mathcal{A} \otimes P_0) = \mathcal{A} \otimes (Q_0 P_0) = 0$, we find

$$(h_n I_s \otimes (Q_0 G^{-1}B))Z = (h_n I_s \otimes Q_0)Z = 0, \quad \text{thus } (I_s \otimes Q_0)Z = 0,\ (I_s \otimes P_0)Z = Z.$$

Writing $\mathcal{A} \otimes P_0 = (\mathcal{A} \otimes I_m)(I_s \otimes P_0)$ we arrive at

$$(\mathcal{A} \otimes I_m + h_n I_s \otimes (G^{-1}B))Z = 0,$$

and hence, if the stepsize h_n is sufficiently small, $Z = 0$ follows. This shows that the coefficient matrix $\mathfrak{A}_n$ is nonsingular for sufficiently small stepsizes $h_n > 0$. In the case of general index-1 DAEs (4.1), feasibility can be proved in a similar way, but by stronger technical effort.

Observe that, by construction, the stage approximations lie in the obvious constraint set,

$$X_{ni} \in \mathcal{M}_0(t_{ni}), \quad i = 1,\ldots,s,$$

which reflects the corresponding solution property $X_*(t_{ni}) \in \mathcal{M}_0(t_{ni})$, $i = 1,\ldots,s$. Consequently, $x_n \in \mathcal{M}_0(t_n)$ for IRK(DAE) methods due to $x_n = X_{ns}$.

Finally, the IRK(DAE) methods applied to DAEs (4.1) are given by

$$x_n := X_{ns}, \tag{5.35}$$

$$f\Big(\frac{1}{h_n}\sum_{j=1}^{s} \alpha_{ij}(D(t_{nj})X_{nj} - D(t_{n-1})x_{n-1}), X_{ni}, t_{ni}\Big) = 0, \qquad i = 1,\ldots,s. \tag{5.36}$$

5.3.3 General linear method

We apply a general linear method to the DAE (4.1),

$$f((Dx)'(t), x(t), t) = 0,$$

as given by

$$f([DX]'_{ni}, X_{ni}, t_{ni}) = 0, \qquad i = 1, \ldots, s, \tag{5.37}$$

$$[DX]_n = h(\mathcal{A} \otimes I_n)[DX]'_n + (\mathcal{U} \otimes I_n)[Dx]^{[n-1]}, \tag{5.38}$$

$$[Dx]^{[n]} = h(\mathcal{B} \otimes I_n)[DX]'_n + (\mathcal{V} \otimes I_n)[Dx]^{[n-1]}\,. \tag{5.39}$$

The stages X_{ni} are approximations to the exact solution values $x_*(t_{ni})$ at the intermediate time points $t_{n1}, \ldots, t_{ns}$. The super-vectors $[DX]_n$ and $[DX]'_n$ are given by

$$[DX]_n = \begin{bmatrix} D(t_{n1})X_{n1} \\ \vdots \\ D(t_{ns})X_{ns} \end{bmatrix}, \qquad [DX]'_n = \begin{bmatrix} [DX]'_{n1} \\ \vdots \\ [DX]'_{ns} \end{bmatrix},$$

where $[DX]'_{ni}$ approximates $(Dx_*)'(t_{ni})$. The input vector

$$[Dx]^{[n-1]} \approx \begin{bmatrix} (Dx_*)(t_{n-1}) \\ h(Dx_*)'(t_{n-1}) \\ \vdots \\ h^{r-1}(Dx_*)^{(r-1)}(t_{n-1}) \end{bmatrix}$$

is assumed to be a Nordsieck vector. Observe that only information regarding the solution's D-component is passed on from step to step. Hence errors in this component are the only ones that are possibly propagated.
Again we rule out all methods having a singular coefficient matrix $\mathcal{A}$.

Because of the nonsingularity of $\mathcal{A}$, the relation (5.38) is equivalent to

$$[DX]'_n = \frac{1}{h}(\mathcal{A}^{-1} \otimes I_n)([DX]_n - (\mathcal{U} \otimes I_n)[Dx]^{[n-1]}). \tag{5.40}$$

The same arguments as used for Runge–Kutta methods apply now to show that this method is feasible.

The quantity x_n that estimates the true solution value $x_*(t_n)$, $t_n = t_{n-1} + h$ has to be obtained by means of a linear combination of the internal stages X_{ni}. Every stage X_{ni} satisfies the obvious constraints, such that

$$X_{ni} \in \mathcal{M}_0(t_{ni}),\ i = 1, \ldots, s.$$

It is a desirable feature for the numerical solution to satisfy $x_n \in \mathcal{M}_0(t_n)$ as well. As we know from the previous section, IRK(DAE) (stiffly accurate) methods guaran-

tee this situation. We will therefore restrict our attention to *stiffly accurate general linear methods*, which means

$$M = \begin{bmatrix} \mathcal{A} & \mathcal{U} \\ \mathcal{B} & \mathcal{V} \end{bmatrix} \text{ with } \mathcal{A} \text{ nonsingular}, \quad e_s^{\mathrm{T}}\mathcal{A} = e_1^{\mathrm{T}}\mathcal{B}, \quad e_s^{\mathrm{T}}\mathcal{U} = e_1^{\mathrm{T}}\mathcal{V}, \quad c_s = 1,$$

such that the last row of $[\mathcal{A},\mathcal{U}]$ coincides with the first row of $[\mathcal{B},\mathcal{V}]$. This implies that x_n coincides with the last stage X_{ns} and hence x_n to belong to $\mathcal{M}_0(t_n)$.

Similarly as in the case of Runge–Kutta methods, since the matrix $\mathcal{A}$ is nonsingular, making use of the relation (5.40) we may reformulate the GLM (5.37)–(5.39) as

$$f\left(\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}\Big(D(t_{nj})X_{nj} - \sum_{\ell=1}^{r}\mu_{j\ell}[Dx]_{\ell}^{[n-1]}\Big), X_{ni}, t_{ni}\right) = 0, \quad i = 1,\ldots,s, \tag{5.41}$$

and $[Dx]^{[n]}$ is given recursively by

$$\begin{bmatrix} [Dx]_1^{[n]} \\ \vdots \\ [Dx]_r^{[n]} \end{bmatrix} = (\mathcal{B}\otimes I_n)(\mathcal{A}^{-1}\otimes I_n)([DX]_n - (\mathcal{U}\otimes I_n)[Dx]^{n-1}) + (\mathcal{V}\otimes I_n)[Dx]^{[n-1]}. \tag{5.42}$$

The coefficients $\alpha_{j\ell}$ and $\mu_{j\ell}$ are the entries of the coefficient matrices $\mathcal{A}^{-1}$ and $\mathcal{U}$.

5.4 When do decoupling and discretization commute?

In contrast to index-1 DAEs in standard formulation, a DAE (4.1) with properly involved derivative holds a natural inherent ODE, the IERODE, which is uniquely determined by the problem data (Proposition 4.7). It is reasonable to ask whether numerical methods being applied to the DAE reach the IERODE unchanged. It would be quite favorable to know which kind of method works on the inner dynamical part. As Example 5.1 demonstrates, it may actually happen that a method arrives at the inner dynamic part just in an essentially converted version, and this may cause serious additional stepsize restrictions.

As before, equation (4.1) is assumed to be an index-1 DAE. We consider BDF methods, IRK(DAE) methods, and GLMs as described in Section 5.3. Recall that x_n generated by the BDF method, by an IRK(DAE) method or by a stiffly stable GLM satisfies the obvious constraint $x_n \in \mathcal{M}_0(t_n)$. Despite the different approaches, we are able to merge all methods into one framework

$$f([D\hat{x}]_n', \hat{x}_n, \hat{t}_n) = 0, \tag{5.43}$$

by introducing the expressions

$$\hat{x}_n := \begin{cases} x_n, & \text{for BDF,} \\ X_{ni}, & i = 1,\dots,s, \text{ for RK and GLM,} \end{cases}$$

$$\hat{t}_n := \begin{cases} t_n, & \text{for BDF,} \\ t_{ni}, & i = 1,\dots,s, \text{ for RK and GLM,} \end{cases}$$

as well as

$$[D\hat{x}]'_n := \begin{cases} \frac{1}{h_n}\sum_{i=0}^{k} \alpha_{ni} D(t_{n-i}) x_{n-i}, & \text{for BDF,} \\ \frac{1}{h_n}\sum_{j=1}^{s} \alpha_{ij}(D(t_{nj})X_{nj} - D(t_{n-1})x_{n-1}), & i = 1,\dots,s, \text{ for RK,} \\ \frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}(D(t_{nj})X_{nj} - \sum_{\ell=1}^{r} \mu_{j\ell}[D\hat{x}]_{\ell}^{[n-1]}), & i = 1,\dots,s, \text{ for GLM} \end{cases}$$

with

$$[D\hat{x}]^{[n-1]} = (\mathcal{B} \otimes I_n)[D\hat{x}]'_{n-1} + (\mathcal{V} \otimes I_n)[D\hat{x}]^{[n-2]}$$

for GLMs. Along with equation (5.43), we give special attention to the perturbed equation

$$f([\widetilde{D\hat{x}}]'_n, \tilde{\hat{x}}_n, \hat{t}_n) = q_n, \tag{5.44}$$

with

$$\tilde{\hat{x}}_n := \begin{cases} \tilde{x}_n, & \text{for BDF,} \\ \tilde{X}_{ni}, & i = 1,\dots,s, \text{ for RK and GLM,} \end{cases}$$

and

$$[\widetilde{D\hat{x}}]'_n := \begin{cases} \frac{1}{h_n}\sum_{i=0}^{k} \alpha_{ni} D(t_{n-i}) \tilde{x}_{n-i}, & \text{for BDF,} \\ \frac{1}{h_n}\sum_{j=1}^{s} \alpha_{ij}(D(t_{nj})\tilde{X}_{nj} - D(t_{n-1})\tilde{x}_{n-1}), & i = 1,\dots,s, \text{ for RK,} \\ \frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}(D(t_{nj})\tilde{X}_{nj} - \sum_{\ell=1}^{r} \mu_{j\ell}[\widetilde{Dx}]_{\ell}^{[n-1]}), & i = 1,\dots,s, \text{ for GLM,} \end{cases}$$

with

$$\begin{bmatrix} \tilde{X}_n \\ [\widetilde{Dx}]^{[n]} \end{bmatrix} = \begin{bmatrix} \mathcal{A} & \mathcal{U} \\ \mathcal{B} & \mathcal{V} \end{bmatrix} \begin{bmatrix} h[\widetilde{D\hat{x}}]'_{n-1} \\ [\widetilde{Dx}]^{[n-1]} \end{bmatrix}$$

for GLMs. The perturbation q_n stands for rounding errors, and for possible errors from a nonlinear solver. It is supposed to be sufficiently small. The tilde symbol indicates possible errors in the quantities under consideration. Furthermore, form

$$\hat{u}_n := D(\hat{t}_n)\hat{x}_n, \quad \tilde{\hat{u}}_n := D(\hat{t}_n)\tilde{\hat{x}}_n, \quad [\hat{u}]'_n := [D\hat{x}]'_n, \quad \tilde{\hat{u}}'_n := [\widetilde{D\hat{x}}]'_n. \tag{5.45}$$

Aiming to achieve better insight, we make use of the decoupling procedure described in Section 4.2, in particular of the implicitly defined function ω^{pert}. Assume the definition domain of this function to be as spacious as needed and the time interval to be $\mathcal{I}_c = [t_0, T]$. Applying Lemma 4.9 for $(\tilde{\hat{u}}, q_n) \in B_\rho(u_*(\hat{t}_n), 0)$, we derive

the expression

$$\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n)=\omega^{\mathrm{pert}}(R(\hat{t}_n)\tilde{\hat{u}}_n,\hat{t}_n,q_n)=D(\hat{t}_n)^{-}[\widetilde{D\hat{x}}]'_n+Q_0(\hat{t}_n)\tilde{\hat{x}}_n \tag{5.46}$$

from (5.44). Multiplying (5.46) by $D(\hat{t}_n)$ and $Q_0(\hat{t}_n)$ yields

$$D(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n)=R(\hat{t}_n)[\widetilde{D\hat{x}}]'_n=R(\hat{t}_n)[\tilde{\hat{u}}]'_n,$$

and

$$Q_0(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n)=Q_0(\hat{t}_n)\tilde{\hat{x}}_n,$$

respectively. In consequence, the solution $\tilde{\hat{x}}_n$ of (5.44) can be represented as

$$\tilde{\hat{x}}_n=D(\hat{t}_n)^{-}D(\hat{t}_n)\tilde{\hat{x}}_n+Q_0(\tilde{t}_n)\tilde{\hat{x}}_n=D(\hat{t}_n)^{-}\tilde{\hat{u}}_n+Q_0(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n),$$

with $\tilde{\hat{u}}_n$ satisfying

$$D(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n)=R(\hat{t}_n)[\tilde{\hat{u}}]'_n.$$

By introducing the quantities $\tilde{U}_{ni}:=D(t_{ni})\tilde{X}_{ni}$, $\tilde{u}_{n-i}:=D(t_{n-i})\tilde{x}_{n-i}$, and $[\tilde{u}]_\ell^{[n-1]}:=[\widetilde{Dx}]_\ell^{[n-1]}$, we may rewrite $[\tilde{\hat{u}}]'_n$ as

$$[\tilde{\hat{u}}]'_n=\begin{cases}\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{ni}\tilde{u}_{n-i}, & \text{for BDF,}\\ \frac{1}{h_n}\sum_{j=1}^{s}\alpha_{ij}(\tilde{U}_{nj}-\tilde{u}_{n-1}), & i=1,\dots,s,\ \text{for RK,}\\ \frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}(\tilde{U}_{nj}-\sum_{\ell=1}^{r}\mu_{j\ell}[\tilde{u}]_\ell^{[n-1]}), & i=1,\dots,s,\ \text{for GLM.}\end{cases} \tag{5.47}$$

To obtain the corresponding relations concerning the case $q_n=0$ we simply drop the tildes, in particular,

$$[\hat{u}]'_n=\begin{cases}\frac{1}{h_n}\sum_{i=0}^{k}\alpha_{ni}u_{n-i}, & \text{for BDF,}\\ \frac{1}{h_n}\sum_{j=1}^{s}\alpha_{ij}(U_{nj}-u_{n-1}), & i=1,\dots,s,\ \text{for RK,}\\ \frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}(U_{nj}-\sum_{\ell=1}^{r}\mu_{j\ell}[u]_\ell^{[n-1]}), & i=1,\dots,s,\ \text{for GLM.}\end{cases} \tag{5.48}$$

For later use, the following proposition summarizes the last lines.

Proposition 5.2. *Let equation (4.1) be an index-1 DAE and $x_*:\mathcal{I}_c\to\mathbb{R}^m$ be a solution of (4.1). Then, the solution $\hat{x}_n$ of (5.43) can be represented as*

$$\hat{x}_n=D(\hat{t}_n)^{-}\hat{u}_n+Q_0(\hat{t}_n)\omega^{\mathrm{pert}}(\hat{u}_n,\hat{t}_n,0),$$

with $\hat{u}_n$ satisfying the equation

$$R(\hat{t}_n)[\hat{u}]'_n=D(\hat{t}_n)\omega^{\mathrm{pert}}(\hat{u}_n,\hat{t}_n,0) \tag{5.49}$$

supposing $\hat{u}_n\in B_\rho((Dx_*)(t_n))$. The solution $\tilde{\hat{x}}_n$ of (5.44) can be represented as

$$\tilde{\hat{x}}_n=D(\hat{t}_n)^{-}\tilde{\hat{u}}_n+Q_0(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n,\hat{t}_n,q_n),$$

with $\tilde{\hat{u}}_n$ satisfying the equation

$$R(\hat{t}_n)[\tilde{\hat{u}}]'_n = D(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n, \hat{t}_n, q_n) \tag{5.50}$$

supposing $(\tilde{\hat{u}}_n, q_n) \in B_\rho((Dx)(t_n), 0)$. The function ω^{pert} is a continuous function provided by Lemma 4.9 in the neighborhood $\{(u,t,q)\,|\,t \in \mathcal{I}_c,\ (u,q) \in B_\rho((Dx)_*(t), 0)\}$.

Next we reconsider, for a moment, the perturbed DAE

$$f((Dx)'(t), x(t), t) = q(t),$$

and its decoupled form (cf. Theorem 4.5)

$$x(t) = D(t)^- u(t) + Q_0(t)\omega^{\mathrm{pert}}(u(t), t, q(t)),$$

with $u(\cdot)$ being a solution of the IERODE (4.24)

$$u'(t) = R'(t)u(t) + D(t)\omega^{\mathrm{pert}}(u(t), t, q(t)).$$

Applying the given integration method to this IERODE, we obtain the discretized inherent ODE

$$[\tilde{\hat{u}}]'_n = R'(\hat{t}_n)\tilde{\hat{u}}_n + D(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n, \hat{t}_n, q(\hat{t}_n)). \tag{5.51}$$

Here, the quantity $\tilde{\hat{u}}_n$ describes the numerical approximation of $u(\hat{t}_n)$, and $[\hat{u}]'_n$ estimates $u'(\hat{t}_n)$ in accordance with formula (5.48). Obviously, formula (5.50) coincides with (5.51) if and only if

$$R(\hat{t}_n)[\tilde{\hat{u}}]'_n = [\tilde{\hat{u}}]'_n - R'(\hat{t}_n)\tilde{\hat{u}}_n \quad \text{and} \quad q(\hat{t}_n) = q_n. \tag{5.52}$$

In other words, discretization and decoupling commute if and only if (5.52) is satisfied (see Figure 5.3).

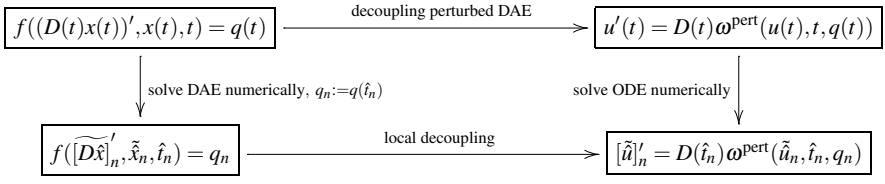


Fig. 5.3 For an index-1 DAE with a properly involved derivative, discretization and local decoupling commute if and only if (5.52) is satisfied. A time-invariant subspace $\mathrm{im}\,D(t)$ is a sufficient condition for (5.52) to be fulfilled.

The commutativity of discretization and decoupling in this sense is of great importance: it ensures that the integration methods approved for explicit ODEs retain

their essential properties if applied to DAEs. The following theorem provides a sufficient condition which guarantees commutativity.

Theorem 5.3. *Let the index-1 DAE (4.1) have a subspace* $\operatorname{im} D(t)$ *independent of* t *and* $x_* : \mathcal{I}_c \to \mathbb{R}^m$ *be a solution of (4.1). Suppose the components* $[Dx]_i^{[0]}$ *of the starting vector* $[Dx]^{[0]}$ *for the GLM belong to* $\operatorname{im} D(t_0)$ *for all* $i = 1, \ldots, r$. *Then, the discretization and the decoupling procedure commute for BDF, IRK(DAE) and stiffly accurate GLM. The discretized IERODE has the form*

$$[\tilde{\hat{u}}]'_n = D(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n, \hat{t}_n, q_n) \tag{5.53}$$

with $[\tilde{\hat{u}}]'_n$ *defined in (5.47) and* $q_n := q(\hat{t}_n)$, *independently of the order, decoupling before discretization or discretization before decoupling. Additionally, in both cases, the estimate* $\tilde{\hat{x}}_n$ *of* $x(\hat{t}_n)$ *is given by*

$$\tilde{\hat{x}}_n = D(\hat{t}_n)^- \tilde{\hat{u}}_n + Q_0(\hat{t}_n)\omega^{\mathrm{pert}}(\tilde{\hat{u}}_n, \hat{t}_n, q_n).$$

Thereby, ω^{pert} *is a continuous function in the neighborhood*

$$\{(u,t,q) \,|\, t \in \mathcal{I}_c,\ (u,q) \in B_\rho((Dx)_*(t), 0)\}$$

for a fixed radius $\rho > 0$, *provided by Lemma 4.9.*

Proof. If $\operatorname{im} D(t)$ is constant then we find a constant projector R_c with

$$\operatorname{im} D(t) = \operatorname{im} R_c = \operatorname{im} R(t), \quad R(t)R_c = R_c, \quad \text{for all } t \in \mathcal{I}.$$

Since $[\tilde{\hat{u}}]'_n \in \operatorname{im} R_c$ by (5.47), we conclude

$$R(\hat{t}_n)[\tilde{\hat{u}}]'_n = R(\hat{t}_n)R_c[\tilde{\hat{u}}]'_n = R_c[\tilde{\hat{u}}]'_n = [\tilde{\hat{u}}]'_n$$

in formula (5.50). On the other hand, the values $\tilde{\hat{u}}_n$ appearing in (5.51) also belong to $\operatorname{im} R_c$ due to (5.45). This implies

$$R'(\hat{t}_n)\hat{u}_n = R'(\hat{t}_n)R_c\hat{u}_n = (R_nR_c)'\hat{u}_n = R'_c\hat{u}_n = 0$$

in formula (5.51). Consequently, (5.53) follows and the proof is complete. □

The discretized IERODE in Theorem 5.3 reflects the special form of the IERODE given by Proposition 4.7 for the case of a time-invariant $\operatorname{im} D(t)$.

Fortunately, DAEs can often be reformulated such that $\operatorname{im} D(t)$ is independent of t. It is even so that one usually finds a formulation with a full row rank matrix $D(t)$, $\operatorname{im} D(t) = \mathbb{R}^n$.

5.5 Convergence on compact intervals and error estimations

We continue analyzing the BDF, IRK(DAE) methods and stiffly stable GLMs applied to DAEs (4.1) with a properly stated derivative and time-invariant $\operatorname{im} D(t)$. In this section we make use of Theorem 5.3 which allows us easily to transfer convergence and stability properties known for these methods, when they are applied to explicit ODEs, to the case of DAEs.

It should be mentioned that the numerical integration methods may show convergence also for the case that $\operatorname{im} D(t)$ varies with t. However, one is then not aware of the method which actually integrates the IERODE. Additional stepsize restrictions might be a consequence.

5.5.1 Backward differentiation formula

We apply the BDF to the DAE (4.1) on a partition π of the compact interval $\mathcal{I}_c = [t_0, T] \in \mathcal{I}_f$. Regarding rounding errors and defects in the nonlinear equations, the BDF methods are given by (cf. (5.44))

$$f\Big(\frac{1}{h_n}\sum\nolimits_{i=0}^{k} \alpha_{ni} D(t_{n-i}) x_{n-i}, x_n, t_n\Big) = q_n, \quad n \geq k. \tag{5.54}$$

The partitions π are assumed to have the following properties:

$$\begin{aligned} \pi &: t_0 < t_1 < \cdots < t_N = T, \\ 0 < h_{min} \leq h_n &:= t_n - t_{n-1} \leq h_{max}, \quad \kappa_1 \leq \frac{h_{n-1}}{h_n} \leq \kappa_2, \quad n \geq 1, \end{aligned} \tag{5.55}$$

where κ_1, κ_2, h_{min} and h_{max} are suitable constants such that the BDFs are stable for explicit ODEs, see [99, 100, 34].

Theorem 5.4. *Let the DAE (4.1) be regular with index 1, and let the subspace* $\operatorname{im} D(t)$ *be independent of* t. *Let* $x_* \in \mathcal{C}_D^1(\mathcal{I}_c, \mathbb{R}^m)$ *be a solution of the DAE (4.1). If the deviations in the starting values* $|D(t_n)x_n - D(t_n)x_*(t_n)|$, $0 \leq n < k$, *and the perturbations* q_n, $k \leq n \leq N$, *are sufficiently small, then the k-step BDF method* ($k \leq 6$) *is feasible for all partitions (5.55) with* h_{max} *being sufficiently small. There exist numerical solution values* x_n *fulfilling (5.44) for each* n *with* $k \leq n \leq N$. *Furthermore, there is a constant* $c > 0$ *such that*

$$\max_{1\leq n\leq N} |x_n - x_*(t_n)| \leq c\Big(\max_{n<k} |D(t_n)x_n - D(t_n)x_*(t_n)| + \max_{n\geq k} |q_n| + \max_{n\geq k} |L_n|\Big)$$

with L_n *being the local error*

$$L_n = f([Dx_*]'_n, x_*(t_n), t_n), \quad k \leq n \leq N,$$

and

$$[Dx_*]'_n := \frac{1}{h_n}\sum_{i=0}^{k} \alpha_{ni}(Dx_*)(t_{n-i}).$$

Proof. From Corollary 4.10 we know that the solution x_* can be written as

$$x_*(t) = D(t)^- u_*(t) + Q_0(t)\omega^{\mathrm{pert}}(u_*(t),t,0), \qquad t \in \mathcal{I}_c, \tag{5.56}$$

with u_* being the solution of the IVP

$$u'_*(t) = D(t)\omega^{\mathrm{pert}}(u_*(t),t,0), \qquad u_*(t_0) = D(t_0)x_*(t_0).$$

The term $R'(t)u_*(t)$ vanishes since we have assumed $\operatorname{im} D(t)$ to be time-invariant, cf. Proposition 4.7(3). The continuous function ω^{pert} is implicitly defined on

$$\{(u,t,q)\,|\,(u,q) \in B_\rho((Dx_*)(t),0),\ t \in \mathcal{I}_c\}$$

by Lemma 4.9. It has continuous partial derivatives ω_u^{pert} and ω_p^{pert}. Proposition 5.2 and (5.53) tell us that

$$x_n = D(t_n)^- u_n + Q_0(t_n)\omega^{\mathrm{pert}}(u_n,t_n,q_n), \quad k \le n \le N, \tag{5.57}$$

is the solution of (5.54), if u_n is the solution of

$$[u]'_n = D(t_n)\omega^{\mathrm{pert}}(u_n,t_n,q_n), \quad k \le n \le N, \quad u_n := D(t_n)x_n, \quad 0 \le n < k, \tag{5.58}$$

with

$$[u]'_n = \frac{1}{h_n}\sum_{i=0}^{k} \alpha_{ni}u_{n-i}.$$

Introducing

$$q_n^u := D(t_n)\omega^{\mathrm{pert}}(u_n,t_n,q_n) - D(t_n)\omega^{\mathrm{pert}}(u_n,t_n,0),$$

we see that u_n solves equation (5.58) if and only if it satisfies the equation

$$\frac{1}{h_n}\sum_{i=0}^{k} \alpha_{ni}u_{n-i} = D(t_n)\omega^{\mathrm{pert}}(u_n,t_n,0) + q_n^u;$$

however, this is the BDF for the explicit ODE

$$u'(t) = D(t)\omega^{\mathrm{pert}}(u(t),t,0), \quad t \in \mathcal{I}_c,$$

with a perturbation q_n^u of the right-hand side in each step. Standard ODE theory ensures the existence of the approximate BDF solutions u_n supposing the errors in the starting phase $|u_n - u_*(t_n)|$, $0 \le n < k$, and $|q_n^u|$, $k \le n \le N$ are sufficiently small. Both demands are satisfied since

$$|u_n - u_*(t_n)| = |D(t_n)x_n - D(t_n)x_*(t_n)|, \quad \text{for } 0 \le n < k,$$

and there is a constant $c_1 > 0$ such that

$$|q_n^u| = |D(t_n)\omega^{\text{pert}}(u_n, t_n, q_n) - D(t_n)\omega^{\text{pert}}(u_n, t_n, 0)| \le c_1 |q_n| \tag{5.59}$$

for $k \le n \le N$. The existence of a continuous partial derivative ω_q^{pert} ensures that (5.59) is valid. Consequently, the existence of a BDF solution x_n satisfying (5.54) is ensured as long as $u_n \in B_\rho(u_*(t_n))$. Applying standard arguments, we find the error estimation

$$\max_{n\ge k} |u_n - u_*(t_n)| \le c_2 \Big(\max_{n<k} |u_n - u_*(t_n)| + \max_{n\ge k} |q_n^u| + \max_{n\ge k} |L_n^u| \Big) \tag{5.60}$$

with a constant $c_2 > 0$, L_n^u being the local error

$$L_n^u = [u_*]_n' - D(t_n)\omega^{\text{pert}}(u_*(t_n), t_n, 0) = [u_*]_n' - u_*'(t_n), \quad k \le n \le N$$

and

$$[u_*]_n' = \frac{1}{h_n} \sum_{i=0}^{k} \alpha_{ni} u_*(t_{n-i}).$$

Defining

$$L_n^w := D(t_n)^- [u_*]_n' + Q_0(t_n) x_*(t_n) = D(t_n)^- [Dx_*]_n' + Q_0(t_n) x_*(t_n), \tag{5.61}$$

we get

$$f(D(t_n)L_n^w, D(t_n)^- u_*(t_n) + Q_0(t_n)L_n^w, t_n) = f([Dx_*]_n', x_*(t_n), t_n) = L_n,$$

thus, by means of the function ω^{pert},

$$L_n^w = \omega^{\text{pert}}(u_*(t_n), t_n, L_n).$$

Regarding (5.61), we see that

$$D(t_n)\omega^{\text{pert}}(u_*(t_n), t_n, L_n) = D(t_n)L_n^w = R(t_n)[u_*]_n' = [u_*]_n'.$$

Here, we again used that $\operatorname{im} D(t)$ is time-invariant. We have

$$L_n^u = D(t_n)\omega^{\text{pert}}(u_*(t_n), t_n, L_n) - D(t_n)\omega^{\text{pert}}(u_*(t_n), t_n, 0).$$

Since ω_q^{pert} exists and is continuous, we find a constant $c_3 > 0$ satisfying

$$|L_n^u| \le c_3 |L_n|, \quad k \le n \le N. \tag{5.62}$$

Inserting (5.62) and (5.59) into (5.60), there is a constant $c_4 > 0$ such that

$$\max_{n\ge k} |u_n - u_*(t_n)| \le c_4 \Big(\max_{n<k} |D(t_n)x_n - D(t_n)x_*(t_n)| + \max_{n\ge k} |q_n| + \max_{n\ge k} |L_n| \Big).$$

Obviously, $u_n \in B_\rho(u_*(t_n))$ is always satisfied if the deviations in the initial values, the perturbations q_n and the local error are sufficiently small. The latter condition can be fulfilled by sufficiently small stepsizes. Hence, the existence of a BDF solution x_n satisfying (5.44) is proved. Regarding the solution representations (5.56), (5.57) and the fact that ω_u^{pert} and ω_q^{pert} exist and are continuous, we obtain the desired estimation

$$\max_{n\geq k}|x_n - x_*(t_n)| \leq c\Big(\max_{n<k}|D(t_n)x_n - D(t_n)x_*(t_n)| + \max_{n\geq k}|q_n| + \max_{n\geq k}|L_n|\Big).$$

with a constant $c > 0$. □

Corollary 5.5. *(Convergence) Let the assumptions of Theorem 5.4 be satisfied. Suppose the errors in the initial values and the perturbations q_n, for $k \leq n \leq N$, have order $O(h_{max}^k)$. Assume Dx_* to be k-times continuously differentiable. Then, the k-step BDF method (5.43) is convergent and globally accurate of order $O(h_{max}^k)$.*

Proof. Following the proof of Theorem 5.4, we see that the error estimation

$$\max_{n\geq k}|x_n - x_*(t_n)| \leq c\Big(\max_{n<k}|D(t_n)x_n - D(t_n)x_*(t_n)| + \max_{n\geq k}|q_n| + \max_{n\geq k}|L_n^u|\Big)$$

is satisfied for a constant $c > 0$ and

$$L_n^u = [u_*]_n' - u_*'(t_n) = \frac{1}{h_n}\sum_{i=0}^{k}\alpha_{ni}u_*(t_{n-i}) - u_*'(t_n) = O(h_{max}^k).$$

Now, the assertion is clear. □

5.5.2 IRK(DAE) method

Before formulating perturbation estimates for IRK(DAE) methods applied to DAEs we present estimations for IRK(DAE) methods applied to ODEs regarding the influence of perturbations caused by rounding and linear/nonlinear solvers.

An IRK(DAE) method for ODEs

$$u'(t) = f(u(t),t) \tag{5.63}$$

can be formulated as

$$\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}(U_{nj} - u_{n-1}) = f(U_{ni},t_{ni}) + q_{ni}, \qquad i = 1,\ldots,s, \tag{5.64}$$

with α_{ij} being the coefficients of $\mathcal{A}^{-1}$, $\mathcal{A}$ being the Runge–Kutta coefficient matrix and q_{ni} reflecting perturbations caused by rounding errors and defects of nonlinear

solvers. The numerical solution u_n at time t_n is given by U_{ns} since $c_s = 1$ and $b_i = a_{si}$ for $i = 1, \ldots, s$ for IRK(DAE) methods.

Lemma 5.6. *Let $u_* \in \mathcal{C}^1(\mathcal{I}_c, \mathbb{R}^m)$ be a solution of the ODE (5.63). Assume f to have the continuous partial derivative f_u. If the deviation in the starting value $|u_0 - u_*(t_0)|$, the stepsize h, the perturbations q_{ni} and the local errors*

$$L_{ni} := \frac{1}{h} \sum_{j=1}^{s} \alpha_{ij}(u_*(t_{nj}) - u_*(t_{n-1})) - u_*'(t_{ni}), \qquad i = 1, \ldots, s, \quad 1 \le n \le N,$$

are sufficiently small, then the IRK(DAE) methods provide numerical solutions U_{ni} fulfilling (5.64). Furthermore, there is a constant $c > 0$ such that

$$\max_{1 \le n \le N} |u_n - u_*(t_n)| \le c \Big(|u_0 - u_*(t_0)| + \max_{1 \le n \le N} \max_{i=1,\ldots,s} (|q_{ni}| + |L_{ni}|) \Big).$$

Proof. Regarding (5.64), we get

$$\begin{aligned} \frac{1}{h} \sum_{j=1}^{s} \alpha_{ij}(U_{nj} - u_{n-1}) - \frac{1}{h} \sum_{j=1}^{s} \alpha_{ij}(u_*(t_{nj}) - u_*(t_{n-1})) \\ = f(U_{ni}, t_{ni}) + q_{ni} - L_{ni} - f(u_*(t_{ni}), t_{ni}) \end{aligned}$$

for $i = 1, \ldots, s$ since u_* is a solution of (5.63). Regarding f_u to be continuous, the implicit function theorem ensures the existence of solutions U_{ni} in a neighborhood of $u_*(t_{ni})$ if q_{ni}, L_{ni} and h are sufficiently small for all $i = 1, \ldots, s$. Furthermore,

$$\frac{1}{h} \sum_{j=1}^{s} \alpha_{ij} \Big((U_{nj} - u_*(t_{nj})) - (u_{n-1} - u_*(t_{n-1})) \Big) = H_{ni}(U_{ni} - u_*(t_{ni})) + q_{ni} - L_{ni}$$

for

$$H_{ni} := \int_0^1 f_u(\tau U_{ni} + (1 - \tau) u_*(t_{ni}), t_{ni}) \, \mathrm{d}\tau, \qquad i = 1, \ldots, s.$$

Introducing

$$U_n := \begin{bmatrix} U_{n1} \\ \vdots \\ U_{ns} \end{bmatrix}, \quad U_{*n} := \begin{bmatrix} u_*(t_{n1}) \\ \vdots \\ u_*(t_{ns}) \end{bmatrix}, \quad q_n := \begin{bmatrix} q_{n1} \\ \vdots \\ q_{ns} \end{bmatrix}, \quad L_n := \begin{bmatrix} L_{n1} \\ \vdots \\ L_{ns} \end{bmatrix},$$

we get

$$\frac{1}{h} (\mathcal{A}^{-1} \otimes I)[(U_n - U_{*n}) - \mathbb{1} \otimes (u_{n-1} - u_*(t_{n-1}))] = H_n(U_n - U_{*n}) + q_n - L_n$$

with a bounded

$$H_n := \begin{bmatrix} H_{n1} & & 0 \\ & \ddots & \\ 0 & & H_{ns} \end{bmatrix}$$

for q_n, L_n and h being sufficiently small. Rearranging this equation yields

$$(I - h(\mathcal{A} \otimes I)H_n)(U_n - U_{*n}) = \mathbb{1} \otimes (u_{n-1} - u_*(t_{n-1})) + h(\mathcal{A} \otimes I)(q_n - L_n).$$

Since H_n is bounded, the matrix $I - h(\mathcal{A} \otimes I)H_n$ is nonsingular for sufficiently small stepsizes h and we find constants $c_1 > 0$, $c_2 > 0$ such that

$$\max_{i=1,\dots,s} |U_{ni} - u_*(t_{ni})| \le (1 + c_1 h)|u_{n-1} - u_*(t_{n-1})| + hc_2 \max_{i=1,\dots,s} (|q_{ni}| + |L_{ni}|).$$

By standard arguments, this implies the existence of a constant $c > 0$ satisfying

$$\max_{1 \le n \le N} \max_{i=1,\dots,s} |U_{ni} - u_*(t_{ni})| \le c \left(|u_0 - u_*(t_0)| + \max_{1 \le n \le N} \max_{i=1,\dots,s} (|q_{ni}| + |L_{ni}|) \right).$$

Since $u_n = U_{ns}$ and $t_n = t_{ns}$ the assertion is proven. □

Using the commutativity of the diagram in Figure 5.3, we may conclude feasibility, error estimations and convergence of IRK(DAE) methods for index-1 DAEs with a properly involved derivative and time-invariant $\operatorname{im} D(t)$. They are given by

$$f([DX]'_{ni}, X_{ni}, t_{ni}) = q_{ni}, \qquad i = 1, \dots, s. \tag{5.65}$$

with

$$[DX]'_{ni} = \frac{1}{h_n} \sum_{j=1}^{s} \alpha_{ij} (D(t_{nj}) X_{nj} - D(t_{n-1}) x_{n-1}), \qquad i = 1, \dots, s,$$

α_{ij} being the coefficients of $\mathcal{A}^{-1}$, $\mathcal{A}$ being the Runge–Kutta coefficient matrix and q_{ni} reflecting perturbations caused by rounding errors and defects of nonlinear solvers. Again, the numerical solution x_n at time t_n is given by X_{ns} since $c_s = 1$ and $b_i = a_{si}$ for $i = 1, \dots, s$ for IRK(DAE) methods.

Theorem 5.7. *Let the DAE (4.1) be regular with index 1, and let the subspace* $\operatorname{im} D(t)$ *be independent of* t. *Let* $x_* \in \mathcal{C}^1_D(\mathcal{I}_c, \mathbb{R}^m)$ *be a solution of the DAE (4.1). If the deviation in the starting value* $\left| D(t_0)x_0 - (Dx_*)(t_0) \right|$, *the stepsize* h, *the perturbations* q_{ni} *and the local errors*

$$L_{ni} := \frac{1}{h} \sum_{j=1}^{s} \alpha_{ij} ((Dx_*)(t_{nj}) - (Dx_*)(t_{n-1})) - (Dx_*)'(t_{ni}), \quad i = 1, \dots, s, \quad 1 \le n \le N,$$

are sufficiently small, then IRK(DAE) methods provide numerical solutions X_{ni} *fulfilling (5.65). Furthermore, there is a constant* $c > 0$ *such that*

$$\max_{1 \le n \le N} |x_n - x_*(t_n)| \le c \Big(|D(t_0)x_0 - (Dx_*)(t_0)| + \max_{1 \le n \le N} \max_{i=1,\dots,s} (|q_{ni}| + |L_{ni}|) \Big).$$

Proof. From Corollary 4.10 we know that the solution x_* can be written as

$$x_*(t) = D(t)^- u_*(t) + Q_0(t)\omega^{\text{pert}}(u_*(t),t,0), \qquad t \in \mathcal{I}_c, \tag{5.66}$$

with u_* being the solution of the IVP

$$u_*'(t) = D(t)\omega^{\text{pert}}(u_*(t),t,0), \qquad u_*(t_0) = D(t_0)x_*(t_0).$$

The term $R'(t)u_*(t)$ vanishes since we have assumed $\operatorname{im} D(t)$ to be time-invariant, (cf. Proposition 4.7 (3)). The continuous function ω^{pert} is implicitly defined on

$$\{(u,t,q) \,|\, (u,q) \in B_\rho((Dx_*)(t),0),\ t \in \mathcal{I}_c\}$$

by Lemma 4.9. It has continuous partial derivatives ω_u^{pert} and ω_p^{pert}. From Proposition 5.2 and (5.53), we may conclude

$$X_{ni} = D(t_{ni})^- U_{ni} + Q(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},q_{ni}), \quad 1 \le n \le N, \quad i = 1,\dots,s, \tag{5.67}$$

are the stage solutions of (5.44) if U_{ni} are the stage solutions of

$$[u]_{ni}' = D(t_n)\omega^{\text{pert}}(U_{ni},t_{ni},q_{ni}), \quad 1 \le n \le N, \quad i = 1,\dots,s, \quad u_0 := D(t_0)x_0, \tag{5.68}$$

with

$$[u]_{ni}' = \frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}(U_{nj} - u_{n-1}), \qquad i = 1,\dots,s.$$

Introducing

$$q_{ni}^u := D(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},q_{ni}) - D(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},0),$$

we see that U_{ni}, $i = 1,\dots,s$, solve equations (5.68) if and only if they satisfy the equation

$$\frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}(U_{nj} - u_{n-1}) = D(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},0) + q_{ni}^u.$$

This is exactly the IRK(DAE) method for the explicit ODE

$$u'(t) = D(t)\omega^{\text{pert}}(u(t),t,0), \quad t \in \mathcal{I}_c,$$

with the perturbations q_{ni}^u of the right-hand side in each step. Lemma 5.6 ensures the existence of the approximate IRK(DAE) solutions U_{ni} supposing the initial error $|u_0 - u_*(t_0)|$, the perturbations $|q_{ni}^u|$, $1 \le n \le N$, $i = 1,\dots,s$, and the local errors

$$L_{ni} = \frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}(u_*(t_{nj}) - u_*(t_{n-1})) - u_*'(t_{ni}), \qquad i = 1,\dots,s, \quad 1 \le n \le N,$$

are sufficiently small. All three demands are satisfied due to the assumptions of the theorem since

$$|u_0 - u_*(t_0)| = |D(t_0)x_0 - D(t_0)x_*(t_0)|,$$

and there is a constant $c_1 > 0$ such that

$$|q^u_{ni}| = |D(t_{ni})\omega^{\text{pert}}(U_{ni}, t_{ni}, q_{ni}) - D(t_{ni})\omega^{\text{pert}}(U_{ni}, t_{ni}, 0)| \le c_1|q_{ni}| \tag{5.69}$$

for $1 \le n \le N$. The existence of a continuous partial derivative ω^{pert}_q ensures (5.69) to be valid. Consequently, the existence of a IRK(DAE) solution X_{ni} satisfying (5.65) is ensured as long as $U_{ni} \in B_\rho(u_*(t_{ni}))$. Applying standard arguments, we find the error estimation

$$\max_{1\le n\le N} |u_{ni} - u_*(t_{ni})| \le c_2\Big(|u_0 - u_*(t_0)| + \max_{n\ge k}\max_{i=1,\dots,s}(|q^u_{ni}| + |L_{ni}|)\Big) \tag{5.70}$$

with a constant $c_2 > 0$. Defining

$$L^w_{ni} := D(t_{ni})^-[u_*]'_{ni} + Q_0(t_{ni})x_*(t_n) = D(t_{ni})^-[Dx_*]'_{ni} + Q_0(t_{ni})x_*(t_{ni}), \tag{5.71}$$

we get

$$f(D(t_{ni})L^w_{ni}, D(t_{ni})^- u_*(t_{ni}) + Q_0(t_{ni})L^w_{ni}, t_{ni}) = f([Dx_*]'_{ni}, x_*(t_{ni}), t_{ni}) =: L^f_{ni},$$

and thus, by means of the function ω^{pert},

$$L^w_{ni} = \omega^{\text{pert}}(u_*(t_{ni}), t_{ni}, L^f_{ni}).$$

Regarding (5.71), we see that

$$D(t_{ni})\omega^{\text{pert}}(u_*(t_{ni}), t_{ni}, L^f_{ni}) = D(t_{ni})L^w_{ni} = R(t_n)[u_*]'_{ni} = [u_*]'_{ni}.$$

Here, we again used that $\operatorname{im} D(t)$ is time-invariant. We have

$$\begin{aligned} L^f_{ni} &= f([Dx_*]'_{ni}, x_*(t_{ni}), t_{ni}) - f((Dx_*)'(t_{ni}), x_*(t_{ni}), t_{ni}) \\ &= f(L_{ni} + (Dx_*)'(t_{ni}), x_*(t_{ni}), t_{ni}) - f((Dx_*)'(t_{ni}), x_*(t_{ni}), t_{ni}). \end{aligned}$$

Since f is continuously differentiable with respect to its first argument, we find a constant $c_3 > 0$ satisfying

$$|L^f_{ni}| \le c_3|L_{ni}|, \quad 1 \le n \le N, \quad i = 1, \dots, s. \tag{5.72}$$

Inserting (5.72) and (5.69) into (5.70), there is a constant $c_4 > 0$ such that

$$\begin{aligned} \max_{1\le n\le N}\max_{i=1,\dots,s} |U_{ni} - u_*(t_{ni})| \le c_4\big(&|D(t_0)x_0 - D(t_0)x_*(t_0)| \\ &+ \max_{1\le n\le N}\max_{i=1,\dots,s}(|q_{ni}| + |L_{ni}|)\big). \end{aligned}$$

Obviously, $U_{ni} \in B_\rho(u_*(t_{ni}))$ is always satisfied if the deviations in the initial values, the perturbations q_n and the local error are sufficiently small. Hence, the existence

of an IRK(DAE) solution x_n satisfying (5.65) is proven. Regarding the solution representations (5.66), (5.67) and the fact that ω_u^{pert} and ω_q^{pert} exist and are continuous, we obtain the estimation

$$\max_{n\geq k} |X_{ni} - x_*(t_{ni})| \leq c\big(|D(t_0)x_0 - D(t_0)x_*(t_0)| + \max_{1\leq n\leq N}\max_{i=1,\ldots,s}(|q_{ni}| + |L_{ni}|)\big)$$

with a constant $c > 0$. Since $x_n = X_{ns}$ and $t_n = t_{ns}$, the proof is complete. □

As a direct consequence of Theorem 5.7 the following results:

Corollary 5.8. *(Convergence) Let the assumptions of Theorem 5.7 be satisfied. Suppose the errors in the initial values, the perturbations q_{ni} and the local errors L_{ni}, $1 \leq n \leq N$, $i = 1,\ldots,s$ have order $O(h^k)$. Then, the IRK(DAE) method (5.65) is convergent and globally accurate of order $O(h^k)$.*

We mention further that if the method coefficients satisfy the $C(k)$ condition (see [29])

$$\sum_{j=1}^{s} a_{ij} c_j^{\ell-1} = \frac{1}{\ell} c_i^{\ell}, \qquad i = 1,\ldots,s, \quad \ell = 1,\ldots,k, \tag{5.73}$$

and Dx_* is k-times continuously differentiable, then the local errors L_{ni} are of order $O(h^k)$.

5.5.3 General linear method

We consider the formulation (5.41) of a stiffly stable general linear method applied to index-1 DAEs (4.1):

$$f\left(\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}\Big(D(t_{nj})X_{nj} - \sum_{\ell=1}^{r}\mu_{j\ell}[Dx]_{\ell}^{[n-1]}\Big), X_{ni}, t_{ni}\right) = q_{ni}, \quad i = 1,\ldots,s, \tag{5.74}$$

and $[Dx]^{[n]}$ is given recursively by

$$\begin{bmatrix}[Dx]_1^{[n]}\\ \vdots \\ [Dx]_r^{[n]}\end{bmatrix} = (\mathcal{B}\otimes I_n)(\mathcal{A}^{-1}\otimes I_n)([DX]_n - (\mathcal{U}\otimes I_n)[Dx]^{n-1}) + (\mathcal{V}\otimes I_n)[Dx]^{[n-1]} \tag{5.75}$$

with

$$[DX]_n = \begin{bmatrix} D(t_{n1})X_{n1}\\ \vdots \\ D(t_{ns})X_{ns}\end{bmatrix}.$$

The coefficients $\alpha_{j\ell}$ and $\mu_{j\ell}$ are the entries of the coefficient matrices $\mathcal{A}^{-1}$ and $\mathcal{U}$. The perturbations q_{ni} reflect the rounding errors and defects caused by nonlinear solvers.

Theorem 5.9. *Let the DAE (4.1) be regular with index 1, and let the subspace* $\operatorname{im} D(t)$ *be independent of* t*. Assume* $x_* \in \mathcal{C}^1_D(\mathcal{I}_c, \mathbb{R}^m)$ *to be a solution of the DAE (4.1) and* $x_n = X_{ns}$ *to be the numerical solution of a stiffly stable GLM. Suppose the GLM has stage order* p *for explicit ODEs with perturbations and deviations in the starting values of magnitude* $\mathcal{O}(h^p)$*.*
If $[Dx]^{[0]} - [Dx_*]^{[0]} = \mathcal{O}(h^p)$ *and* $q_{ni} = \mathcal{O}(h^p)$ *for all* $1 \le n \le N$ *and* $i = 1, \dots, s$ *then there is a constant* $c > 0$ *such that* $|x_n - x(t_n)| \le ch^p$ *for all* $1 \le n \le N$*.*

Proof. From Corollary 4.10 we know that the solution x_* can be written as

$$x_*(t) = D(t)^- u_*(t) + Q_0(t)\omega^{\mathrm{pert}}(u_*(t), t, 0), \qquad t \in \mathcal{I}_c, \tag{5.76}$$

with u_* being the solution of the IVP

$$u_*'(t) = D(t)\omega^{\mathrm{pert}}(u_*(t), t, 0), \qquad u_*(t_0) = D(t_0)x_*(t_0).$$

The term $R'(t)u_*(t)$ vanishes since we have assumed $\operatorname{im} D(t)$ to be time-invariant, (cf. Proposition 4.7 (3)). The continuous function ω^{pert} is implicitly defined on

$$\{(u, t, q) \,|\, (u, q) \in B_\rho((Dx_*)(t), 0),\ t \in \mathcal{I}_c\}$$

by Lemma 4.9. It has continuous partial derivatives ω_u^{pert} and ω_p^{pert}. From Proposition 5.2 and (5.53), we may conclude

$$X_{ni} = D(t_{ni})^- U_{ni} + Q(t_{ni})\omega^{\mathrm{pert}}(U_{ni}, t_{ni}, q_{ni}), \quad 1 \le n \le N, \quad i = 1, \dots, s, \tag{5.77}$$

are the stage solutions of (5.44) if U_{ni} are the stage solutions of

$$[u]'_{ni} = D(t_n)\omega^{\mathrm{pert}}(U_{ni}, t_{ni}, q_{ni}), \quad 1 \le n \le N, \quad i = 1, \dots, s, \tag{5.78}$$

with

$$[u]'_{ni} = \frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}\Big(U_{nj} - \sum_{\ell=1}^{r} \mu_{j\ell}[u]_\ell^{[n-1]}\Big), \qquad i = 1, \dots, s$$

and

$$[u]^{[0]} - [u_*]^{[0]} = \mathcal{O}(h^p).$$

Introducing

$$q^u_{ni} := D(t_{ni})\omega^{\mathrm{pert}}(U_{ni}, t_{ni}, q_{ni}) - D(t_{ni})\omega^{\mathrm{pert}}(U_{ni}, t_{ni}, 0),$$

we see that U_{ni}, $i = 1, \dots, s$, solve the equations (5.78) if and only if they satisfy the equation

$$\frac{1}{h}\sum_{j=1}^{s} \alpha_{ij}\Big(U_{nj} - \sum_{\ell=1}^{r} \mu_{j\ell}[u]_\ell^{[n-1]}\Big) = D(t_{ni})\omega^{\mathrm{pert}}(U_{ni}, t_{ni}, 0) + q^u_{ni}.$$

This is exactly the GLM for the explicit ODE

$$u'(t) = D(t)\omega^{\text{pert}}(u(t),t,0), \quad t \in \mathcal{I}_c,$$

with the perturbations q_{ni}^u of the right-hand side in each step. The existence of a continuous partial derivative ω_q^{pert} provides a constant $c_1 > 0$ satisfying

$$|q_{ni}^u| = |D(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},q_{ni}) - D(t_{ni})\omega^{\text{pert}}(U_{ni},t_{ni},0)| \le c_1|q_{ni}|$$

for $1 \le n \le N$. Consequently, $q_{ni}^u = \mathcal{O}(h^p)$. Since the GLM has stage order p for explicit ODEs, we obtain $U_{ni} - u_*(t_{ni}) = \mathcal{O}(h^p)$, $i = 1,\dots,s$. Regarding the solution representations (5.76), (5.77) and the fact that ω_u^{pert} and ω_q^{pert} exist and are continuous, we obtain a constant $c > 0$ such that

$$|X_{ni} - x_*(t_{ni})| \le ch^p, \quad i = 1,\dots,s.$$

Since the GLM is stiffly stable, we have $x_n = X_{ns}$ as well as $t_n = t_{ns}$ and the proof is complete. □

5.6 Notes and references

(1) We do not at all reflect the enormous amount of literature concerning numerical integration methods for explicit ODEs and index-1 DAEs in standard formulation. We have mentioned merely several sources in the corresponding subsections.
Convergence results and perturbation estimations for multistep methods and Runge–Kutta methods applied to index-1 DAEs in standard form have been well-known for a long time (e.g., [90, 96, 25, 103, 105]).

(2) To a large extent, concerning BDF and Runge–Kutta methods applied to index-1 DAEs with properly stated leading term. Our presentation follows the lines of [96, 114, 115]. First convergence proofs and error estimations for quasilinear DAEs can be found in [114, 113, 115], which, in turn, follow the lines of [96].
For general linear methods, we summarize the results of [211, 198, 197].

(3) More detailed results including the existence of numerical solutions, consistency and stability of GLMs for DAEs are given in [211].
In [199, 211], general linear methods have been extended to linear DAEs with properly stated leading term, and to implicit nonlinear DAEs of the form

$$y'(t) = f(y(t),z'(t)), \quad z(t) = g(y(t),t)$$

with $I - f_{z'}g_y$ being nonsingular. Such a nonlinear DAE has differentiation index 1, but perturbation index 2. Notice that this DAE does not meet the form (4.1) we are interested in. The slightly modified version

$$y'(t) = f(y(t),(g(y(t),t))'), \quad z(t) = g(y(t),t)$$

meets the conditions discussed in Chapter 3, and it is a regular DAE with tractability index 1 (cf. Example 3.54, Definition 3.28). Here, we follow the approach of [199, 211], and apply it to the general index-1 DAE of the form (4.1)

$$f((Dx)'(t), x(t), t) = 0.$$

(4) Runge–Kutta methods with $c_s = 1$ and $b_i = a_{si}$ for $i = 1, \ldots, s$, are known as *stiffly accurate* methods [29] since they are well suited for stiff explicit ODEs. DAEs are frequently considered as infinitely stiff systems since one can describe them, under certain conditions, as a limit of singular perturbed systems with a small parameter ε tending to zero. However, the solutions of the resulting DAE need not have a stiff behavior. As a simple example, we regard the system

$$x_1'(t) = -x_2(t), \quad 0 = 3x_1(t) - x_2(t).$$

It can be considered as a limit of the singular perturbed system

$$x_1'(t) = -x_2(t), \quad \varepsilon x_2'(t) = 3x_1(t) - x_2(t)$$

for $\varepsilon \to 0$. Whereas the singular perturbed system is stiff if $0 < \varepsilon \ll 1$, the DAE solution

$$x_1(t) = e^{3(t_0 - t)} x_1(t_0), \quad x_2(t) = 3e^{3(t_0 - t)} x_1(t_0)$$

does not show any stiffness. Nevertheless, stiffly accurate Runge–Kutta methods are particularly suited for DAEs for the reasons explained before. The solution x_n is enforced to fulfill the constraints of the system. Since we want to stress this essential property, we follow here the notation of [96] and call them *IRK(DAE) methods*. An IRK(DAE) is an implicit Runge–Kutta method that is particularly suitable for DAEs because of the nonsingular $\mathcal{A}$, and the conditions $c_s = 1$ and $b_i = a_{si}$, $i = 1, \ldots, s$.

(5) A first implementation of the BDF method for DAEs with a properly involved derivative in C++ is presented in [76]. Furthermore, it is used in commercial software packages, for instance in the in-house simulator TITAN of Infineon Technologies and in the multiphysics network simulator MYNTS developed at the University of Cologne together with the Fraunhofer Institute SCAI.

(6) Special GLMs described in Subsection 5.3.3 are implemented in FORTRAN as the software package GLIMDA, see [210].

(7) Often DAEs can be reformulated such that $\operatorname{im} D(t)$ is independent of t. It is even so that one finds a formulation with a full row rank matrix $D(t)$, $\operatorname{im} D(t) = \mathbb{R}^n$. For a detailed discussion of possible reformulations, we refer to [115] (see also Section 3.11). Some benefits of certain constant subspaces were already observed in [96, 83] for DAEs in standard form, and in particular, contractivity results are obtained in this way.

(8) In [96], dealing with standard form DAEs, it is pointed out that one can benefit from a constant leading nullspace $\ker \mathfrak{f}_{x^1}$, such that the given integration method arrives unchanged at an inherent ODE and one obtains, for instance, B-stability.

Applying modified Runge–Kutta methods to linear standard form DAEs, in [83, 112] similar advantages are obtained from a time-invariant range $\operatorname{im} E(t)$ of the leading coefficient. Altogether, these properties led us to the notion of *numerically qualified* formulations of index-1 DAEs. A regular index-1 DAE

$$f((D(t)x(t))',x(t),t) = 0,$$

with properly stated leading term is said to be in numerically qualified form, if $\operatorname{im} D(t)$ is actually time-invariant, see [115].

Chapter 6
Stability issues

Here we consider DAEs and their numerical approximations on infinite time intervals. We discuss contractivity, dissipativity and stability in Lyapunov's sense. With the help of this structural insight, we show that it is reasonable to formulate the DAE itself in a numerically qualified manner, which means, one should have a DAE with properly stated leading term with, additionally, $\operatorname{im} D(t)$ being independent of t. Then the integration methods are passed unchanged to the inherent ODE, and the numerical approximations reflect the qualitative solution behavior as well as in the case of explicit ODEs and one avoids additional stepsize restrictions.
Section 6.1 describes the basic notions for explicit ODEs. Sections 6.2 and 6.3 comprise the notions of contractivity and dissipativity, as well as flow properties of contractive and dissipative DAEs. Then, it is discussed how these properties are reflected by numerical approximations. Section 6.4 presents a stability criterion by means of linearization, also including solvability results on the infinite time interval.

6.1 Preliminaries concerning explicit ODEs

Stable systems play their role in theory and applications. As a stable system one usually has in mind a linear ODE

$$x'(t) = Cx(t) + q(t), \tag{6.1}$$

the coefficient matrix C of which has exclusively eigenvalues with nonpositive real parts, and the purely imaginary eigenvalues are nondefective. Each pair of solutions $x(\cdot)$, $\bar{x}(\cdot)$ of such an ODE satisfies the inequality

$$|x(t) - \bar{x}(t)| \leq \mathrm{e}^{-\beta(t-t_0)}|x(t_0) - \bar{x}(t_0)|, \quad \text{for all } t \geq t_0, \tag{6.2}$$

in a suitable norm, and with a nonnegative constant β. The norm as well as the value of β are determined by the coefficient matrix C (cf. Appendix C, Lemma C.3).

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_6,

Namely, to each arbitrary $m \times m$ matrix C with real entries, which has the prescribed properties, there exist a value $\beta \geq 0$ and an inner product $\langle \cdot, \cdot \rangle$ inducing the $\mathbb{R}^m$-norm $|\cdot|$ such that

$$\langle C(x-\bar{x}), x-\bar{x} \rangle \leq -\beta |x-\bar{x}|^2, \quad \text{for all } x, \bar{x} \in \mathbb{R}^m. \tag{6.3}$$

In other words, the function $|\cdot|^2$ is a Lyapunov function for the homogeneous ODE $x'(t) = Cx(t)$.
If C has no eigenvalues on the imaginary axis, then the constant β is strictly positive, and the system is said to be *asymptotically stable*.
Each arbitrary solution of a stable ODE (6.1) is *stable in the sense of Lyapunov* (e.g., Definition 6.3 below), and each arbitrary solution of such an asymptotically stable homogeneous ODE is asymptotically stable in Lyapunov's sense.

Given a stepsize $h > 0$, as well as grid points $t_n = t_0 + nh$, each solution pair of a stable ODE (6.1) satisfies the inequalities

$$|x(t_n) - \bar{x}(t_n)| \leq \mathrm{e}^{-\beta h} |x(t_{n-1}) - \bar{x}(t_{n-1})|, \quad \text{for all } n > 0, \tag{6.4}$$

no matter how large the stepsize is chosen. Among the step-by-step numerical integration methods, the *absolutely stable (A-stable)* ones (e.g., [104], [29]) are intended to reflect this solution property devoid of restrictions on the stepsize h for this reason.

There are various popular generalizations of stability (e.g. [204]) for nonlinear ODEs

$$x'(t) = g(x(t), t). \tag{6.5}$$

Typically, the function g is continuous, with the continuous partial derivative g_x on $\mathbb{R}^m \times [0, \infty)$. In particular, *contractivity* and *dissipativity* generalize stability as global properties of the ODE, applying to all solutions. In contrast, *(asymptotical) stability in the sense of Lyapunov* applies just locally to a reference solution.

Definition 6.1. The ODE (6.5) is named *contractive*, if there are a constant $\beta \geq 0$ and an inner product $\langle \cdot, \cdot \rangle$, such that

$$\langle g(x,t) - g(\bar{x},t), x-\bar{x} \rangle \leq -\beta |x-\bar{x}|^2, \quad \text{for all } x, \bar{x}, \in \mathbb{R}^m,\ t \geq 0. \tag{6.6}$$

If $\beta > 0$, then we speak of *strongly contractive* ODEs.

Contractivity means a contractive flow, although the formal definition is given in terms of g via the so-called one-sided Lipschitz condition. All solutions of a contractive ODE exist on the infinite interval, and for each pair of them, the inequality (6.2) holds true. A strongly contractive ODE has at most one stationary solution (see Proposition C.2). For linear ODEs, contractivity is equivalent to stability, and strong contractivity is the same as asymptotical stability.

Regarding numerical integration, so-called *B-stable Runge–Kutta methods* (e.g., [104], [29]) reflect contractivity devoid of stepsize restrictions by

$$|x_n - \bar{x}_n| \le |x_{n-1} - \bar{x}_{n-1}|, \quad \text{for all } n > 0.$$

Dissipative ODEs (6.5) are those which possess an absorbing set, i.e., a bounded, positively invariant set sucking in all solutions. A large class of dissipative ODEs is characterized in terms of g by the dissipativity inequality below that ensures that the solutions exist on the infinite interval, and to be absorbed in balls around the origin, with radius $\varepsilon + \alpha/\beta$, for any $\varepsilon > 0$.

Definition 6.2. The ODE (6.5) satisfies the *dissipativity inequality*, if there are constants $\beta > 0$, $\alpha \ge 0$, and an inner product $\langle \cdot, \cdot \rangle$, such that

$$\langle g(x,t), x \rangle \le \alpha - \beta |x|^2, \quad \text{for all } x \in \mathbb{R}^m,\ t \ge 0. \tag{6.7}$$

Notice that asymptotically stable homogeneous linear ODEs (6.1) satisfy the inequality (6.7) with $\alpha = 0$. The origin is the only stationary solution of these systems, and balls around the origin serve as absorbing sets.

In contrast to the previous global system properties, the next stability notion is tied up with a reference solution.

Definition 6.3. A solution $x_* \in \mathcal{C}^1([0,\infty), \mathbb{R}^m)$ of the ODE (6.5) is said to be

(1) *stable in the sense of Lyapunov*, if for every $\varepsilon > 0$, $t_0 \ge 0$, there is a $\delta(\varepsilon, t_0) > 0$ such that

$$|x_*(t_0) - x_0| < \delta(\varepsilon, t_0)$$

implies the existence of a solution $x(\cdot; t_0, x_0)$ on the entire infinite interval as well as the estimation

$$|x_*(t) - x(t; t_0, x_0)| < \varepsilon \ \text{ for } t \ge t_0,$$

(2) *asymptotically stable*, if for each $\varepsilon > 0$, $t_0 \ge 0$, there is a $\delta(\varepsilon, t_0) > 0$ such that

$$|x_*(t_0) - x_0| < \delta(\varepsilon, t_0)$$

implies the existence of a solution $x(\cdot; t_0, x_0)$ on the infinite interval as well as the limit

$$|x_*(t) - x(t; t_0, x_0)| \xrightarrow[t\to\infty]{} 0.$$

Making use of the inherent structure of index-1 DAEs, as provided in Section 4.2, we modify the stability notions to become reasonable for nonlinear index-1 DAEs (4.1) in the next section. Here we add slight generalizations of contractivity and dissipativity for this aim.

Definition 6.4. Let the ODE (6.5) have the possibly time-dependent invariant subspace $\mathfrak{G}(t) \subseteq \mathbb{R}^m, t \ge 0$, that is, if for an ODE solution $x_*(\cdot)$ there is a $t_0 \ge 0$ such that $x_*(t_0) \in \mathfrak{G}(t_0)$, then $x_*(t)$ belongs to $\mathfrak{G}(t)$ all time.

(1) The ODE is named *contractive on* $\mathfrak{G}(\cdot)$, if there are a constant $\beta \ge 0$ and an inner product $\langle \cdot, \cdot \rangle$, such that

$$\langle g(x,t)-g(\bar{x},t),x-\bar{x}\rangle \le -\beta|x-\bar{x}|^2, \text{ for all } x,\bar{x}\in\mathfrak{G}(t)\subseteq\mathbb{R}^m,\ t\ge 0. \quad (6.8)$$

If $\beta<0$, then we speak of *strong contractivity on* $\mathfrak{G}$.

(2) The ODE satisfies the *dissipativity inequality on* $\mathfrak{G}(\cdot)$, if there are constants $\beta>0$, $\alpha\ge 0$, and an inner product $\langle\cdot,\cdot\rangle$, such that

$$\langle g(x,t),x\rangle \le \alpha-\beta|x|^2, \quad \text{for all } x\in\mathfrak{G}(t),\ t\ge 0. \quad (6.9)$$

All solutions of an ODE being contractive on $\mathfrak{G}(\cdot)$, which start in $\mathfrak{G}(t_0)$ at time $t_0\ge 0$, exist on the infinite interval, and for each pair of them, the inequality (6.2) is valid.

6.2 Contractive DAEs and B-stable Runge–Kutta methods

In this part, the definition domain of the index-1 DAE (4.1) is assumed to be such that $\mathcal{I}_f=[0,\infty)$, and the function $D\omega$ is supposed to be given on the domain $\operatorname{dom} D\omega=\mathbb{R}^n\times\mathcal{I}_f$.

The flow of an explicit ODE (6.5) takes over the entire space $\mathbb{R}^m$, and this state space is, trivially, independent of t. In contrast, the flow of a DAE is restricted to the obvious constraint set $\mathcal{M}_0(t)$, which in turn may move with time (e.g., Examples 3.7, 4.8). This is the view from outside. The inner structure of an index-1 DAE, as described in Section 4.2, shows the IERODE flow within $\operatorname{im}D(\cdot)$ that is somehow wrapped up to become the DAE flow. In the case of linear DAEs, the wrapping is given by means of the canonical projector function Π_{can}. In the light of Example 2.57, we direct our interest to the contractivity of the IERODE, as well as to a wrapping which does not amplify the IERODE flow unboundedly.

In the nonlinear index-1 case, the canonical projector function Π_{can} projects onto $N=\ker f_yD=\ker D$ along $S=\{z\in\mathbb{R}^n:\ f_xz\in\operatorname{im}f_y=\operatorname{im}f_yD\}$ (cf. page 320). By Lemma A.10, Π_{can} can be described as $\Pi_{can}=I-Q_0G^{-1}f_x$. We make use of this representation, when calculating the difference of two solutions.

Applying the solution representation provided by Theorem 4.5, for any two solutions $x(\cdot),\bar{x}(\cdot)\in\mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, defined on an interval $\mathcal{I}\subseteq\mathcal{I}_f$, and $u(\cdot):=D(\cdot)x(\cdot)$, $\bar{u}(\cdot):=D(\cdot)\bar{x}(\cdot)$, we describe the difference

$$\begin{aligned}
x(t)-\bar{x}(t) &= D(t)^-(u(t)-\bar{u}(t))+Q_0(t)(\omega(u(t),t)-\omega(\bar{u}(t),t))\\
&= \int_0^1 [D(t)^- + Q_0(t)\omega_u'(su(t)+(1-s)\bar{u}(t),t)]\,\mathrm{d}s\,(u(t)-\bar{u}(t))\\
&= \int_0^1 [D(t)^- - Q_0(t)(G^{-1}f_x)(\eta_{(u(t),\bar{u}(t),t)}(s))D(t)^-]\,\mathrm{d}s\,(u(t)-\bar{u}(t))\\
&= \int_0^1 \Pi_{can}(\eta_{(u(t),\bar{u}(t),t)}(s))D(t)^-\,\mathrm{d}s\,(u(t)-\bar{u}(t)), \qquad (6.10)
\end{aligned}$$

with

$$\eta_{(u(t),\bar{u}(t),t)}(s) := \big(D(t)\omega(su(t)+(1-s)\bar{u}(t),t),$$
$$D(t)^-(su(t)+(1-s)\bar{u}(t))+Q_0(t)\omega(su(t)+(1-s)\bar{u}(t),t),\ t\big),$$

which suggests that we trace back the contractivity question for the DAE to that for the IERODE

$$u'(t) = R'(t)u(t)+D(t)\omega(u(t),t) \tag{6.11}$$

on its invariant subspace $\operatorname{im} D(\cdot)$, but then to suppose that the product $\Pi_{can}(\cdot)D(\cdot)^-$ remains bounded.

One could believe that the matrix function $\Pi_{can}(\cdot)D(\cdot)^-$ depends on the choice of the projector function $P_0(\cdot)$, however this is not the case. Namely, if there are two different P_0 and $\tilde{P}_0$, as well as the corresponding D and $\tilde{D}$, one can compute $\Pi_{can}\tilde{D}^- = \Pi_{can}P_0\tilde{D}^- = \Pi_{can}D^-D\tilde{D}^- = \Pi_{can}D^-R = \Pi_{can}D^-$.

The IERODE (6.11) is contractive on $\operatorname{im} D(\cdot)$ (Definition 6.4), if there exist an inner product and a constant $\beta \geq 0$ such that

$$\langle D(t)(\omega(u,t)-\omega(\bar{u},t))+R'(t)(u-\bar{u}),u-\bar{u}\rangle \leq \beta|u-\bar{u}|^2, \text{ for } u,\bar{u}\in\operatorname{im} D(t),\ t\geq 0. \tag{6.12}$$

Since the inherent ODE is given implicitly only, we look for a criterion in terms of the original DAE.

Definition 6.5. The regular index-1 DAE (4.1) is said to be *contractive* if there are an inner product $\langle\cdot,\cdot\rangle$ and a value $\beta \geq 0$ such that

$$\langle y-\bar{y},D(t)(x-\bar{x})\rangle + \langle R'(t)D(t)(x-\bar{x}),D(t)(x-\bar{x})\rangle \leq -\beta|D(t)(x-\bar{x})|^2, \tag{6.13}$$

for all $x,\bar{x}\in\mathcal{M}_0(t)$, $y,\bar{y}\in\operatorname{im} D(t)$, $t\geq 0$, satisfying $f(y,x,t)=0$, $f(\bar{y},\bar{x},t)=0$.
If $\beta > 0$, then we speak of *strong contractivity*.

The next theorem confirms this definition in view of reasonable solution properties. We mention at this point that Definition 6.5 is a straight generalization of Definition 6.1 given for explicit ODEs. Namely, letting $f(y,x,t) = y-g(x,t)$, $n=m$, $D(t)=I$ it results that $R=I$, $\mathcal{M}_0(t)=\mathbb{R}^m$, and the inequality (6.13) says nothing but (6.6). A straightforward check makes it clear that an index-1 DAE (4.1) is (strongly) contractive, if its IERODE is so on $\operatorname{im} D$.
For a linear constant coefficient index-1 DAE

$$A(Dx(t))'+Bx(t) = q(t)$$

the condition (6.13) simplifies to

$$\langle -DG^{-1}BD^-z,z\rangle \leq -\beta|z|^2, \quad \text{for all } z\in\operatorname{im} D.$$

This DAE is contractive, if the finite eigenvalues of the matrix pair $\{AD,B\}$ have nonnegative real parts, and the eigenvalues on the imaginary axis are nondefective. It is strongly contractive, if the finite eigenvalues of $\{AD,B\}$ have positive real parts (cf. Section 1.4).

In the theory of (explicit) ODEs and dynamical systems, stationary solutions play an important role. From the viewpoint of DAEs, it seems to be reasonable to consider also solutions having a stationary IERODE component only. A particular such case can be found in Example 4.8, with $c = 0$.

Definition 6.6. A solution $x_* \in \mathcal{C}_D^1(\mathcal{J}, \mathbb{R}^m)$ of the index-1 DAE (4.1) is called a *solution with stationary core*, if it has the form

$$x_*(t) = D(t)^- c + Q_0(t)\omega(c,t),\ t \in \mathcal{I},$$

where $c \in \operatorname{im} D(t)$, $t \in \mathcal{I}$, is a stationary solution of the IERODE.

Theorem 6.7. *A contractive index-1 DAE (4.1), with $\mathcal{I}_f = [0,\infty)$, $\operatorname{dom}_{D\omega} = \mathbb{R}^n \times \mathcal{I}_f$, features the following:*

(1) *The IERODE (6.11) is contractive on* $\operatorname{im} D$.

(2) *For each arbitrary $t_0 \in \mathcal{I}_f$, $x_0 \in \mathcal{M}_0(t_0)$, the DAE has a unique solution satisfying the condition $x(t_0) = x_0$. This solution exists on the entire infinite interval $\mathcal{I} = \mathcal{I}_f$.*

(3) *Any pair of solutions fulfills the inequalities*

$$|D(t)(x(t) - \bar{x}(t))| \le e^{-\beta(t-t_0)} |D(t_0)(x(t_0) - \bar{x}(t_0))|, \quad t \ge t_0,$$

and

$$|x(t) - \bar{x}(t)| \le K_{(x,\bar{x})}(t) |D(t)(x(t) - \bar{x}(t))|, \quad t \ge t_0,$$

with

$$\begin{aligned} K_{(x,\bar{x})}(t) :&= \max_{s\in[0,1]} \max_{z \in \operatorname{im} D(t),\, |z|=1} |\Pi_{can}(\eta_{((Dx)(t),(D\bar{x})(t),t)}(s)) D(t)^- z| \\ &\le \max_{s\in[0,1]} |\Pi_{can}(\eta_{((Dx)(t),(D\bar{x})(t),t)}(s)) D(t)^-|. \end{aligned}$$

(4) *If the matrix function $\Pi_{can} D^-$ is uniformly bounded by the constant K, then it follows that*

$$\begin{aligned} |x(t) - \bar{x}(t)| &\le K\, |D(t)(x(t) - \bar{x}(t))| \\ &\le K e^{-\beta(t-t_0)} |D(t_0)(x(t_0) - \bar{x}(t_0))|, \quad t \ge t_0. \end{aligned}$$

(5) *If $\beta > 0$, then the DAE possesses one solution with a stationary core at the most.*

Proof. (1) We show the contractivity of the IERODE (6.11) on $\operatorname{im} D$. For $t \ge 0$, $u, \bar{u} \in \operatorname{im} D(t)$, we form

$$\begin{aligned} x &:= Q_0(t)\omega(u,t) + D(t)^- u \in \mathcal{M}_0(t), & y &:= D(t)\omega(u,t) \in \operatorname{im} D(t), \\ \bar{x} &:= Q_0(t)\omega(\bar{u},t) + D(t)^- \bar{u} \in \mathcal{M}_0(t), & \bar{y} &:= D(t)\omega(\bar{u},t) \in \operatorname{im} D(t), \end{aligned}$$

which implies $D(t)x = u$, $D(t)\bar{x} = \bar{u}$, $f(y,x,t) = 0$, $f(\bar{y},\bar{x},t) = 0$ due to the construction of the function ω. The contractivity assumption (6.13) gives

$$\langle D(t)(\omega(u,t) - \omega(\bar{u},t)), u - \bar{u}\rangle + \langle R'(t)(u-\bar{u}), u-\bar{u}\rangle \leq -\beta|u-\bar{u}|^2,$$

and hence, the inherent ODE (6.11) is contractive on $\operatorname{im} D$ with the same inner product and constant β as in (6.13).

(2) Theorem 4.11, item (1), provides the local existence of a unique solution passing through x_0 at time t_0. Due to the contractivity of the IERODE on $\operatorname{im} D$, this solution can be continued to the entire infinite interval.

(3) Let two solutions $x(\cdot)$ and $\bar{x}(\cdot)$ be given. The components $u(\cdot) := D(\cdot)x(\cdot)$ and $\bar{u}(\cdot) := D(\cdot)\bar{x}(\cdot)$ satisfy the IERODE, such that

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u(t)-\bar{u}(t)|^2 &= \langle u'(t)-\bar{u}'(t), u(t)-\bar{u}(t)\rangle \\ &= \langle R(t)(u'(t)-\bar{u}'(t)) + R'(t)(u(t)-\bar{u}(t)), u(t)-\bar{u}(t)\rangle \\ &= \langle D(t)(\omega(u(t),t)-\omega(\bar{u}(t),t)) + R'(t)(u(t)-\bar{u}(t)), u(t)-\bar{u}(t)\rangle,\end{aligned}$$

and regarding the contractivity we conclude that

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u(t)-\bar{u}(t)|^2 \leq -\beta|u(t)-\bar{u}(t)|^2.$$

Now the Gronwall lemma leads to the first inequality in assertion (3), that is

$$|u(t)-\bar{u}(t)| \leq e^{-\beta(t-t_0)}|u(t_0)-\bar{u}(t_0)|, \quad t \geq t_0.$$

The second inequality follows from (6.10), since (6.10) provides us with the estimate

$$\begin{aligned}|x(t)-\bar{x}(t)| &\leq \int_0^1 |\Pi_{can}(\eta_{(u(t),\bar{u}(t),t)}(s))D(t)^-(u(t)-\bar{u}(t))|\,\mathrm{d}s \\ &\leq K_{(x,\bar{x})}(t)|(Dx)(t)-(D\bar{x})(t)|.\end{aligned}$$

(4) If $\Pi_{can}(\cdot)D(\cdot)^-$ is uniformly bounded by the constant K, then assertion (4) results from (3).
(5) If c and $\bar{c}$ are stationary solutions of the IERODE, then it follows from contractivity that $|c-\bar{c}| \leq e^{-\beta(t-t_0)}|c-\bar{c}|$, therefore $c = \bar{c}$. Since the IERODE has at most one stationary solution, due to Theorem 4.5 describing the structure of the DAE solution, the DAE has at most one solution with a stationary core. □

Example 6.8 (Contractive DAE). We continue to investigate the DAE from Example 4.8,

$$\begin{aligned}x_1'(t) + \beta x_1(t) &= 0,\\ x_1(t)^2 + x_2(t)^2 - 1 &= \gamma(t),\end{aligned}$$

which is contractive for $\beta \geq 0$. The inner product is the usual product of real numbers, and the given β applies.
The solutions of the DAE related to the region $\mathcal{G}_+$ are

$$x_1(t) = \mathrm{e}^{-\beta(t-t_0)}x_{01},$$
$$x_2(t) = (1+\gamma(t)+\mathrm{e}^{-2\beta(t-t_0)}x_{01}^2)^{\frac{1}{2}}.$$

Using the canonical projector function Π_{can} given in Example 4.8 we find that

$$\Pi_{can}(y,x,t)D(t)^- = \begin{bmatrix} 1 \\ -\frac{x_1}{x_2} \end{bmatrix}$$

is not globally bounded. Compute, additionally, the difference of the second components of two solutions corresponding to the initial data x_{01} and $\bar{x}_{01}$,

$$\begin{aligned} x_2(t) - \bar{x}_2(t) &= (1+\gamma(t)+\mathrm{e}^{-2\beta(t-t_0)}x_{01}^2)^{\frac{1}{2}} - (1+\gamma(t)+\mathrm{e}^{-2\beta(t-t_0)}\bar{x}_{01}^2)^{\frac{1}{2}} \\ &= \frac{1}{2}(1+\gamma(t)+\mathrm{e}^{-2\beta(t-t_0)}c)^{-\frac{1}{2}}\mathrm{e}^{-2\beta(t-t_0)}(x_{01}^2 - \bar{x}_{01}^2), \end{aligned}$$

with any $c \in [x_{01}^2, \bar{x}_{01}^2]$, which confirms the preciseness of the estimations given by Theorem 6.7 (3). One can benefit from a bounded product $\Pi_{can}D^-$, if both solutions reside on a bounded set with respect to x_1, as well as bounded away from the border $x_2 = 0$. It strongly depends on the function γ whether this is possible or not. In general, a disadvantageous function γ might force the solutions, and the differences of solutions, to grow. □

Next, we discuss to what extent numerical solutions generated by Runge–Kutta methods reflect the contractive behavior of the DAE solutions. For explicit ODEs, this question is well-known to be answered by the B-stability concept. In particular, *algebraically stable Runge–Kutta methods* are *B-stable* ([30], [54]), which means that every two sequences of numerical solutions generated step-by-step along a given grid

$$t_0 < t_1 < \cdots < t_{n-1} < t_n < \cdots$$

by an algebraically stable Runge–Kutta method, if applied to a contractive explicit ODE (6.5) satisfy the inequality

$$|x_n - \bar{x}_n| \leq |x_{n-1} - \bar{x}_{n-1}|, \quad \text{for all } n \geq 1, \tag{6.14}$$

and this inequality is a correct counterpart of the true solution property

$$|x(t_n) - \bar{x}(t_n)| \leq |x(t_{n-1}) - \bar{x}(t_{n-1})|, \quad \text{for all } n \geq 1. \tag{6.15}$$

The point is here that (6.14) reflects this contractivity behavior independently of the chosen stepsizes of the grid. No stepsize restrictions whatsoever are caused by this reason. For instance, the implicit Euler method and the RADAU IIA methods are

algebraically stable. At the same time, these methods belong to the class IRK(DAE).

Turn back to DAEs. As we have seen in Example 5.1, in general, we cannot expect that algebraically stable Runge–Kutta methods, in particular the implicit Euler method, preserve the decay behavior of the exact DAE solution without strong stepsize restrictions, not even when we restrict the class of DAEs to linear ones. This depends on how the DAE is formulated. If the DAE has a properly involved derivative formulated in such a way that the image space of $D(t)$ is independent of t, then we already know (cf. Section 5.4) that the IRK(DAE) applied to the index-1 DAE reaches the IERODE unchanged. If the IRK(DAE) is algebraically stable, and the DAE is contractive, then we are sure to reflect the true solution properties

$$|D(t_n)x(t_n) - D(t_n)\bar{x}(t_n)| \le |D(t_{n-1})x(t_{n-1}) - D(t_{n-1})\bar{x}(t_{n-1})|, \tag{6.16}$$

$$\begin{aligned} |x(t_n) - \bar{x}(t_n)| &\le K_{(x,\bar{x})}(t_n)|D(t_n)x(t_n) - D(t_n)\bar{x}(t_n)| \\ &\le K_{(x,\bar{x})}(t_n)|D(t_{n-1})x(t_{n-1}) - D(t_{n-1})\bar{x}(t_{n-1})|, \end{aligned} \tag{6.17}$$

with no restrictions on the stepsizes. The next theorem confirms this fact.
At this point, we emphasize once again the specific structure of DAE solutions that leads to (6.16), (6.17) instead of (6.15) given in the case of explicit ODEs. For DAEs, one cannot expect an inequality (6.15), and so one should no longer try for the strong condition (6.14).

Theorem 6.9. *Assume the index-1 DAE (4.1), with $\mathcal{I}_f = [0,\infty)$, $\mathrm{dom}_{D\omega} = \mathbb{R}^n \times \mathcal{I}_f$, to be contractive, and $\mathrm{im}\, D(t)$ to be independent of t.*
Then, an algebraically stable IRK(DAE) method starting with any two values $x_0, \bar{x}_0 \in \mathcal{M}_0(t_0)$ yields sequences of numerical solutions x_n and $\bar{x}_n$ satisfying

$$|D(t_n)x_n - D(t_n)\bar{x}_n| \le |D(t_{n-1})x_{n-1} - D(t_{n-1})\bar{x}_{n-1}|, \quad \textit{for all } n \ge 1, \tag{6.18}$$

$$\begin{aligned} |x_n - \bar{x}_n| &\le \tilde{K}_{(x_n,\bar{x}_n)}(t_n)\ |D(t_n)x_n - D(t_n)\bar{x}_n| \\ &\le \tilde{K}_{(x_n,\bar{x}_n)}(t_n)\ |D(t_{n-1})x_{n-1} - D(t_{n-1})\bar{x}_{n-1}|,\ n \ge 1, \end{aligned}$$

where

$$\begin{aligned} \tilde{K}_{(x_n,\bar{x}_n)}(t_n) &:= \max_{s\in[0,1]} \max_{z\in \mathrm{im}\, D(t_n),\,|z|=1} |\Pi_{can}(\eta_{(u_n,\bar{u}_n,t_n)}(s))D(t_n)^- z|, \\ \eta_{(u_n,\bar{u}_n,t_n)}(s) &:= \big(D(t)\omega(su_n + (1-s)\bar{u}_n, t_n), \\ &\qquad D(t_n)^-(su_n + (1-s)\bar{u}_n) + Q_0(t_n)\omega(su_n + (1-s)\bar{u}_n, t_n),\ t_n\big). \end{aligned}$$

If K is a global bound of the matrix function $\Pi_{can}(\cdot)D(\cdot)^-$, then it holds further that

$$\begin{aligned} |x_n - \bar{x}_n| &\le K\ |D_n x_n - D_n \bar{x}_n| \\ &\le K\ |D_{n-1}x_{n-1} - D_{n-1}\bar{x}_{n-1}|, \quad \textit{for all } n \ge 1. \end{aligned} \tag{6.19}$$

Proof. By Theorem 6.7, the IERODE (6.11) is contractive on $\operatorname{im} D(t)$. Since this subspace does not depend on t, we know (cf. Theorem 5.3, Proposition 5.2) that the numerical solutions satisfy

$$x_n = D(t_n)^- u_n + Q_0(t_n)\omega(u_n, t_n), \quad \bar{x}_n = D(t_n)^- \bar{u}_n + Q_0(t_n)\omega(\bar{u}_n, t_n)$$

with $u_n = D(t_n)x_n$ and $\bar{u}_n := D(t_n)\bar{x}_n$ fulfilling the discretized IERODE

$$[u]'_n = D_n\omega(u_n, t_n), \quad [\bar{u}]'_n = D_n\omega(\bar{u}_n, t_n).$$

Therefore, u_n and $\bar{u}_n$ are at the same time the numerical solutions generated by the given algebraically stable IRK(DAE) method being directly applied to the contractive IERODE (6.11). Algebraically stable Runge–Kutta methods are B-stable and, thus,

$$|u_n - \bar{u}_n| \le |u_{n-1} - \bar{u}_{n-1}|,$$

which proves the first inequality of the theorem.
Analogously to (6.10), we derive the second inequality from

$$\begin{aligned} x_n - \bar{x}_n &= D(t_n)^-(u_n - \bar{u}_n) + Q_0(t_n)(\omega(u_n, t_n) - \omega(\bar{u}_n, t_n)) \\ &= \int_0^1 \Pi_{can}(\eta_{(u_n, \bar{u}_n, t_n)}(s))D(t_n)^- \,\mathrm{d}s\,(u_n - \bar{u}_n). \end{aligned}$$

If the matrix function $\Pi_{can}(\cdot)D(\cdot)^-$ is bounded by the constant K, then we get the remaining part of the assertion:

$$|x_n - \bar{x}_n| \le K\,|D(t_n)x_n - D(t_n)\bar{x}_n| \le K\,|D(t_{n-1})x_{n-1} - D(t_{n-1})\bar{x}_{n-1}|.$$

□

We emphasize once more that the solutions of a contractive DAE with bounded product $\Pi_{can}(\cdot)D(\cdot)^-$ are not expected to satisfy the inequality (6.15) as the solutions of a contractive explicit ODE do. Instead, the inequality (6.17) is natural for DAEs. Also, the numerical solutions generated by an algebraically stable IRK(DAE) applied to the DAE do not fulfill the inequality (6.14), as for explicit ODEs, but instead (6.19).

6.3 Dissipativity

A further popular qualitative property of dynamical systems described by explicit ODEs is dissipativity, where an absorbing set sucks up all solutions. First, we have to clarify what this could mean for DAEs (cf. [115]). In contrast to an explicit ODE the flow of which extends within its entire constant state space $\mathbb{R}^m$, the flow corresponding to a regular index-1 DAE (4.1) is restricted to the proper subset $\mathcal{M}_0(t) \subset \mathbb{R}^m$, which in turn may move in $\mathbb{R}^m$ with time t. We think, then, that it makes sense to

allow the absorbing set itself to vary with time, also.
As in the previous subsection, we assume the DAE (4.1) to be regular with index 1, $\mathcal{I}_f = [0,\infty)$ and $\operatorname{dom} D\omega = R^n \times \mathcal{I}_f$. We denote the solution of the index-1 DAE (4.1) passing at time t_+ through $x_+ \in \mathcal{M}_0(t_+)$ by $x(t;t_+,x_+)$.

Definition 6.10. Consider a regular index-1 DAE (4.1) the solutions of which exist on the entire infinite interval $\mathcal{I}_f = [0,\infty)$.

(1) A possibly time-dependent set $\mathcal{B}(t) \subset \mathcal{M}_0(t)$, $t \geq 0$, is called a *positively invariant set* of the DAE if $x_+ \in \mathcal{B}(t_+)$ implies $x(t;t_+,x_+) \in \mathcal{B}(t)$ for all $t > t_+$.
(2) A positively invariant set $\mathcal{B}(t)$, $t \geq 0$, is called an *absorbing set* of the DAE, if, for any $t_+ \in [0,\infty)$ and any bounded set $E \subset \mathcal{M}_0(t_+)$, there is a time $t_{(E,t_+)} \geq t_+$ such that $x_+ \in E$ implies $x(t,t_+,x_+) \in \mathcal{B}(t)$ for $t \geq t_{(E,t_+)}$.
(3) The DAE (4.1) is said to be *dissipative* if it has a bounded absorbing set.

In the next proposition we formulate an inequality in terms of the DAE (4.1) generalizing the well-known dissipativity inequality for explicit ODEs (Definition 6.2). This is actually a sufficient dissipativity condition for the IERODE on its invariant subspace $\operatorname{im} D(\cdot)$, and also for DAEs with bounded matrix functions $\Pi_{can}(\cdot)D(\cdot)^-$ and $Q_0(\cdot)\omega(0,\cdot)$. In the case of an autonomous DAE, $Q_0(\cdot)\omega(0,\cdot)$, Q_0 and ω are independent of t, and this expression is trivially bounded.

Proposition 6.11. *Assume (4.1) to be a regular index-1 DAE with $\mathcal{I}_f = [0,\infty)$ and $\operatorname{dom} D\omega = R^n \times \mathcal{I}_f$. Let an inner product $\langle\cdot,\cdot\rangle$ and constants $\alpha \geq 0$, $\beta > 0$ exist such that the inequality*

$$\langle y, D(t)x\rangle + \langle R'(t)D(t)x, D(t)x\rangle \leq \alpha - \beta |D(t)x|^2 \tag{6.20}$$

is satisfied for all $x \in \mathcal{M}_0(t)$, $y \in \operatorname{im} D(t)$, $f(y,x,t) = 0$, $t \geq 0$.

(1) *Then the IERODE (4.12) is dissipative on $\operatorname{im} D(t)$, and*

$$\mathcal{B}_{\mathrm{IERODE}}(t) := \left\{ v \in \operatorname{im} D(t) : |v|^2 \leq \frac{\alpha}{\beta} + \varepsilon \right\}$$

is an absorbing set for each $\varepsilon > 0$.
(2) *All DAE solutions can be continued to exist on the infinite interval.*
(3) *If, additionally, $\Pi_{can}(\cdot)D^-(\cdot)$ is uniformly bounded by a constant K and $Q_0(\cdot)\omega(0,\cdot)$ is bounded by a constant K_Q, then the DAE (4.1) is dissipative with the absorbing sets*

$$\mathcal{B}(t) = \left\{ x \in \mathcal{M}_0(t) : |x| \leq K\left(\frac{\alpha}{\beta} + \varepsilon\right)^{1/2} + K_Q \right\}, t \geq 0,\ \varepsilon > 0.$$

Proof. Recall that the function $\omega(u,t)$ in (4.12) is implicitly given by means of

$$f(D(t)\omega(u,t), D(t)^- u + Q_0(t)\omega(u,t), t) = 0 \quad u \in \mathbb{R}^n, t \in [0,\infty).$$

For each arbitrary $t \geq 0$, $u \in \operatorname{im} D(t)$, we introduce $y := D(t)\omega(u,t)$, $x := D(t)^- u + Q_0(t)\omega(u,t)$, which gives $f(y,x,t) = 0$, and hence, by (6.20),

$$\langle D(t)\omega(u,t), u\rangle + \langle R'(t)u, u\rangle \leq \alpha - \beta|u|^2. \tag{6.21}$$

The IERODE (4.12) satisfies the dissipativity inequality on the invariant subspace $\operatorname{im} D(\cdot)$. For any solution $u(\cdot)$ of the IERODE that belongs to $\operatorname{im} D(\cdot)$, (6.21) yields

$$\begin{aligned}\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}|u(t)|^2 &= \langle u'(t), u(t)\rangle = \langle R'(t)u(t) + D(t)\omega(u(t),t), u(t)\rangle \\ &\leq \alpha - \beta|u(t)|^2.\end{aligned}$$

For any $t_+ \geq 0$, $u_+ \in D(t_+)$, the solution $u(\cdot)$ of the corresponding IVP exists on a certain interval $\mathcal{I}_+ \ni t_+$, and satisfies there the inequality

$$|u(t)|^2 \leq \frac{\alpha}{\beta} + e^{-2\beta(t-t_+)}\left(|u_+|^2 - \frac{\alpha}{\beta}\right),\ t \geq t_+ \in \mathcal{I}_+.$$

Since the solution $u(\cdot)$ is bounded, it can be continued, and hence it exists on the entire interval $[t_+, \infty)$. Then,

$$x(t) := D(t)^- u(t) + Q_0(t)\omega(u(t),t), \quad t \in [t_+, \infty),$$

is the related DAE solution on the infinite interval, and assertion (2) is verified. Furthermore, the inequality

$$|u(t)| \leq \max\left\{|u_+|, \left(\frac{\alpha}{\beta}\right)^{1/2}\right\},\ t \geq t_+,$$

results, and this shows the set $\mathcal{B}_{\mathrm{IERODE}}(t)$ to be positively invariant for the IERODE. We check if it absorbs the solutions. Let a bounded set $E_u \subset \operatorname{im} D(t_+)$ be given. Denote $r := \sup\{|v| : v \in E_u\}$. For all $u_+ \in E_u$, the resulting IVP solutions satisfy

$$|u(t)| \leq \frac{\alpha}{\beta} + e^{-2\beta(t-t_+)}\left(r^2 - \frac{\alpha}{\beta}\right),\ t \geq t_+.$$

Choosing $\bar{t} = \bar{t}(E_u, t_+)$ so that $e^{-2\beta(\bar{t}-t_+)}\left(r^2 - \frac{\alpha}{\beta}\right) \leq \varepsilon$, we obtain $|u(t)| \leq \frac{\alpha}{\beta} + \varepsilon$ for all $u_+ \in E_u$ and $t \geq \bar{t}$. In other words, the set $\mathcal{B}_{\mathrm{IERODE}}(t)$ indeed absorbs the solutions. Consider now an arbitrary bounded set $E \subset \mathcal{M}_0(t_+)$, $t_+ \geq 0$, and put $E_u := D(t_+)E$. For each arbitrary $x_+ \in E$, we know that the IVP solution $x(t) = x(t; t_+, x_+)$ has the representation

$$x(t) = D(t)^- u(t) + Q_0(t)\omega(u(t),t),$$

whereby $u(t) = D(t)x(t)$ satisfies the IERODE (4.12) as well as the initial condition $u(t_+) = u_+ := D(t_+)x_+ \in E_u$. Due to (1), it holds that $|u(t)| \leq \sqrt{\frac{\alpha}{\beta} + \varepsilon}$ for all $t \geq \tilde{t}$, and uniformly for all $u_+ \in E_u$. In consequence,

$$
\begin{aligned}
|x(t)| &= |D(t)^- u(t) + Q_0(t)\omega(u(t),t) - Q_0(t)\omega(0,t) + Q_0(t)\omega(0,t)| \\
&\le \left| \int_0^1 \left(I - Q_0(t)\omega_u'(\sigma u(t),t)\right) \mathrm{d}s\, D(t)^- u(t) \right| + |Q_0(t)\omega(0,t)| \\
&\le K|u(t)| + \gamma \;\le\; K\left(\frac{\alpha}{\beta} + \varepsilon\right)^{1/2} + K_Q \quad \text{for } t \ge \bar{t}.
\end{aligned}
$$

□

For what concerns numerical integration methods, we refer once more to the fact that the integration method reaches the IERODE unchanged, if the index-1 DAE (4.1) is given in such a way that $\operatorname{im} D(t)$ does not at all vary with t. Then, the results about the numerical integration of dissipative explicit ODEs can be carried over to hold for the DAE (4.1), too. For instance, [204] shows that the backward Euler method reflects dissipativity without any stepsize restriction, whereas general algebraically stable Runge–Kutta methods reflect the dissipative flow under certain stepsize restrictions. We adopt the result for the implicit Euler method here.

Proposition 6.12. *Let the conditions of Proposition 6.11 be given, and, additionally, let* $\operatorname{im} D(\cdot)$ *be constant. Then the implicit Euler method reflects the dissipativity behavior properly without any stepsize restriction. The absorbing sets of the discretized DAE are the same as described in Proposition 6.11.*

Proof. Since $\operatorname{im} D(t)$ is constant, discretization and decoupling commute (see Subsection 5.4). If we apply the corresponding result for explicit ODEs (e.g., [204, Theorem 5.5.3]) and match the components as in Proposition 6.11, we obtain the desired result. □

6.4 Lyapunov stability

If we want to apply Lyapunov stability to DAEs then we have to consider the neighboring solutions of a reference solution. More precisely, we have to identify these neighboring solutions by consistent initial values or by appropriate initial conditions. For regular index-1 DAEs we know (see Theorem 4.11) that the set of consistent initial values at time t_0 is given by

$$\mathcal{M}_0(t_0) = \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : f(y,x,t_0) = 0\}.$$

If we are given a reference solution $x_*(\cdot)$, and in particular the value $x_*(t_0) \in \mathcal{M}_0(t_0)$, the values of all neighboring solutions have to belong to $\mathcal{M}_0(t_0)$, too. Since $x_0 \in \mathcal{M}_0(t_0)$ can be expressed as

$$x_0 = D(t_0)^- D(t_0)x_0 + Q_0(t_0)\omega(D(t_0)x_0, t_0),$$

an consistent value x_0 of an index-1 DAE is fully determined by its component $P_0(t_0)x_0$ or equivalently by $D(t_0)x_0$.
In contrast, Theorem 4.11 states the initial condition as

$$D(t_0)x(t_0) = D(t_0)x^0, \quad \text{with } x^0 \in \mathbb{R}^m.$$

Thereby, x^0 is not necessarily consistent, and it simply holds only that $D(t_0)x(t_0) = D(t_0)x^0$. No information regarding the component $Q_0(t_0)x^0$ slips in. This leads to the following equivalent possibilities to figure out the neighboring solutions by means of initial conditions.

(a) $x_0 \in \mathcal{M}_0(t_0)$, $|x_0 - x_*(t_0)| < \tau_a$, $x(t_0) = x_0$,
(b) $x^0 \in \mathbb{R}^m$, $|D(t_0)(x^0 - x_*(t_0))| < \tau_b$, $D(t_0)(x(t_0) - x^0) = 0$,
(c) $x^0 \in \mathbb{R}^m$, $|D(t_0)(x^0 - x_*(t_0))| < \tau_c$, $x(t_0) - x^0 \in \ker D(t_0)$,
(d) $x^0 \in \mathbb{R}^m$, $|x^0 - x_*(t_0)| < \tau_d$, $x(t_0) - x^0 \in \ker D(t_0)$.

In essence, the definition below coincides with the one already given in [96] for standard form index-1 DAEs. While [96] applies version (d), we now make use of version (a).

Definition 6.13. Let the DAE (4.1) be regular with index 1, and $\mathcal{I}_f = [0, \infty)$.
The solution $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ is said to be

(1) *stable in the sense of Lyapunov* if, for each $\varepsilon > 0$, $t_0 \in \mathcal{I}$, there is a $\delta(\varepsilon, t_0) > 0$ such that

$$|x_*(t_0) - x_0| < \delta(\varepsilon, t_0), \quad x_0 \in \mathcal{M}_0(t_0)$$

imply the existence of a solution $x(t; t_0, x_0)$ on $[t_0, \infty)$ as well as the estimation

$$|x_*(t) - x(t; t_0, x_0)| < \varepsilon \text{ for } t \geq t_0,$$

(2) *asymptotically stable* if for every $\varepsilon > 0$, $t_0 \in \mathcal{I}$, there is a $\delta(\varepsilon, t_0) > 0$ such that

$$|x_*(t_0) - x_0| < \delta(\varepsilon, t_0), \quad x_0 \in \mathcal{M}_0(t_0)$$

imply the existence of a solution $x(t; t_0, x_0)$ on $[t_0, \infty)$ as well as the limit

$$|x(t) - x(t; t_0, x_0)| \xrightarrow[t \to \infty]{} 0.$$

By Theorem 6.7, each solution of a strongly contractive index-1 DAE, with bounded product $\Pi_{can} D^-$, is asymptotically stable. However, a general index-1 DAE (4.1) may have stable and unstable solutions at the same time.

Example 6.14 (Stable periodic solution). Consider the index-1 DAE given by $m = 3, n = 2$,

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

There is the asymptotically stable solution (see [141] for a proof via Floquet theory)

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

as well as the unstable stationary solution

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

This example is rather too simple with its time-invariant constraint set $\mathcal{M}_0 = \{x \in \mathbb{R}^3 : x_1^2 + x_2^2 = 1 - x_3\}$. Figure 6.1 shows the flow on the constraint set. □

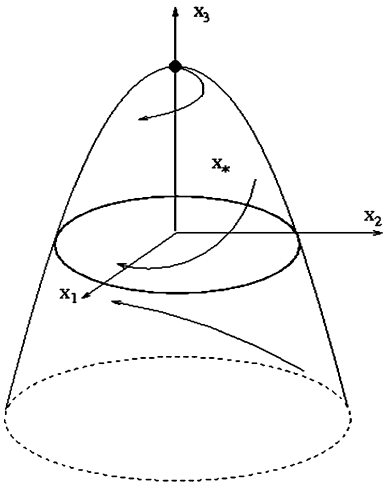


Fig. 6.1 Flow on the constraint set

The situation in the next example is less transparent although the DAE is also semi-explicit and has the same dimensions $m = 3, n = 2$.

Example 6.15 (Voltage doubling network). The DAE

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

describes the voltage doubling network from Figure 6.2, where

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

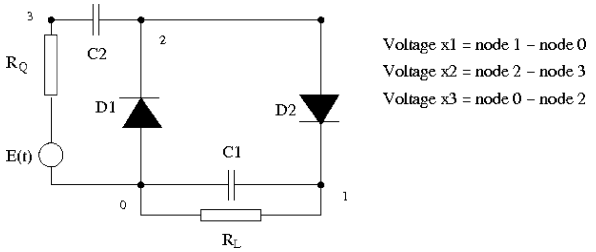


Fig. 6.2 Voltage doubling network

and

$$C_1 = C_2 = 2.75\text{nF}, \quad G_L = \frac{1}{R_L}, \quad R_Q = 0.1\text{M}\Omega, \quad R_L \in [1,\infty).$$

For $R_L = 10$, there is an asymptotically stable T-periodic solution, which is displayed in Figure 6.3. It can be provided numerically, only. In [141], stability is checked via the eigenvalues of the monodromy matrix $X_*(T,0)$, where $X_*(\cdot,0)$ denotes the fundamental solution of the linearized DAE normalized at $t = 0$. In our context, this Floquet procedure consists of an equivalent periodic reduction to a strongly contractive constant coefficient DAE. □

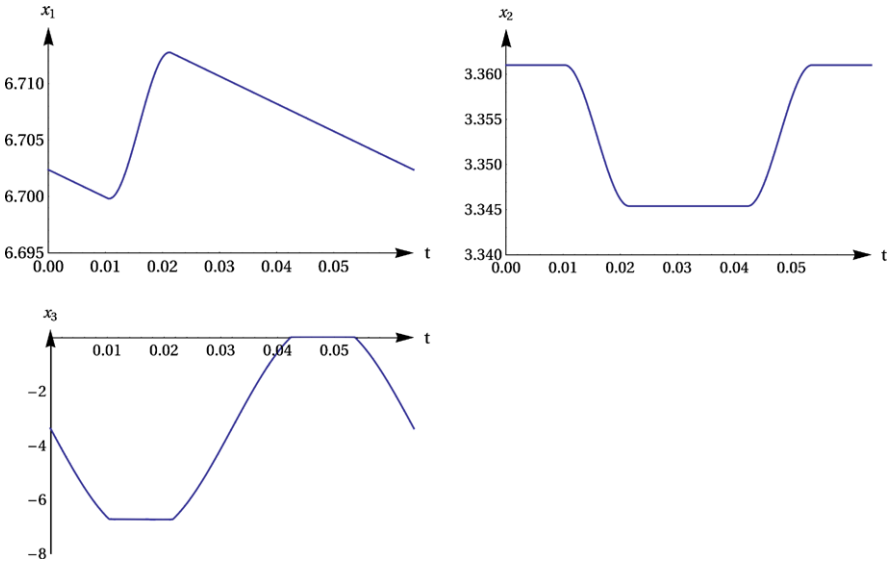


Fig. 6.3 T-periodic solution

Theorem 6.16. *Let the DAE (4.1) be regular with index 1, and $\mathcal{I}_f = [0,\infty)$. Additionally to the basic assumptions (cf. Assumption 4.1, Definition 4.3) we suppose f to feature continuous second partial derivatives f_{yy}, f_{yx}, f_{xx}.*

Let $x_* \in \mathcal{C}^1_D([0,\infty),\mathbb{R}^m)$ be a solution of the DAE, and let the DAE linearized along x_*

$$A_*(t)(D(t)x(t))' + B_*(t)x(t) = 0, \tag{6.22}$$

with

$$A_*(t) := f_y((D(t)x_*(t))', x_*(t), t),\ B_*(t) := f_x((D(t)x_*(t))', x_*(t), t), \quad t \in [0,\infty),$$

be strongly contractive.
Let the given first and second partial derivatives as well as G^{-1} be bounded in a neighborhood of the graph of the reference solution. Then $u_* := Dx_*$ is an asymptotically stable solution of the IERODE with respect to $\operatorname{im} D$.
If, additionally, the product $\Pi_{can}D^-$ remains bounded, then x_* is asymptotically stable.

Proof. The linear DAE (6.22) has, as its origin (4.1), a properly involved derivative, and it is regular with index 1, thus, $G_*(t) = A_*(t)D(t) + B_*(t)Q_0(t)$ remains nonsingular. Moreover, due to the strong contractivity, there are a value $\beta > 0$ and an inner product such that, for all $[0,\infty)$,

$$\langle y - \bar{y}, D(t)(x-\bar{x})\rangle + \langle R'(t)D(t)(x-\bar{x}), D(t)(x-\bar{x})\rangle \le -\beta |D(t)(x-\bar{x})|^2,$$

$\forall x, \bar{x} \in \mathbb{R}^m$, with $A_*(t)y + B_*(t)x = 0,\ A_*(t)\bar{y} + B_*(t)\bar{x} = 0,\ y = R(t)y,\ \bar{y} = R(t)\bar{y}$. This implies

$$y - \bar{y} = -D(t)G_*(t)^{-1}B_*(t)D(t)^-D(t)(x-\bar{x}),$$

and therefore the inequality

$$\langle (R'(t) - D(t)G_*(t)^{-1}B_*(t)D(t)^-)D(t)(x-\bar{x}), D(t)(x-\bar{x})\rangle \le -\beta |D(t)(x-\bar{x})|^2,$$

holds for all $x, \bar{x} \in \mathbb{R}^m$, and, equivalently

$$\langle (R'(t) - D(t)G_*(t)^{-1}B_*(t)D(t)^-)v, v\rangle \le -\beta |v|^2, \quad \text{for all } v \in \operatorname{im} D(t). \tag{6.23}$$

Turn to the IERODE, and to the explicit ODE

$$v'(t) = (R'(t) + D(t)\omega(u_*(t),t))v(t) + h(v(t),t) \tag{6.24}$$

resulting from the IERODE by the translation $v(t) = u(t) - u_*(t)$. The function h is defined to be

$$h(v,t) := D(t)\big(\omega(v + u_*(t),t) - \omega(u_*(t),t) - \omega_u(u_*(t),t)\big).$$

Lemma 4.4 provides us with

$$D(t)\omega(u,t) = -(DG^{-1}f_xD^-)(D(t)\omega(u,t), D(t)^-u + Q_0(t)\omega(u,t), t),$$
$$D(t)\omega(u_*(t),t) = -D(t)G_*(t)^{-1}B_*(t)D(t)^-.$$

The function ω has a continuous second partial derivative ω_{uu} due to the smoothness of f. Then, the function h is continuous, and has continuous partial derivatives h_v and h_{vv}, in particular, it holds that

$$h(0,t) = 0,\ h_v(0,t) = 0,\ h_{vv}(v,t) = D(t)\omega_{uu}(v + u_*(t),t).$$

Let K_h be a bound of the second partial derivative h_{vv} such that

$$|h(v,t)| \le K_h|v|^2, \quad \text{for all sufficiently small } |v|.$$

Such a K_h is available, since the involved partial derivatives of f and G^{-1} are locally bounded around the reference solution. Choose a value $\varepsilon > 0$ such that $\beta - \varepsilon K_h =: \tilde{\beta} > 0$, and fix a $t_0 \ge 0$. The IVP for (6.24), and the initial condition $v(t_0) = v_0 \in \operatorname{im} D(t_0)$, $|v_0| \le \varepsilon$, has a unique solution $v(\cdot)$, say on the interval $[t_0, t_0 + T)$. With regard to (6.23) we derive

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}|v(t)|^2 &= \langle v'(t), v(t)\rangle \\
&= 2\langle (R'(t) - D(t)G_*(t)^{-1}B_*(t)D(t)^-)v(t), v(t)\rangle + 2\langle h(v(t),t), v(t)\rangle \\
&\le -2\beta|v(t)|^2 + 2|h(v(t),t)||v(t)| \\
&\le (-2\beta + \varepsilon K_h)|v(t)|^2 = -2\tilde{\beta}|v(t)|^2, \ \text{for } t \in [t_0, t_0 + T).
\end{aligned}$$

By Gronwall's lemma, it follows that

$$|v(t)| \le e^{\tilde{\beta}(t-t_0)}|v(t_0)| = e^{\tilde{\beta}(t-t_0)}|v_0| \le \varepsilon,\ t \in [t_0, t_0 + T).$$

Now it is evident that $v(\cdot)$ can be continued to exist on the entire interval $[t_0, \infty)$, and

$$|v(t)| \le e^{\tilde{\beta}(t-t_0)}|v(t_0)| = e^{\tilde{\beta}(t-t_0)}|v_0| \le \varepsilon,\ t \in [t_0, \infty).$$

The existence of $v(\cdot)$ corresponds to the existence of the function $u(\cdot) = u_*(\cdot) + v(\cdot)$ which satisfies the IERODE, and meets the condition $u(t_0) = u_*(t_0) + v_0$. In summary, we have the following: To each $t_0 \ge 0$ and $\varepsilon > 0$, there is a $\delta(\varepsilon, t_0) := \varepsilon > 0$ such that, for each $u_0 \in \operatorname{im} D(t_0)$, $|u_0 - u_*(t_0)| \le \delta(\varepsilon, t_0)$, the IERODE has a solution on the infinite interval, that meets the initial condition $u(t_0) = u_0$. The difference $|u(t; t_0, u_0) - u_*(t)|$ tends to zero, if t tends to infinity. This means, that in fact, $u_*(\cdot)$ is an asymptotically stable solution of the IERODE with respect to $\operatorname{im} D(\cdot)$.

For $t_0 \ge 0$ and $\varepsilon > 0$, we consider the initial condition

$$D(t_0)(x(t_0) - x^0) = 0,\ x^0 \in \mathbb{R}^m,\ |D(t_0)(x^0 - x_*(t_0)| \le \delta(\varepsilon, t_0),$$

for the nonlinear DAE (4.1). By means of the IERODE solution $u(\,\cdot\,; t_0, D(t_0)x^0)$ we build

$$x(t; t_0, x^0) := D(t)^- u(t; t_0, D(t_0)x^0) + Q_0(t)\omega(u(t; t_0, D(t_0)x^0), t),\ t \in [t_0, \infty),$$

which is a solution of the DAE, with $D(t_0)x(t_0;t_0,x^0) = D(t_0)D(t_0)^- D(t_0)x^0 = D(t_0)x^0$. Regarding the expression (6.10) for differences of DAE solutions, and the boundedness of the product $\Pi_{can}D^-$ by the constant K, we obtain

$$|x(t;t_0,x^0) - x_*(t)| \le K|u(t;t_0,D(t_0)x^0) - u_*(t)| \le Ke^{-\tilde{\beta}(t-t_0)}|D(t_0)x^0 - D(t_0)x_*(t_0)|.$$

This proves the assertion. □

Example 6.17. We turn once again to the DAE in Example 4.8. Assume $\beta > 0$. We consider the reference solution

$$x_*(t) = \begin{bmatrix} 0 \\ (1+\gamma(t))^{\frac{1}{2}} \end{bmatrix},$$

which has a stationary core. The DAE (6.22) linearized along x_* reads

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

This linear DAE is strongly contractive with the constant β and the standard product in $\mathbb{R}$. If additionally, the function γ fulfills the condition

$$1 + \gamma(t) \ge \alpha > 0, \quad \text{for all} \quad t \in [0,\infty),$$

then, by the above theorem, x_* is asymptotically stable.
If γ vanishes identically, the nonlinear DAE is autonomous, the reference solution becomes a stationary one, and the linearized DAE has constant coefficients. The related matrix pencil is

$$\left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \beta & 0 \\ 0 & 2 \end{bmatrix} \right\}.$$

The pencil has the only finite eigenvalue $-\beta < 0$, and hence asymptotical stability is once more confirmed by the next corollary. □

Corollary 6.18. *Let the autonomous DAE*

$$f((Dx(t))', x(t)) = 0 \tag{6.25}$$

be regular with index 1, and let f belong to class $\mathcal{C}^2$.
Let $x_*(t) = c$ be a stationary solution. If all finite eigenvalues of the matrix pair $\{f_y(0,c)D, f_x(0,c)\}$ are strictly negative, then c is an asymptotically stable stationary solution.

Proof. Since the finite spectrum of the matrix pair $\{f_y(0,c)D, f_x(0,c)\} =: \{A_*, B_*\}$ lies in $\mathbb{C}^-$, the linear constant coefficient DAE $A_*(Dx(t))' + B_*x(t) = 0$ is contractive, and the assertion follows from Theorem 6.16. □

Having an appropriate stability notion with regard to the exact solutions of the DAE (4.1), the question arises, to what extent do numerical methods generate stable nu-

merical solutions if the exact DAE solution is stable. What about A-stable integration methods?

The DAE given by (5.6)–(5.7) shows that a general positive answer for DAEs cannot be expected. Recall that also in the case of explicit ODEs, A-stable methods show the required property without stepsize restrictions just for linear time-invariant systems and autonomous ODEs with weak nonlinearities. Already in the case of time-varying explicit ODEs extra stepsize restrictions for stability reasons may occur. Expecting better results for DAEs would be naive as DAEs incorporate explicit ODEs. In general, the situation in the case of DAEs is worse. The time dependencies play their role.

An A-stable integration method preserves its benefit, if it is applied to a DAE which has a linear constant coefficient IERODE and a time-invariant $\operatorname{im} D(t)$ such as the DAE in Example 4.8.

6.5 Notes and references

(1) To a large extent, the presentation concerning contractivity and dissipativity follows the lines of [96, 115]. In particular, the contractivity and dissipativity notions as well as Theorem 6.7 take up corresponding results given in [115] for quasi-linear DAEs. The presentation of stability in the sense of Lyapunov generalizes and modifies those in [96] given there for standard form DAEs.

(2) It seems that also Floquet theory can be appropriately adapted following the lines of [141], see Example 6.15.

(3) Stable and asymptotically stable linear DAEs are introduced in Chapter 2, Definition 2.53, as generalizations of the respective notions for explicit ODEs, which exploits the DAE structure in a reasonable manner. Looking back once again at Example 2.57, we recognize the role of the IERODE and that of the canonical projector function Π_{can} wrapping up the IERODE flow to the DAE flow. Choosing there $\alpha > 0$, $\beta = 0$, the IERODE becomes stable. However, taking a look at the fundamental solution matrix we see that, even if the IERODE is stable, the DAE may have unbounded solutions. This happens in fact, if certain entries of the canonical projector function Π_{can} grow unboundedly. In contrast, if all entries of Π_{can} are bounded, then the stability of the IERODE is passed over to the DAE. In our view, the dominance of the wrapping over the IERODE flow is somewhat beside the point. In the present chapter, regarding nonlinear DAEs, we concentrate on problems featuring uniformly bounded canonical projector functions. For an extended discussion of the boundedness conditions see [178].

Roughly speaking, if the canonical projector function remains bounded, then an index-1 DAE is contractive or dissipative, if its IERODE is so. Moreover, in essence, numerical integration methods preserve their A- and B-stability for the DAE supposing the matrix function $D(\cdot)$ has a constant range. In this case, the integration methods reaches the IERODE unchanged (cf. Example 5.1). Otherwise, serious stepsize

restrictions are necessitated.

(4) The following is worth mentioning: If in the standard form DAE

$$E(t)x'(t) + F(t)x(t) = q(t),$$

the matrix function E has a time-invariant nullspace, $\ker E(t) = N_E$, taking a projector P_E, with $\ker P_E = N_E$, the DAE can be written as

$$E(t)(P_E x(t))' + F(t)x(t) = q(t).$$

If $\ker E(t)$ varies with time, but $E(t)$ has a constant range, $\operatorname{im} E(t) = R_E$, then we can write

$$V_E(E(t)x(t))' + F(t)x(t) = q(t),$$

where V_E is a projector such that $\operatorname{im} V_E = R_E$. In both cases, the reformulation is a DAE with properly involved derivative, and the subspace corresponding to $\operatorname{im} D(t)$ is independent of t. This confirms and explains the former contractivity results in [96, for nonlinear DAEs, with constant nullspace] and [83, linear DAEs, with constant range]. Of course, a Runge–Kutta method applied to the different formulations provides different results, for instance, the implicit Euler method reads in the first case

$$E(t_n)\frac{1}{h}(x_n - x_{n-1}) + F(t_n)x_n = q(t_n),$$

and in the second case

$$\frac{1}{h}(E(t_n)x_n - E(t_{n-1})x_{n-1}) + F(t_n)x_n = q(t_n).$$

However, we would not like to speak here of *different methods*, but we emphasize that the same given method is *applied to different DAE forms*. And we emphasize the benefit of trying to find a *numerically qualified DAE formulation* which features a constant subspace $\operatorname{im} D(t)$.

Part III
Computational aspects

Part III is mainly devoted to computational aspects of the practical preparation of all ingredients of admissible matrix function sequences and the associated projectors. In particular one has to carry out matrix factorizations, rank calculations, and determinations of generalized inverses. Chapter 7 provides several versions to accomplish the basic step of the matrix function sequence from one level to the next. Moreover, a special more involved algorithm is developed for regular DAEs. The characteristic values arise as byproducts of matrix factorizations. From the numerical viewpoint, the widely orthogonal projector functions are favorably.

The second chapter sheds light on aspects of the direct numerical treatment of higher index DAEs, index monitoring, consistent initialization and numerical integration. Not surprisingly, the integration methods approved for regular index-1 DAEs not longer perform well or fail, if they are applied in the same way to general higher index DAEs, for instance to time-varying linear index-3 DAEs. This is due to the ill-posed character of the DAE solutions with respect to perturbations. Fortunately, exploiting special structural peculiarities, one can often create special methods for restricted classes of DAE.

Chapter 7
Computational linear algebra aspects

Originally, the tractability index concept was designed rather for the theoretical investigation of DAEs. However, the resulting clear index criteria by rank conditions let us trust that it also has practical meaning. Moreover, the projectors prove their value when characterizing the different solution components, when looking for consistent initial values and formulating appropriate initial conditions as well. And these are good arguments to implement the associated matrix function sequences.
The algorithmic realization of a matrix function sequence (2.5)–(2.8), (see also (3.19)–(3.21))

$$\begin{aligned} G_{i+1} &= G_i + B_i Q_i, \\ B_{i+1} &= B_i P_i - G_{i+1} D^- (D\Pi_{i+1} D^-)' D\Pi_i \end{aligned}$$

requires the computation of the involved generalized inverse D^- and the admissible projectors Q_i (cf. Definitions 1.10, 2.6, 2.25, 3.21).
For a DAE that has the leading term $A(t)(D(t)x(t))'$, it is also important to check whether this leading term is actually properly stated by testing the transversality condition

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n.$$

The last question is considered in Section 7.2, whereas the basics of the computation of nullspace and image projectors associated with matrices are collected in Section 7.1. At this point we also bring to mind the detailed Appendix A on linear algebra issues. Methods of computing a suitable generalized inverse D^- are described in Section 7.1. In Section 7.3 we deal with the basic step of the construction of admissible matrix functions, that is, with the step from level i to level $i+1$ by the computation of an appropriate projector. After that, in Section 7.4, sequences of matrices with admissible projectors are delivered, first level by level on the background of Section 7.3, and then by a strongly involved version only for the regular case.
We stress that all the computation are more or less related to matrix decompositions and rank calculations, and, naturally, one has to expect to inherit all the related numerical problems.

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_7,

7.1 Image and nullspace projectors

For a given $G \in \mathbb{R}^{k\times m}$, with $\operatorname{rank} G = r$, any matrix $Q \in \mathbb{R}^{m\times m}$ that satisfies

$$GQ = 0, \quad Q^2 = Q, \quad \operatorname{rank} Q = m - r$$

is a projector onto $\ker G$. Any matrix $W \in \mathbb{R}^{k\times k}$ that satisfies

$$WG = 0, \quad W^2 = W, \quad \operatorname{rank} W = k - r$$

is a projector along $\operatorname{im} G$.

Clearly, having a basis of the subspace in question, a required projector can immediately be described by these basis elements (cf. Lemma A.7). In particular, if $n_1, \ldots, n_{m-r} \in R^m$ form a basis of $\ker G$ and $\Gamma := \begin{bmatrix} n_1 & \cdots & n_{m-r} \end{bmatrix}$, then $Q = \Gamma(\Gamma^*\Gamma)^{-1}\Gamma^*$ represents the orthogonal projector onto this nullspace. If the $n_1, \ldots, n_{m-r}$ form an orthonormal basis, the expression simplifies to

$$Q = \Gamma\Gamma^* = \sum_{i=1}^{m-r} n_i n_i^*.$$

In other words, knowledge of an orthonormal basis can immediately be used to form an orthogonal projector as the sum of the dyadic product of the basis vectors. For problems of limited dimension a formula manipulation system like *Mathematica*® or *Maple*® can be used to compute a basis. The command in *Mathematica* is `NullSpace`[G] and in *Maple* `nullspace`(G).
However, to provide a basis of the nullspace of a given matrix one usually has to carry out a factorization, for instance a singular value decomposition (SVD).

If a generalized reflexive inverse G^- (cf. Appendix A.2) is known, we gain at the same time the nullspace projector $Q = I - G^-G$ and the projector along the image $W = I - GG^-$. To compute a generalized inverse of the given matrix G, again a factorization of that matrix serves as an appropriate tool.
Each decomposition

$$G = \mathcal{U} \begin{bmatrix} S & \\ & 0 \end{bmatrix} \mathcal{V}^{-1}, \tag{7.1}$$

with nonsingular $S \in \mathbb{R}^{r\times r}$, $\mathcal{U} =: \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{k\times k}$ and $\mathcal{V} =: \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{m\times m}$, and $U_1 \in \mathbb{R}^{k\times r}$, $V_2 \in \mathbb{R}^{m\times(m-r)}$, immediately delivers the bases $\ker G = \operatorname{span} V_2$ and $\operatorname{im} G = \operatorname{span} U_1$ as well as (7.1) the family of reflexive generalized inverses of G by

$$G^- = \mathcal{V} \begin{bmatrix} S^{-1} & M_2 \\ M_1 & M_1 S M_2 \end{bmatrix} \mathcal{U}^{-1}, \tag{7.2}$$

with the free parameter matrices M_1 and M_2 (see Appendix A.13). The resulting projectors are

$$Q = \mathcal{V}\begin{bmatrix} 0 \\ -M_1 S\ I \end{bmatrix}\mathcal{V}^{-1} \text{ and } \quad W = \mathcal{U}\begin{bmatrix} 0 & -SM_2 \\ & I \end{bmatrix}\mathcal{U}^{-1}.$$

If we are looking for orthogonal projectors, we have to ensure symmetry, that is $\mathcal{U}^{-1} = \mathcal{U}^*$, $\mathcal{V}^{-1} = \mathcal{V}^*$, $M_1 \equiv 0$ and $M_2 \equiv 0$.

There are different ways to generate matrix decompositions (7.1). Applying the SVD one delivers orthogonal matrices $\mathcal{U}$ and $\mathcal{V}$, and the orthogonal projector Q is given by

$$Q = \begin{bmatrix} \mathcal{V}_1 & \mathcal{V}_2 \end{bmatrix}\begin{bmatrix} 0 & \\ & I \end{bmatrix}\begin{bmatrix} \mathcal{V}_1^* \\ \mathcal{V}_2^* \end{bmatrix} = \mathcal{V}_2\mathcal{V}_2^*. \tag{7.3}$$

Also the Householder method is suitable for computing a decomposition (7.1). The Householder decomposition needs less computational work than the SVD. For a singular matrix G, a Householder decomposition with column pivoting is needed. We obtain

$$GI_{per} = U\begin{bmatrix} R_1 & R_2 \\ 0 & 0 \end{bmatrix}$$

with a column permutation matrix I_{per}, an orthogonal matrix U and a nonsingular upper triangular matrix R_1. The required decomposition (7.1) then has the structure

$$G = U\begin{bmatrix} R_1 & \\ & 0 \end{bmatrix}\underbrace{\begin{bmatrix} I & R_1^{-1}R_2 \\ & I \end{bmatrix}I_{per}^*}_{=:\mathcal{V}^{-1}}, \tag{7.4}$$

and hence the nullspace projector

$$Q = I_{per}\begin{bmatrix} I & -R_1^{-1}R_2 \\ & I \end{bmatrix}\begin{bmatrix} 0 & \\ -M_1R_1 & I \end{bmatrix}\begin{bmatrix} I & R_1^{-1}R_2 \\ & I \end{bmatrix}I_{per}^*$$

and the projector

$$W = U\begin{bmatrix} 0 & -R_1M_2 \\ & I \end{bmatrix}U^*$$

along the image of G results. The free parameter matrices M_1 and M_2 can be used to provide special properties of the projectors as, for instance, we do in Section 7.4. Since the Householder method provides an orthogonal matrix U, choosing $M_2 = 0$ we arrive at an orthoprojector W. If we apply the Householder method to G^* instead of G, we also deliver an orthogonal nullspace projector for G.

In principle also an LU decomposition of G using the Gaussian method with scaling and pivoting yields a decomposition (7.1). With a row permutation matrix I_{per} we obtain

$$I_{per}G = LU = \begin{bmatrix} L_1 & \\ L_2 & I \end{bmatrix}\begin{bmatrix} R_1 & R_2 \\ & 0 \end{bmatrix}$$

and the decomposition

$$G = \underbrace{I_{per}^{*} \begin{bmatrix} L_1 \\ L_2 \; I \end{bmatrix}}_{=:\mathcal{U}} \begin{bmatrix} R_1 & \\ & 0 \end{bmatrix} \underbrace{\begin{bmatrix} I \; R_1^{-1} R_2 \\ & I \end{bmatrix}}_{=:\mathcal{V}^{-1}}.$$

It is well-known that rank determination by the Gaussian method is not as robust as it is by the Householder method or SVD (cf. [93]), which is confirmed by our practical tests. We do not recommend this method here.

7.2 Matters of a properly stated leading term

Having a pair of matrices A and D one might be interested in making sure whether they are *well matched* in the sense of Definition 1.36. For instance, if a DAE with a quasi-proper leading term is given, one can check pointwise if the DAE even has a proper leading term (see Definitions 2.72, 3.2). In this way critical points can be indicated and eventual programming errors in handwritten subroutines as well. Moreover, when generating the basic matrix function sequences starting pointwise from the given coefficients A, D, and B, the reflexive generalized inverses D^{-} and the border projector R play their role.

Let the two matrices $A \in \mathbb{R}^{k\times n}$ and $D \in \mathbb{R}^{n\times m}$ be given and $G := AD$. Then the inclusions $\operatorname{im} G \subseteq \operatorname{im} A$ and $\ker D \subseteq \ker G$ are valid. Owing to Lemma A.4, A and D are well matched, exactly if

$$\operatorname{rank} A = \operatorname{rank} G = \operatorname{rank} D, \tag{7.5}$$

$$\operatorname{im} G = \operatorname{im} A, \tag{7.6}$$

$$\ker D = \ker G. \tag{7.7}$$

The failure of one of these three conditions indicates that A and D miss the mark. Since, in turn, (7.6), (7.7) imply the rank condition (7.5), these two conditions already ensure the well-matchedness.

Let G^{-} denote a reflexive generalized inverse of G, e.g., provided by a decomposition (7.2). Then the conditions (7.6), (7.7) can be written as

$$(I - GG^{-})A = 0, \tag{7.8}$$

$$D(I - G^{-}G) = 0, \tag{7.9}$$

and these conditions are also useful for testing the well-matchedness.

Next we suppose A and D to be well matched, and hence (7.8) and (7.9) to be valid. Then

$$D^{-} := G^{-}A, \quad A^{-} := DG^{-} \tag{7.10}$$

are reflexive generalized inverses of D and A, and

$$R := DD^{-} = DG^{-}A = A^{-}A$$

is nothing else than the projector matrix onto $\operatorname{im} D$ along $\ker A$. Namely, it holds that

$$\begin{aligned} DD^-D &= DG^-AD = DG^-G = D, \\ D^-DD^- &= G^-ADG^-A = G^-A = D^-, \\ AA^-A &= ADG^-A = GG^-A = A, \\ A^-AA^- &= DG^-ADG^- = DG^- = A^-. \end{aligned}$$

It turns out that, decomposing G delivers at the same time a reflexive generalized inverse G^- such that one can first check the conditions (7.8) and (7.9), and then, supposing they hold true, form the generalized inverses D^-, A^- and the border projector R.
We stress at this point that an orthogonal projector is often preferable. It can be reached by a SVD applied to G or a Householder factorization applied to G^* (Section 7.1).

An alternative way to test well-matchedness of A and D and then to provide D^- and R uses factorizations of both matrices A and D. This makes sense, if the factorizations of A and D are given or easily available.
Suppose the decompositions (cf. (7.1)) of A and D are

$$A = U_A \begin{bmatrix} S_A \\ 0 \end{bmatrix} V_A^{-1} \text{ and } D = U_D \begin{bmatrix} S_D \\ 0 \end{bmatrix} V_D^{-1}. \tag{7.11}$$

We can now check the rank conditions $\operatorname{rank} S_A = \operatorname{rank} S_D$ which are necessary for well-matchedness (see (7.5)). Also $AD = G$ has to have the same rank. The decompositions yield

$$AD = U_A \begin{bmatrix} S_A \\ 0 \end{bmatrix} V_A^{-1} U_D \begin{bmatrix} S_D \\ 0 \end{bmatrix} V_D^{-1} \tag{7.12}$$

and, denoting $V_A^{-1}U_D =: H = \begin{bmatrix} H_1 & H_2 \\ H_3 & H_4 \end{bmatrix}$, the necessary rank condition is satisfied iff H_1 remains nonsingular.

The generalized inverses of D and A are not independent of each other, but they have to satisfy the relation $DD^- = A^-A$. Using the given decompositions (7.11) the reflexive generalized inverses are immediately found (see (A.13)) as

$$A^- = V_A \begin{bmatrix} S_A^{-1} & M_{2,A} \\ M_{1,A} & M_{1,A}S_AM_{2,A} \end{bmatrix} U_A^{-1} \text{ and } D^- = V_D \begin{bmatrix} S_D^{-1} & M_{2,D} \\ M_{1,D} & M_{1,D}S_DM_{2,D} \end{bmatrix} U_D^{-1},$$

which leads to

$$DD^- = U_D \begin{bmatrix} I & S_DM_{2,D} \\ 0 & 0 \end{bmatrix} U_D^{-1}, \quad A^-A = V_A \begin{bmatrix} I & 0 \\ M_{1,A}S_A & 0 \end{bmatrix} V_A^{-1}.$$

Using again the notation $U_D = V_AH$, the relation $DD^- = A^-A$ becomes equivalent to

$$H \begin{bmatrix} I & S_D M_{2,D} \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} I & 0 \\ M_{1,A} S_A & 0 \end{bmatrix} H.$$

This fixes two of the free parameter matrices, namely

$$M_{2,D} = S_D^{-1} H_1^{-1} H_2 \tag{7.13}$$

and

$$M_{1,A} = H_3 H_1^{-1} S_A^{-1}.$$

The other two parameter matrices $M_{1,D}$ and $M_{2,A}$ can be used to ensure further properties.

Finally in this section, we briefly turn to standard form DAEs given with a leading term of the form $Gx'(t)$. A factorization (7.1) is then adjuvant in determining a properly stated leading term version (cf. Section 1.5). We can define A and D as

(a) $A = \mathcal{U} \begin{bmatrix} S & \\ & 0 \end{bmatrix}$, $D = \mathcal{V}^{-1}$,

(b) $A = \mathcal{U}$, $D = \begin{bmatrix} S & \\ & 0 \end{bmatrix} \mathcal{V}^{-1}$,

and with $\mathcal{U} =: [\mathcal{U}_1, \mathcal{U}_2]$ and $\mathcal{V}^{-1} =: \begin{bmatrix} (\mathcal{V}^{-1})_1 \\ (\mathcal{V}^{-1})_2 \end{bmatrix}$

(c) $A = \mathcal{U}_1 S$, $D = (\mathcal{V}^{-1})_1 \in \mathbb{R}^{r \times m}$,

(d) $A = \mathcal{U}_1$, $D = S(\mathcal{V}^{-1})_1$.

The cases (c) and (d) provide the splitting with full rank matrices A and D, which is advantageous, e.g., because the border projector is simply $R = I$.
Analogously one can proceed in the case of time-varying term coefficients $G(t)$, but then one needs a continuous matrix decomposition and a continuously differentiable $D(\cdot)$ as well as its derivative.
Notice that often standard form DAEs are given with separated derivative-free equations such that a continuous projector function $I - W(t)$ onto $\operatorname{im} G(t)$ is available at the beginning. Then one can make use of this situation and put $A(t) := I - W(t)$, $D(t) := G(t)$ (cf. Chapter 2, Note (7)).

7.3 The basic step of the sequence

Now we consider the basic part of the determination of an admissible matrix function sequence, that is the step from G_i to G_{i+1} (cf. for the constant coefficient case (1.10), for variable coefficients (2.6)–(2.8), and in the nonlinear case (3.21)). Let a projector $\Pi_i := P_0 \cdots P_i$ be already computed. We are looking for the next admissible projector Q_{i+1}. An admissible projector must satisfy the required properties (cf.

Definitions 1.10, 2.6, and 3.21). Not only its image $\ker G_{i+1}$, but also a part of the kernel is fixed such that $\ker \Pi_i \subseteq \ker \Pi_i Q_{i+1}$ is valid.

If we are dealing with matrix functions, the determinations are carried out pointwise for frozen arguments.

In the following we suppress the step index i. G complies with $G_{i+1}(z)$ and Π with $\Pi_i(z)$, where z is an arbitrary frozen argument.

For a given matrix $G \in \mathbb{R}^{k\times m}$ with $\operatorname{rank} G = r$ and a given projector $\Pi \in \mathbb{R}^{m\times m}$ with $\operatorname{rank}\Pi = \rho$, we seek a new projector matrix Q such that

$$\begin{aligned} \operatorname{im} Q &= \ker G, \\ \ker Q &\supseteq X \text{ (cf. (1.13))}, \end{aligned}$$

and X is any complement of $\widehat{N} := \ker\Pi \cap \operatorname{im} Q$ in $\ker\Pi$ (cf. (1.12)), which means that Q has to satisfy (cf. Proposition 1.13 (3)) the conditions

$$GQ = 0,\ \operatorname{rank} Q = m - r, \tag{7.14}$$

$$\Pi Q(I - \Pi) = 0. \tag{7.15}$$

Owing to Lemma A.7 such a projector Q exists. Denote $N := \ker G$ and $K := \ker\Pi = \operatorname{im}(I-\Pi)$. Condition (7.14) implies $\operatorname{im} Q = N$. If N and K intersect only trivially, i.e., $K \cap N = \{0\}$, which we call the *regular case*, we can form Q to satisfy $X = K \subseteq \ker Q$, and then condition (7.15) holds. In general the computation of a representation of X is needed. We have to fix a set $X \subseteq K$ such that $K = X \oplus N$. Notice that X is not uniquely defined. An example illustrates the situation.

Example 7.1. For $\Pi = \begin{bmatrix} 0 & & \\ & 0 & \\ & & 1 \end{bmatrix}$ and $G = \begin{bmatrix} 0 & -1 & 1 \\ 0 & 1 & -1 \\ & & 0 \end{bmatrix}$, $m = 3$, we obtain $K = \ker\Pi = \operatorname{span}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$ and $N = \ker G = \operatorname{span}\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$, further $K \cap N = \operatorname{span}\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ and $K \oplus N = \mathbb{R}^m$. Any plane given by $(K\cap N)^c := \operatorname{span}\begin{bmatrix} \cos\alpha & 0 \\ \sin\alpha & -\cos\alpha \\ 0 & \beta \end{bmatrix}$, with fixed $\alpha \in (0,\pi)$, and $\beta \neq 0$, is a complement of $K \cap N$ in $\mathbb{R}^m$. A possible subspace X can be given as $X = K \cap (K\cap N)^c = \operatorname{span}\begin{bmatrix} \cos\alpha \\ \sin\alpha \\ 0 \end{bmatrix}$. As we can see by the different choices of α and β, the complement $(K\cap N)^c$ as well as X are not unique. For reasons of dimensions, in this example, since $\dim(K+N) = m$, the projector onto N along X is uniquely determined as

$$Q = \begin{bmatrix} 1 & -\frac{\cos\alpha}{\sin\alpha} & \frac{\cos\alpha}{\sin\alpha} \\ & 0 & 1 \\ & & 1 \end{bmatrix}.$$

Figure 7.1 shows this case. In general, $R^m = \underbrace{N \oplus X}_{K+N} \oplus (K+N)^c$ holds with a nontrivial complement $(K+N)^c$, which shows that fixing X does not completely fix the projector. It is worth mentioning that always restricting the choice to orthogonal complements we arrive at the so-called widely orthogonal projectors, and those are uniquely determined. This case corresponds here to the choice $\alpha = \frac{\pi}{2}$ and $\beta = 1$.

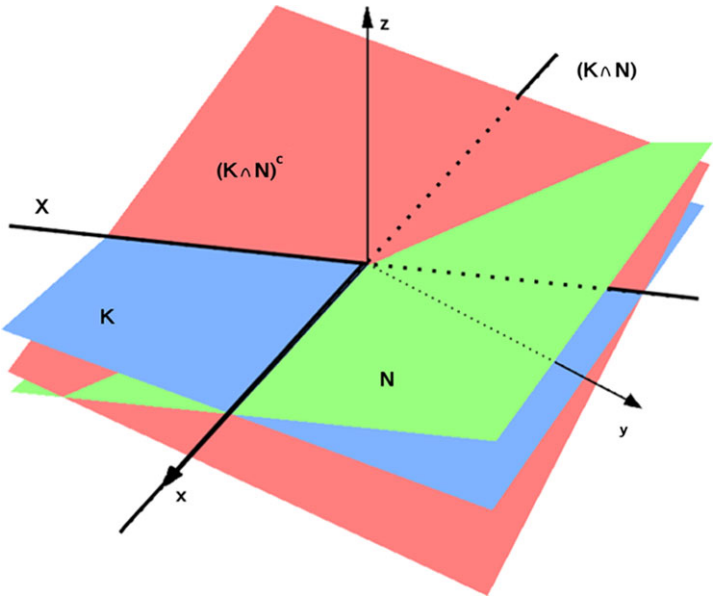


Fig. 7.1 Decomposition of $\mathbb{R}^3$

Now we start to discuss several methods of constructing projectors Q.

7.3.1 Basis representation methods

If a basis $n_1,\dots,n_{m-r}$ of N and a basis $\chi_1,\dots,\chi_\sigma$ of a suitable X are available, $X\cap N=0$, we immediately form a projector Q onto N satisfying $X\subseteq\ker Q$ we are looking for as (cf. Lemma A.7)

$$Q=H\begin{bmatrix}I\\0\end{bmatrix}H^-,$$

whereas $\mathcal{N}=\begin{bmatrix}n_1\dots n_{m-r}\end{bmatrix}$, $\mathcal{X}=\begin{bmatrix}\chi_1\dots\chi_\sigma\end{bmatrix}$ and $H:=\begin{bmatrix}\mathcal{N},\mathcal{X}\end{bmatrix}$ have full column rank, and H^- is any reflexive generalized inverse of H. Consider different ways of generating suitable bases, and at the same time, a suitable subspace X.

A basis of N is delivered by decomposition (7.1). We have to provide a basis of a suitable subspace X. Recall that for any matrix $\mathfrak{A}$ the relation $\ker\mathfrak{A}^*\mathfrak{A}=\ker\mathfrak{A}$ is true. Therefore, because of

$$\widehat{N}=N\cap K=\ker\begin{bmatrix}G\\\Pi\end{bmatrix}=\ker\left(G^*G+\Pi^*\Pi\right),$$

by means of a decomposition of $\begin{bmatrix} G \\ \Pi \end{bmatrix} \in \mathbb{R}^{k+m,m}$ or of $(G^*G + \Pi^*\Pi) \in \mathbb{R}^{m,m}$ we can design a projector Z onto $\widehat{N}$. The choice of this projector also fixes a possible complement $\widehat{N}^c := \operatorname{im} Z$ of $\widehat{N}$. By means of Z we compute a basis of X by one of the relations

$$\ker \begin{bmatrix} Z \\ \Pi \end{bmatrix} = \ker (Z^*Z + \Pi^*\Pi) = (N \cap K)^c \cap K = X.$$

This method of providing the projector Q needs three decompositions including those of matrices with $k+m$, respectively $2m$ rows as well as the computation of expressions like $(G^*G + \Pi^*\Pi)$.

An alternative possibly cheaper way to construct an admissible projector Q is suggested by Lemma A.5. Decomposing

$$G = U_G \begin{bmatrix} S_G \\ 0 \end{bmatrix} V_G^{-1}, \quad V_G =: [V_{G,1}, V_{G,2}] \tag{7.16}$$

we obtain $N = \ker G = \operatorname{im} V_{G,2}$, that is, a basis of N. Then, in order to apply Lemma A.5, we decompose

$$\Pi V_{G,2} = U_{\Pi N} \begin{bmatrix} S_{\Pi N} \\ 0 \end{bmatrix} V_{\Pi N}^{-1}, \quad V_{\Pi N} =: [V_{\Pi N,1}, V_{\Pi N,2}],$$

and hence $\ker \Pi V_{G,2} = \operatorname{im} V_{\Pi N,2}$ is valid. Then, owing to Lemma A.5, $Y := V_{G,2} V_{\Pi N,2} \in \mathbb{R}^{m \times q}$ represents a basis of $\ker G \cap \ker \Pi = N \cap K$. Having the basis Y of $N \cap K$ we could, as before, compute a projector Z onto $N \cap K$, and put $(N \cap K)^c = \ker Z$, but here we actually do not compute Z, but provide a basis of the nullspace of Z in a different way. We decompose

$$Y = U_Y \begin{bmatrix} S_Y \\ 0 \end{bmatrix}, \quad U_Y =: [U_{Y,1}, U_{Y,2}],$$

with nonsingular U_Y, S_Y. Now, $U_{Y,2} \in \mathbb{R}^{m \times (m-q)}$ serves as a basis of a complement $(N \cap K)^c = \ker Z$, which means $\ker Z = \operatorname{im} U_{Y,2}$. To apply Lemma A.5 once more we compute a basis of $\ker \Pi U_{Y,2}$ by the further decomposition

$$\Pi U_{Y,2} = U_X \begin{bmatrix} S_X \\ 0 \end{bmatrix} V_X^{-1}, \quad V_X =: [V_{X,1}, V_{X,2}],$$

yielding $\ker \Pi U_{Y,2} = \operatorname{im} V_{X,2}$. This finally leads to

$$X = (N \cap K)^c \cap K = \ker Z \cap \ker \Pi = \operatorname{im} U_{Y,2} V_{X,2}.$$

Here, four lower-dimensional matrix decompositions are needed to compute the admissible projector Q.

7.3.2 Basis representation methods—Regular case

In the regular case (cf. Definition 2.6), if

$$K \cap N = \{0\}, \tag{7.17}$$

equation (7.15) simplifies to

$$Q(I - \Pi) = 0. \tag{7.18}$$

Condition (7.17) implies $m - \rho \leq r$. With the background of the decomposition (7.16) of G, each projector onto N has the form

$$Q = V_G \begin{bmatrix} 0 & 0 \\ -M_1 S_G & I_{m-r} \end{bmatrix} V_G^{-1}. \tag{7.19}$$

A basis of $\operatorname{im}(I - \Pi) = \ker \Pi$ can be computed by means of the decomposition

$$\Pi = U_\Pi \begin{bmatrix} S_\Pi & \\ & 0 \end{bmatrix} V_\Pi^{-1}, \quad S_\Pi \in \mathbb{R}^{\rho \times \rho} \text{ nonsingular}, \quad V_\Pi =: [V_{\Pi,1}, V_{\Pi,2}]$$

yielding $\operatorname{im}(I - \Pi) = \operatorname{im} V_{\Pi,2}$, $\operatorname{rank} V_{\Pi,2} = m - \rho$. Now condition (7.18) means $QV_{\Pi,2} = 0$, or $V_{\Pi,2} = PV_{\Pi,2}$, with $P := I - Q$. This leads to

$$V_{\Pi,2} = PV_{\Pi,2} = V_G \begin{bmatrix} I & 0 \\ M_1 S_G & 0 \end{bmatrix} \underbrace{V_G^{-1} V_{\Pi,2}}_{=: \begin{bmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \end{bmatrix}} = V_G \begin{bmatrix} I & 0 \\ M_1 S_G & I \end{bmatrix} \begin{bmatrix} \mathcal{V}_1 \\ 0 \end{bmatrix}, \tag{7.20}$$

which shows that $\operatorname{rank} \mathcal{V}_1 = \operatorname{rank} V_{\Pi,2} = m - \rho$, i.e., $\mathcal{V}_1 \in \mathbb{R}^{r, m-\rho}$ has full column rank.
The requirement $QV_{\Pi,2} = 0$ results in the condition $-M_1 S_G \mathcal{V}_1 + \mathcal{V}_2 = 0$, which determines M_1. The choice

$$M_1 = \mathcal{V}_2 \mathcal{V}_1^- S_G^{-1} \tag{7.21}$$

satisfies this relation with an arbitrary generalized reflexive inverse $\mathcal{V}_1^-$, since $\mathcal{V}_1^- \mathcal{V}_1 = I$.
If Π is symmetric, V_Π is orthogonal and $\mathcal{V}_1^+$ is the Moore–Penrose inverse, then the choice

$$M_1 = \mathcal{V}_2 \mathcal{V}_1^+ S_G^{-1} \tag{7.22}$$

generates the widely orthogonal projector Q, which is shown at the end of Subsection 7.3.3.

7.3.3 Projector representation method

Now we build the projector Q without using subspace bases. We again apply the decomposition (7.16) and the general projector representation (7.19), that is

$$Q = V_G \begin{bmatrix} 0 & 0 \\ -M_1 S_G & I_{m-r} \end{bmatrix} V_G^{-1}.$$

Introducing $\tilde{\Pi} := V_G^{-1} \Pi V_G$ we derive the expression

$$\begin{bmatrix} V_G^{-1} & \\ & V_G^{-1} \end{bmatrix} \begin{bmatrix} \Pi \\ I-Q \end{bmatrix} V_G = \begin{bmatrix} V_G^{-1} \Pi V_G \\ I - V_G^{-1} Q V_G \end{bmatrix} = \begin{bmatrix} \tilde{\Pi}_{11} & \tilde{\Pi}_{12} \\ \tilde{\Pi}_{21} & \tilde{\Pi}_{22} \\ I_r & 0 \\ M_1 S_G & 0 \end{bmatrix} \begin{matrix} \} r \\ \\ \\ \\ \end{matrix} . \tag{7.23}$$

From $\ker \begin{bmatrix} \Pi \\ I-Q \end{bmatrix} = K \cap N = \widehat{N}$ and $u = \dim \widehat{N}$ it follows that $\dim \ker \begin{bmatrix} \Pi \\ I-Q \end{bmatrix} = m-u$. Regarding this we conclude from (7.23) that the rank condition $\operatorname{rank} \begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix} = m-u-r$ is valid.

Lemma 7.2. *Given a projector Π and the decomposition (7.1) of a matrix G, $\operatorname{rank} G = r$, then the projector Q defined by (7.19) satisfies the properties (7.14) and (7.15), supposing one of the following three conditions is satisfied:*

(1) $M_1 = -\begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix}^- \begin{bmatrix} \tilde{\Pi}_{11} \\ \tilde{\Pi}_{21} \end{bmatrix} S_G^{-1}$.

(2) $M_1 = -\tilde{\Pi}_{22}^- \tilde{\Pi}_{21} S_G^{-1}$, *and the reflexive generalized inverse* $\tilde{\Pi}_{22}^-$ *satisfies* $\tilde{\Pi}_{12} = \tilde{\Pi}_{12} \tilde{\Pi}_{22}^- \tilde{\Pi}_{22}$.

(3) $\Pi = \Pi^*$, V_G *is orthogonal, and* $M_1 = -\tilde{\Pi}_{22}^- \tilde{\Pi}_{21} S_G^{-1}$.

Moreover, the special choice of the Moore–Penrose inverse in case (3),

$$M_1 = -\tilde{\Pi}_{22}^+ \tilde{\Pi}_{21} S_G^{-1},$$

provides a symmetric ΠQ and a widely orthogonal Q.

Proof. (1) Condition (7.14) is always given by the construction and it remains to verify (7.15). We let $M := M_1 S_G = -\begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix}^- \begin{bmatrix} \tilde{\Pi}_{11} \\ \tilde{\Pi}_{21} \end{bmatrix}$ and compute

$$
\begin{aligned}
V_G^{-1}\Pi Q(I-\Pi)V_G &= \begin{bmatrix}\tilde\Pi_{11} & \tilde\Pi_{12}\\ \tilde\Pi_{21} & \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}0 & 0\\ -M & I\end{bmatrix}\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\\ -\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}\\
&= \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-M & I\end{bmatrix}\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\\ -\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}\\
&= \left[-\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix} M\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\right]\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\\ -\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}.
\end{aligned}
$$

The relation $0=\tilde\Pi(I-\tilde\Pi)$ provides

$$
\begin{bmatrix}\tilde\Pi_{11}\\ \tilde\Pi_{21}\end{bmatrix}\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\end{bmatrix} = -\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix},
$$

and hence

$$
M\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\end{bmatrix} = \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}^{-}\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}.
$$

Regarding this we finally find

$$
\begin{aligned}
V_G^{-1}\Pi Q(I-\Pi)V_G &= \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix} M\begin{bmatrix}I-\tilde\Pi_{11} & -\tilde\Pi_{12}\end{bmatrix} + \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}\\
&= \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}^{-}\left(-\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}\right)\\
&\qquad + \begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\begin{bmatrix}-\tilde\Pi_{21} & I-\tilde\Pi_{22}\end{bmatrix}\\
&= 0.
\end{aligned}
$$

(2) If we are aware of a $(m-r)\times(m-r)$ submatrix of $\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}\in\mathbb{R}^{m\times(m-r)}$, which has $\operatorname{rank} m-r-u$, a generalized reflexive inverse of $\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}$ can be computed, i.e., by a Householder decomposition. We assume without loss of generality that $\operatorname{rank}\tilde\Pi_{22}=m-r-u$. If the submatrix is distributed over the rows of the matrix, a row permutation leads to the same assumption but at the end the factor U contains row permutations.

Decompose $\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix} = \left[\begin{array}{c|c}Z_{12} & 0\\ \hline Z_{22,1} & 0\\ Z_{22,2} & 0\end{array}\right] U$ with nonsingular $Z_{22,1}\in\mathbb{R}^{m-r-u,m-r-u}$ and orthogonal U, and fix the reflexive generalized inverse $\tilde\Pi_{22}^{-}=U^*\begin{bmatrix}Z_{22,1}^{-1} & 0\\ 0 & 0\end{bmatrix}$. Below we show that a generalized reflexive inverse is given by

$$
\begin{bmatrix}\tilde\Pi_{12}\\ \tilde\Pi_{22}\end{bmatrix}^{-} := \begin{bmatrix}0 & \tilde\Pi_{22}^{-}\end{bmatrix}. \tag{7.24}
$$

Applying (1) and this special structure of the inverse provides

$$M_1 = -\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}^{-}\begin{bmatrix}\tilde{\Pi}_{11}\\ \tilde{\Pi}_{21}\end{bmatrix}S_G^{-1} = \begin{bmatrix}0 & -\tilde{\Pi}_{22}^{-}\end{bmatrix}\begin{bmatrix}\tilde{\Pi}_{11}\\ \tilde{\Pi}_{21}\end{bmatrix}S_G^{-1} = -\tilde{\Pi}_{22}^{-}\tilde{\Pi}_{21}S_G^{-1},$$

which verifies the assertion. It remains to verify that (7.24) in fact serves as a reflexive generalized inverse. The condition

$$\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}\begin{bmatrix}0 & -\tilde{\Pi}_{22}^{-}\end{bmatrix}\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix} = \begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}$$

is valid because of our assumption concerning the generalized inverse $\tilde{\Pi}_{22}^{-}$, namely

$$\tilde{\Pi}_{12} = \tilde{\Pi}_{12}\tilde{\Pi}_{22}^{-}\tilde{\Pi}_{22} \text{ or equivalently } \operatorname{im}(I - \tilde{\Pi}_{22}^{-}\tilde{\Pi}_{22}) = \ker\tilde{\Pi}_{22} \subseteq \ker\tilde{\Pi}_{12}. \quad (7.25)$$

(3) The symmetry of Π and $\tilde{\Pi}$ (with orthogonal V_G) yields

$$\tilde{\Pi}_{22} = \begin{bmatrix}\tilde{\Pi}_{21} & \tilde{\Pi}_{22}\end{bmatrix}\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix} = \begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}^{*}\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}$$

and therefore $\operatorname{rank}\tilde{\Pi}_{22} = m-r-u$ and $\ker\tilde{\Pi}_{22} = \ker\begin{bmatrix}\tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\end{bmatrix}$, i.e., $\ker\tilde{\Pi}_{22} \subseteq \ker\tilde{\Pi}_{12}$. Considering (7.25), assertion (3) is shown to be a consequence of (2).

Finally we have to verify that taking the Moore–Penrose inverse in case (3) one delivers a widely orthogonal projector Q. By Definition 1.12 a widely orthogonal projector Q projects onto N along $(K+N)^{\perp} \oplus X$ with $X = \widehat{N^{\perp}} \cap K$. Lemma A.7 (7) describes sufficient conditions. Put $M = \tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21}$ and derive

$$\Pi Q = V_G\tilde{\Pi}V_G^{*}V_G\begin{bmatrix}0 & 0\\ -M & I\end{bmatrix}V_G^{*} = V_G\begin{bmatrix}\tilde{\Pi}_{12}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21} & \tilde{\Pi}_{12}\\ \tilde{\Pi}_{22}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21} & \tilde{\Pi}_{22}\end{bmatrix}V_G^{*}.$$

The symmetry of $\tilde{\Pi}$ implies the symmetry of $\tilde{\Pi}_{11}, \tilde{\Pi}_{22}$ and $\tilde{\Pi}_{12} = \tilde{\Pi}_{21}^{*}$. The Moore–Penrose inverse of a symmetric matrix is symmetric itself, therefore $\tilde{\Pi}_{22}\tilde{\Pi}_{22}^{+} = \tilde{\Pi}_{22}^{+}\tilde{\Pi}_{22}$. We consider the matrix blocks of ΠQ which seemingly derange the symmetry,

$$(\tilde{\Pi}_{12}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21})^{*} = \tilde{\Pi}_{21}^{*}(\tilde{\Pi}_{22}^{+})^{*}\tilde{\Pi}_{12}^{*} = \tilde{\Pi}_{12}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21} \text{ and}$$

$$\tilde{\Pi}_{22}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{21} = (\tilde{\Pi}_{22}\tilde{\Pi}_{22}^{+})^{*}\tilde{\Pi}_{12}^{*} = (\tilde{\Pi}_{12}\tilde{\Pi}_{22}\tilde{\Pi}_{22}^{+})^{*} = (\tilde{\Pi}_{12}\tilde{\Pi}_{22}^{+}\tilde{\Pi}_{22})^{*} \overset{(7.25)}{=} \tilde{\Pi}_{12}^{*} = \tilde{\Pi}_{21},$$

but this shows the symmetry of ΠQ and naturally ΠP with $P = I - Q$.
The last properties we have to show are the symmetry of $P(I-\Pi)$ and the condition $Q\Pi P = 0$ as well. Derive

$$P(I-\Pi) = (I-Q)(I-\Pi) = V_G \begin{bmatrix} I & 0 \\ M & 0 \end{bmatrix} V_G^* V_G (I-\tilde{\Pi}) V_G^* \\ = V_G \begin{bmatrix} I-\tilde{\Pi}_{11} & -\tilde{\Pi}_{12} \\ M(I-\tilde{\Pi}_{11}) & -M\tilde{\Pi}_{12} \end{bmatrix} V_G^*,$$

further

$$M(I-\tilde{\Pi}_{11}) = -\tilde{\Pi}_{22}^+ \underbrace{\tilde{\Pi}_{21}(I-\tilde{\Pi}_{11})}_{\tilde{\Pi}_{22}\tilde{\Pi}_{21}} = -\tilde{\Pi}_{21} = -\tilde{\Pi}_{12}^*$$
$$M\tilde{\Pi}_{12} = -\tilde{\Pi}_{22}^+ \tilde{\Pi}_{21}\tilde{\Pi}_{12} = -\tilde{\Pi}_{22}^+ \tilde{\Pi}_{22}(I-\tilde{\Pi}_{22})$$

which shows the symmetry of $P(I-\Pi)$. Next we compute

$$Q\Pi P = V_G \begin{bmatrix} 0 & 0 \\ -M & I \end{bmatrix} V_G^* V_G \begin{bmatrix} \tilde{\Pi}_{11} & \tilde{\Pi}_{12} \\ \tilde{\Pi}_{21} & \tilde{\Pi}_{22} \end{bmatrix} V_G^* V_G \begin{bmatrix} I & 0 \\ M & 0 \end{bmatrix} V_G^* \\ = V_G \begin{bmatrix} 0 & 0 \\ -M\tilde{\Pi}_{11} + \tilde{\Pi}_{21} + (-M\tilde{\Pi}_{12} + \tilde{\Pi}_{22})M & 0 \end{bmatrix} V_G^*,$$

and

$$-M\tilde{\Pi}_{11} + \tilde{\Pi}_{21} + (-M\tilde{\Pi}_{12} + \underbrace{\tilde{\Pi}_{22})M}_{-\tilde{\Pi}_{21}} = \tilde{\Pi}_{22}^+ (- \underbrace{\tilde{\Pi}_{21}\tilde{\Pi}_{11}}_{(I-\tilde{\Pi}_{22})\tilde{\Pi}_{21}} + \underbrace{\tilde{\Pi}_{21}\tilde{\Pi}_{12}}_{\tilde{\Pi}_{22}(I-\tilde{\Pi}_{22})} \tilde{\Pi}_{22}^+ \tilde{\Pi}_{21}) = 0,$$

and hence $Q\Pi P = 0$. Now the assertion follows from Lemma A.7 (7). □

We are especially interested in the regular case, where $\widehat{N} = \{0\}$. Lemma 7.2 suggests a practical way to compute a widely orthogonal projector for that case. Since $\widehat{N} = \{0\}$ and $u = \dim \widehat{N} = 0$ the matrix $\begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix}$ in (7.23) has full column rank. Then $\tilde{\Pi}_{22} = \begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix}^* \begin{bmatrix} \tilde{\Pi}_{12} \\ \tilde{\Pi}_{22} \end{bmatrix}$ is nonsingular and $\tilde{\Pi}_{22}^+ = \tilde{\Pi}_{22}^{-1}$. Moreover, $\tilde{\Pi}_{22}$ is not only nonsingular but positive definite as the following lemma proves.

Lemma 7.3. *Let the symmetric projector $\Pi = \begin{bmatrix} \Pi_{11} & \Pi_{12} \\ \Pi_{21} & \Pi_{22} \end{bmatrix}$ have a nonsingular block Π_{22}. Then this block Π_{22} is positive definite.*

Proof. Π is a projector and therefore $\Pi_{22} = \Pi_{21}\Pi_{12} + \Pi_{22}^2$. It holds that $\Pi_{21} = \Pi_{12}^*$ and $\Pi_{22} = \Pi_{22}^*$. We consider $\langle \Pi_{22}x, x\rangle$ for $x \neq 0$.

$$\begin{aligned} \langle \Pi_{22}x, x\rangle &= \langle (\Pi_{21}\Pi_{12} + \Pi_{22}^2)x, x\rangle \\ &= \langle \Pi_{21}\Pi_{12}x, x\rangle + \langle \Pi_{22}^2 x, x\rangle \\ &= \langle \Pi_{12}x, \Pi_{12}x\rangle + \langle \Pi_{22}x, \Pi_{22}x\rangle \\ &\geq \langle \Pi_{22}x, \Pi_{22}x\rangle > 0. \end{aligned}$$

□

Lemma 7.3 suggests to decompose $\tilde{\Pi}_{22}$ by Cholesky decomposition when computing

$$M_1 = -\tilde{\Pi}_{22}^{-1}\tilde{\Pi}_{21}S_G^{-1} \tag{7.26}$$

for widely orthogonal projectors.
Next we show that, in the regular case, formula (7.22) provides exactly the same M_1 as formula (7.26), and hence an additional way to compute the widely orthogonal projectors.
The projector Π is symmetric and has the decomposition

$$\Pi = V_\Pi \begin{bmatrix} I & \\ & 0 \end{bmatrix} V_\Pi^T =: \begin{bmatrix} V_{\Pi,1} & V_{\Pi,2} \end{bmatrix} \begin{bmatrix} I & \\ & 0 \end{bmatrix} \begin{bmatrix} V_{\Pi,1}^T \\ V_{\Pi,2}^T \end{bmatrix}$$

with an orthogonal matrix V_Π. We obtain $\Pi = I - V_{\Pi,2}V_{\Pi,2}^T$, which leads to

$$\begin{aligned} \tilde{\Pi} = V_G^{-1}\Pi V_G &= I - V_G^{-1}V_{\Pi,2}V_{\Pi,2}^T V_G \quad \text{(cf. (7.20))} \\ &= I - \begin{bmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \end{bmatrix} \begin{bmatrix} \mathcal{V}_1 \\ \mathcal{V}_2 \end{bmatrix}^T \\ &= \begin{bmatrix} I - \mathcal{V}_1\mathcal{V}_1^T & -\mathcal{V}_1\mathcal{V}_2^T \\ -\mathcal{V}_2\mathcal{V}_1^T & I - \mathcal{V}_2\mathcal{V}_2^T \end{bmatrix}. \end{aligned}$$

Applying (7.26) we obtain

$$M_1 = \underbrace{(I - \mathcal{V}_2\mathcal{V}_2^T)^{-1}\mathcal{V}_2}_{=\mathcal{V}_2(\mathcal{V}_1^T\mathcal{V}_1)^{-1}} \mathcal{V}_1^T S_G^{-1} = \mathcal{V}_2 \underbrace{(\mathcal{V}_1^T\mathcal{V}_1)^{-1}\mathcal{V}_1^T}_{=\mathcal{V}_1^+} S_G^{-1}, \tag{7.27}$$

which coincides with (7.21).

7.4 Matrix function sequences

7.4.1 Stepping level by level

The admissible sequences of matrix functions are constructed pointwise. We start with matrices (standing for matrix functions with frozen arguments) A, D and B. We compute a generalized inverse of $G_0 := AD$ and fix in that way a projector $Q_0 := I - P_0 = I - G_0^- G_0$ onto $\ker G_0$. The starting matrices of the sequence are $G_0, G_0^-, B_0 := B, \Pi_0 := P_0$.
Let us assume that we have determined the sequence up to level i, which means, G_i, the admissible projectors $Q_j, j = 1, \ldots, i$, and the projectors $\Pi_j = P_0 \ldots P_j$ are already computed. Since they are admissible, the condition (cf. (7.15))

$$\Pi_{j-1}Q_j(I-\Pi_{j-1})=0$$

holds for every level $j=1,\dots,i$. We have to build $G_{i+1}=G_i+B_iQ_i$ and a nullspace projector Q_{i+1} onto $\ker G_{i+1}$ satisfying

$$X_{i+1}=(N_0+\cdots+N_i)\ominus\widehat{N}_i\subset\ker Q_{i+1} \tag{7.28}$$

(cf. (2.45)), or equivalently,

$$\Pi_iQ_{i+1}(I-\Pi_i)=0. \tag{7.29}$$

The decomposition

$$G_{i+1}=\mathcal{U}_{i+1}\begin{bmatrix}S_{i+1}&\\&0\end{bmatrix}\mathcal{V}_{i+1}^{-1} \tag{7.30}$$

provides the reflexive generalized inverse

$$G_{i+1}^{-}=\mathcal{V}_{i+1}\begin{bmatrix}S_{i+1}^{-1}&M_{2,i+1}\\M_{1,i+1}&M_{1,i+1}S_{i+1}M_{2,i+1}\end{bmatrix}\mathcal{U}_{i+1}^{-1}$$

and the nullspace projector

$$Q_{i+1}=\mathcal{V}_{i+1}\begin{bmatrix}0&0\\-M_{1,i+1}S_{i+1}&I\end{bmatrix}\mathcal{V}_{i+1}^{-1}.$$

The entry $M_{1,i+1}$ can be computed by means of one of the proposals in Section 7.3 and $M_{2,i+1}$ can be set to zero.

Since we proceed pointwise with frozen arguments, to ensure continuity of the nullspace projector and then that of the next matrix function, it is recommended to apply widely orthogonal projectors. For widely orthogonal projectors we need an orthogonal matrix $\mathcal{V}_{i+1}$ (see Lemma 7.2 (3)), which requires a decomposition of G_{i+1} by an SVD or by the Householder method (decomposition of G_{i+1}^*). After having generated G_{i+1} and the nullspace projector Q_{i+1} we have to provide also the next

$$B_{i+1}=B_iP_i-G_{i+1}D^{-}(D\Pi_{i+1}D^{-})'D\Pi_i \quad \text{(cf. (2.8))}$$

or, in the invariant case,

$$B_{i+1}=B_iP_i.$$

The latter case does not present any difficulty; however, in general the involved derivative of $D\Pi_{i+1}D^{-}$ represents a serious challenge. In [137] finite differences are used to approximate this derivative, which delivers quite accurate results in lower index cases and if the relevant subspaces are invariant. A more accurate approximation of the derivative by automatic differentiation (cf. [98]) is done in [143]. The application of automatic differentiation needs higher smoothness assumptions as needed for the tractability index concept itself. In Section 8.1 the index determination for nonlinear DAEs is discussed.

7.4.2 Involved version for the regular case

A complete new decomposition of G_{i+1} at each level appears to be expensive. In the regular case, a possibility to make better use of results obtained in previous steps is developed in [137]. We use the representation

$$G_{i+1} = G_i + B_i Q_i = (G_i + W_i B_0 Q_i) F_{i+1}$$

with the projector W_i along $\operatorname{im} G_i$ and the nonsingular matrix $F_{i+1} = I + G_i^- B_i Q_i$ (cf. Proposition 2.5 (3)). For the matrix $G_j,\ j = 0, \ldots, i$ we already have the decomposition

$$G_j = \mathcal{U}_j \begin{bmatrix} S_j & \\ & 0 \end{bmatrix} \mathcal{V}_j^{-1}$$

with $\mathcal{U}_j$, S_j and $\mathcal{V}_j$ nonsingular matrices. The other components are given for $j = 0, \ldots, i$ by

$$G_j^- = \mathcal{V}_j \begin{bmatrix} S_j^{-1} & M_{2,j} \\ M_{1,j} & M_{1,j} S_j M_{2,j} \end{bmatrix} \mathcal{U}_j^{-1},$$

$$W_j = I - G_j G_j^- = \mathcal{U}_j \begin{bmatrix} 0 & -S_j M_{2,j} \\ & I \end{bmatrix} \mathcal{U}_j^{-1} = \mathcal{U}_j T_{u,j}^{-1} \begin{bmatrix} 0 & \\ & I \end{bmatrix} \mathcal{U}_j^{-1}, \tag{7.31}$$

$$Q_j = I - G_j^- G_j = \mathcal{V}_j \begin{bmatrix} 0 & \\ -M_{1,j} S_j & I \end{bmatrix} \mathcal{V}_j^{-1} = \mathcal{V}_j \begin{bmatrix} 0 & \\ & I \end{bmatrix} T_{l,j}^{-1} \mathcal{V}_j^{-1} \tag{7.32}$$

with the upper and lower triangular matrices

$$T_{u,j} := \begin{bmatrix} I & S_j M_{2,j} \\ & I \end{bmatrix} \text{ and } T_{l,j} := \begin{bmatrix} I & \\ M_{1,j} S_j & I \end{bmatrix}.$$

Using the detailed structure of the various matrices we find

$$G_{i+1} = \mathcal{U}_i T_{u,i}^{-1} \left(\begin{bmatrix} S_i & \\ & 0 \end{bmatrix} + \begin{bmatrix} 0 & \\ & I \end{bmatrix} \underbrace{\mathcal{U}_i^{-1} B_0 \mathcal{V}_i}_{\bar{B}_i} \begin{bmatrix} 0 & \\ & I \end{bmatrix} \right) T_{l,i}^{-1} \mathcal{V}_i^{-1} F_{i+1}.$$

If we write $\bar{B}_i = \begin{bmatrix} B_{11}^i & B_{12}^i \\ B_{21}^i & B_{22}^i \end{bmatrix}$ and decompose $B_{22}^i = \tilde{U}_{i+1} \begin{bmatrix} \tilde{S}_{i+1} & \\ & 0 \end{bmatrix} \tilde{V}_{i+1}^{-1}$, we can use this decomposition and obtain

$$G_{i+1} = \underbrace{\mathcal{U}_i T_{u,i}^{-1} \begin{bmatrix} I & \\ & \tilde{U}_{i+1} \end{bmatrix}}_{=:\mathcal{U}_{i+1}} \begin{bmatrix} S_i & & \\ & \tilde{S}_{i+1} & \\ & & 0 \end{bmatrix} \underbrace{\begin{bmatrix} I & \\ & \tilde{V}_{i+1}^{-1} \end{bmatrix} T_{l,i}^{-1} \mathcal{V}_i^{-1} F_{i+1}}_{=:\mathcal{V}_{i+1}^{-1}}. \tag{7.33}$$

Defining $S_{i+1} := \begin{bmatrix} S_i & \\ & \tilde{S}_{i+1} \end{bmatrix}$ we now have the required decomposition of

$$G_{i+1} = \mathcal{U}_{i+1} \begin{bmatrix} S_{i+1} & \\ & 0 \end{bmatrix} \mathcal{V}_{i+1}^{-1} \tag{7.34}$$

and

$$G_{i+1}^{-} = \mathcal{V}_{i+1} \begin{bmatrix} S_{i+1}^{-1} & M_{2,j} \\ M_{1,i+1} & M_{1,i+1} S_{i+1} M_{2,i+1} \end{bmatrix} \mathcal{U}_{i+1}^{-1}.$$

The projector

$$Q_{i+1} = I - G_{i+1}^{-} G_{i+1} = \mathcal{V}_{i+1} \begin{bmatrix} 0 & 0 \\ -M_{1,i+1} S_{i+1} & I \end{bmatrix} \mathcal{V}_{i+1}^{-1}.$$

is a nullspace projector for each $M_{1,i+1}$.
To fix the projector, the different entries $M_{1,i+1}$ can be determined as described in Section 7.3, where Π is replaced by Π_i. The computation of M_1 by (7.21) goes better with the step-by-step computation. Widely orthogonal projectors are computed using the Moore–Penrose inverse of $\mathcal{V}_1$ (see (7.27)).
The advantage of the involved step-by-step computation of the sequence is that, at each step, we decompose only the matrix $\bar{B}_{22}^{i}$, whose dimension reduces from step to step.
After having computed G_{i+1} and Q_{i+1} we have to provide

$$B_{i+1} = B_i P_i - G_{i+1} D^{-} (D \Pi_{i+1} D^{-})' D \Pi_i \quad \text{(cf. (2.8))}.$$

Here again, the challenge is the differentiation of $D\Pi_{i+1}D^{-}$.

7.4.3 Computing characteristic values and index check

The characteristic values of the DAE under consideration, which are (see Definition 2.9) the values

$$r_i = \operatorname{rank} G_i,\ u_i = \dim \widehat{N}_i,\ \widehat{N}_i = \ker \begin{bmatrix} \Pi_{i-1} \\ I - Q_i \end{bmatrix} = \ker \begin{bmatrix} \Pi_{i-1} \\ G_i \end{bmatrix} = \ker \left[G_i^* G_i + \Pi_{i-1}^* \Pi_{i-1} \right]$$

are rank values arising as byproducts within the factorizations when generating the matrix sequences as described in the previous subsection.
If one meets a nonzero value u_i, the given DAE fails to be regular, which makes the question

$$\text{“}u_i = 0 \text{ ?”}$$

serve as a regularity test.
The determination of the tractability index of a regular DAE requires the determination of the matrix sequence up to a nonsingular matrix G_μ. We concentrate on the regular case.
At every level G_i, $i = 0, 1, \ldots$, the characteristic value $r_i = \operatorname{rank} G_i$ is determined by checking the nonsingularity of G_i. The step-by-step algorithm of Section 7.4.2

delivers the characteristic values successively starting from $r_0 = \operatorname{rank} G_0$ and $r_{i+1} = r_i + r_B^i$, $i = 0, 1, \ldots$, with $r_B^i := \operatorname{rank} \bar{B}_{22}^i$.

The regularity is implicitly checked by computing an admissible projector at every level. In the case of a critical point, we are faced with a rank drop of $\mathcal{V}_1$ if we use (7.21) or a singular block $\tilde{\Pi}_{22}$ if we apply (7.26).

The computation of B_i, $i > 0$, needs the differentiation of $D\Pi_i D^-$. The factorization $G_{i+1} = (G_i + W_i B_0 Q_i) \underbrace{(I + G_i^- B_i Q_i)}_{\text{nonsingular}}$ allows us to determine $r_{i+1} = \operatorname{rank}(G_i + W_i B_0 Q_i)$, which is easier, since one can do without computing the derivative of $D\Pi_i D^-$. The first level where the derivative of $D\Pi_i D^-$ may influence the sequence matrix occurs for $i = 2$. The check of the index-3 property needs only one differentiation, which is accurately realizable by finite differences.

Algorithmic differentiation (AD) to compute the derivative of $D\Pi_i D^-$ is applied in [144]. Using an AD tool all computations are made by Taylor polynomials and a derivative is reduced to a shift of the Taylor series. The application of this technique requires higher smoothness assumptions.

For time-invariant linear DAEs the tractability index coincides with the Kronecker index (cf. Theorem 1.31), i.e., the numerical determination of the characteristic values discloses the inner structure of the DAE.

For a further discussion of the numerical index determination of nonlinear DAEs see Section 8.1.

Chapter 8
Aspects of the numerical treatment of higher index DAEs

At the beginning of the numerical treatment of DAEs, several experiments with the integration of initial value problems of higher index DAEs were done. The results were usually not satisfactorily. One could observe instabilities and numerical difficulties, in particular when integrating index-3 DAEs arising from rigid body mechanics (cf. [63]). Meanwhile several stabilizing techniques (see, e.g., [88], [23]) have been introduced for problems with a special structure (e.g., DAEs in Hessenberg form) to counteract these problems. Reading only the titles of papers (e.g., [50], [132]) one could think that it is no longer a challenge to solve higher index DAEs. But one should be aware that derivative arrays of DAEs are used there to reduce the higher index DAE to an index-0 or index-1 DAE before performing any integration method.

This chapter advises the reader of various troubles arising when numerical methods are applied directly to higher index DAEs. Before demonstrating this, we present a procedure for the practical calculation of the index and make a few remarks on consistent initialization in the higher index case. This is of importance for users of DAE solver packages since they usually require knowledge about the DAE index.

8.1 Practical index calculation

The calculation of the index of a DAE now coincides, in the fully nonlinear case, with the determination of a regular point of the DAE. Two ways are appropriate. The theoretical index investigation using structure, smoothness and maybe additional assumptions determining the resulting index for a class of problems as it is done, e.g., for Hessenberg systems in Section 3.5 or for DAEs simulating electrical circuits in Section 3.6. The other way is a pointwise numerical determination. We choose a time point t and a jet $(x, x^1, x^2, \ldots, x^\nu)$ with $\nu < \mu \leq m$. At this point we compute the matrix sequence (cf. Section 3.2) and determine the index of the DAE.
The tractability index depends on the jet (see Definition 3.28). The fact that different solutions of a DAE may indicate different values of the differentiation index (cf. [5],

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_8,

p. 235) or the structural index (cf. [186], Example 2.3) is a known property, as we will see in the example where we compare the differentiation, the structural, and the tractability index.

Example 8.1 (Index dependence on jet variables (cf. [144])). We consider the DAE

$$x_2' + x_1 - t = 0, \tag{8.1}$$
$$x_2' + x_3' + \gamma x_1 x_2 + \eta x_2 - 1 = 0, \tag{8.2}$$
$$x_2(1 - \frac{x_2}{2}) + x_3 = 0. \tag{8.3}$$

The proper formulation reads

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 0 \end{bmatrix}}_{A} \left(\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \right)' + \begin{bmatrix} x_1 - t \\ \gamma x_1 x_2 + \eta x_2 - 1 \\ x_2(1 - \frac{x_2}{2}) + x_3 \end{bmatrix} = 0. \tag{8.4}$$

The differentiation index is based upon the derivative array, which is given up to order 2 by (8.1)–(8.3), and the differentiated equations (8.5)–(8.8)

$$x_2'' + x_1' - 1 = 0, \tag{8.5}$$
$$x_2'' + x_3'' + \gamma(x_1' x_2 + x_1 x_2') + \eta x_2' = 0, \tag{8.6}$$
$$x_2'(1 - x_2) + x_3' = 0, \tag{8.7}$$
$$- -$$
$$x_2''' + x_1'' = 0,$$
$$x_2''' + x_3''' + \gamma(x_1'' x_2 + 2x_1' x_2' + x_1 x_2'') + \eta x_2'' = 0,$$
$$x_2'' + x_3'' - x_2'' x_2 - (x_2')^2 = 0. \tag{8.8}$$

The differentiation index requires us to filter an ODE system from the derivative array.
We form the ODE system by (8.1) (for x_2'), (8.7) (for x_3') and (8.6) (for x_1'). Replacing $x_2'' + x_3''$ in (8.6) we use (8.8) and, finally, replacing x_2'' and x_2' we need (8.5) and (8.1). The system we thus obtain reads

$$\begin{aligned} x_1' \underline{x_2(\gamma - 1)} + x_2 + (t - x_1)(t - x_1 + \gamma x_1 + \eta) &= 0, \\ x_2' + x_1 - t &= 0, \\ x_3' + (t - x_1)(1 - x_2) &= 0. \end{aligned} \tag{8.9}$$

Hence, the DAE (8.1)–(8.3) has differentiation index $\mu_d = 2$ if and only if the condition $x_2(\gamma - 1) \neq 0$ is satisfied.

The structural index is based on quantities deduced from the DAE. We apply the definition given in [186]. We have to compute the signature matrix Σ, the equation offsets c, the variable offsets d, and the system Jacobian J with

$$J_{ij} = \begin{cases} \frac{\partial f_i}{\partial((d_j - c_i)\text{th derivative of } x_j)} & \text{if this derivative is present in } f_i \\ 0 & \text{otherwise incl. } d_j - c_i < 0, \end{cases}$$

which has to be nonsingular. In this case the structural index is defined by

$$\mu_s = \max_i c_i + \begin{cases} 0 & \text{if all} \quad d_j > 0 \\ 1 & \text{if some} \quad d_j = 0. \end{cases}$$

For the DAE (8.1)–(8.3) we obtain

$$\Sigma = \begin{array}{c} \\ d \end{array} \begin{array}{c} \begin{bmatrix} 0 & 1 & -\infty \\ 0 & 1 & 1 \\ -\infty & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix} \end{array} \begin{array}{c} c \\ 0 \\ 0 \\ 1 \end{array}$$

and the related system Jacobian matrix

$$J = \begin{bmatrix} 1 & 1 & 0 \\ \gamma x_2 & 1 & 1 \\ 0 & 1 - x_2 & 1 \end{bmatrix}.$$

J is nonsingular if $x_2(1-\gamma) \neq 0$ and the structural index $\mu_s = 2$.

The tractability index matrix sequence as defined in (3.21) starts for (8.4) with the matrices

$$G_0 = AD = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \; B = \begin{bmatrix} 1 & 0 & 0 \\ \gamma x_2 & \gamma x_1 + \eta & 0 \\ 0 & 1 - x_2 & 1 \end{bmatrix},$$

where (x, x') denotes the chosen point in the jet space. A nullspace projector Q_0 onto $\ker G_0$ and the next sequence matrix G_1, and a nullspace projector Q_1 with $Q_1 Q_0$ onto $\ker G_1$ are given by

$$Q_0 = \begin{bmatrix} 1 & & \\ & 0 & \\ & & 0 \end{bmatrix}, \; G_1 = G_0 + BQ_0 = \begin{bmatrix} 0 & 1 & 0 \\ \gamma x_2 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \; Q_1 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & \gamma x_2 - 1 & 0 \end{bmatrix}.$$

From (3.21) it follows that $B_1 = BP_0 - G_1 D^- (DP_1 D^-)' D$ with

$$D^- = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \; DP_1 D^- = \begin{bmatrix} 1 & 0 \\ \gamma x_2 & 0 \end{bmatrix},$$

and we obtain

$$B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \gamma(x_1 + x_2') + \eta & 0 \\ 0 & x_2(\gamma - 1) & 0 \end{bmatrix}, \; G_2 = G_1 + B_1 Q_1 = \begin{bmatrix} 1 & 1 & 0 \\ \gamma x_2 & 1 + \gamma(x_1 + x_2') + \eta & 1 \\ 0 & x_2(\gamma - 1) & 0 \end{bmatrix}.$$

$\det G_2 = x_2(\gamma - 1)$ and the DAE has tractability index $\mu_t = 2$ iff the point belongs to a region where $x_2(\gamma - 1) \neq 0$.

This shows that, in the considered index definitions, the DAE (8.4) has the same index under the same restriction. If we assume that $\gamma \neq 1$, then the DAE has a critical point at $x_2 = 0$ or, more precisely, $x_2 = 0$ separates two regularity regions. Therefore, the index obviously depends on the chosen point.
DAE (8.4) was considered for $\gamma = 1$ in Example 3.60. Unfortunately, in that case the structural index is not defined, because the matrix J becomes singular.
However, the DAE has tractability index 3 if additionally $x_1 + \eta + x_2^1 \neq 0$ (cf. Example 3.60). The determination of the differentiation index requires the same condition and leads to the index-3 property. □

The complete characterization of a DAE, that is to figure out all regularity regions including the characteristic values by numerical methods, seems to be too extensive in the general case. What we can do is to check several necessary regularity conditions (cf. Theorem 3.33), e.g., monitoring the index in the points computed during an integration of an IVP.
In Example 3.34 a solution of DAE (3.27) crosses in $t = \frac{1}{2}$ a critical point leading to bifurcations (see Figure 3.4) which might not be discovered by the integration method.
An index monitor should supervise the characteristic values (which includes the index) and properties of the sequence matrices G_i, $i = 0, \ldots, \mu$. During an integration we may pass a critical point only. Here a monitoring of the condition of G_μ is helpful. This could be done by a check of pivot elements of the applied decompositions when the characteristic values are computed.

We now sketch an algorithm to calculate and test the characteristic values by means of the matrix sequence (3.21):
Fix a time point t and a jet $(x, x^1, \ldots, x^\nu)$, $\nu \leq \mu - 1 < m$, or choose a linearization function $x(\cdot)$, which is sufficiently smooth to compute the required derivatives at t.

1. Compute the initialization matrices A, D, B (cf. (3.13)–(3.15)), check the well-matched condition of A and D (cf. Section 7.2), compute D^-, set $i = 0$.
2. If $i == 0$
   set $G_0 = AD$, $B_0 = B$, $Q_0 = I - D^- D$, $r_0 = \operatorname{rank} G_0$.
   else
   compute $G_i = G_{i-1} + B_{i-1} Q_{i-1}$ (cf. (3.21) and Section 7.4), $r_i = \operatorname{rank} G_i$.
3. If $r_i == m \Rightarrow$ tractability index $\mu = i$, SUCCESS.
4. If $i == m \Rightarrow$ no regular matrix sequence $\Rightarrow$ critical point, STOP.
5. If $i > 0$ compute an admissible projector Q_i projecting onto $\ker G_i$ (cf. Definition 3.21).
   If no projector exists,
   i.e., $M_{1,i}$ (cf. Section 7.3) not calculable $\Rightarrow$ critical point, STOP.
6. Compute $P_i = I - Q_i$,
   if $i == 0$
   set $\Pi_0 = P_0$

else

compute $\Pi_i = \Pi_{i-1}P_i$, $(D\Pi_i D^-)'$ (cf. Section 7.4.3) and B_i (cf. (3.21)).

7. Set $i := i+1$, GOTO 2.

This algorithm is implemented using widely orthogonal projectors and the differentiation of $(D\Pi_i D^-)'$ is done by algorithmic differentiation using the MATLAB AD-tool INTLAB (cf. [195]).

Example 8.2 (Robotic arm ([57])). The DAE describes a prescribed path control of a two-link, flexible-joint, planar robotic arm as presented in [43].

$$\begin{bmatrix} I_6 \\ 0 \end{bmatrix} \left(\begin{bmatrix} I_6 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ u_1 \\ u_2 \end{bmatrix} \right)' - \begin{bmatrix} x_4 \\ x_5 \\ x_6 \\ f_4(x_2,x_3,x_4,x_6) + a(x_3)(u_1-u_2) \\ f_5(x_2,x_3,x_4,x_6) - a(x_3)(u_1-u_2) + u_2 \\ f_6(x_2,x_3,x_4,x_6) - (a(x_3)+b(x_3))(u_1-u_2) \\ \cos x_1 + \cos(x_1+x_3) - p_1(t) \\ \sin x_1 + \sin(x_1+x_3) - p_2(t) \end{bmatrix} = 0.$$

Set $t = 1$. We determine the index and characteristic values r_i at $x = (-1.72,\ 0.39,\ 1.718,\ -2.72,\ 4.29,\ 1.72,\ 13.59,\ 14.33)$. The QR decomposition of G_i provides the diagonal elements R_{r_i,r_i} of the upper triangular matrix for a rank decision. We observe in the next table that the gap between $R_{r_i,r_i} > \mathit{threshold} = 10^{-12}$ and the next diagonal element R_{r_i+1,r_i+1} allows a robust rank determination. We obtain

i	$\det G_i$	$r_i = \operatorname{rank} G_i$	$\lvert R_{r_i,r_i}\rvert$	$\lvert R_{r_i+1,r_i+1}\rvert$
0	0	6	1	0
1	0	6	1	0
2	0	6	9.6556e–1	4.0164e–17
3	−5.3724e–17	7	1.0968e–1	3.2092e–17
4	−1.0724e–15	7	1.3430e–1	6.7654e–17
5	−2.4783e+00	8	5.4233e–2	

The same DAE was investigated by other authors using different index concepts (cf. [42, 185]).

	dimension	index
differentiation ([42])	40 (27)	5
structural ([185]), *derivatives of 2nd order*	5	3
manually modified	9	5
tractability	8	5

The differentiation index needs to investigate a derivative array of dimension 40 or, if one knows in advance which equations are to be differentiated, at least dimension 27. The determined index equals 5, (cf. [42]).

The structural index applied to a modified DAE version with second derivatives determines the index 3. A manually modified DAE of dimension 9 delivers the index

5, (cf. [185]).
The DAE has no degrees of freedom, i.e., there is no dynamics within the system, which leads to $\Pi_4 \equiv 0$. We can use this property to check the accuracy of the numerical results. We obtain

$$\max_{i,j} |(D\Pi_4 D^-)_{ij}| = 1.005\text{e}{-15}, \quad \max_{i,j} |(D\Pi_4 D^-)'_{ij}| = 1.354\text{e}{-15}.$$

Also the accuracy of the projector calculation lies near the machine precision:

Projector property	$\max_i \lvert Q_i^2 - Q_i \rvert = 4.022\text{e}{-15}$
Admissibility	$\max_{i>j} \lvert Q_i Q_j \rvert = 3.075\text{e}{-15}$

□

8.2 Consistent initialization

An initial value problem of a nonlinear DAE of index μ is described by

$$f((d(x,t))', x, t) = 0,\ t \in \mathcal{I}, \tag{8.10}$$

$$C(x(t_0) - x^0) = 0. \tag{8.11}$$

The choice of the matrix C is in the nonlinear higher index case a nontrivial task. In the linear case we can take any matrix C with $\ker C = N_{can}(t_0)$ or equivalently $C = C\Pi_{\mu-1}(t_0)$ (cf. Theorem 3.66). In the nonlinear case the projector $\Pi_{\mu-1}$ may depend on the solution in t_0 up to its $(\mu-1)$th derivatives (cf. Section 3.2) and is therefore in general not available. But in most cases taking advantage of the structure of the given DAE or using numerical computations of $\Pi_{\mu-1}$ a matrix C is available. It is important to fix with (8.11) directly or indirectly the components of the inherent ODE only.
To start an integration of an index-μ DAE we have to compute consistent initial values (cf. Definition 3.6) at t_0. The $\Pi_{\mu-1}x(t_0)$ component is fixed by the initial condition (8.11), e.g., we have to compute the $(I - \Pi_{\mu-1})x(t_0)$ component, which is fixed by the obvious constraint and in the case of higher index $\mu > 1$ additionally the hidden constraints and we have to compute a value $y_0 = d'(x(t_0), t_0)$ such that $f(y_0, x(t_0), t_0) = 0$. The pair y_0, x_0 is also called *consistent initialization*.
We illustrate the situation by the next examples.

Example 8.3 (Consistent initialization of an index-1 DAE). Let us consider the DAE

$$x_1'(t) - x_1(t)\, t - 1 = 0, \tag{8.12}$$

$$x_1(t)^2 + x_2(t)^2 - 1 = 0. \tag{8.13}$$

The proper formulation is realized by $A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$, $D = \begin{bmatrix} 1 & 0 \end{bmatrix}$ and we have $AD = G_0 =: P_0$. This DAE has index 1 for $x_2 \neq 0$. If we choose $C := P_0$ we can established the DAE with the initial condition

$$P_0(x(t_0) - x^0) = 0.$$

At t_0 we have equations (8.12), (8.13) and the initial condition

$$x_1'(t_0) - x_1(t_0)\, t_0 - 1 = 0, \tag{8.14}$$
$$x_1(t_0)^2 + x_2(t_0)^2 - 1 = 0, \tag{8.15}$$
$$x_1(t_0) = x_1^0. \tag{8.16}$$

This leads directly to $x_1(t_0) = x_1^0$ and with (8.14) we obtain $x_1'(t_0) = x_1^0\, t_0 + 1$ and from (8.15) also the last component $x_2(t_0) = \sqrt{1 - (x_1^0)^2}$ is determined. We discover that $x(t_0) \in \mathcal{M}_0(t_0)$. □

For a general procedure for index-1 DAEs we refer to Section 4.3.

Example 8.4 (Consistent initialization of an index-2 DAE). We consider the DAE

$$x_1'(t) - x_1(t)\, t - 1 = 0, \tag{8.17}$$
$$x_2(t)x_2'(t) - x_3(t) = 0, \tag{8.18}$$
$$x_1(t)^2 + x_2(t)^2 - 1 = 0. \tag{8.19}$$

The matrix sequence starts with $A = \begin{bmatrix} 1 \\ x_2 \\ 0 \end{bmatrix}$, $D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$, $B = \begin{bmatrix} -t & 0 & 0 \\ 0 & x_2^1 & 1 \\ 2x_1 & 2x_2 & 0 \end{bmatrix}$. This leads to the sequence

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \text{ if } x_2 \neq 0$$

$$G_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -x_2 & \end{bmatrix}, G_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & x_2 + x_2^1 & 1 \\ 0 & 2x_2 & 0 \end{bmatrix} \text{ with } \det G_2 = -2x_2.$$

The DAE (8.17)–(8.19) has index 2 if $x_2 \neq 0$. We choose

$$C := \Pi_1 = P_0 P_1 = (I - Q_0)(I - Q_1) = \begin{bmatrix} 1 & & \\ & 0 & \\ & & 0 \end{bmatrix}.$$

While Q_1 depends on x_2 the projector Π_1 is constant which is advantageous for the choice of C. We establish the DAE with the initial condition

$$\Pi_1(x(t_0) - x^0) = 0.$$

$x_1(t_0) = x_1^0$ is given by the initial condition, from (8.19) we obtain $x_2(t_0)$, and we compute $x_1'(t_0) = x_1^0\, t_0 + 1$ from (8.17). But we need an additional equation to determine $x_3(t_0)$ and $x_2'(t_0)$. The projector $\mathcal{W}_1 = \operatorname{diag}(0,0,1)$ along $\operatorname{im} G_1$ tells us (cf. also Section 2.10.3) that we have to differentiate (8.19) and obtain, after replacing the derivatives in the point $t = t_0$,

$$x_3(t_0) + x_1(t_0)(x_1(t_0)\, t_0 + 1) = 0.$$

This equation describes the hidden constraint $\mathcal{H}(t_0)$, see Figure 8.1, and we have $x_3(t_0) = -x_1(t_0)(x_1(t_0)\, t_0 + 1)$ and also $x_2'(t_0)$ can be determined. It is obvious that $x(t_0) \in \mathcal{M}_1(t_0)$ as discussed also in Example 3.8. □

8.3 Numerical integration

Here, we discuss the direct integration of higher index DAEs without performing any preliminary index reduction steps. In contrast, the integration procedures for higher index DAEs as proposed in [50], [132] use derivative arrays and reduce the original DAE to DAEs of index 0 or index 1 previous to the integration steps.

IVPs resulting from higher index DAEs above all are ill-posed problems (see Example 1.5 and Theorem 3.66), and hence, in view of numerical integration, we have to look out for serious difficulties. In particular, it may well happen that an integration code seemingly works; however, it generates wrong results. For this reason, tools for monitoring the DAE structure would be very useful.

Essentially, we demonstrate various troubles associated with a direct integration of higher index DAEs. The difficulties are mostly method independent but problem dependent. This motivates us to restrict our demonstrations to BDF methods. Analyzing other methods in the same manner will show the same trouble. First we demonstrate by a simple example which is a slight generalization of a linear Hessenberg size-3 DAE, that even the direct numerical integration of linear index-3 DAEs with variable coefficients is somewhat hopeless; not only order reductions but also fatal error accumulations happen.

Example 8.5 (Index-3 example, [160]). Consider the DAE

$$\begin{bmatrix} 0 & 1 & 0 \\ 0 & t\eta & 1 \\ 0 & 0 & 0 \end{bmatrix} x'(t) + \begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta + 1 & 0 \\ 0 & t\eta & 1 \end{bmatrix} x(t) = q(t) \tag{8.20}$$

with a right-hand side q such that the solution is given by

$$x_1(t) = \mathrm{e}^{-t} \sin t, \quad x_2(t) = \mathrm{e}^{-2t} \sin t, \quad x_3(t) = \mathrm{e}^{-t} \cos t.$$

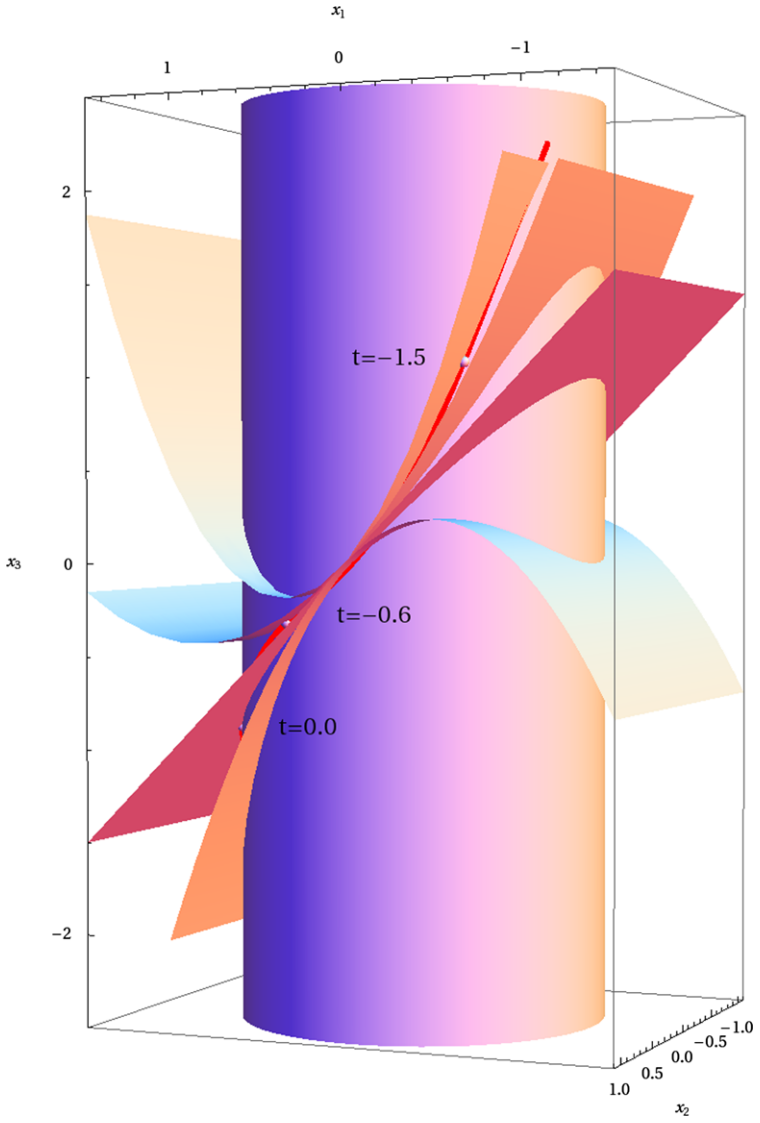


Fig. 8.1 Obvious and hidden constraint of Example 8.4 with $x_1(0) = 0.9$

The leading coefficient matrix in (8.20) has constant nullspace and constant image space and a properly stated representation is given by

$$\underbrace{\begin{bmatrix} 1 & 0 \\ \eta t & 1 \\ 0 & 0 \end{bmatrix}}_{A} \Big(\underbrace{\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{D} x(t) \Big)' + \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \eta+1 & 0 \\ 0 & t\eta & 1 \end{bmatrix}}_{B} x(t) = q(t). \tag{8.21}$$

An admissible matrix function sequence for (8.21) is given by

$$G_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & t\eta & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 1 & \\ & 0 \\ & & 0 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & t\eta & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & -t\eta & 0 \end{bmatrix}$$

$$(D\Pi_1 D^-)' = \begin{bmatrix} 0 & 0 \\ \eta & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & t\eta & 1 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1+t\eta & 1 \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q_2 = \begin{bmatrix} 0 & t\eta & 1 \\ 0 & -t\eta & -1 \\ 0 & t\eta(1+t\eta) & 1+t\eta \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1+t\eta & 1 \\ 0 & t\eta & 1 \end{bmatrix}, \quad \det G_3 = 1.$$

The DAE, in both versions, is regular with index 3 independent of η.

Here, the BDF applied to the standard form DAE (8.20) and the BDF applied to the properly formulated version (8.21) result in the same formulas.

Table 8.1, taken from [160], shows the error accumulation for different parameter values η. For the starting phase in each case consistent values are used. Except for the case $\eta = 0$ (constant coefficient DAE) the results are no longer acceptable. The results in Table 8.1 were rechecked with different methods in standard and properly stated form with consistent initial values but starting values computed as usual, but the results were worse than those presented in the table. Notice that the subspaces $N_1 = \ker G_1$, $N_2 = \ker G_2$, and $\operatorname{im} P_0 Q_1 Q_2$ move with time (cf. [158]). □

We now investigate linear index-μ DAEs with constant coefficients

$$Ex'(t) + Fx(t) = q(t). \tag{8.22}$$

Applying BDF methods of order $k \le 6$ with constant stepsize h yields

$$\frac{1}{h} \sum_{l=0}^{k} \alpha_l E x_{n-l} + F x_n = q_n - \delta_n \tag{8.23}$$

with $q_n := q(t_n)$ and δ_n summarizing the rounding errors and possible defects when solving (8.23) by a linear iterative solver. We introduce the local discretization error τ_n in t_n as

$$\frac{1}{h} \sum_{l=0}^{k} \alpha_l E x(t_{n-l}) + F x(t_n) - q_n =: \tau_n. \tag{8.24}$$

	$\eta = -0.5$		$\eta = 0$		$\eta = 2.0$	
h	P_0x	Q_0x	P_0x	Q_0x	P_0x	Q_0x
BDF-2						
2.5e–02	3e–02	3e+00	3e–04	4e–03	1e-04	5e-02
3.1e–03	3e+07	2e+10	5e–06	6e–05	1e–05	2e–05
7.8e–04	1e+43	3e+46	3e–07	4e–06	1e–06	2e–06
3.9e–04	-	-	8e–08	1e–06	2e–07	5e–07
1.9e–04	-	-	2e–08	2e–07	6e–08	1e–07
9.7e–05	-	-	5e–09	9e–09	1e–08	3e–08
BDF-3						
2.5e–02	-	-	1e–05	2e–02	1e–03	8e–02
3.1e–03	-	-	3e–08	1e–07	3e–04	1e–02
7.8e–04	-	-	4e–10	2e–09	1e–01	1e+02
3.9e–04	-	-	5e–11	1e–08	3e+03	4e+06
1.9e–04	-	-	5e–12	3e–08	6e+12	2e+16
9.7e–05	-	-	9e–13	1e–07	1e+32	1e+36
BDF-6						
3.1e–03	-	-	2e–13	3e–09	-	-
7.8e–04	-	-	5e–13	2e–08	-	-
3.9e–04	-	-	4e–12	3e–08	-	-
1.9e–04	-	-	2e–12	2e–07	-	-
9.7e–05	-	-	4e–12	4e–06	-	-

Table 8.1 Error of the BDF solution of (8.20) for different stepsizes

The difference of (8.24) and (8.23) results in

$$\frac{1}{h}\sum_{l=0}^{k}\alpha_l E(x(t_{n-l})-x_{n-l})+F(x(t_n)-x_n)=\tau_n+\delta_n. \tag{8.25}$$

Performing a complete decoupling of (8.25) as presented in Section 1.2.3 we obtain, as a version for (8.25), the decoupled system

$$\left[\begin{array}{c|ccc} I & & & \\ \hline 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\mu-1} \\ & \ddots & \ddots & \vdots \\ & & \ddots & \mathcal{N}_{\mu-2,\mu-1} \\ & & & 0 \end{array}\right] \left[\begin{array}{c} \frac{1}{h}\sum_{l=0}^{k}\alpha_l(u(t_{n-l})-u_{n-l}) \\ \hline 0 \\ \frac{1}{h}\sum_{l=0}^{k}\alpha_l(v_1(t_{n-l})-v_{1,n-l}) \\ \vdots \\ \frac{1}{h}\sum_{l=0}^{k}\alpha_l(v_{\mu-1}(t_{n-l})-v_{\mu-1,n-l}) \end{array}\right] \tag{8.26}$$

$$+\left[\begin{array}{c|ccc} \mathcal{W} & & & \\ \hline 0 & I & & \\ \vdots & & \ddots & \\ \vdots & & & \ddots \\ 0 & & & I \end{array}\right] \left[\begin{array}{c} u(t_n)-u_n \\ \hline v_0(t_n)-v_{0,n} \\ \vdots \\ \vdots \\ v_{\mu-1}(t_n)-v_{\mu-1,n} \end{array}\right] = \left[\begin{array}{c} \mathcal{L}_d \\ \mathcal{L}_0 \\ \vdots \\ \vdots \\ \mathcal{L}_{\mu-1} \end{array}\right] (\tau_n+\delta_n)$$

with

$$v_{0,n} := Q_0 x_n,\ \ v_{i,n} := \Pi_{i-1} Q_i x_n,\ \ i = 1, \ldots, \mu - 1,\ \ u_n := \Pi_{\mu-1} x_n,$$

and

$$x_n = v_{0,n} + v_{1,n} + \cdots + v_{\mu-1,n} + u_n.$$

Consider the stepwise integration on the compact interval $[t_0, T]$. For sufficiently small stepsizes $h > 0$, we find a constant $c > 0$ such that

$$\begin{aligned}
|u(t_n) - u_n| &\le c \max_{0\le l\le k-1} (|u(t_l) - u_l| + \max_{k\le l\le n} |\tau_l + \delta_l|), \text{ for } n \ge k \\
|v_{\mu-1}(t_n) - v_{\mu-1,n}| &\le c|\tau_n + \delta_n|, \text{ for } n \ge k \\
|v_{\mu-2}(t_n) - v_{\mu-2,n}| &\le c\frac{1}{h} \max_{n-k\le l\le n} |\tau_l + \delta_l|, \text{ for } n \ge 2k \\
&\vdots \\
|v_0(t_n) - v_{0,n}| &\le c\frac{1}{h^{\mu-1}} \max_{n-(\mu-1)k\le l\le n} |\tau_l + \delta_l|, \text{ for } n \ge \mu k.
\end{aligned}$$

We can conclude the following proposition.

Proposition 8.6. *The BDF (8.23) applied to the regular index-μ DAE (8.22) on $[t_0, T]$ generates values x_n, $\mu k \le n \le \frac{T-t_0}{h}$, which satisfy*

$$|x(t_n) - x_n| \le C\Big(\max_{0\le l\le k-1} |x(t_l) - x_l| + \max_{k\le l\le n} |\tau_l + \delta_l| + \sum_{i=1}^{\mu-1} \frac{1}{h^i} \max_{n-ik\le l\le n} |\tau_l + \delta_l| \Big)$$

with a constant $C > 0$, supposing h is sufficiently small.
If all errors δ_l vanish and the starting values are exact $x(t_l) = x_l$, $l = 0, \ldots, k-1$, then the estimation

$$|x(t_n) - x_n| \le C\Big(\max_{k\le l\le n} |\tau_l| + \sum_{i=1}^{\mu-2} \frac{1}{h^i} \max_{n-ik\le l\le n} |\tau_l| \Big)$$

becomes valid.

Proof. It remains to verify the second estimation. First, we note that $\mathcal{L}_{\mu-1}\tau_n = 0$ because

$$\tau_n = \frac{1}{h}\sum_{l=0}^{k} \alpha_l E x(t_{n-l}) + F x(t_n) - q_n = \frac{1}{h}\sum_{l=0}^{k} \alpha_l E x(t_{n-l}) - E x'(t_n) \in \operatorname{im} E.$$

□

The amplifying factors $\frac{1}{h^i}$ are caused by the differentiations involved in higher index problems. In the worst case, parts of the defects δ_l are amplified by $\frac{1}{h^{\mu-1}}$ for index-μ DAEs.

The estimations of Proposition 8.6 are somewhat coarse; however they remain valid in the case of variable stepsize BDFs, if the ratio of the adjacent steps is kept bounded and the stability for explicit ODEs is preserved. Then, neglecting the errors δ_l and supposing appropriate starting values one obtains approximations

$$x_n = x(t_n) + O(h_{max}^q), \quad n \geq \mu k,$$

with the order $q := \min\{k, k - \mu + 2\}$. This confirms well-known results, see [25, p. 45]. In particular, the numerical integration of an index-3 DAE by the implicit Euler method may lead to $\mathcal{O}(1)$ errors!

Example 8.7 (Variable stepsize integration problem). We consider the simplest index-3 DAE

$$\begin{aligned} x_1' - x_2 &= 0, \\ x_2' - x_3 &= 0, \\ x_1 &= g(t), \end{aligned}$$

which has the only solution

$$x_1(t) = g(t), \quad x_2(t) = g'(t), \quad x_3(t) = g''(t).$$

The implicit Euler method (with dropped errors δ_j) yields, after three integration steps,

$$\begin{aligned} x_{1,n} &= g(t_n), \\ x_{2,n} &= \frac{1}{h_n}(g(t_n) - g(t_{n-1})), \\ x_{3,n} &= \frac{1}{h_n}\left(\frac{1}{h_n}(g(t_n) - g(t_{n-1})) - \frac{1}{h_{n-1}}(g(t_{n-1}) - g(t_{n-2}))\right). \end{aligned}$$

If $h_n = h_{n-1}$, the solution component $x_{3,n}$ converges to $g''(t_n)$ with an error of $\mathcal{O}(h_n)$, but it blows up as $h_n \to 0$ and h_{n-1} fixed. For instance, if $h_{n-1} = 2h_n$ then the resulting $x_{3,n} = \frac{3}{2}g''(t_n) + \mathcal{O}(h_n)$ fails to approximate $x_3(t_n) = g''(t_n)$.

This phenomenon was already described in [25, p. 57]. It is closely related to the fact that the implicit Euler method is not suitable to start an integration of DAEs with index $\mu \geq 3$. Namely, even with a consistent initial value,

$$x_{1,0} = g(t_0), \quad x_{2,0} = g'(t_0), \quad x_{3,0} = g''(t_0),$$

one arrives at

$$
\begin{aligned}
x_{1,1} &= g(t_1),\\
x_{2,1} &= \frac{1}{h_1}(g(t_1)-g(t_0)),\\
x_{3,1} &= \frac{1}{h_1}\left(\frac{1}{h_1}(g(t_1)-g(t_0))-g'(t_0)\right)\\
&= \frac{1}{h_1^2}\left(g(t_0+h_1)-g(t_0)-h_1 g'(t_0)\right) = \frac{1}{2}g''(\theta),
\end{aligned}
$$

with a mean value θ. Obviously, $\frac{1}{2}g''(\theta)$ cannot be seen as an approximation of $g(t_1)$, even if g is a second-degree polynomial. □

The bad results in the previous example reflect that derivatives of order higher than 1 are not correctly approximated by BDF methods with variable stepsize.

On the other hand, Example 8.7 indicates order preservation in the case of constant stepsizes. In fact, again neglecting the errors δ_l we obtain from (8.26) that

$$
\begin{aligned}
v_{\mu-1}(t_n) - v_{\mu-1,n} &= \mathcal{L}_{\mu-1}\tau_n = 0, && n \geq k,\\
v_{\mu-2}(t_n) - v_{\mu-2,n} &= \mathcal{L}_{\mu-2}\tau_n, && n \geq 2k,\\
v_{\mu-3}(t_n) - v_{\mu-3,n} &= \mathcal{L}_{\mu-3}\tau_n - \mathcal{N}_{\mu-3,\mu-2}\frac{1}{h}\sum_{l=0}^{k}\alpha_l \mathcal{L}_{\mu-2}\tau_{n-l}, && n \geq 3k.
\end{aligned}
$$

Supposing a sufficiently smooth solution, we have

$$
\frac{1}{h}\sum_{l=0}^{k}\alpha_l \mathcal{L}_{\mu-2}\tau_{n-l} = \mathcal{O}(h^k)
$$

and hence $v_{\mu-3}(t_n) - v_{\mu-3,n} = \mathcal{O}(h^k)$. The further rigorous analysis of the recursion (8.26) and of the starting phase yields (cf. [25, Theorem 3.1.1])

$$
x(t_n) - x_n = \mathcal{O}(h^k),\ n \geq \mu k - (k-1).
$$

The last order result remains also valid in the case of nontrivial defects $\delta_l = \mathcal{O}(h^{k+\mu-1})$; however, we emphasize that neither the errors δ_l can be neglected or supposed to be smooth in practical computations—see the case $\eta = 0$ in Table 8.1—nor the starting steps can be skipped. Therefore, these order results are of limited meaning.

Next we investigate the direct numerical integration of linear index-2 DAEs. Already in [84], it has been shown that the BDF methods may fail completely for standard form index-2 DAEs

$$
E(t)x'(t) + F(t)x(t) = q(t),
$$

as the next example illustrates.

Example 8.8 (Standard form index-2 DAE with variable coefficients). The DAE from [84]

$$\begin{aligned} x_1' + \eta t x_2' + (1+\eta) x_2 &= 0, \\ x_1 + \eta t x_2 &= g(t) \end{aligned} \tag{8.27}$$

is regular with index 2 independent of the value of the parameter η. It has the unique solution

$$\begin{aligned} x_2 &= -g', \\ x_1 &= g - \eta t x_2. \end{aligned} \tag{8.28}$$

Figure 8.2 shows the numerical solution for $g = g(t) = \mathrm{e}^{-t}$ for different methods and parameter values. One can see that all the tested methods fail for certain pa-

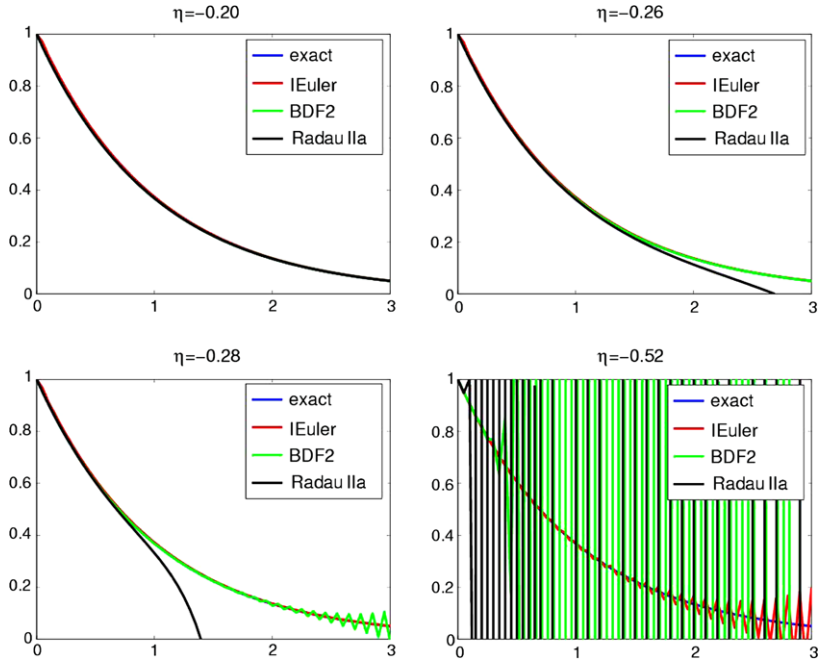


Fig. 8.2 The solutions (second component) of (8.27) for various parameter values η, the constant stepsize $h = 10^{-1.5}$ and the consistent initial value $x^0 = (1,1)^T$. The different curves represent the exact solution of the problem (exact) and the numerical solutions by the implicit Euler method (IEuler), by the two-step BDF method (BDF2) and by the two-stage RADAU IIA method (RADAU IIa).

rameter values. For instance, the BDF methods are no longer feasible for $\eta = -1$. They are feasible but unstable and nonconvergent for $\eta < -0.5$ (see [90]). This is really an alarming behavior since all the methods used (implicit Euler, two-step BDF, RADAU IIA) are approved for explicit ODEs and index-1 DAEs. □

However, the situation becomes better when reformulating the previous example for a DAE with a properly stated leading term

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t). \tag{8.29}$$

The BDF now applied to (8.29) yields the recursion

$$A(t_n)\frac{1}{h}\sum_{l=0}^{k}\alpha_l D(t_{n-l})x_{n-l} + B(t_n)x_n = q(t_n) - \delta_n. \tag{8.30}$$

Again, the term δ_n represents rounding errors and defects. Fortunately, BDF methods and IRK(DAE) methods (see Chapter 5) work quite well for index-2 DAEs with properly stated leading term.

Example 8.9 (Properly stated index-2 DAE with variable coefficients). The index-2 DAE

$$\begin{aligned} x_1' + (\eta t x_2)' + x_2 &= 0, \\ x_1 + \eta t x_2 &= g(t), \end{aligned} \tag{8.31}$$

or in compact form

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} \left(\begin{bmatrix} 1 & \eta t \end{bmatrix} x(t) \right)' + \begin{bmatrix} 0 & 1 \\ 1 & \eta t \end{bmatrix} x(t) = \begin{bmatrix} 0 \\ g(t) \end{bmatrix},$$

has, obviously, the same unique solution (8.28) as in Example 8.8. Figure 8.3 shows the numerical approximation. In contrast to the integration of the problem in standard formulation, one can see that now all tested methods work well for all parameter values. □

To formulate a convergence result for general linear index-2 DAEs, we introduce the local discretization error

$$\begin{aligned} \tau_n &:= A(t_n)\frac{1}{h}\sum_{l=0}^{k}\alpha_l D(t_{n-l})x(t_{n-l}) + B(t_n)x(t_n) - q(t_n) \\ &= A(t_n)\left\{ \frac{1}{h}\sum_{l=0}^{k}\alpha_l D(t_{n-l})x(t_{n-l}) - (Dx)'(t_n) \right\}, \; n \geq k, \end{aligned}$$

and set for the starting phase

$$\delta_l := G_2(t_l)(x_l - x(t_l)), \quad l = 0, \ldots, k-1.$$

Proposition 8.10. *Let the DAE (8.29) be regular with index 2 on $[t_0, T]$, then:*

(1) *For sufficiently small stepsizes $h > 0$, the BDF (8.30), $k \leq 6$, generates values x_n, $k \leq n \leq \frac{T-t_0}{h}$, which satisfy the estimation*

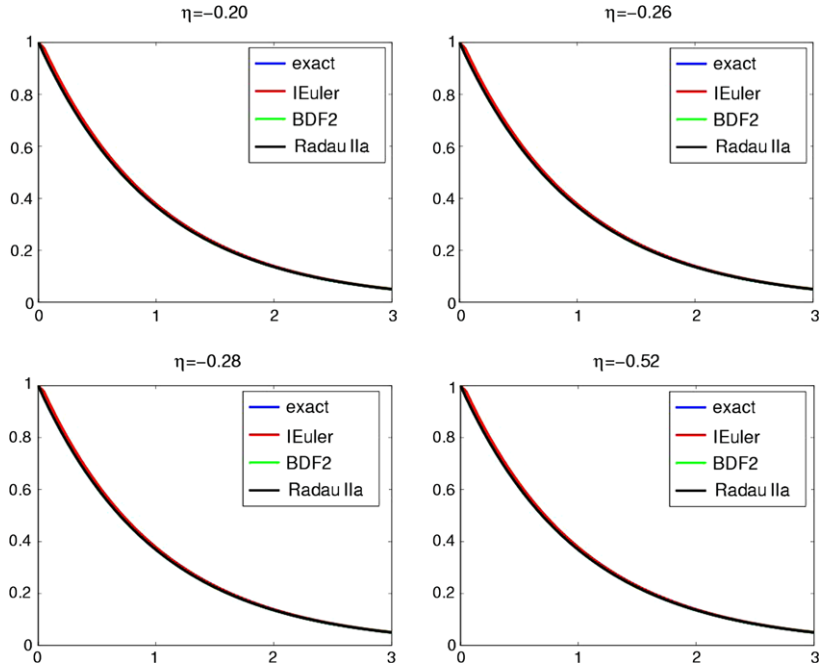


Fig. 8.3 The solutions (second component) of (8.31) for various parameter values η, the constant stepsize $h = 10^{-1.5}$ and the consistent initial value $x^0 = (1,1)^T$. The different curves represent the exact solution of the problem (exact) and the numerical solutions by the implicit Euler method (IEuler), by the two-step BDF method (BDF2) and by the two-stage RADAU IIA method (RADAU IIa).

$$|x(t_n) - x_n| \le C\Big\{ \max_{l=0,\dots,k-1} |D(t_l)(x(t_l) - x_l)| + \max_{l=k,\dots,n} |\tau_n + \delta_n| \qquad (8.32)$$
$$+ \max_{l=0,\dots,n} |\frac{1}{h}(DQ_1G_2^{-1})(t_n)\delta_n| \Big\}$$

with a constant C independent of the stepsize h.

(2) *If, additionally, the errors δ_n, $l \ge k$, vanish and the starting values are exact, $x_l = x(t_l)$, $l = 0,\dots,k-1$, then it follows that*

$$|x(t_n) - x_n| \le C\Big\{ \max_{l=k-1,\dots,n} |\tau_n| \Big\}.$$

Proof. Assertion (1) is obtained in [116] by simultaneous decoupling of the index-2 DAE (8.29) and the BDF recursion (8.30).
Assertion (2) is an immediate consequence of (1). □

The last proposition implies convergence of order k; however in practical computations parts of the errors are amplified by $\frac{1}{h}$. Owing to the index-2 structure and the linearity these errors are not propagated.

Similar results apply to IRK(DAE)s (see, e.g., [59]).

In Subsection 5.4,it is shown that one can benefit from a time-invariant $\operatorname{im} D(t)$. If a regular index-1 DAE is in numerically qualified form, in this sense, then the integration is as smooth as for explicit ODEs. This means that the given method arrives at the IERODE unchanged and there are no additional stepsize restrictions for stability reasons. An analogous situation can be observed in the index-2 case. Now the two subspaces $\operatorname{im} D\Pi_{can}$ and $\operatorname{im} D(I-\Pi_{can})$ have to be time-invariant to ensure that the integration method reaches the IERODE unchanged (see [116]). At this point we mention that the DAE (8.31) is in numerically qualified formulation. The associated canonical projector is $\Pi_{can}=0$, and $\operatorname{im} D\Pi_{can}=0$, $\operatorname{im} D(I-\Pi_{can})=\operatorname{im} D=\mathbb{R}$ are constant. This explains why the integration methods perform so well for the relatively large stepsize $h=10^{-1.5}$ (see Figure 8.3).

The following example illustrates the impact of time-varying subspaces. Note that the refactorization into a numerically qualified (index-2) DAE does no longer show those errors (see [116]).

Example 8.11 (Index-2 DAE with varying $D(I-\Pi_{can})$, [109]). Consider the Hessenberg index-2 DAE

$$\begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix} x'(t) + \begin{bmatrix} \lambda & -1 & -1 \\ \eta t(1-\eta t)-\eta & \lambda & -\eta t \\ 1-\eta t & 1 & 0 \end{bmatrix} x(t) = 0$$

where $\lambda, \eta \in \mathbb{R}$ are constant parameters. If $x_0 \in \mathbb{R}^3$ is a consistent initial value at $t=0$ (i.e., $x_1^0+x_2^0=0,\ x_3^0+x_2^0=0$), the solution of the DAE is

$$x_1(t) = x_1^0 e^{-\lambda t}, \qquad x_2(t) = (\eta t-1)x_1(t), \qquad x_3(t) = -x_2(t).$$

Taking the proper formulation with

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

and the admissible projector sequence

$$Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \qquad Q_1 = \begin{bmatrix} 1-\eta t & 1 & 0 \\ -\eta t(\eta t-1) & \eta t & 0 \\ 1-\eta t & 1 & 0 \end{bmatrix},$$

we obtain

$$\operatorname{im} DQ_1 = \operatorname{im} \begin{bmatrix} 1-\eta t & 1 & 0 \\ -\eta t(\eta t-1) & \eta t & 0 \end{bmatrix}$$

to be dependent on t for $\eta \neq 0$. Notice that

$$\operatorname{im} DQ_1 = DN_1 = D(N_0 \oplus N_1) = \operatorname{im} D(I - \Pi_{can}).$$

For different values of η, Figure 8.4 shows the first component of the numerical solutions calculated using the different integration methods.

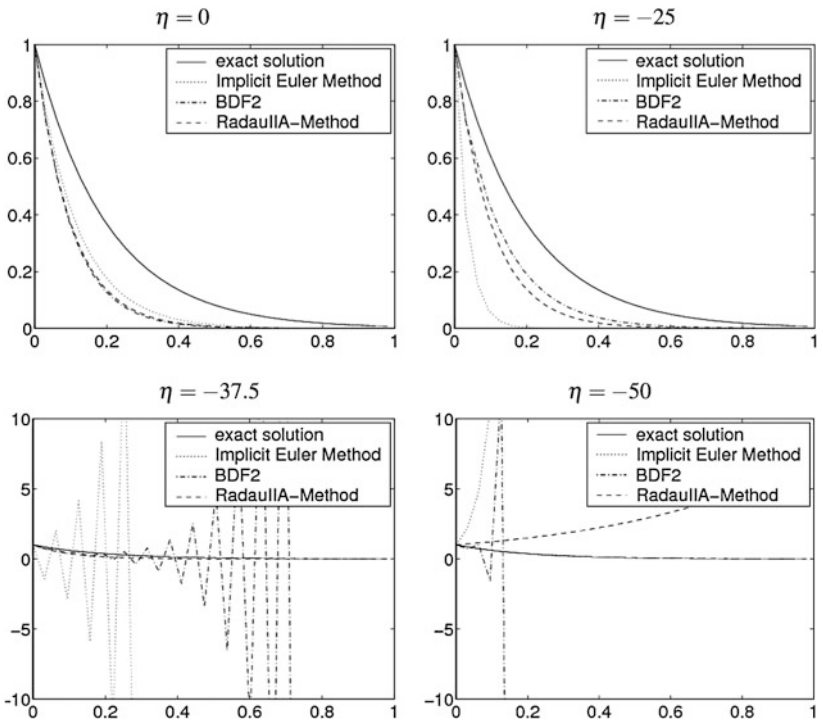


Fig. 8.4 Numerical solutions (first component) for $\lambda = 10$ and $h = 10^{-1.5}$ and the consistent initial value $x^0 = (1, -1, 1)^T$

8.4 Notes and references

(1) The integration of higher index DAEs—i.e., the approximate solution of an ill-posed problem (cf. Theorem 3.66)— leads in general to unsatisfactory results. Therefore the formulation of a DAE model should, if ever possible, result in an at most index-1 DAE.

(2) A large number of papers have investigated the behavior of ODE methods directly applied to DAEs with an index greater than 1, beginning with the monographs

[25] and [103]. Higher index DAEs mostly are restricted to autonomous Hessenberg form DAEs of size 2 or 3.
If the DAE has higher index but a very special structure, suitable methods can be developed. This is especially the case if the relevant DAE components are separated, as it is the case for Hessenberg form DAEs. We mention here [103], [24], [23],

(3) The DAEs of rigid body mechanics have index 3. A huge number of particular methods to compute consistent initial values and to solve IVPs and BVPs have been developed. Often index reduced DAEs of index 0 or 1 are finally solved.
For a comprehensive overview we refer to [119] and [63].

(4) Example 8.4 emphasizes that the computation of consistent initial values, in the higher index case, additionally needs information about the hidden constraints. Compared with (4.30), (4.31) for the index-1 case, the extension to index-2 DAEs comprises an additional equation which contains information about the hidden constraint.
This idea was realized for index-2 DAEs in standard formulation in [68] and [69], for DAEs with properly stated leading term in [137] and [14], and for special structured large dimensional DAEs in [108]. The index-3 case is discussed in [142]. The necessary differentiations are realized numerically by generalized backward differentiation formulas (GBDFs) on nonuniform meshes.

(5) Proposition 8.10 is slightly generalized for linear index-2 DAEs with harmless critical points and possible index changes in [59]. In this paper, one finds an elaborate description of the corresponding decoupling.

(6) In the case of higher index constant coefficient DAEs the errors $\frac{1}{h^i}\delta_l$ are local; they are not propagated. This situation is also given in the case of linear variable coefficient index-2 DAEs with properly stated leading term.

Unfortunately, already in the case of linear index-3 DAEs those bad error terms can be propagated, see Example 8.20. No doubt, in the case of nonlinear DAEs, the situation is even worse.

Only for quite special classes of nonlinear index-2 DAEs we can handle the error term $\frac{1}{h^i}\delta_l$. To our knowledge, the class of index-2 DAEs of the forms

$$\begin{aligned} A(t)x'(t) + b(x(t),t) &= 0 \quad \text{and} \\ A(t)(D(t)x(t))' + b(x(t),t) &= 0 \end{aligned}$$

are the richest ones for which respective proofs are given, see [205], respectively [211]. In both cases certain structural conditions concerning subspaces are supposed.

We refer to the Theorem 3.56 concerning local existence for slightly more general structural conditions.

Part IV
Advanced topics

Part IV of this monograph continues the hierarchical projector based approach to DAEs, which is discussed in Part I, in view of three different aspects. We consider quasi-regular DAEs, nonregular DAEs, and ADAEs (abstract DAEs) in Hilbert spaces. An additional chapter conveys results obtained by the projector based analysis concerning minimization problems with DAE constraints.

The chapter on minimization starts with a discussion of adjoint and self-adjoint DAEs. It contains necessary and sufficient extremal conditions in terms of the original data. Special attention is directed to properties of the optimality DAE as the basis for indirect optimization methods. Further, an appropriate generalization of the Riccati feedback for LQPs is given.

For quasi-regular DAEs, we relax the constant-rank condition supporting the admissible matrix functions and the regularity. If, due to rank changes in a matrix function, a continuous nullspace no longer exists, we use instead a continuous subnullspace. In this way we figure out quasi-regularity. Linear DAEs that are transformable into standard canonical form are quasi-regular. However, the characteristic values characterizing regularity in Part I now lose their meaning and quasi-regularity appears to be somewhat diffuse—similarly to the differentiation index approach.

Nonregular DAEs may comprise a different number of equations and components of the unknown function. Discussing mainly linear DAEs, we emphasize the scope of possible different interpretations. We generalize the tractability index as well as the decouplings to apply also to those equations.

The concept of regular DAEs so far applied to DAEs in finite-dimensional spaces is then, in the chapter on ADAEs, generalized for equations

$$A(t)\frac{\mathrm{d}}{\mathrm{d}t}d(x(t),t)+b(x(t),t)=0,$$

with operators acting in Hilbert spaces. After having briefly discussed various special cases we turn to a class of linear ADAEs which covers parabolic PDEs and index-1 DAEs as well as couplings thereof. We treat this class in detail by means of Galerkin methods yielding solvability of the ADAE.

Chapter 9
Quasi-regular DAEs

The regularity notion in Part I is supported by several constant-rank conditions. In this chapter we relax these rank conditions and allow for certain rank changes. If a matrix function changes the rank, its nullspace is no longer continuous, and also the projector functions onto the nullspace fail to be continuous. Replacing the nullspaces $\ker G_i$ that are no longer continuous by continuous subnullspaces $N_i \subseteq \ker G_i$ we again obtain continuous matrix functions sequences. These modified sequences inherit most of the properties given for the standard sequences in Chapters 1, 2, and 3. This allows us to cover a large class of equations with somewhat harmless index changes. In particular, linear DAEs that are transferable into so-called *standard canonical form* (cf. [39]) are proved to be quasi-regular, whereas regular linear DAEs comprise DAEs transferable into *strong standard canonical form*. However, the price to be paid for this is the loss of useful information concerning local characteristic values and index. From this point of view, quasi-regular DAEs show a certain affinity to DAEs having a well-defined differentiation index.
The chapter is organized as follows. After collecting basics in Section 9.1 and describing quasi-admissible projector functions in Section 9.2, we introduce quasi-regularity in Section 9.3 and show relations to linearizations in Section 9.4. In Section 9.5 we prove quasi-regularity for all linear DAEs that are transferable into SCF. A general decoupling procedure applied to linear quasi-regular DAEs is offered in Section 9.6. In Section 9.7 we touch on difficulties arising with the use of subnullspaces.

9.1 Quasi-proper leading terms

We deal with equations of the form

$$f((d(x(t),t))',x(t),t) = 0, \tag{9.1}$$

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_9,

where $f(y,x,t) \in \mathbb{R}^m$, $d(x,t) \in \mathbb{R}^n$, for $y \in \mathbb{R}^n$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$, and $\mathcal{D}_f \subseteq \mathbb{R}^m$ is open, $\mathcal{I}_f \subseteq \mathbb{R}$ is an interval. The functions f and d are supposed to be continuous on their definition domains together with their partial derivatives f_y, f_x, d_x, d_t. This means that the general Assumption 3.1 is again satisfied. We restrict our interest to systems having an equal number of equations and unknowns, $m = k$.
As agreed upon in Chapter 3, Definition 3.2: A *solution x_* of equation (9.1)* is a continuous function x_* defined on an interval $\mathcal{I}_* \subseteq \mathcal{I}_f$, with values $x_*(t) \in \mathcal{D}_f, t \in \mathcal{I}_*$, such that the function $u_*(.) := d(x_*(.),.)$ is continuously differentiable, and x_* satisfies the DAE (3.1) pointwise on $\mathcal{I}_*$. For convenience, we repeat the definition of a DAE with quasi-proper leading term (cf. 3.80).

Definition 9.1. Equation (9.1) is said to be a DAE with *quasi-proper leading term* on $\mathcal{D}_f \times \mathcal{I}_f$, if $\operatorname{im} d_x$ is a $\mathcal{C}^1$-subspace, $\ker d_x$ is nontrivial and there exists a further $\mathcal{C}^1$-subspace N_A, possibly depending on y,x,t, such that the inclusion

$$N_A(y,x,t) \subseteq \ker f_y(y,x,t), \quad y \in \mathbb{R}^n,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{9.2}$$

and the transversality condition

$$N_A(y,x,t) \oplus \operatorname{im} d_x(x,t) = \mathbb{R}^n, \quad x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \tag{9.3}$$

are valid.

Whereas an attendant property of a properly stated leading term consists of the constant-rank condition for both matrix functions f_y and d_x, in the case of a quasi-proper leading term, for the matrix function f_y rank changes are allowed. In contrast, the subspace $\operatorname{im} d_x$, as a $\mathcal{C}^1$-subspace, always has constant dimension

$$r := \dim \operatorname{im} d_x(x,t), \quad x \in \mathcal{D}_f, t \in \mathcal{I}_f.$$

Then, the continuous matrix function d_x has constant rank r, and its nullspace $\ker d_x$ is a continuous subspace of dimension $m-r$. To ensure that $\ker d_x$ is nontrivial the inequality $m > r$ must be given. Notice that a function d with nonsingular Jacobian d_x, for instance $d(x,t) \equiv x$, is excluded here.
A good idea to create a DAE with quasi-proper leading term is to arrange things in such a way that d_x is rectangular and has full row rank $r = n$ with $n \leq m-1$. In this case, the trivial subspace $N_A = \{0\}$ satisfies both conditions (9.2), (9.3).
If a standard form DAE

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

is given, one often finds a singular square matrix D_{inc}, with entries being zeros or ones, such that $\mathfrak{f}(x^1,x,t) = \mathfrak{f}(D_{inc}x^1,x,t)$ holds for all arguments; further $\operatorname{im} D_{inc} \cap \ker D_{inc} = \{0\}$. By such an *incidence matrix*, the standard form DAE can be written as

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

and letting $N_A := \ker D_{inc}$ we attain a DAE with quasi-proper leading term.

Example 9.2 (Quasi-proper by an incidence matrix). The nonlinear system

$$\begin{aligned} \alpha(x_2(t),x_3(t),t)\,x_2'(t)+x_1(t)\;-\;q_1(t)&=0,\\ \beta(x_3(t),t)\,x_3'(t)+x_2(t)\;-\;q_2(t)&=0,\\ x_3(t)\;-\;q_3(t)&=0, \end{aligned}$$

with smooth functions $\alpha,\beta : \mathcal{D}\times\mathcal{I}\to\mathbb{R}$, which do not vanish identically on the definition domain, but which have zeros or vanish identically on subsets, cannot be written as a DAE with proper leading term. However, choosing $D_{inc} = \operatorname{diag}(0,1,1)$ we obtain a DAE with quasi-proper leading term. □

It is a typical phenomenon of DAEs with quasi-proper leading term—in contrast to DAEs with properly stated leading term—that there might be *natural local reformulations* on subdomains, which show a lower rank $\tilde{d}_x$ and a larger subspace $\tilde{N}_A$ or are even properly stated versions (cf. Examples 3.81, 3.76).

It may happen that there is no singular incidence matrix in the given standard form DAE, since seemingly all derivative components are involved. Nevertheless, a quasi-proper formulation might be found as the next example shows.

Example 9.3 (Quasi-proper DAE with nonlinear d). The system

$$\begin{aligned} x_1'(t)-x_3(t)\,x_2'(t)+x_2(t)\,x_3'(t)-q_1(t)&=0,\\ x_2(t)-q_2(t)&=0,\\ x_3(t)-q_3(t)&=0, \end{aligned}$$

has standard form $\mathfrak{f}(x'(t),x(t),t)=0$, with

$$\mathfrak{f}(x',x,t)=\begin{bmatrix}1&-x_3&x_2\\0&0&0\\0&0&0\end{bmatrix}x'+\begin{bmatrix}0\\x_2\\x_3\end{bmatrix}-q(t),$$

and $\ker \mathfrak{f}_{x'}$ depends on x. This example is the one created by Ch. Lubich in [103] to show that the differentiation index might be 1 while the perturbation index is 2. We rearrange this system as

$$\begin{aligned} (x_1(t)+x_2(t)\,x_3(t))'-2x_3(t)\,x_2'(t)-q_1(t)&=0,\\ x_2(t)-q_2(t)&=0,\\ x_3(t)-q_3(t)&=0, \end{aligned}$$

i.e., in the form (9.1) with $m=3$, $n=2$,

$$d(x,t)=\begin{bmatrix}x_1+x_2x_3\\x_2\end{bmatrix},\quad f(y,x,t)=\begin{bmatrix}y_1-2x_3y_2-q_1(t)\\x_2-q_2(t)\\x_3-q_3(t)\end{bmatrix}.$$

The partial Jacobian

$$d_x(x,t)=\begin{bmatrix}1&x_3&x_2\\0&1&0\end{bmatrix}$$

has full row rank n, as well as a one-dimensional nullspace. Choosing $N_A = \{0\}$ we obtain a quasi-proper leading term.
For every continuous q with a continuously differentiable component q_2, $\bar{t} \in \mathcal{I}$, $\bar{c} \in \mathbb{R}$, this DAE has the unique solution

$$x_1(t) = -q_2(t)q_3(t) + \bar{c} + q_2(\bar{t})q_3(\bar{t}) + \int_{\bar{t}}^{t} (q_1(s) + 2q_3(s)q_2'(s))ds,$$
$$x_2(t) = q_2(t),$$
$$x_3(t) = q_3(t), \quad t \in \mathcal{I}.$$

Later on, in Example 9.17, this DAE is shown to be quasi-regular with a nonsingular matrix function on level 2. □

In Section 2.9, by means of several examples, it is pointed out that a rank change of the coefficient $A(t)$ of a linear DAE

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \tag{9.4}$$

indicates a critical solution behavior. In contrast, in the context of quasi-proper leading terms those critical points are no longer indicated at this level, but they are hidden.

Example 9.4 (Hidden critical point). The system

$$x_1(t)x_1'(t) - x_2(t) = 0,$$
$$x_1(t) - x_2(t) = 0,$$

possesses the solutions $x_{*1}(t) = x_{*2}(t) = t + c$, where c denotes an arbitrary real constant. Additionally, the identically vanishing function $\bar{x}_*(t) = 0$ is also a solution. Through every point on the diagonal line $x_1 = x_2$, except for the origin, there passes exactly one solution. However, two different solutions satisfy the initial condition $x(0) = 0$, which characterizes the origin as a critical point. Writing the DAE in the form (9.1) with $n = 1$, $m = k = 2$, $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = \mathbb{R}$,

$$f(y,x,t) = \begin{bmatrix} x_1 y - x_2 \\ x_1 - x_2 \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}, \quad d(x,t) = x_1, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 \end{bmatrix},$$

one arrives at a DAE which fails to have a properly stated leading term. However, the leading term is quasi-proper, which emphasizes that the constant-rank condition (and the proper leading term setting) is helpful in indicating critical points. □

Example 9.5 (Hidden singular ODE). The system

$$t^k x_1'(t) + M x_2(t) = q_1(t), \quad x_1(t) + x_2(t) = q_2(t),$$

written as

$$\begin{bmatrix} t^k \\ 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 \end{bmatrix} x(t) \right)' + \begin{bmatrix} M & 0 \\ 1 & 1 \end{bmatrix} = q(t),$$

with $k > 0$, has a quasi-proper leading term on intervals $\ni 0$. However, no doubt, depending on k and M, the critical point $t_* = 0$ may indicate serious singularities concerning flow and solvability. □

The constant-rank requirement for f_y of properly stated leading terms rules out those singularities, but it also rules out certain harmless rank changes that are perceivable only when looking for rigorous low smoothness solvability. Those harmless critical points can be found in Example 9.2. Each zero of one of the functions α and β yields such a harmless critical points.

The *obvious constraint set* of a DAE with quasi-proper leading term possesses the same form as that for DAEs with properly involved derivative (see Definition 3.9), namely

$$\mathcal{M}_0(t) := \{x \in \mathcal{D}_f : \exists y \in \mathbb{R}^n : y - d_t(x,t) \in \operatorname{im} d_x(x,t),\ f(y,x,t) = 0\},\ t \in \mathcal{I}_f. \quad (9.5)$$

Whereas Proposition 3.10 guarantees the uniqueness of the associated y-values for given $x \in \mathcal{M}_0(t)$ for DAEs with properly involved derivatives, this fact is no longer true for DAEs with quasi-proper leading terms. Now there might be several values y corresponding to a fixed $x \in \mathcal{M}_0(t)$, $t \in \mathcal{I}_f$, as the following example shows.

Example 9.6 (Nonuniqueness of y). The system

$$\begin{aligned} x_1'(t)^2 - x_2(t)^2 - \gamma(t) &= 0, \\ x_2(t) - \varphi(x_1(t),t) &= 0, \end{aligned}$$

has the form (9.1) with $n = 1$, $m = k = 2$, $\mathcal{D}_f = \mathbb{R}^2$, $\mathcal{I}_f = \mathbb{R}$,

$$f(y,x,t) = \begin{bmatrix} y^2 - x_2^2 - \gamma(t) \\ x_2 - \varphi(x_1,t) \end{bmatrix}, \quad f_y(y,x,t) = \begin{bmatrix} 2y \\ 0 \end{bmatrix}, \quad d(x,t) = x_1, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 \end{bmatrix}.$$

This DAE has a quasi-proper leading term, and a full row rank matrix function d_x on the definition domain. The obvious constraint is given by

$$\mathcal{M}_0(t) = \{x \in \mathbb{R}^2 : x_2 - \varphi(x_1,t) = 0,\ y^2 = x_2^2 + \gamma(t)\}.$$

To each $x \in \mathcal{M}_0(t)$, with $x_2^2 + \gamma(t) > 0$, there are associated two values $y = \pm\sqrt{x_2^2 + \gamma(t)}$.
Observe that the matrix function f_y has a rank drop at $y = 0$. Again, this might be interpreted as a critical point. □

With the structural characterization of quasi-regular DAEs in mind we construct a matrix function sequence and accompanying projector functions in a quite similar way as in Chapter 3. For this aim we introduce once again the auxiliary coefficient functions

$$\begin{aligned} D(x,t) &:= d_x(x,t), \\ A(x^1,x,t) &:= f_y(D(x,t)x^1 + d_t(x,t),x,t), \\ B(x^1,x,t) &:= f_x(D(x,t)x^1 + d_t(x,t),x,t), \quad \text{for } x^1 \in \mathbb{R}^m,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f, \end{aligned} \tag{9.6}$$

to be used throughout this chapter. These coefficients A, D and B are continuous. If the DAE has a quasi-proper leading term, then conditions (9.2) and (9.3) imply the inclusion

$$N_A\big(D(x,t)x^1 + d_t(x,t), x, t\big) \subseteq \ker A(x^1,x,t),$$

and the decomposition

$$N_A\big(D(x,t)x^1 + d_t(x,t), x, t\big) \oplus \operatorname{im} D(x,t) = \mathbb{R}^n, \quad \text{for } x^1 \in \mathbb{R}^m,\ x \in \mathcal{D}_f,\ t \in \mathcal{I}_f.$$

We introduce the projector valued function R pointwise by

$$\operatorname{im} R(x^1,x,t) = \operatorname{im} D(x,t), \quad \ker R(x^1,x,t) = N_A\big(D(x,t)x^1 + d_t(x,t), x, t\big). \tag{9.7}$$

R is continuous by construction. If $N_A(y,x,t)$ is independent of y, then $R(x^1,x,t) = R(0,x,t)$ is independent of x^1, and R is a continuously differentiable projector function in (x,t).
If $N_A(y,x,t)$ varies with y, we enforce the projector function R to be continuously differentiable by additionally supposing continuous second partial derivatives for d. We summarize the basic assumptions for later use.

Assumption 9.7. (Basic assumption for (9.1))

(a) *The function f is continuous on $\mathbb{R}^n \times \mathcal{D}_f \times \mathcal{I}_f$ together with its first partial derivatives f_y, f_x. The function d is continuously differentiable on $\mathcal{D}_f \times \mathcal{I}_f$.*
(b) *The DAE (9.1) has a quasi-proper leading term.*
(c) *If $N_A(y,x,t)$ depends on y, then d is supposed to have continuous second partial derivatives.*

Assumption 9.7 is the counterpart of Assumption 3.16 in Part I.

9.2 Quasi-admissible matrix function sequences and admissible projector functions

We suppose the DAE (9.1) satisfies Assumption 9.7. Aiming for a projector based structural analysis we proceed quite similarly as in Part I. The following construction of matrix function sequences and involved projector functions is close to the construction for regular DAEs. The only difference is that *now we deal with subspaces N_i* of the nullspaces $\ker G_i$ instead of these nullspaces themselves. Denote

$$N_0(x,t) := \ker D(x,t), \quad \text{for } x \in \mathcal{D}_f,\ t \in \mathcal{I}_f,$$

and introduce pointwise projectors $Q_0(x,t), P_0(x,t) \in L(\mathbb{R}^m)$, such that

$$Q_0(x,t)^2 = Q_0(x,t), \quad \operatorname{im} Q_0(x,t) = N_0(x,t), \quad P_0(x,t) = I - Q_0(x,t).$$

Since $D(x,t)$ has constant rank r, its nullspace has constant dimension $m-r$. This allows for choosing Q_0, P_0 to be continuous, and we do so.
We emphasize again that, while for properly stated leading terms the nullspaces $\ker AD$ and $\ker D = N_0$ coincide, N_0 is now just a *subspace* of $\ker AD$. In contrast to the case of proper leading terms, the intersection $\ker A \cap \operatorname{im} D$ might be nontrivial. Actually, $\ker AD$ is not necessarily a continuous subspace in the present chapter.

Next we provide the generalized inverse $D(x^1,x,t)^- \in L(\mathbb{R}^n, \mathbb{R}^m)$ by means of the four conditions

$$\begin{aligned} D(x,t)D(x^1,x,t)^-D(x,t) &= D(x,t), \\ D(x^1,x,t)^-D(x,t)D(x^1,x,t)^- &= D(x^1,x,t)^-, \\ D(x,t)D(x^1,x,t)^- &= R(x^1,x,t), \\ D(x^1,x,t)^-D(x,t) &= P_0(x,t), \end{aligned} \tag{9.8}$$

for $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$. $D(x^1,x,t)^-$ is uniquely determined by (9.8), and it represents a continuous function (Proposition A.17).
Owing to the quasi-proper leading term, one has available the identities

$$\begin{aligned} A(x^1,x,t) &= A(x^1,x,t)R(x^1,x,t), \\ \operatorname{im} A(x^1,x,t) &= \operatorname{im} A(x^1,x,t)D(x^1,x,t), \end{aligned}$$

which we already know from proper leading terms. Namely, the first identity is a simple consequence of (9.2), $\ker A \supseteq N_A = \operatorname{im}(I-R)$, $A(I-R) = 0$. The second one follows from $\operatorname{im} A \subseteq \operatorname{im} AD = \operatorname{im} ADD^- = \operatorname{im} AR = \operatorname{im} A$.

We compose matrix function sequences for the DAE (9.1) starting from

$$G_0 := AD, \quad B_0 := B, \quad \Pi_0 := P_0, \tag{9.9}$$

which are defined pointwise, for all $x^1 \in \mathbb{R}^m$, $x \in \mathcal{D}_f$, $t \in \mathcal{I}_f$.
Recall that $N_0 = \ker D$ is a continuous subnullspace of G_0. We build, for $i \geq 0$, as long as the expressions exist,

$$G_{i+1} := G_i + B_i Q_i, \tag{9.10}$$

then choose a continuous subspace N_{i+1},

$$N_{i+1} \subseteq \ker G_{i+1}, \tag{9.11}$$

as well as projectors P_{i+1}, Q_{i+1} such that $P_{i+1} := I - Q_{i+1}$, $\operatorname{im} Q_{i+1} = N_{i+1}$, and put

$$\Pi_{i+1} := \Pi_i P_{i+1}, \tag{9.12}$$

$$B_{i+1} := B_i P_i - G_{i+1} D^- (D\Pi_{i+1} D^-)' D\Pi_i. \tag{9.13}$$

Denote $r_i := \operatorname{rank} G_i$, $i \geq 0$.

Again, all these definitions are meant pointwise for all arguments. $(D\Pi_{i+1}D^-)'$ stands for the total derivative with respect to the jet variables.
As explained in Section 3.2, at each level i, a new jet variable x^i comes in. While G_1 has the arguments x^1, x, t, G_2 depends on x^2, x^1, x, t, and so on.
The matrix function sequence (9.9)–(9.13) looks formally like those introduced in Subsection 2.2.2 and Section 3.2, with the only difference that now, N_{i+1} is possibly a continuous *subspace* of $\ker G_{i+1}$ and also $N_0 = \ker D$ belongs to $\ker G_0$ but does not necessarily coincide.
Working with continuous subspaces of the corresponding nullspaces offers the possibility to dispense with continuous nullspaces and to allow for rank changes in the matrix functions G_j.

By construction, it holds that

$$G_{i+1} P_i = G_i P_i = G_i, \quad G_{i+1} Q_i = B_i Q_i,$$

hence,

$$\operatorname{im} G_0 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq \operatorname{im} G_{i+1},$$

$$r_0 \leq \cdots \leq r_i \leq r_{i+1}.$$

As previously, we try for continuous matrix functions G_i. In contrast to Section 3.2, G_i *may now change its rank*, and r_i is an integer valued function. Owing to the arbitrariness of possible subnullspaces (cf. Section 9.7), the rank values r_i lose their meaning of characteristic values for the DAE.

To show further properties we use projector functions $\mathcal{W}_i : \mathcal{I} \to L(\mathbb{R}^m))$ and generalized inverses $G_i^- : \mathcal{I} \to L(\mathbb{R}^m))$ of G_i with

$$G_i G_i^- G_i = G_i, \quad G_i^- G_i G_i^- = G_i^-, \quad G_i G_i^- = I - \mathcal{W}_i. \tag{9.14}$$

$\mathcal{W}_i$ projects along $\operatorname{im} G_i$. In contrast to Section 3.2, since now G_i may change its rank, both G_i^- and $\mathcal{W}_i$ are no longer necessarily continuous.
The subspace N_i is part of the nullspace $\ker G_i$, and we can decompose

$$\ker G_i = N_i \oplus (\ker G_i \cap \operatorname{im} P_i),$$

and choose the generalized inverse G_i^- such that

$$\operatorname{im} G_i^- G_i \subseteq \operatorname{im} P_i. \tag{9.15}$$

Remember that, in contrast, if N_i coincides with $\ker G_i$, then also $\operatorname{im} P_i$ and $\operatorname{im} G_i^- G_i$ coincide.
Many of the properties given in Propositions 2.5 and 3.20 for standard sequences are maintained by now.

Proposition 9.8. *For the matrix function sequence (9.9)–(9.13) the following relations are satisfied:*

(1) $\ker \Pi_i \subseteq \ker B_{i+1}$,
(2) $\mathcal{W}_{i+1}B_{i+1} = \mathcal{W}_{i+1}B_i = \cdots = \mathcal{W}_{i+1}B_0 = \mathcal{W}_{i+1}B$, $\mathcal{W}_{i+1}B_{i+1} = \mathcal{W}_{i+1}B_0\Pi_i$,
(3) $\operatorname{im} G_{i+1} = \operatorname{im} G_i \oplus \operatorname{im} \mathcal{W}_i B Q_i$,
(4) $N_i \cap \ker B_i = N_i \cap \ker G_{i+1}$,
(5) $\ker G_i \cap \operatorname{im} P_i \subseteq \ker G_{i+1}$.

Proof. (1)–(3) We use the same arguments as in Proposition 2.5. Then $F_{i+1} := I + G_i^- B_i Q_i$ is nonsingular because of (9.15).
(4) From $z = Q_i z$ we conclude $G_{i+1}z = G_i z + B_i Q_i z = B_i z$, which implies the assertion.
(5) $G_i z = 0$, $Q_i z = 0$ yields $G_{i+1}z = G_i z + B_i Q_i z = 0$. □

We restrict the variety of projector functions in an analogous way as we did in Part I concerning standard sequences.

Definition 9.9. Suppose we are given a DAE (9.1) which satisfies Assumption 9.7. Let $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$ be open.
Let the projector function Q_0 onto N_0 be continuous. $P_0 = I - Q_0$. The generalized inverse D^- is given by $DD^-D = D$, $D^-DD^- = D^-$, $DD^- = R$, $D^-D = P_0$.
For a given level $\kappa \in \mathbb{N}$, we call the sequence $G_0, \ldots, G_\kappa$ a *quasi-admissible matrix function sequence* associated to the DAE on $\mathcal{G}$, if it is built by the rule

Set $G_0 := AD$, $B_0 := B$, $N_0 := \ker G_0$.
For $i \geq 1$:

$$G_i := G_{i-1} + B_{i-1}Q_{i-1},$$
$$B_i := B_{i-1}P_{i-1} - G_i D^-(D\Pi_i D^-)' D\Pi_{i-1}$$

fix a subspace $N_i \subseteq \ker G_i$, $\quad \widehat{N}_i := (N_0 + \cdots + N_{i-1}) \cap N_i$,

fix a complement X_i such that $N_0 + \cdots + N_{i-1} = \widehat{N}_i \oplus X_i$,
choose a projector Q_i such that $\operatorname{im} Q_i = N_i$ and $X_i \subseteq \ker Q_i$,
set $P_i = I - Q_i$, $\Pi_i = \Pi_{i-1}P_i$

and additionally

(a) the subspace N_i is continuous, $i = 1, \ldots, \kappa$,
(b) Π_i is continuous and $D\Pi_i D^-$ is continuously differentiable, $i = 1, \ldots, \kappa$.

The projector functions $Q_0, \ldots, Q_\kappa$ associated to a quasi-admissible matrix function sequence are said to be *quasi-admissible*, too.

With quasi-admissible projector functions $Q_0, \ldots, Q_\kappa$ we are sure to have continuous matrix functions $G_0, \ldots, G_\kappa$ and $G_{\kappa+1}$.

Definition 9.10. If quasi-admissible projector functions $Q_0,\dots,Q_\kappa$ are chosen such that

$$\ker Q_0 = N_0^\perp, \quad \ker Q_i = [N_0+\cdots+N_i]^\perp \oplus X_i, \quad X_i := [N_0+\cdots+N_{i-1}]\cap \widehat{N}_i^{\,\perp},$$

for $i=1,\dots,\kappa$, then we speak of *widely orthogonal quasi-admissible* projector functions.

Proposition 9.11. *If $Q_0,\dots,Q_\kappa$ are quasi-admissible projectors, then the following relations are valid*

(1) $\ker \Pi_i = N_0+\cdots+N_i,\ i=0,\dots,\kappa,$
(2) *the products $\Pi_i = P_0\cdots P_i$, $\Pi_{i-1}Q_i = P_0\cdots P_{i-1}Q_i$, $D\Pi_i D^- = DP_0\cdots P_iD^-$ and $D\Pi_{i-1}Q_iD^- = DP_0\cdots P_{i-1}Q_iD^-$ are projectors for $i=1,\dots,\kappa$,*
(3) $N_0+\cdots+N_{i-1}\subseteq \ker \Pi_{i-1}Q_i,\ i=1,\dots,\kappa,$
(4) $G_{i+1}Q_j = B_jQ_j,\ 0\le j\le i,\ i=1,\dots,\kappa,$
(5) $D(N_0+\cdots+N_i) = \operatorname{im} D\Pi_{i-1}Q_i \oplus \operatorname{im} D\Pi_{i-2}Q_{i-1}+\cdots+\operatorname{im} D\Pi_0Q_1,$ $i=1,\dots,\kappa,$
(6) $\widehat{N}_i \subseteq \ker G_j,\ 1\le i<j\le \kappa+1.$

Proof. (1)–(5) We use the same arguments as for Proposition 2.7.

(6) $z\in\widehat{N}_i$ means $z\in N_i$ and so $z=Q_iz$. Property (1) implies $\Pi_{i-1}z=0$. This allows us to conclude that $G_{i+1}z = G_iz+B_iQ_iz=0$. For $s=2,\dots,\kappa+1-i$, we have

$$G_{i+s}z = G_{i+s-1}z+B_{i+s-1}Q_{i+s-1}z = G_{i+s-1}z+B_{i+s-1}Q_{i+s-1}\Pi_{i-1+s-1}z = G_{i+s-1}z.$$

Therefore, we find $G_{i+2}z=0,\dots,G_{\kappa+1}=0$. □

We reconsider the above examples featuring different kinds of critical points and describe associated matrix functions.

Example 9.12 (Hidden critical point). The DAE from Example 9.4

$$\begin{aligned} x_1(t)x_1'(t)-x_2(t)&=0,\\ x_1(t)-x_2(t)&=0,\end{aligned}$$

yields

$$f_y(y,x,t)=\begin{bmatrix}x_1\\0\end{bmatrix},\quad d_x(x,t)=\begin{bmatrix}1&0\end{bmatrix},\quad G_0=\begin{bmatrix}x_1&0\\0&0\end{bmatrix},\quad f_x(y,x,t)=\begin{bmatrix}y&-1\\1&-1\end{bmatrix},$$

further

$$Q_0=\begin{bmatrix}0&0\\0&1\end{bmatrix},\quad G_1=\begin{bmatrix}x_1&-1\\0&-1\end{bmatrix}.$$

There is no nontrivial continuous subnullspace N_1, since the matrix function G_1 is nonsingular except for $x_1=0$. □

Example 9.13 (Nonuniqueness of y). For the DAE from Example 9.6,

$$\begin{aligned} x_1'(t)^2 - x_2(t)^2 - \gamma(t) &= 0, \\ x_2(t) - \varphi(x_1(t),t) &= 0, \end{aligned}$$

with

$$f_y(y,x,t) = \begin{bmatrix} 2y \\ 0 \end{bmatrix}, \quad d_x(x,t) = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad f_x(y,x,t) = \begin{bmatrix} 0 & -2x_2 \\ -\varphi_{x_1}(x_1,t) & 1 \end{bmatrix},$$

we find

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 2x_1^1 & 2x_2 \\ 0 & 1 \end{bmatrix}.$$

There is no nontrivial continuous subnullspace N_1 such that the sequence cannot be continued. The last matrix function is nonsingular for $x_1^1 \neq 0$ but it has a rank drop at $x_1^1 = 0$. □

Example 9.14 (Hidden singular ODE). For the DAE from Example 9.5

$$\begin{bmatrix} t^k \\ 0 \end{bmatrix} (\begin{bmatrix} 1 & 0 \end{bmatrix} x(t))' + \begin{bmatrix} M & 0 \\ 1 & 1 \end{bmatrix} = q(t),$$

we generate

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} t^k & 0 \\ 0 & 1 \end{bmatrix}.$$

Again there is no nontrivial continuous subnullspace N_1, since the matrix G_1 is nonsingular if $t \neq 0$. □

Example 9.15 (Harmless critical point). We revisit Example 2.70 from Section 2.9, which shows a harmless critical point. The DAE

$$\begin{bmatrix} 0 & \alpha(t) \\ 0 & 0 \end{bmatrix} (\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} x)' + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x = q, \tag{9.16}$$

has a quasi-proper leading term. We choose

$$Q_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \Pi_0 = P_0 = R = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and compute

$$G_0 = A = \begin{bmatrix} 0 & \alpha \\ 0 & 0 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & \alpha \\ 0 & 0 \end{bmatrix}, \quad N_1 = \{z \in \mathbb{R}^2 : z_1 + \alpha z_2 = 0\}.$$

Next, we choose

$$Q_1 = \begin{bmatrix} 0 & -\alpha \\ 0 & 1 \end{bmatrix}, \text{ thus, } \quad P_0 P_1 = \Pi_1 = 0, \quad B_1 = BP_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & \alpha \\ 0 & 1 \end{bmatrix}.$$

Observe that N_1 equals $\ker G_1$, and G_2 remains nonsingular on the entire interval independently of how the function α behaves. □

Observe that the quasi-admissible matrix function sequences do not at all indicate harmless critical points, but they indicate other critical points in later stages of the sequence.

9.3 Quasi-regularity

In general, a nonsingular matrix function G_κ indicates quasi-regularity by the following definition.

Definition 9.16. A DAE (9.1) which satisfies Assumption 9.7 is said to be *quasi-regular* on the open connected set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$, if there is a quasi-admissible matrix function sequence $G_0, \ldots, G_\kappa$ such that G_κ is nonsingular.
Then the set $\mathcal{G}$ is called a *quasi-regularity region*.
A point $(\bar{x}, \bar{t}) \in \mathcal{D}_f \times \mathcal{I}_f$ is named *quasi-regular* if there is a quasi-regularity region $\mathcal{G} \ni (\bar{x}, \bar{t})$.

Definition 9.16 generalizes Definition 3.28. In the same way one can generalize Definition 3.62 which allows for a localization of the jet variables, too. Clearly, if a DAE is regular, then it is also quasi-regular, but the opposite is not true. The class of quasi-regular DAEs is much wider.
The following example continues the discussion of Example 9.3 and shows a quasi-regular DAE involving a nonlinear function d.

Example 9.17 (Lubich's DAE, [103]). The DAE

$$\begin{aligned} (x_1(t) + x_2(t)x_3(t))' - 2x_3(t)x_2'(t) &= q_1(t), \\ x_2(t) &= q_2(t), \\ x_3(t) &= q_3(t), \end{aligned}$$

has the form (9.1) with $m = 3$, $n = 2$,

$$d(x,t) = \begin{bmatrix} x_1 + x_2 x_3 \\ x_2 \end{bmatrix}, \quad f(y,x,t) = \begin{bmatrix} y_1 - 2x_3 y_2 - q_1(t) \\ x_2 - q_2(t) \\ x_3 - q_3(t) \end{bmatrix}.$$

The matrix function

$$G_0 = AD = \begin{bmatrix} 1 & -x_3 & x_2 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

has the two-dimensional nullspace $\ker G_0 = \{z \in \mathbb{R}^3 : z_1 - x_3 z_2 + x_2 z_3 = 0\}$, while $N_0 = \ker D = \{z \in \mathbb{R}^3 : z_2 = 0, z_1 + x_2 z_3 = 0\}$ has dimension one only. With

$$Q_0 = \begin{bmatrix} 0 & 0 & -x_2 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad P_0 = \begin{bmatrix} 1 & 0 & x_2 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & -2x_2^1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we obtain

$$G_1 = \begin{bmatrix} 1 & -x_3 & x_2 - 2x_2^1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Choosing now $N_1 = \{z \in \mathbb{R}^3 : z_3 = 0,\ z_1 - x_3 z_2 = 0\}$ and

$$Q_1 = \begin{bmatrix} 0 & x_3 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \text{ we realize } Q_1 Q_0 = 0, \text{ and further } \mathcal{W}_1 B Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

such that G_2 must have constant rank $r_2 = 3$. Therefore, Lubich's DAE is quasi-regular with $r_2 = 3$. □

Example 9.18 (G_3 is nonsingular). We investigate the DAE from Example 9.2,

$$\begin{bmatrix} 0 & \alpha & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} x(t) \right)' + x(t) - q(t) = 0.$$

We derive

$$G_0 = \begin{bmatrix} 0 & \alpha & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 1 & \alpha_{x_2} x_2^1 & \alpha_{x_3} x_2^1 \\ 0 & 1 & \beta_{x_3} x_3^1 \\ 0 & 0 & 1 \end{bmatrix},$$

$$D^- = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad P_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & \alpha & 0 \\ 0 & 0 & \beta \\ 0 & 0 & 0 \end{bmatrix}.$$

Because of the zeros of the function β, the matrix function G_1 changes its rank. This makes the nullspace of G_1 discontinuous. Choose the continuous subnullspace $N_1 = \{z \in \mathbb{R}^3 : z_1 + \alpha z_2 = 0, z_3 = 0\}$ as well as the projector

$$Q_1 = \begin{bmatrix} 0 & -\alpha & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

such that $Q_1 Q_0 = 0$, and

$$P_0 Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad G_2 = \begin{bmatrix} 1 & \alpha + \alpha_{x_2} x_2^1 & 0 \\ 0 & 1 & \beta \\ 0 & 0 & 0 \end{bmatrix}.$$

Now G_2 has constant rank 2, and we may use the nullspace $N_2 = \ker G_2 = \{z \in \mathbb{R}^3 : z_1 + (\alpha + \alpha_{x_2} x_2^1) z_2 = 0,\ z_2 + \beta z_3 = 0\}$ as well as the nullspace projec-

tor

$$Q_2 = \begin{bmatrix} 0 & 0 & (\alpha + \alpha_{x_2} x_2^1)\beta \\ 0 & 0 & -\beta \\ 0 & 0 & 1 \end{bmatrix}.$$

This yields

$$\Pi_1 Q_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Pi_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 & \alpha + \alpha_{x_2} x_2^1 & 0 \\ 0 & 1 & \beta \\ 0 & 0 & 1 \end{bmatrix}.$$

Consequently, the given DAE is quasi-regular independently of the behavior of the functions α and β. G_2 has constant rank, and G_3 remains nonsingular. Observe that there might be open subsets of $\mathcal{D} \times \mathcal{I}$ where natural reformulations of the DAE are regular with tractability index 1, 2 or 3. This kind of more precise information is not available within the framework of quasi-regular DAEs (cf. Section 3.12). □

The class of quasi-regular DAEs seems to be close to the class of DAEs with well-defined differentiation index (cf. Section 3.10). In both versions, the local nature of the DAE may essentially differ from the global one, however, the framework does not comprise the precise local characterization.

By Proposition 9.11 (6), we can never reach a nonsingular G_κ if one of the intersections $\widehat{N}_i$, $i = 1, \ldots, \kappa - 1$, is nontrivial. Further, from Proposition 9.8 (5) we conclude that, for the nonsingularity of G_κ, the relation

$$N_{\kappa-1} = \ker G_{\kappa-1}$$

is necessary. In particular, if N_0 is in fact a proper subspace of $\ker G_0$, it can never happen that G_1 is nonsingular.

As mentioned before, if G_κ is nonsingular, then all intersections $\widehat{N}_1, \ldots, \widehat{N}_{\kappa-1}$, must be trivial and the sums $N_0 + \cdots + N_i$, $i = 1, \ldots, \kappa - 1$, are direct sums. Owing to the construction, $X_i = N_0 + \cdots + N_{i-1} \subseteq \ker Q_i$, $i = 1, \ldots, \kappa - 1$, is valid for quasi-regular DAEs and we then have

$$Q_i Q_j = 0, \quad 0 \le j \le i - 1, \ \ i = 1, \ldots, \kappa.$$

Furthermore, Proposition 9.11 (4) implies $G_\kappa Q_j = B_j Q_j$. In consequence, for quasi-regular DAEs the associated projector functions $Q_j = G_\kappa^{-1} B_j \Pi_{j-1} Q_j$, $j = 1, \ldots, \kappa - 1$, are continuous in all their components.

We show in Section 9.6 that a similar decoupling, as approved for regular linear DAEs, is possible also for quasi-regular DAEs.

As pointed out before, in general, the structural relevance of the rank values $r_i(t)$ gets lost. However, there are special situations if they keep their meaning. Suppose the matrix function sequence under consideration is built by means of widely orthogonal quasi-admissible projector functions. If the subspaces N_i for all i coincide

with the nullspaces $\ker G_i$ on a dense subset $\mathcal{G}_{dense}$ of the open set $\mathcal{G}$, then the admissible projector functions are actually continuous extensions of the related nullspace projector functions defined on $\mathcal{G}_{dense}$, and the ranks r_i are constants on $\mathcal{G}_{dense}$. In this case, the ranks r_i regain a structural meaning.

9.4 Linearization

Supposing the DAE (9.1) satisfies Assumption 9.7, for any reference function $x_* \in \mathcal{C}(\mathcal{I}_*, \mathbb{R}^m)$, $\mathcal{I}_* \subseteq \mathcal{I}_f$, with values in $\mathcal{D}_f$, i.e., $x_*(t) \in \mathcal{D}_f$, $t \in \mathcal{I}_*$, and such that $d(x_*(\cdot),\cdot) \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^n)$, the linear DAE

$$A_*(t)(D_*(t)x(t))' + B_*(t)x(t) = q(t), \quad t \in \mathcal{I}_*, \tag{9.17}$$

with coefficients given by

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

is called a *linearization of the nonlinear DAE* (9.1) along the reference function x_* (cf. Definition 3.12).
The linear DAE (9.17) has continuous coefficients. Owing to the quasi-proper leading term of the nonlinear DAE (9.1), the subspace $\ker D_*$ is a nontrivial $\mathcal{C}$-subspace, and it holds that

$$N_A((d(x_*(t),t))',x_*(t),t) \oplus \operatorname{im} D_*(t) = \mathbb{R}^n, \quad \operatorname{im}(A_*(t)D_*(t)) = \operatorname{im} A_*(t), \quad t \in \mathcal{I}_*. \tag{9.18}$$

If the reference function $x_*(.)$ is continuously differentiable, these subspaces are $\mathcal{C}^1$-subspaces, D_* is continuously differentiable, and hence the linearization (9.17) inherits the quasi-proper leading term from the nonlinear DAE.
Depending on where the graph of the reference function x_* is located, it may well happen that the linearization (9.17) possesses a more precise reformulation in the sense that it has a properly stated leading term or at least a higher dimensional subspace $\ker \tilde{D}_*$.

In case of smooth reference functions x_*, the coefficients of the DAE (9.17) can be described using (9.6) as

$$A_*(t) = A(x_*'(t),x_*(t),t), \tag{9.19}$$
$$D_*(t) = D(x_*(t),t), \tag{9.20}$$
$$B_*(t) = B(x_*'(t),x_*(t),t), \quad t \in \mathcal{I}_*. \tag{9.21}$$

It is due to the construction of the matrix function sequences that linearizations of quasi-regular DAEs inherit the quasi-regularity. This constitutes the content of the

following theorem. However, we cannot expect the reverse assertion to be valid, and this makes a serious difference to the class of regular DAEs (see Theorem 3.33). The case studies in Examples 9.2 and 9.18 illustrate that a linearization does not necessarily inherit the values r_i.

Theorem 9.19. *If the DAE (9.1) satisfies Assumption 9.7 and is quasi-regular on the open set $\mathcal{G} \subseteq \mathcal{D}_f \times \mathcal{I}_f$, then, for each arbitrary reference function $x_* \in \mathcal{C}^\nu(\mathcal{I}_*, \mathbb{R}^m)$, with $(x_*(t),t) \in \mathcal{G}, t \in \mathcal{I}_* \subseteq \mathcal{I}_f$, $\nu \in \mathbb{N}$ sufficient large, the linear DAE (9.17) is also quasi-regular.*

Proof. Since x_* is at least continuously differentiable, we may use the expressions (9.19), (9.20), (9.21) for the coefficients of the linearized DAE (9.17). Then the assertion becomes an immediate consequence of the structure of the matrix function sequences. In particular, $Q_{*,0}(t) := Q_0(t)$, $Q_{*,1}(t) := Q_1(x_*'(t), x_*(t), t)$. Further, for $i \geq 2$, we see that $Q_{*,i}(t) = Q_0(x_*^{(i)}(t), \ldots, x_*'(t), x_*(t), t)$ represent quasi-admissible projector functions for the linear DAE (9.17). If G_κ is nonsingular, so is $G_{*,\kappa}$. □

Linearizations of a quasi-regular DAE may actually be regular with different characteristic values as demonstrated by the next example.

Example 9.20 (Different linearizations). Put $\alpha(s) = \begin{cases} s^2 & \text{for } s > 0 \\ 0 & \text{for } s \leq 0 \end{cases}$ and let $a \in \mathbb{R}$ be a constant. The DAE

$$\begin{aligned} x_1'(t) - x_2(t) &= 0, \\ x_2'(t) + x_1(t) &= 0, \\ \alpha(x_1(t))\, x_4'(t) + x_3(t) &= 0, \\ x_4(t) - a &= 0, \end{aligned} \tag{9.22}$$

satisfies Assumption 9.7 with $\mathcal{D}_f = \mathbb{R}^4$, $\mathcal{I}_f = \mathbb{R}$, $n = 3$ and

$$d(x,t) = \begin{bmatrix} x_1 \\ x_2 \\ x_4 \end{bmatrix}, f(y,x,t) = \begin{bmatrix} y_1 - x_2 \\ y_2 + x_1 \\ \alpha(x_1)y_3 + x_3 \\ x_4 - a \end{bmatrix}, G_0 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 0 & \alpha(x_1) \\ & & & 0 \end{bmatrix}.$$

Notice that the DAE (9.22) has the two maximal regularity regions $\mathcal{G}_+ := \{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x_1 > 0\}$ and $\mathcal{G}_- := \{(x,t) \in \mathcal{D}_f \times \mathcal{I}_f : x_1 < 0\}$. The latter is associated with a proper reformulation

$$d_{new}(x,t) = \begin{bmatrix} x_1 \\ x_2 \\ 0 \end{bmatrix}, \; f_y(y,x,t) = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; R_{new} = \begin{bmatrix} 1 & & \\ & 1 & \\ & & 0 \end{bmatrix}.$$

It is easy to check that the DAE is actually regular with index 1 on $\mathcal{G}_-$ and regular with index 2 on $\mathcal{G}_+$.

Set

$$Q_0 = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 0 \end{bmatrix}, \; N_0 = \operatorname{im} Q_0 = \ker d_x,$$

and compute

$$G_1 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \alpha(x_1) \\ & & & 0 \end{bmatrix}, Q_1 = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & 0 & -\alpha(x_1) \\ & & & 1 \end{bmatrix}, G_2 = \begin{bmatrix} 1 & & & \\ & 1 & & \\ & & 1 & \alpha(x_1) \\ & & & 1 \end{bmatrix},$$

which proves quasi-regularity on $\mathcal{D}_f \times \mathcal{I}_f$.
The linearization along a reference function $x_*(\cdot)$ has the form

$$\begin{bmatrix} 1 & 0 & & 0 \\ 0 & 1 & & 0 \\ 0 & 0 & \alpha(x_{*1}(t)) & \\ 0 & 0 & & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} x(t) \right)' + \begin{bmatrix} & 0 & & -1 & 0 & 0 \\ & 1 & & 0 & 0 & 0 \\ \alpha'(x_{*1}(t))\, x'_{*4}(t) & & & 0 & 1 & 0 \\ & 0 & & 0 & 0 & 1 \end{bmatrix} x(t) = q(t). \tag{9.23}$$

In particular, if $x_*(t) = [0,\ 0,\ 0,\ a]^T$, which is a stationary solution of the original DAE, the linear DAE (9.23) is actually regular with index 1.
If $x_*(t) = [\sin t, \cos t, 0, a]^T$, which is a periodic solution of the nonlinear DAE, the linearization is in detail

$$\begin{aligned} x'_1(t) - x_2(t) &= q_1(t), \\ x'_2(t) + x_1(t) &= q_2(t), \\ \alpha(\sin t)\, x'_4(t) + x_3(t) &= q_3(t), \\ x_4(t) &= q_4(t). \end{aligned}$$

This linear DAE has in turn index 2 and index 1 on its regularity intervals $(0, \pi)$, $(\pi, 2\pi)$, $(2\pi, 3\pi)$, and so on. □

9.5 A DAE transferable into SCF is quasi-regular

We show in this section that if a DAE is transformable into standard canonical form (SCF) then it is quasi-regular.

By Definition 2.77 a DAE is already in SCF if it is in the form

$$\begin{bmatrix} I & 0 \\ 0 & N(t) \end{bmatrix} x'(t) + \begin{bmatrix} W(t) & 0 \\ 0 & I \end{bmatrix} x(t) = q(t), \quad t \in \mathcal{I}, \tag{9.24}$$

where N is strictly upper or lower triangular.

We stress that N does not need to have constant rank nor uniform nilpotency index. We use N as the strictly upper triangular $\kappa \times \kappa$ matrix function

$$N = \begin{bmatrix} 0 & n_{12} & \cdots & n_{1\kappa} \\ & \ddots & \ddots & \vdots \\ & & \ddots & n_{\kappa-1,\kappa} \\ 0 & & & 0 \end{bmatrix}. \tag{9.25}$$

With the $m \times m$ matrix functions

$$A = \left[\begin{array}{c|c} I & \\ \hline & N \end{array}\right], \quad D = \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & 1 & \\ & & & \ddots \\ & & & & 1 \end{array}\right], \quad B = \left[\begin{array}{c|c} W & \\ \hline & I \end{array}\right],$$

the DAE (9.24) becomes a DAE with quasi-proper leading term.

Lemma 9.21. *Each DAE of the form (9.24), (9.25) is quasi-regular, more precisely, there is a sequence of quasi-admissible matrix functions $G_0, \ldots, G_\kappa$ resulting in a nonsingular G_κ, whereby κ denotes the size of the nilpotent matrix N in (9.24).*

Proof. We start constructing the sequence with $P_0 := D$, $D^- = P_0 = \Pi_0$,

$$G_0 = AD = A, \qquad Q_0 = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & 1 & & \\ & & 0 & \\ & & & \ddots \\ & & & & 0 \end{array}\right],$$

$$G_1 = \left[\begin{array}{c|cccc} I & & & & \\ \hline & 1 & n_{12} & \cdots & n_{1\kappa} \\ & & 0 & \ddots & \vdots \\ & & & \ddots & n_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right], \qquad Q_1 = \left[\begin{array}{c|ccccc} 0 & & & & & \\ \hline & 0 & -n_{12} & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{array}\right].$$

This choice fulfills $Q_1 Q_0 = 0$. Further,

$$P_0 P_1 = \Pi_1 = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline & 0 & & & & \\ & & 0 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{array}\right], \quad B_1 = B_0 P_0 = \left[\begin{array}{c|ccccc} W & & & & & \\ \hline & 0 & & & & \\ & & 1 & & & \\ & & & \ddots & & \\ & & & & & 1 \end{array}\right].$$

Since the entries $n_{i,i+1}$ may have zeros or also vanish identically on subintervals, the nullspaces $\ker G_0$ and $\ker G_1$ can be much larger than $N_0 = \operatorname{im} Q_0$ and $N_1 = \operatorname{im} Q_1$, respectively. Next we derive

$$G_2 = \left[\begin{array}{c|cccc} I & & & & \\ \hline & 1 & n_{12} & & n_{1\kappa} \\ & & 1 & n_{23} & \\ & & & 0 & \vdots \\ & & & & \ddots \; n_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right]$$

and choose

$$Q_2 := \left[\begin{array}{c|cccc} 0 & & & & \\ \hline & 0 & q_1^{(2)} & & \\ & & 0 & q_2^{(2)} & \\ & & & 1 & \\ & & & 0 & 0 \\ & & & \vdots & \ddots \\ & & & 0 & 0 \end{array}\right] \text{ implying } \Pi_2 = \left[\begin{array}{c|cccc} I & & & & \\ \hline & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 1 \\ & & & & \ddots \\ & & & & 1 \end{array}\right],$$

with $q_2^{(2)\cdot} := -n_{23}$, $q_1^{(2)} := -n_{13} + n_{12}n_{23}$, and $Q_2Q_1 = 0$, $Q_2Q_0 = 0$. Since $G_2Q_2 = 0$ the subspace $N_2 := \operatorname{im} Q_2$ belongs to $\ker G_2$. Proceeding further in this way we arrive at

$$\Pi_{i-1} = \left[\begin{array}{c|cccc} I & & & & \\ \hline & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ & & & & 1 \\ & & & & \ddots \\ & & & & 1 \end{array}\right] \left.\begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \end{array}\right\} i, \quad G_i = \left[\begin{array}{c|cccc} I & & & & \\ \hline & 1 & n_{12} & & n_{1\kappa} \\ & & \ddots & & \\ & & & 1 \; n_{i,i+1} & \vdots \\ & & & 0 & \\ & & & & \ddots \; n_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right],$$

and choose

$$Q_i = \left[\begin{array}{c|ccccc} 0 & & & & & \\ \hline 0 & q_1^{(i)} & & & & \\ & \ddots & \vdots & & & \\ & & 0 & q_i^{(i)} & & \\ & & & 1 & & \\ & & & 0 & 0 & \\ & & & \vdots & & \ddots \\ & & & 0 & & 0 \end{array}\right] \text{ with } q_i^{(i)} = -n_{i,i+1}, \text{ etc.}$$

This leads to

$$\Pi_i = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline 0 & & & & & \\ & \ddots & & & & \\ & & 0 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{array}\right] \Bigg\} i+1 \quad , \quad B_i = \left[\begin{array}{c|c} W & 0 \\ \hline 0 & I \end{array}\right] \Pi_{i-1}.$$

Further,

$$G_{i+1} = \left[\begin{array}{c|ccccc} I & & & & & \\ \hline 1 & n_{12} & & & & n_{1\kappa} \\ & \ddots & \ddots & & & \\ & & 1 & n_{i+1,i+2} & & \vdots \\ & & & 0 & \ddots & \\ & & & & \ddots & n_{\kappa-1,\kappa} \\ & & & & & 0 \end{array}\right],$$

and so on. We end up with the nonsingular matrix function

$$G_\kappa = \left[\begin{array}{c|ccc} I & & & \\ \hline & 1 & n_{12} & n_{1\kappa} \\ & & 1 & \ddots \\ & & & \ddots \quad n_{\kappa-1,\kappa} \\ & & & 1 \end{array}\right],$$

which completes the proof. □

The SCF-DAE (9.24), (9.25) represents at the same time the two special cases

$$N = 0 \quad \text{and} \quad N = \begin{bmatrix} 0 & 1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{bmatrix}.$$

In both these instances, the DAE is regular in its natural proper reformulation. The index is 1 in the first case, and κ in the second case. This makes it clear that the number κ does not contain any further essential information about the DAE structure. If we knew more about the entries of N, we could possibly find higher dimensional subspaces N_i and obtain a nonsingular matrix function $G_{\bar{\kappa}}$ with $\bar{\kappa} \leq \kappa$.

Turn now to general linear DAEs in standard form

$$E(t)x'(t) + F(t)x(t) = q(t), \quad t \in \mathcal{I}. \tag{9.26}$$

Definition 9.22. The linear DAE (9.26) in standard form is said to be *quasi-regular* if there is a reformulation with quasi-proper leading term, such that the reformulated DAE is quasi-regular.

Let the DAE (9.26) be equivalent to a DAE in SCF. Then nonsingular matrix functions $L \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m))$ and $K \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^m))$ exist such that premultiplication of (9.26) by $L(t)$ and the transformation $x(t) = K(t)\bar{x}(t)$ yield a DAE in SCF (9.24). Denote

$$A := LEK = \begin{bmatrix} I & 0 \\ 0 & N \end{bmatrix}, \quad B := LFK + LEK' = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}$$

and introduce the incidence matrix

$$D := \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & 1 & \\ & & & \ddots \\ & & & 1 \end{array}\right]$$

such that $A = AD$ holds and the transformed SCF-DAE

$$A(t)\bar{x}'(t) + B(t)\bar{x}(t) = L(t)q(t), \quad t \in \mathcal{I}, \tag{9.27}$$

can be written as a DAE with quasi-proper leading term

$$A(t)(D\bar{x}(t))' + B(t)\bar{x}(t) = L(t)q(t), \quad t \in \mathcal{I}. \tag{9.28}$$

Let $Q_0, \ldots, Q_{\kappa-1}$ be the quasi-admissible projector functions provided by Lemma 9.21 for equation (9.28). Put $\tilde{A} := L^{-1}A$, $\tilde{D} := DK^{-1}$, $\tilde{B} := L^{-1}BK^{-1}$, and consider the DAE with quasi-proper leading term

$$\tilde{A}(t)(\tilde{D}x(t))' + \tilde{B}(t)x(t) = q(t), \quad t \in \mathcal{I}, \tag{9.29}$$

which arises from scaling (9.28) by $L(t)^{-1}$ and letting $\bar{x}(t) = K(t)^{-1}x(t)$. By construction, $\tilde{N}_0 = \ker \tilde{D} = K\ker D = KN_0$ is a continuous subnullspace of $\tilde{G}_0 = \tilde{A}\tilde{D} = L^{-1}ADK^{-1} = L^{-1}G_0K^{-1}$ and $\tilde{Q}_0 := KQ_0K^{-1}$ is a quasi-admissible projector function for (9.29). Choosing at each level $\tilde{Q}_i := KQ_iK^{-1}$, and then deriving $\tilde{G}_{i+1} = L^{-1}G_{i+1}K^{-1}$, we find quasi-admissible projector functions $\tilde{Q}_0, \ldots, \tilde{Q}_{\kappa-1}$ yielding a nonsingular $\tilde{G}_\kappa$. This means that the SCF-DAE (9.27) and the transformed DAE (9.29) are quasi-regular at the same time.

Theorem 9.23. *If a linear DAE (9.26) can be transformed into SCF, then it is quasi-regular.*

Proof. Regarding Lemma 9.21, it remains to show that the transformed DAE(9.29) is actually a quasi-proper reformulation of the original DAE (9.26). Applying the property $AD = A$ we derive

$$\tilde{A}\tilde{D} = L^{-1}ADK^{-1} = L^{-1}AK^{-1} = E,$$

and

$$\begin{aligned}\tilde{B} + \tilde{A}\tilde{D}' &= L^{-1}BK^{-1} + L^{-1}AD(K^{-1})' = L^{-1}BK^{-1} + L^{-1}A(K^{-1})' \\ &= L^{-1}(LFK + LEK')K^{-1} + L^{-1}LEK(K^{-1})' \\ &= F + E\big(K'K^{-1} + K(K^{-1})'\big) = F.\end{aligned}$$

Then, the DAE (9.26) can be written as $\tilde{A}\tilde{D}x' + (\tilde{B} + \tilde{A}\tilde{D}')x = q$, which verifies the assertion. □

We show in the next section (cf. Proposition 9.25) that the structure of each quasi-regular linear DAE can be uncovered by means of quasi-admissible projectors. Therefore, Theorem 9.23 actually implies solvability results concerning DAEs being transformable into SCF. More precisely, supposing the coefficients E, F are smooth enough, for each inhomogeneity q that is also sufficiently smooth, and each $t_0 \in \mathcal{I}$, $a \in \mathbb{R}^m$, there is exactly one function $x_* \in \mathcal{C}_{\tilde{D}}^1(\mathcal{I}, \mathbb{R}^m)$ which satisfies the DAE as well as the initial condition

$$\tilde{D}(t_0)\tilde{\Pi}_{\kappa-1}(t_0)(x(t_0) - a) = 0.$$

Under certain additional smoothness requirements, x_* is continuously differentiable and satisfies the original DAE (9.26). This confirms once again the well-known solvability results given in [41].

9.6 Decoupling of quasi-regular linear DAEs

Here we deal with linear DAEs (9.4), i.e.,

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t),$$

that have a quasi-proper leading term and quasi-admissible projectors $Q_0, \dots, Q_{\kappa-1}$ such that the associated matrix function G_κ is nonsingular on the given interval $\mathcal{I} \subseteq \mathbb{R}$.

We decouple the DAEs by means of the same arguments and tools as used for regular DAEs in Section 2.4. Due to $A = ADD^-$, $D^-D = P_0$, the DAE (9.4) is

$$\begin{aligned} G_0 D^- (Dx)' + B_0 x &= q, \\ G_0 D^- (Dx)' + B_0 P_0 x + B_0 Q_0 x &= q, \end{aligned} \tag{9.30}$$

and, with $G_1 D^- = G_0 D^-$, $G_1 Q_0 = B_0 Q_0$,

$$G_1 D^- (Dx)' + B_0 P_0 x + G_1 Q_0 x = q. \tag{9.31}$$

Expressing

$$\begin{aligned} P_1 D^- (Dx)' &= P_0 P_1 D^- (Dx)' + Q_0 P_1 D^- (Dx)' = D^- D P_0 P_1 D^- (Dx)' + Q_0 P_1 D^- (Dx)' \\ &= D^- (D P_0 P_1 x)' - D^- (D P_0 P_1 D^-)' Dx + Q_0 P_1 D^- (Dx)' \\ &= D^- (D\Pi_1 x)' - D^- (D\Pi_1 D^-)' Dx - (I - \Pi_0) Q_1 D^- (Dx)' \\ &= D^- (D\Pi_1 x)' - D^- (D\Pi_1 D^-)' Dx - (I - \Pi_0) Q_1 D^- (D\Pi_0 Q_1 x)' \\ &\qquad + (I - \Pi_0) Q_1 D^- (D\Pi_0 Q_1 D^-)' Dx \end{aligned}$$

we obtain

$$\begin{aligned} G_1 D^- (D\Pi_1 x)' + B_1 x + G_1 Q_0 x - G_1 (I - \Pi_0) Q_1 D^- (D\Pi_0 Q_1 x)' \\ - G_1 (I - \Pi_0) Q_1 D^- (D\Pi_1 D^-)' D\Pi_0 x = q. \end{aligned} \tag{9.32}$$

Proceeding analogously, by means of $G_1(I - \Pi_0) = G_2(I - \Pi_0)$, $B_1 Q_1 = G_2 Q_2$,

$$\begin{aligned} G_1 D^- &= G_2 P_2 P_1 D^- = G_2 (\Pi_1 P_2 P_1 D^- + (I - \Pi_1) P_2 P_1 D^-) \\ &= G_2 (\Pi_2 D^- + (I - \Pi_1) P_1 D^- - (I - \Pi_1) Q_2 D^-) \\ &= G_2 (D^- D \Pi_2 D^- + (I - \Pi_1) P_1 D^- - (I - \Pi_1) Q_2 D^- D \Pi_1 Q_2 D^-) \end{aligned}$$

we arrive at

$$\begin{aligned} G_2 D^- (D\Pi_2 x)' + B_2 x + G_2 \sum_{j=0}^{1} Q_j x \\ - G_2 \sum_{j=0}^{1} (I - \Pi_j) Q_{j+1} D^- (D\Pi_j Q_{j+1} x)' + G_2 \sum_{j=0}^{1} V_j D\Pi_j x = q \end{aligned} \tag{9.33}$$

with $V_j := (I - \Pi_j)\{P_j D^- (D\Pi_j D^-)' - Q_{j+1} D^- (D\Pi_{j+1} D^-)'\} D\Pi_j D^-$.

Formally, these expressions look like those given in Section 2.4 (e.g., (2.39)). However, since now the subspaces N_0, N_1, N_2 do not necessarily coincide with

$\ker G_0$, $\ker G_1$ and $\ker G_2$, and so on, we are confronted with a different meaning of the coefficients.

Let us go into detail for the case when G_2 is nonsingular on $\mathcal{I}$, but G_1 is not. Then, evidently we have $Q_2 = 0$, $P_2 = I$, $\Pi_1 = \Pi_2$, and (9.33) simplifies to

$$\begin{aligned} G_2 D^-(D\Pi_1 x)' + B_2\Pi_1 x + G_2(Q_0 x + Q_1 x) - G_2(I-\Pi_0)Q_1 D^-(D\Pi_0 Q_1 x)' \\ + G_2\Big\{ -(I-\Pi_0)Q_1 D^-(D\Pi_1 D^-)' D\Pi_0 \\ + (I-\Pi_1)P_1 D^-(D\Pi_1 D^-)' D\Pi_1 \Big\} x = q. \end{aligned} \tag{9.34}$$

Notice that due to Proposition 9.8(5) the relation $N_1 = \ker G_1$ must be valid on $\mathcal{I}$. Scaling (9.34) by G_2^{-1} and then splitting by $I = D^-D\Pi_1 + (I-\Pi_1)$ leads to the system consisting of the ODE

$$(D\Pi_1 x)' - (D\Pi_1 D^-)' D\Pi_1 x + D\Pi_1 G_2^{-1} B_2 D^- D\Pi_1 x = D\Pi_1 G_2^{-1} q, \tag{9.35}$$

and the relation

$$\begin{aligned} (I-\Pi_1)D^-(D\Pi_1 x)' + (I-\Pi_1)G_2^{-1}B_2\Pi_1 x + Q_0 x + Q_1 x \\ -(I-\Pi_0)Q_1 D^-(D\Pi_0 Q_1 x)' - (I-\Pi_0)Q_1 D^-(D\Pi_1 D^-)' D(\Pi_0 Q_1 + \Pi_1)x \\ + (I-\Pi_1)P_1 D^-(D\Pi_1 D^-)' D\Pi_1 x = (I-\Pi_1)G_2^{-1} q. \end{aligned}$$

Due to $Q_1 D^-(D\Pi_1 D^-)' D\Pi_0 Q_1 = 0$, the latter reads as

$$\begin{aligned} Q_0 x + Q_1 x - (I-\Pi_0)Q_1 D^-(D\Pi_0 Q_1 x)' \\ = (I-\Pi_1)G_2^{-1} q - (I-\Pi_1)G_2^{-1} B_0\Pi_1 x + Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 x. \end{aligned} \tag{9.36}$$

In turn equation (9.36) decouples to

$$\Pi_0 Q_1 x = \Pi_0 Q_1 G_2^{-1} q - \Pi_0 Q_1 G_2^{-1} B_0 \Pi_1 x, \tag{9.37}$$

$$\begin{aligned} Q_0 x - Q_0 Q_1 D^-(D\Pi_0 Q_1 x)' \\ = Q_0 P_1 G_2^{-1} q - Q_0 P_1 G_2^{-1} B_0 \Pi_1 x + Q_0 Q_1 D^-(D\Pi_1 D^-)' D\Pi_1 x. \end{aligned} \tag{9.38}$$

It becomes clear that any quasi-regular DAE (9.4) with $\kappa = 2$ is equivalent via the decomposition $x = D^-D\Pi_1 x + \Pi_0 Q_1 x + Q_0 x$ to the system (9.35), (9.37), (9.38). Equation (9.35) is an explicit ODE that determines the component $D\Pi_1 x$, thus $\Pi_1 x$. Equation (9.37) describes the algebraic components $\Pi_0 Q_1 x$ as in the case of regular DAEs. Equation (9.38) looks like a differentiation problem, however, now, in contrast to the decoupling of regular DAEs, it may happen that $Q_0 Q_1$ vanishes on $\mathcal{I}$ or on subintervals. A lot of different situations are integrated in the framework of quasi-regular DAEs.

Example 9.24 (G_2 is nonsingular). The linear system with continuous coefficients

$$\begin{aligned} A_{11}x_1' + x_1 &= q_1 \quad \} m_1, \\ A_{23}x_3' + x_2 &= q_2 \quad \} m_2, \\ x_3 &= q_3 \quad \} m_3, \end{aligned} \tag{9.39}$$

has a quasi-proper leading term with

$$A = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & 0 & A_{23} \\ 0 & 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \quad B = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

$D^- = D, R = D, P_0 = D, G_0 = AD = A, Q_0 = I - P_0$. Let the entry A_{11} be nonsingular on the given interval $\mathcal{I}$. Compute

$$G_1 = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & I & A_{23} \\ 0 & 0 & 0 \end{bmatrix}, \; Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -A_{23} \\ 0 & 0 & I \end{bmatrix}, \; \Pi_0 Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix}, \; G_2 = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & I & A_{23} \\ 0 & 0 & I \end{bmatrix}.$$

The projectors Q_0, Q_1 are quasi-admissible. G_2 is nonsingular on $\mathcal{I}$ independently of the rank of A_{23}, thus the DAE is quasi-regular with $\kappa = 2$. Then it holds that $\ker G_1 = N_1$. With

$$D\Pi_1 D^- = \begin{bmatrix} I & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad D\Pi_1 G_2^{-1} B_2 D^- = \begin{bmatrix} A_{11}^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix}, \quad D\Pi_1 G_2^{-1} = \begin{bmatrix} A_{11}^{-1} & & \\ & 0 & \\ & & 0 \end{bmatrix},$$

$$\Pi_0 Q_1 G_2^{-1} = \begin{bmatrix} 0 & & \\ & 0 & \\ & & I \end{bmatrix}, \quad \Pi_0 Q_1 G_2^{-1} B_0 \Pi_1 = 0, \quad Q_0 Q_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -A_{23} \\ 0 & 0 & 0 \end{bmatrix},$$

$$Q_0 P_1 G_2^{-1} = \begin{bmatrix} 0 & & \\ & I & \\ & & 0 \end{bmatrix}, \quad 32 Q_0 P_1 G_2^{-1} B_0 \Pi_1 = 0,$$

the decoupled system (9.35), (9.37), (9.38) is of the form (omitting the zeros)

$$\begin{aligned} x_1' + A_{11}^{-1} x_1 &= A_{11}^{-1} q_1, \\ x_3 &= q_3, \\ x_2 + A_{23} x_3' &= q_2, \end{aligned} \tag{9.40}$$

which is, of course, not a surprise. Notice that the third equation may contain further derivative-free equations. Namely, with any generalized inverse A_{23}^-, it holds that $(I - A_{23}A_{23}^-)x_2 = (I - A_{23}A_{23}^-)q_2$, and these equations disappear (are trivial ones) only if A_{23} has full row rank m_2 on $\mathcal{I}$. Different situations arise depending on A_{23}.

Case 1: $\ker A_{23} = \{0\}$, $m_2 \geq m_3$.

Here, the DAE actually has a proper leading term, $\ker G_0 = \ker D = N_0$, the projectors Q_0, Q_1 are admissible ones, and the DAE is regular with tractability index 2. All

components of q_3 have to be differentiated. Observe that $\operatorname{rank} Q_0Q_1 = \operatorname{rank} A_{23} = m_3$. The decoupled system (9.35), (9.37), (9.38) is exactly what we have already obtained in Subsection 2.4.3.2.

Case 2: $A_{23} = 0$.

Obviously, this system (9.39) is in fact regular with tractability index one. However, as we now cover only a part of $\ker G_0$ by the quasi-admissible projector Q_0, the other part is included in N_1. Because of $Q_0Q_1 = 0$, equation (9.38) de facto does not contain any term with a differentiation problem.
This example makes it obvious that, initially, κ has nothing to do with the index.

Case 3: $\operatorname{rank} A_{23}$ varies on $\mathcal{I}$.

Here we may find local index-2 problems on subintervals $\mathcal{I}_2 \subset \mathcal{I}$ with $\ker A_{23} = \{0\}$. On other subintervals $\mathcal{I}_1 \subset \mathcal{I}$ there might be regular index-1 problems. The index changes in between are harmless. Even clusters of critical points, for instance at $m_2 = m_3 = 1$, $A_{23}(t) = t \sin \frac{1}{t}$, are allowed.

Case 4: $\operatorname{rank} A_{23} = m_2$, $m_3 > m_2$.

In contrast to Case 1, even though A has constant rank, $N_0 = \ker D$ is only a proper subspace of $\ker G_0$. The third equation in the system (9.40) does not contain any derivative-free part; however, it is not necessary to differentiate all components of x_3, which becomes obvious, e.g., in the special case $A_{23} = \begin{bmatrix} I & 0 \end{bmatrix}$. A possible reformulation with proper leading term would read

$$\tilde{A} = A, \quad \tilde{D} = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & P_{23} \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & -A_{23}P'_{23} \\ 0 & 0 & I \end{bmatrix}, \quad \tilde{G}_0 = \tilde{A}\tilde{D} = A,$$

with $P_{23} := A_{23}^- A_{23}$. An admissible sequence would be, for instance,

$$\tilde{Q}_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I - P_{23} \end{bmatrix}, \quad \tilde{G}_1 = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & I & A_{23}(I - P'_{23}) \\ 0 & 0 & I - P_{23} \end{bmatrix},$$

$$\tilde{Q}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -A_{23}(I - P'_{23})P_{23} \\ 0 & 0 & P_{23} \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -A_{23} \\ 0 & 0 & P_{23} \end{bmatrix}, \quad \tilde{D}\tilde{\Pi}_1\tilde{D}^- = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

$$\tilde{G}_2 = \begin{bmatrix} A_{11} & 0 & 0 \\ 0 & I & A_{23}(I - P'_{23}) \\ 0 & 0 & I \end{bmatrix}.$$

This means that the DAE is regular with index 2, with characteristic values

$$\tilde{r}_0 = m_1 + m_2, \ \tilde{r}_1 = m_1 + m_2 + (m_3 - m_2), \ \tilde{r}_2 = m_1 + m_2 + m_3 = m, \ \mu = 2.$$

Comparing, for the latter case, the given quasi-admissible and admissible projectors, reveals that the loss of a component of $\ker G_0$ is, one might say, compensated for in the next step. □

In general, any quasi-regular DAE decouples into an inherent explicit regular ODE that determines the solution component $D\Pi_\kappa x$ and equations defining the other components, possibly with inherent differentiation problems. This seems to be the same as for regular DAEs, but as demonstrated by the previous example, the meaning might differ essentially.

Proposition 9.25. *If the linear DAE (9.4) is quasi-regular with G_κ being nonsingular, then it can be rewritten as*

$$G_\kappa D^-(D\Pi_{\kappa-1}x)' + B_\kappa\Pi_{\kappa-1}x + G_\kappa\sum_{j=0}^{\kappa-1}\Big\{Q_jx-(I-\Pi_j)Q_{j+1}D^-(D\Pi_jQ_{j+1}x)'+V_jD\Pi_jx\Big\} = q, \tag{9.41}$$

and decoupled via

$$\begin{aligned} &x = D^-u + v_0 + \cdots + v_{\kappa-1},\\ &u = D\Pi_{\kappa-1}x,\ v_0 = Q_0x,\ v_i = \Pi_{i-1}Q_ix,\ i = 1,\ldots,\kappa-1, \end{aligned}$$

into the system

$$u' - (D\Pi_{\kappa-1}D^-)'u + D\Pi_{\kappa-1}G_\kappa^{-1}B_\kappa D^-u = D\Pi_{\kappa-1}G_\kappa^{-1}q, \tag{9.42}$$

$$\begin{bmatrix} 0 & \mathcal{N}_{01} & \cdots & \mathcal{N}_{0,\kappa-1}\\ & 0 & & \vdots\\ & & \ddots & \mathcal{N}_{\kappa-2,\kappa-1}\\ & & & 0 \end{bmatrix}\begin{bmatrix} 0\\ Dv_1\\ \vdots\\ Dv_{\kappa-1}\end{bmatrix}' + \begin{bmatrix} I & \mathcal{M}_{01} & \cdots & \mathcal{M}_{0,\kappa-1}\\ & I & & \vdots\\ & & \ddots & \mathcal{M}_{\kappa-2,\kappa-1}\\ & & & I \end{bmatrix}\begin{bmatrix} v_0\\ v_1\\ \vdots\\ v_{\kappa-1}\end{bmatrix} + \begin{bmatrix} \mathcal{H}_0\\ \mathcal{H}_1\\ \vdots\\ \mathcal{H}_{\kappa-1}\end{bmatrix}u = \begin{bmatrix} \mathcal{L}_0\\ \mathcal{L}_1\\ \vdots\\ \mathcal{L}_{\kappa-1}\end{bmatrix}q \tag{9.43}$$

with continuous coefficients for $i = 1,\ldots,\kappa-2$,

$$V_j = (I-\Pi_j)\Big\{P_jD^-(D\Pi_jD^-)' - Q_{j+1}D^-(D\Pi_{j+1}D^-)'\Big\}D\Pi_jD^-,$$
$$j = 0,\dots\kappa-1,$$
$$\mathcal{N}_{01} = -Q_0Q_1D^-, \qquad \mathcal{N}_{0j} = -Q_0P_1\cdots P_jQ_jD^-, \qquad j = 2,\dots,\kappa-1,$$
$$\mathcal{N}_{i,i+1} = -\Pi_{i-1}Q_iQ_{i+1}D^-, \quad \mathcal{N}_{ij} = -\Pi_{i-1}Q_iP_{i+1}\cdots P_{j-1}Q_jD^-, \; j = i+2,\dots,\kappa-1,$$
$$\mathcal{M}_{0j} = Q_0P_1\cdots P_{\kappa-1}\mathcal{M}_jD\Pi_{j-1}Q_j, \qquad j = 1,\dots,\kappa-1,$$
$$\mathcal{M}_{ij} = \Pi_{i-1}Q_iP_{i+1}\cdots P_{\kappa-1}\mathcal{M}_jD\Pi_{j-1}Q_j, \qquad j = i+1,\dots,\kappa-1,$$
$$\mathcal{M}_j = \sum_{s=0}^{j-1} V_sD\Pi_{j-1}Q_jD^-, \qquad j = 1,\dots,\kappa-1,$$
$$\mathcal{L}_0 = Q_0P_1\cdots P_{\kappa-1}G_\kappa^{-1}, \qquad \mathcal{L}_i = \Pi_{i-1}Q_iP_{i+1}\cdots P_{\kappa-1}G_\kappa^{-1},$$
$$\mathcal{L}_{\kappa-1} = \Pi_{\kappa-2}Q_{\kappa-1}G_\kappa^{-1},$$
$$\mathcal{H}_0 = Q_0P_1\cdots P_{\kappa-1}\mathcal{K}\Pi_{\kappa-1}, \quad \mathcal{H}_i = \Pi_{i-1}Q_iP_{i+1}\cdots P_{\kappa-1}\mathcal{K}\Pi_{\kappa-1},$$
$$\mathcal{H}_{\kappa-1} = \Pi_{\kappa-2}Q_{\kappa-1}\mathcal{K}\Pi_{\kappa-1},$$
$$\mathcal{K} = (I-\Pi_{\kappa-1})G_\kappa^{-1}B_{\kappa-1}\Pi_{\kappa-1} + \sum_{\ell=0}^{\kappa-1} V_\ell D\Pi_{\kappa-1}.$$

Proof. Formula (9.41) is proved by the induction arguments used in Proposition 2.23 for (2.39). Thereby we take into account that $Q_\kappa = 0$, $P_\kappa = I$, $\Pi_{\kappa-1} = \Pi_\kappa$. The further decoupling will be obtained by scaling (9.41) with G_κ^{-1} and then splitting by $D\Pi_{\kappa-1}D^-$ and $I-\Pi_{\kappa-1}$ and repeating the respective arguments from Section 2.4.2 (cf. (2.48), (2.50) and (2.52)). □

We stress once again that the decoupling formulas look like those obtained for regular DAEs but they have a different meaning. The following solvability assertion constitutes a simple benefit of the above decoupling.

Corollary 9.26. *If the linear DAE (9.4) is quasi-regular and it possesses sufficiently smooth coefficients, then it is solvable for each arbitrary sufficiently smooth excitations.*

9.7 Difficulties arising with the use of subnullspaces

The use of continuous subnullspaces and quasi-admissible matrix functions instead of nullspaces and admissible matrix functions offers more flexibility concerning the DAE analysis, but it is less rigorous in some sense. Any regular DAE is also quasi-regular, and hence quasi-regularity can be thought as a concept generalizing regularity. However, the meaning of the rank functions r_i to indicate characteristic values of the DAE gets lost.

It remains open how one should choose the continuous subnullspaces, and this makes the approach somewhat precarious. Compared with the case of admissible projector functions, the closest way is to look for continuous subnullspaces with

maximal possible dimension, i.e., for maximal subnullspaces. If, at the level i, the matrix function G_i has constant rank, then the maximal continuous subnullspace N_i coincides with the nullspace of G_i itself, and if G_i has constant rank on a dense subset of its definition domain, then the maximal continuous subnullspace represents the continuous extension of the nullspace restricted to the dense subset. However, in general, if the rank function r_i varies arbitrarily, as demonstrated before, it heavily depends on the locality whether a continuous subnullspace is maximal, and maximality may get lost on subsets. Hence, global maximality does not imply local maximality.

The opposite way to build continuous subnullspaces is to take one-dimensional ones. This seems to be easier in practice. For regular DAEs having the tractability index μ, the quasi-admissible sequences resulting this way need more levels, and end up with a nonsingular G_κ, $\kappa \geq \mu$. The following examples indicate once again the loss of information formerly contained in the ranks r_i.

Example 9.27 (Different sequences). The linear constant coefficient DAE

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} x(t) \right)' + x(t) = q(t)$$

is regular with tractability index 2 and characteristic values $r_0 = 2, r_1 = 2, r_2 = 4$. We describe two different matrix sequences built with one-dimensional subnullspaces starting from

$$G_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Case A:

$$Q_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; \mathcal{W}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

further,

$$Q_2 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$B_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Case B:

$$Q_1 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \; \mathcal{W}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \; B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

further,

$$Q_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad G_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

and

$$B_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

In both cases, we have the same rank values $r_0 = 2$, $r_1 = 2$, $r_3 = 3$ and $r_4 = 4$, except for r_2 which equals 2 in Case A and 3 in Case B. □

Example 9.28 (Loss of meaning). The linear constant coefficient DAE

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} x(t) \right)' + x(t) = q(t)$$

is regular with tractability index 3. The characteristic values are $r_0 = 2$, $r_1 = 3$, $r_2 = 3$, $r_3 = 4$. We describe a quasi-admissible sequence corresponding to one-dimensional subnullspaces:

$$G_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_0 = I,$$

$$G_1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_1 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G_3 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad Q_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathcal{W}_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G_4 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

The rank sequence $r_0 = 2$, $r_1 = 2$, $r_2 = 2$, $r_3 = 3$, $r_4 = 4$ for this index-3 DAE is exactly the same as the corresponding sequence in Case A of the previous example which comprises an index-2 DAE. □

9.8 Notes and references

(1) If there is some structure available to be exploited, the construction of quasi-admissible projector functions might be easier than the construction of admissible ones. For instance, while the regularity proof for linear DAEs with well-defined strangeness index is quite difficult because of the complex structure of the relevant nullspaces (cf. Section 2.10), quasi-regularity is much easier to show as we are going to do now. The DAE

$$\underbrace{\left[\begin{array}{c|cccc} I & 0 & K_2 & \cdots & K_\kappa \\ \hline & 0 & N_{12} & \cdots & N_{1\kappa} \\ & & 0 & & \vdots \\ & & & \ddots & N_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right]}_{A} x' + \underbrace{\left[\begin{array}{c|cccc} W & & & & \\ \hline & I & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & I \end{array}\right]}_{B} x = q \tag{9.44}$$

with full row rank blocks $N_{i,i+1} \in L(\mathbb{R}^{l_{i+1}}, \mathbb{R}^{l_i})$, $i = 1, \ldots, \kappa - 1$ and

$$0 < l_1 \leq \cdots \leq l_i \leq l_{i+1} \leq \cdots \leq l_\kappa$$

is called a DAE in *S-canonical form* (cf. Section 2.10). Each square DAE $\tilde{A}\tilde{x}' + \tilde{C}\tilde{x} = \tilde{q}$ with sufficiently smooth coefficients, for which the strangeness index is well-defined and the undetermined part vanishes, can be brought into the form (9.44) by scalings and transformations of the unknown function (cf. [130]).
With the incidence matrix

$$D = \left[\begin{array}{c|ccc} I & & & \\ \hline & 0 & & \\ & & I & \\ & & & \ddots \\ & & & I \end{array}\right],$$

the DAE (9.44) can be rewritten as DAE (9.4) with quasi-proper leading term. By constructing quasi-admissible projectors $Q_0, \ldots, Q_{\kappa-1}$ providing a nonsingular G_κ we show that this DAE is quasi-regular. Note once again that this DAE is even proved to be regular in Section 2.10 and the aim here is just to demonstrate the comfort of quasi-admissible sequences.
We start with $G_0 = AD = A$, $D^- = D$, $R = D$ and choose

$$Q_0 = \left[\begin{array}{c|ccc} 0 & & & \\ \hline & I & & \\ & & 0 & \\ & & & \ddots \\ & & & 0 \end{array}\right] \text{ implying } G_1 = \left[\begin{array}{c|cccc} I & 0 & K_2 & \cdots & K_\kappa \\ \hline & I & N_{12} & \cdots & N_{1\kappa} \\ & & 0 & \ddots & \vdots \\ & & & \ddots & N_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right].$$

Next, we choose

$$Q_1 = \left[\begin{array}{c|c|ccc} 0 & 0 & -K_2 & & \\ \hline & 0 & -N_{12} & & \\ & & I & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & 0 \end{array}\right]$$

providing

$$\Pi_0 Q_1 = \left[\begin{array}{c|cccc} 0 & 0 & -K_2 & & \\ \hline & 0 & & & \\ & & I & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & 0 \end{array}\right], \quad \Pi_1 = \left[\begin{array}{c|cccc} 0 & 0 & K_2 & & \\ \hline & 0 & & & \\ & & 0 & & \\ & & & I & \\ & & & & \ddots \\ & & & & I \end{array}\right],$$

and

$$B_1 Q_1 = \left[\begin{array}{c|cccc} 0 & 0 & \Gamma_2 & & \\ \hline & 0 & 0 & & \\ & 0 & I & & \\ & & & 0 & \\ & & & & \ddots \\ & & & & 0 \end{array}\right], \quad G_2 = \left[\begin{array}{c|cccc} I & 0 & K_2^{(2)} & & K_\kappa \\ \hline & I & N_{12} & \cdots \ \cdots & N_{1\kappa} \\ & & I & & \vdots \\ & & & 0 & \vdots \\ & & & \ddots & N_{\kappa-1,\kappa} \\ & & & & 0 \end{array}\right],$$

with $K_2^{(2)} = \Gamma_2 + K_2$, $\Gamma_2 = -WK_2 + K_2'$. Using

$$Q_2 := \left[\begin{array}{c|cccc} 0 & & * & & \\ \hline & 0 & * & & \\ & & 0 & -N_{23} & \\ & & & I & \\ & & & & 0 \\ & & & & \ddots \\ & & & & 0 \end{array}\right]$$

we find

$$\Pi_1 Q_2 = \left[\begin{array}{c|cccc} 0 & & * & & \\ \hline & 0 & & & \\ & & 0 & & \\ & & & I & \\ & & & & 0 \\ & & & & \ddots \\ & & & & 0 \end{array}\right], \quad \Pi_2 = \left[\begin{array}{c|cccc} I & 0 & * & * & \\ \hline & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & I \\ & & & & \ddots \\ & & & & I \end{array}\right].$$

For any $i \leq \kappa - 1$ we obtain

$$G_i = \left[\begin{array}{c|cccc} I & 0 & * & \cdots & * \\ \hline & I & N_{12} & \cdots & N_{1\kappa} \\ & & I & & \\ & & & \ddots & \\ & & & & I \\ & & & & 0 \\ & & & & & \ddots \\ & & & & & & N_{\kappa-1,\kappa} \\ & & & & & & 0 \end{array}\right]$$

when choosing

$$Q_i = \left[\begin{array}{c|ccccc} 0 & & & & * & \\ \hline & 0 & & & * & \\ & & \ddots & & \vdots & \\ & & & 0 & * & \\ & & & & I & \\ & & & & & 0 \\ & & & & & & \ddots \\ & & & & & & & 0 \end{array}\right].$$

Finally,

$$G_\kappa = \left[\begin{array}{c|cccc} I & 0 & * & \ldots & * \\ \hline & I & N_{12} & \ldots & N_{1\kappa} \\ & & \ddots & \ddots & \\ & & & \ddots & N_{\kappa-1,\kappa} \\ & & & & I \end{array}\right].$$

Observing the relations

$$(I-\Pi_{i-1})Q_i = \left[\begin{array}{c|ccccc} 0 & & & & * & \\ \hline & 0 & & & \vdots & \\ & & \ddots & & * & \\ & & & N_{i,i+1} & & \\ & & & & 0 & \\ & & & & & \ddots \\ & & & & & & 0 \end{array}\right],$$

for $i = 1, \ldots, \kappa-1$ we find $\operatorname{rank}((I-\Pi_{i-1})Q_i) \geq \operatorname{rank} N_{i,i+1} = l_1$. This means that all differentiation terms appear in the corresponding decoupled system.

(2) Quasi-regularity supposes the existence of nontrivial continuous subspaces N_i of $\ker G_i$ on each level i until a nonsingular G_κ is reached. In particular, $N_0 = \ker D \subseteq G_0$ must be continuous and nontrivial. It may happen that, although G_i has an everywhere nontrivial nullspace, there is no *continuous* subnullspace of

dimension greater than or equal to one. For instance (cf. [25]), for given functions $\alpha, \beta \in \mathcal{C}^1(\mathcal{I}, \mathbb{R})$ such that $\alpha(t)\beta(t) = 0$, $\alpha'(t)\beta'(t) - 1 \neq 0$, $t \in \mathcal{I}$, the DAE in standard form

$$\underbrace{\begin{bmatrix} 0 & \alpha \\ \beta & 0 \end{bmatrix}}_{=N(t)} x'(t) + x(t) = q(t), \quad t \in \mathcal{I}, \tag{9.45}$$

is solvable, with the solution

$$x(t) = (I - N(t)')^{-1}\big(q(t) - (N(t)q(t))'\big), \quad t \in \mathcal{I}.$$

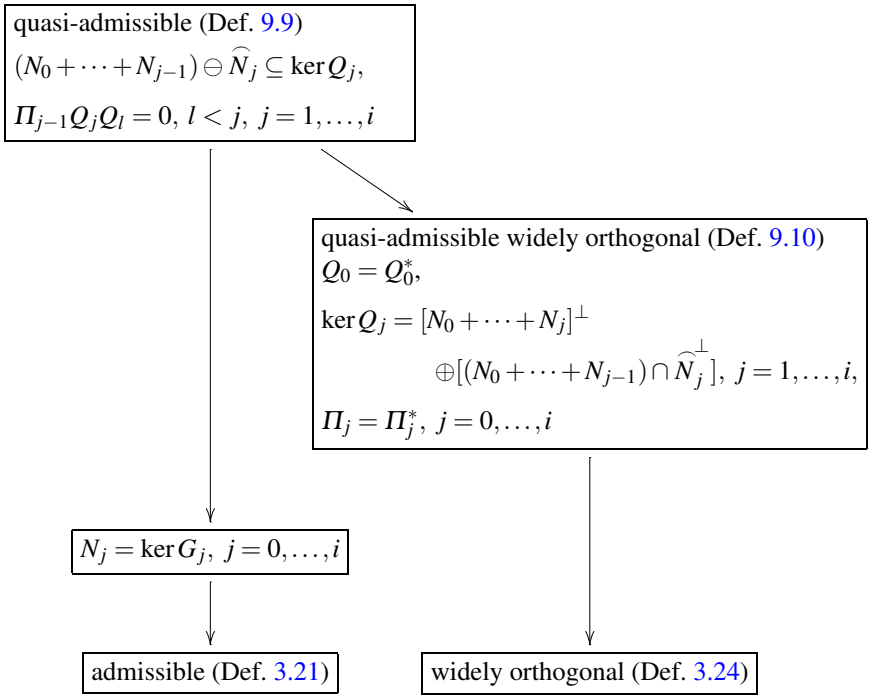
If one of the functions α or β vanishes identically, this DAE is in SCF, and hence quasi-regular.
In general, $N(t)$ has rank 1 where $\alpha(t)^2 + \beta(t)^2 \neq 0$, and rank 0 elsewhere. The nullspace $\ker G_0(t)$ is spanned by $\begin{bmatrix} 1 & 0 \end{bmatrix}^{\mathrm{T}}$, if $\alpha(t) \neq 0$, and spanned by $\begin{bmatrix} 0 & 1 \end{bmatrix}^{\mathrm{T}}$ if $\beta(t) \neq 0$. If $\alpha(t) = \beta(t) = 0$, then $\ker N(t)$ coincides with $\mathbb{R}^2$. Thus, if neither α nor β vanishes identically on $\mathcal{I}$, then there is no nontrivial continuous subnullspace N_0, and the DAE cannot be put into a quasi-proper formulation.
In this example, one could turn to families of subintervals of $\mathcal{I}$ where the DAE is quasi-regular. This would closely correspond to the approach discussed in [41], and the references therein, to put the DAE into SCF on subintervals.

(3) The framework of quasi-regular DAEs was first addressed in [48]. In [59], a comprehensive analysis of so-called *index-2 DAEs with harmless critical points* is given by means of the framework of quasi-regular DAE yielding an admissible matrix function such that G_2 is nonsingular. Those DAEs comprise regularity intervals with index 1 and index 2. The quasi-regular decoupling is applied for obtaining error estimations for numerical integration methods.

(4) In contrast to investigations of singularities by means of smooth projector extensions (e.g., [173, 174], cf. Proposition 2.76) in the more general quasi-regular DAEs different rank values on open subsets are admitted. Then, continuous extensions of the nullspace projector functions do not exist.

9.9 Hierarchy of quasi-admissible projector function sequences for general nonlinear DAEs

The matrices $Q_0, \ldots, Q_i$ are projectors, where Q_j projects onto $N_j = \ker G_j$, $j = 0, \ldots, i$, with $P_0 := I - Q_0$, $\Pi_0 := P_0$ and $P_j := I - Q_j$, $\Pi_j := \Pi_{j-1} P_j$, $\widehat{N}_j := (N_0 + \cdots + N_{j-1}) \cap N_j$, $j = 1, \ldots, i$.


quasi-admissible (Def. 9.9)
$(N_0+\cdots+N_{j-1})\ominus\widehat{N}_j\subseteq\ker Q_j$,
$\Pi_{j-1}Q_jQ_l=0,\ l<j,\ j=1,\ldots,i$
quasi-admissible widely orthogonal (Def. 9.10)
$Q_0=Q_0^*$,
$\ker Q_j=[N_0+\cdots+N_j]^\perp$
$\oplus[(N_0+\cdots+N_{j-1})\cap\widehat{N}_j^\perp],\ j=1,\ldots,i$,
$\Pi_j=\Pi_j^*,\ j=0,\ldots,i$
$N_j=\ker G_j,\ j=0,\ldots,i$
admissible (Def. 3.21)
widely orthogonal (Def. 3.24)

Chapter 10
Nonregular DAEs

We deal with DAEs of the form

$$f((D(t)x(t))',x(t),t) = 0,$$

and, in particular, with linear DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t).$$

These DAEs have properly involved derivatives and comprise k equations and m unknowns. It should be emphasized once again that the prior purpose of this monograph is the detailed analysis of *regular* DAEs. In particular, we aim for practicable and rigorous regularity criteria, and, in this way we would like to assist in modeling regular DAEs in applications, and in avoiding DAE models that fail to be regular.

On the other hand, several authors lay particular emphasis on the importance of so-called rectangular DAEs and spend much time investigating those DAEs (cf. [130]). In our view, this class of problems is less interesting, so we touch on this topic just slightly.

At times one speaks of *overdetermined* systems, if $k > m$, but of *underdetermined* systems, if $k < m$. However, this notion does not say very much; it simply indicates the relation between the numbers of equations and unknown functions. It seems to be more appropriate to speak of *nonregular DAEs*, that is, of DAEs not being regular. This option also includes the square systems (with $m = k$) which may also contain free variables and consistency conditions if the regularity conditions are violated. We speak only then of an *overdetermined* DAE, if the DAE shows the structure of a regular DAE subject to an additional constraint. In contrast, an *underdetermined* DAE comprises a component to be chosen freely, such that, if this component is fixed, the resulting DAE is regular.

First of all, in Section 10.1, we illustrate the latitude for interpretations concerning nonregular DAEs by case studies. In Section 10.2, we provide a projector based decoupling of linear DAEs and generalize the tractability index to apply also to nonregular DAEs. Section 10.3 is devoted to underdetermined nonlinear DAEs.

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_10,

10.1 The scope of interpretations

The simple constant coefficient system

$$x' + x = q_1, \tag{10.1}$$
$$x = q_2, \tag{10.2}$$

represents an overdetermined DAE with $k = 2$ equations and $m = 1$ unknowns. How should we interpret these equations? If one emphasizes the algebraic equation $x = q_2$, one is led to a differentiation of q_2 as well as to a consistency condition coming from the first equation, namely

$$q_2' + q_2 - q_1 = 0.$$

On the other hand, if one puts emphasis on the differential equation $x' + x = q_1$ one can solve this equation with a constant x_0 for

$$x(t) = e^{-t}\left(x_0 + \int_0^t e^s q_1(s)ds\right)$$

and then consider the second equation to be responsible for consistency. This leads to the consistency condition

$$e^{-t}\left(x_0 + \int_0^t e^s q_1(s)ds\right) - q_2(t) = 0.$$

At a first glance this consistency condition looks quite different, but differentiation immediately yields again $q_2 - q_1 + q_2' = 0$.
The last interpretation is oriented to solve differential equations rather than to solve algebraic equations and then to differentiate. We adopt this point of view.

A further source for deliberation can be seen when regarding the equation

$$(x_1 + x_2)' + x_1 = q, \tag{10.3}$$

which represents a DAE comprising just a single equation, $k = 1$, and $m = 2$ unknowns. This is an underdetermined DAE; however, should we choose x_1 or x_2 to be free? One can also think of writing

$$(x_1 + x_2)' + (x_1 + x_2) - x_2 = q, \tag{10.4}$$

or

$$(x_1 + x_2)' + \frac{1}{2}(x_1 + x_2) + \frac{1}{2}(x_1 - x_2) = q. \tag{10.5}$$

As described in Subsection 2.4.1, the special structure of an admissible matrix function sequence (cf. Definition 2.6) allows a systematic rearrangement of terms in general linear DAEs with properly stated leading terms, among them also nonregular ones. Subsection 2.4.1 ends up with a first glance at DAEs whose initial matrix

function $G_0 = AD$ already shows maximal rank. We resume this discussion noting that, in the above two examples, we have the constant matrix functions

$$G_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \text{resp.} \quad G_0 = \begin{bmatrix} 1 & 1 \end{bmatrix},$$

and both already have maximal possible rank. Recall that, in this case, the DAE is equivalent to the structured system (2.42), that is to

$$(Dx)' - R'Dx + DG_0^- B_0 D^- Dx + DG_0^- B_0 Q_0 x = DG_0^- q, \tag{10.6}$$

$$\mathcal{W}_0 B_0 D^- Dx = \mathcal{W}_0 q, \tag{10.7}$$

whose solution decomposes as $x = D^- Dx + Q_0 x$.
For the overdetermined system (10.1), (10.2), we have in detail: $D = D^- = R = 1$, $Q_0 = 0$,

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \; G_0 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; G_0^- = \begin{bmatrix} 1 & 0 \end{bmatrix}, \; \mathcal{W}_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \; DG_0^- B_0 D^- = 1.$$

Inserting these coefficients, we see that equation (10.6) coincides with the ODE (10.1), whereas the second equation (10.7) is nothing else than (10.2). This emphasizes the interpretation of the DAE to be primarily the explicit ODE (10.1) subject to the consistency condition (10.2).

For the underdetermined DAE (10.3), one has $A = 1, D = \begin{bmatrix} 1 & 1 \end{bmatrix}, R = 1, B = \begin{bmatrix} 1 & 0 \end{bmatrix}$, $\mathcal{W}_0 = 0$, and the equation (10.7) disappears. Various projectors Q_0 are admissible, and different choices lead to different ODEs

$$(Dx)' + DG_0^- B_0 D^- Dx + DG_0^- B_0 Q_0 x = DG_0^- q, \tag{10.8}$$

as well as solution representations $x = D^- Dx + Q_0 x$. We consider three particular cases:

(a) Set and compute

$$D^- = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}, \; P_0 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \; Q_0 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix}, \; G_0^- = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix},$$

and further $DG_0^- B_0 D^- = \frac{1}{2}$, $DG_0^- = 1$, $DG_0^- B_0 Q_0 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \end{bmatrix}$, and we see that the corresponding ODE (10.8) coincides with (10.5).

(b) Set and compute

$$D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; P_0 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & -1 \\ 0 & 1 \end{bmatrix}, \; G_0^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$DG_0^- B_0 D^- = 1$, $DG_0^- = 1$, $DG_0^- B_0 Q_0 = \begin{bmatrix} 0 & -1 \end{bmatrix}$. Now equation (10.8) coincides with (10.4).

(c) Set and compute

$$D^- = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \; P_0 = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 1 & 0 \\ -1 & 0 \end{bmatrix}, \; G_0^- = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$DG_0^- B_0 D^- = 0$, $DG_0^- = 1$, $DG_0^- B_0 Q_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}$, and equation (10.8) coincides with the version (10.3).

Observe that the eigenvalues of $DG_0^- B_0 D^-$ depend on the choice of the projector Q_0: the eigenvalues are $\frac{1}{2}$, 1 and 0, in the three cases, respectively.
Each case corresponds to a particular choice of the free variable. Which is the right one? One could restrict the variety of projectors and take just the widely orthogonal ones. In our example this corresponds to item (a). However, this would be an arbitrary action.
At this point it seems to be worth mentioning that the *inherent explicit regular ODE* of a *regular DAE* is uniquely defined by the problem data. In particular, it is independent of the choice of the fine decoupling projectors. Obviously, the nonregular case is more subtle.

Admissible matrix sequences (cf. Definition 1.10) can be constructed in the same way also for arbitrary ordered matrix pairs $\{E, F\} = \{AD, B\}$, and we expect the sequence of matrices G_j to become stationary as in Example 1.9. Let us have a look at further simple special cases of constant coefficient DAEs.

We revisit the nonregular DAE

$$\begin{aligned} (x_1 + x_2)' + x_2 &= q_1, \\ x_4' &= q_2, \\ x_3 &= q_3, \\ x_3' &= q_4, \end{aligned} \tag{10.9}$$

discussed in Examples 1.9 and 1.11. The solutions as well as an admissible matrix sequence are described there. Again, as in the previous two examples (10.1), (10.2) and (10.3), the matrix G_0 already has maximal rank three, and hence the subspaces $\operatorname{im} G_i$ are stationary beginning with $i = 0$. In contrast, the sequence itself becomes stationary at level two, which means $G_i = G_2$ holds for all $i > 2$.
Now, we compare again three different projectors Q_0, starting the sequence with

$$G_0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}, \quad B_0 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

(a) Choose and compute

$$Q_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

(b) Choose and compute

$$Q_0 = \begin{bmatrix} 0 & -1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi_0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

(c) Choose and compute

$$Q_0 = \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \quad \Pi_0 = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} \frac{1}{2} & \frac{3}{2} & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}.$$

In all three cases, G_1 has rank $r_1 = r_0 = 3$ and the orthoprojector along $\operatorname{im} G_1 = \operatorname{im} G_0$ is simply

$$\mathcal{W}_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

How should one interpret this DAE which fails to be regular? There are different possibilities. The one which we prefer is the following: Consider the equation picked up by the projector $\mathcal{W}_0$, that is the third equation, as a consistency condition. The remaining system of three equations can then be seen as an explicit ODE for the components marked by the projectors Π_0, i.e., for $x_1 + x_2$, x_3, and x_4. The component recorded by the projector Q_0 can be considered as an arbitrary continuous function. This means that the DAE (10.9) is interpreted as having index zero (the level μ where the maximal subspace $\operatorname{im} G_\mu$ is reached first). However, again, different projectors Q_0 fix different components to be free, in the above three cases, x_1, x_2 and $\frac{1}{2}(x_1 - x_2)$, respectively. Furthermore, the resulting inherent ODE is affected by this.

In contrast to our view, the questions of which variable should be the free one and which equations should actually represent consistency conditions, can be answered in a different way. Considering the fourth equation of system (10.9) as the consistency condition and choosing x_2 to be free, the remaining three equations look like a regular index-1 DAE for the components $x_1 + x_2$,x_3, and x_4.
Furthermore, the last two equations of (10.9) somehow remain an index-2 problem, which is mirrored by the strangeness index (cf. [130]) of (10.9), which equals 1.

Consider now the underdetermined DAE

$$\begin{aligned} x_2' + x_1 &= q_1, \\ x_3' + x_2 &= q_2, \\ x_4' + x_3 &= q_3, \end{aligned} \tag{10.10}$$

with

$$G_0 = \begin{bmatrix} 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}, \; B_0 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 1\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \\ 0\,0\,0\,0 \end{bmatrix}, \; \Pi_0 = \begin{bmatrix} 0\,0\,0\,0 \\ 0\,1\,0\,0 \\ 0\,0\,1\,0 \\ 0\,0\,0\,1 \end{bmatrix}.$$

This matrix G_0 has full row rank, and no equation should be seen as a consistency condition. Again, the ranges $\operatorname{im} G_i$ are stationary at the beginning. The matrix sequence itself becomes stationary at level 3. We treat this DAE as an index-0 DAE for the components indicated by the projector Π_0; here x_2, x_3, x_4, and we see the variable indicated by Q_0, here x_1, to be free. Note that any other choice of the projector includes the variable x_1, too.

In contrast, choosing instead x_4 in (10.10) to be the free component, one arrives at a regular index-3 DAE for x_1, x_2, x_3.

Next take a look at the overdetermined DAE

$$\begin{aligned} x_1 &= q_1, \\ x_1' + x_2 &= q_2, \\ x_2' + x_3 &= q_3, \\ x_3' &= q_4 \end{aligned} \tag{10.11}$$

for which the first matrix G_0 is injective, and thus the matrix sequence is stationary at the beginning. Seeing the first equation in (10.11), which means the equation indicated by the projector $\mathcal{W}_0 = \operatorname{diag}(1,0,0,0)$, as a consistency condition, the other three equations in (10.11) can be treated as a regular index-0 DAE for the components x_1, x_2, x_3.

On the other hand, considering the last equation of (10.11) to be the consistency condition one arrives at a regular index-3 DAE for x_1, x_2, x_3. Note that the DAE (10.11) has strangeness index 3, while—as we will see in Section 10.2—the tractability index equals 0.
We stress once again the large scope of possible interpretations.

10.2 Linear DAEs

10.2.1 Tractability index

As in Chapter 2, we consider equations of the form

$$A(Dx)' + Bx = q, \tag{10.12}$$

with continuous coefficients

$$A \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^n, \mathbb{R}^k)), \quad D \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^n)), \quad B \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^k)),$$

and excitations $q \in \mathcal{C}(\mathcal{I}, \mathbb{R}^k)$. $\mathcal{I} \in \mathbb{R}$ is an interval. A solution of such an equation is a function belonging to the function space

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

which satisfies the DAE in the classical sense, that is, pointwise on the given interval.
The leading term in equation (10.12) is supposed to be *properly stated* on the interval $\mathcal{I}$. This means (cf. Definition 2.1) that the transversality condition

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{R}^n, \quad t \in \mathcal{I}, \tag{10.13}$$

is valid and $\operatorname{im} D$ and $\ker A$ are $\mathcal{C}^1$-subspaces.
Furthermore, the matrix function AD is assumed to have constant rank. Denote

$$r := \operatorname{rank} D(t), \quad \bar{k} := \operatorname{rank} [A(t)D(t)\ B(t)]\ t \in \mathcal{I}, \quad \rho := r + 1. \tag{10.14}$$

It holds that $\bar{k} \le k$. It would be reasonable to arrange $\bar{k} = k$ when creating a DAE.

The tractability index of a regular DAE is known to be the smallest integer μ such that an admissible matrix function sequence $G_0, \ldots, G_\mu$ associated with the DAE exists and G_μ is nonsingular (Definition 2.25). We intend to adapt this notion to be valid for general, possibly nonregular DAEs (10.12), too.
The construction of admissible matrix functions (cf. Definition 2.6) as well as the properties provided in Sections 2.2 and 2.3 already apply to the general DAE (10.12). Among the basic properties we find the inclusions

$$\operatorname{im} G_0 \subseteq \operatorname{im} G_1 \subseteq \cdots \subseteq \operatorname{im} G_i \subseteq \cdots \subseteq \operatorname{im} [AD\ B] \subseteq \mathbb{R}^k.$$

In the regular index-μ case, one has $k = m$ and $\bar{k} = k$, and G_μ has maximal possible rank m. Furthermore, the admissible matrix function sequence could be continued up to infinity by letting $Q_{\mu+i} = 0$, $G_{\mu+1+i} = G_{\mu+i} = G_\mu$ for $i \ge 0$, which shows that

$$\operatorname{im} G_0 \subseteq \operatorname{im} G_1 \subseteq \cdots \subseteq \operatorname{im} G_{\mu-1} \subset \operatorname{im} G_\mu = \operatorname{im} G_{\mu+1} = \cdots = \operatorname{im} [AD\ B] = \mathbb{R}^k.$$

For general linear DAEs (10.12), we intent to assign the tractability index to the smallest integer μ, for which an admissible matrix function sequence exists and μ is the smallest index such that the image inclusion becomes stationary, or, equivalently, G_μ has the maximal possible rank.

In this sense, all examples in the previous section possess tractability index 0.

How can we practically recognize the maximal rank? The ranks of the admissible matrix functions form a nondecreasing sequence $r = r_0 \le r_1 \le \cdots \le r_i$, but not necessarily a strictly increasing one. It may well happen that the ranks do not change in several consecutive steps. For instance, any Hessenberg size-μ DAE is characterized by the sequence $r_0 = \cdots = r_{\mu-1} < r_\mu$. This feature makes the task of recognizing the maximal rank and stopping the construction of the matrix functions in practice somewhat more subtle than previously thought.

Of course, if $\operatorname{im} G_\mu = \operatorname{im}[AD\ B]$ is reached, or equivalently $\mathcal{W}_\mu B = \{0\}$, then r_μ is maximal, and one can stop. If an injective G_μ is found, then one can stop because of the resulting stationarity $G_\mu = G_{\mu+1} = \cdots = G_{\mu+i}$.

Before we turn to the detailed definition, we provide the following useful assertion.

Proposition 10.1. *Suppose we are given the DAE (10.12) with continuous coefficients, a properly stated leading term, and the constants $r, \bar{k}, \rho$ from (10.14).*

(1) *If there is an admissible matrix function sequence $G_0, \ldots, G_\kappa$, such that*

$$G_\kappa = G_{\kappa+1},$$

then, letting $Q_{\kappa+i} := Q_\kappa$, $G_{\kappa+1+i} := G_\kappa$ for $i \geq 0$, the sequences $G_0, \ldots, G_{\kappa+j}$ are also admissible, and it holds that

$$G_\kappa = G_{\kappa+i}, \quad N_0 + \cdots + N_\kappa = N_0 + \cdots + N_{\kappa+i}.$$

(2) *If $G_0, \ldots, G_{\rho+1}$ is an admissible matrix function sequence, then*

$$G_\rho = G_{\rho+1}. \tag{10.15}$$

Proof. (1) $N_\kappa = N_{\kappa+1}$ implies $N_{\kappa+1} \subseteq N_0 + \cdots + N_\kappa$, $N_0 + \cdots + N_\kappa = N_0 + \cdots + N_{\kappa+1}, N_0 + \cdots + N_\kappa = X_\kappa \oplus N_\kappa = X_\kappa \oplus N_{\kappa+1}$, hence, choosing $X_{\kappa+1} := X_\kappa$, $Q_{\kappa+1} := Q_\kappa$ leads to $u_{\kappa+1} = u_\kappa$, $D\Pi_{\kappa+1}D^- = D\Pi_\kappa D^-$, so that $Q_0, \ldots, Q_\kappa, Q_{\kappa+1}$ are admissible, and further $B_{\kappa+1}Q_{\kappa+1} = B_{\kappa+1}\Pi_\kappa Q_{\kappa+1} = 0$, $G_{\kappa+2} = G_{\kappa+1}$, and so on.
(2) Let $G_0, \ldots, G_{\rho+1}$ be an admissible matrix function sequence. We decompose $N_i = \widehat{N}_i \oplus \mathcal{Y}_i$, which is accompanied by $(N_0 + \cdots + N_{i-1}) \cap \mathcal{Y}_i = \{0\}$. Namely, $z \in (N_0 + \cdots + N_{i-1}) \cap \mathcal{Y}_i$ yields $z \in (N_0 + \cdots + N_{i-1}) \cap N_i = \widehat{N}_i$, thus $z = 0$. It follows that $N_0 + \cdots + N_i = N_0 + \cdots + N_{i-1} + \mathcal{Y}_i = (N_0 + \cdots + N_{i-1}) \oplus \mathcal{Y}_i$, that is, the supplement to $N_0 + \cdots + N_{i-1}$ is exactly the subspace $\mathcal{Y}_i$, and therefore $\dim(N_0 + \cdots + N_i) = \dim(N_0 + \cdots + N_{i-1}) + \dim \mathcal{Y}_i$.
Next, if $\dim \mathcal{Y}_i \geq 1$ for all $j = 1, \ldots, \rho$, then

$$\dim(N_0 + \cdots + N_{\rho-1}) \geq \dim N_0 + r = m - r_0 + r = m.$$

In consequence, the subspaces $N_0 + \cdots + N_{\rho-1}$ and $N_0 + \cdots + N_\rho$ must coincide. This implies $N_\rho \subseteq N_0 + \cdots + N_{\rho-1}$, hence $\Pi_{\rho-1}Q_\rho = 0$, $B_\rho Q_\rho = B_\rho \Pi_{\rho-1} Q_\rho = 0$, $G_{\rho+1} = G_\rho$.

Otherwise, if there is an index $j_* \leq r$ such that $\dim \mathcal{Y}_i = 0$, then we have $N_{j_*} = \widehat{N}_{j_*} = N_{j_*} \cap (N_0 + \cdots + N_{j_*-1})$, and the inclusion $N_{j_*} \subseteq N_0 + \cdots + N_{j_*-1}$ is valid. This leads to $N_0 + \cdots + N_{j_*-1} = N_0 + \cdots + N_{j_*}$, and further to $G_{j_*} = G_{j_*+1}$. □

By Proposition 10.1 it is enough to provide an admissible matrix function sequence at most up to level ρ. We are looking for the smallest index μ such that $r_\mu = \operatorname{rank} G_\mu$ reaches the maximal possible value. Now we recognize the upper bound

$$\mu \le \rho = 1 + \operatorname{rank} D(t). \tag{10.16}$$

This bound is rigorous which is confirmed by Example 2.11 with $m_1 = m_2 = m_3 = 1$, $r_0 = 2$, and $\mu = 3$, i.e., $\mu = r_0 + 1 = \rho$.

Definition 10.2. Let the DAE (10.12) have continuous coefficients, a properly stated leading term and constants r, $\bar{k}$, ρ from (10.14).

(1) The DAE is said to be *tractable on $\mathcal{I}$ with index 0* if either $\operatorname{im} G_0 = \operatorname{im}[AD\ B]$ or there is an admissible matrix function sequence $G_0, \ldots, G_\rho$ such that

$$\operatorname{im} G_0 = \cdots = \operatorname{im} G_\rho.$$

(2) The DAE is said to be *tractable on $\mathcal{I}$ with index $\mu \in \mathbb{N}$*, if either there is an admissible matrix function sequence $G_0, \ldots, G_\mu$ with

$$\operatorname{im} G_\mu = \operatorname{im}[AD\ B],$$

or there is an admissible matrix function sequence $G_0, \ldots, G_\rho$ with

$$\operatorname{im} G_\mu = \cdots = \operatorname{im} G_\rho \subset \operatorname{im}[AD\ B].$$

and μ is the smallest integer of this kind.

(3) The DAE is *regular on $\mathcal{I}$ with tractability index $\mu \in \mathbb{N} \cup \{0\}$*, if it is tractable with index μ, and, additionally $m = k = \bar{k}$ and $\operatorname{im} G_\mu = \mathbb{R}^m$.

This definition generalizes Definition 2.25. Item (3) repeats Definition 2.25 for completeness.
The special examples (10.1), (10.2) and (10.3) show DAEs that are tractable with index zero.
From our point of view one should take care to attain the condition $\operatorname{im}[AD\ B] = \mathbb{R}^k$ during the modeling.

A case of particular interest is given if one meets a matrix functions G_μ that is injective. This can only happen if $k \ge m$. Then, the tractability index is the smallest integer μ such that G_μ is injective, thus $r_\mu = m$. It is worth mentioning that then $u_0 = \cdots = u_{\mu-1} = 0$, i.e., the intersections $\widehat{N}_i$ are trivial.

If the complement subspace X_1 used for the construction of the admissible projector function Q_1 is trivial, then it holds that $G_i = G_0$ for all $i \ge 1$, and the DAE is tractable with index zero and therefore, if $X_1 = \{0\}$, then one can stop. Namely, $X_1 = \{0\}$ means $N_1 \cap N_0 = N_0$. This implies $N_0 \subseteq N_1$, and $N_0 = N_1$ because of the dimensions $\dim N_0 = m - r_0 \ge m - r_1 = \dim N_1$. Choose $Q_1 := Q_0$. The projector functions Q_0, Q_1 are admissible. It follows that $0 = G_1 Q_1 = G_0 Q_1 + B_0 Q_0 Q_1 = B_0 Q_0$, thus $G_1 = G_0$ and $G_2 = G_1 + B_1 Q_1 = G_1 + B_1 P_0 Q_1 = G_1$. Then we set $Q_2 := Q_1$, and so on. In particular, it follows that $X_i = \{0\}$ for all $i \ge 1$.
Notice that, if there is a trivial complement subspace X_κ in a matrix function sequence, then these subspaces X_i must be trivial for all i.

10.2.2 General decoupling

We continue to investigate the rearranged version (2.38) of the DAE (10.12) obtained in Subsection 2.4.1, and we provide a refined form which serves then as a basis of further decouplings.

Proposition 10.3. *Suppose we are given the DAE (10.12) with continuous coefficients and a properly stated leading term such that (10.14).*
Then, if $G_0,\dots,G_{\kappa+1}$ represent an admissible matrix function sequence associated to this DAE, $\kappa\in\mathbb{N}$, the DAE can be rewritten as

$$G_\kappa D^-(D\Pi_\kappa x)'+B_\kappa x+G_\kappa\sum_{l=0}^{\kappa-1}\big\{Q_l x-(I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)' + \mathcal{V}_l D\Pi_l x+\mathcal{U}_l(D\Pi_l x)'\big\}=q \tag{10.17}$$

with coefficients

$$\mathcal{U}_l := -(I-\Pi_l)\big\{Q_l+Q_{l+1}(I-\Pi_l)Q_{l+1}P_l\big\}\Pi_l D^-,$$
$$\mathcal{V}_l := (I-\Pi_l)\big\{(P_l+Q_{l+1}Q_l)D^-(D\Pi_l D^-)'-Q_{l+1}D^-(D\Pi_{l+1}D^-)'\big\}D\Pi_l D^-.$$

Before we verify this assertion, we point out that the coefficients $\mathcal{V}_l$ are caused by variations in time, so that these coefficients vanish in the constant coefficient case.

The coefficients $\mathcal{U}_l$ disappear, if the intersections $\widehat{N}_1,\dots,\widehat{N}_l$ are trivial.

If the intersections $\widehat{N}_1,\dots,\widehat{N}_\kappa$ are trivial, then it follows (cf. Proposition 2.23) that $\mathcal{V}_l=V_l$, $l=1,\dots,\kappa$.

Proof. Proposition 2.23 provides the rearranged version (2.38) of the DAE (10.12), that is

$$G_\kappa D^-(D\Pi_\kappa x)'+B_\kappa x+G_\kappa\sum_{l=0}^{\kappa-1}\big\{Q_l x+(I-\Pi_l)(P_l-Q_{l+1}P_l)D^-(D\Pi_l x)'\big\}=q. \tag{10.18}$$

For $\kappa=1$ we compute

$$\begin{aligned}G_1(I-\Pi_0)(P_0-Q_1P_0)D^-(D\Pi_0x)' &= -G_1(I-\Pi_0)Q_1D^-(D\Pi_0x)'\\ &= -G_1(I-\Pi_0)Q_1D^-(D\Pi_0Q_1x)'-G_1(I-\Pi_0)Q_1D^-(D\Pi_1D^-D\Pi_0x)'\\ &= -G_1(I-\Pi_0)Q_1D^-(D\Pi_0Q_1x)'+G_1\mathcal{V}_0D\Pi_0x+G_1\mathcal{U}_0(D\Pi_0x)'\end{aligned}$$

with

$$\mathcal{U}_0 = -(I-\Pi_0)Q_1\Pi_1D^- = -(I-\Pi_0)\{Q_0+Q_1(I-\Pi_0)Q_1P_0\}\Pi_0D^-,$$
$$\mathcal{V}_0 = -(I-\Pi_0)Q_1D^-(D\Pi_1D^-)'D\Pi_0D^-.$$

Now we assume $\kappa>1$. First we take a closer look at the expressions

$$\mathcal{E}_l := (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l x)', \quad 0 \le l \le \kappa - 1.$$

Compute

$$\begin{aligned}
\mathcal{E}_l &= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-\big((D\Pi_l D^-)'D\Pi_l x + D\Pi_l D^-(D\Pi_l x)'\big)\\
&= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)'D\Pi_l x\\
&\qquad + (I-\Pi_l)(-Q_l - Q_{l+1}P_l)\Pi_l D^-(D\Pi_l x)'\\
&= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)'D\Pi_l x - (I-\Pi_l)Q_l\Pi_l D^-(D\Pi_l x)'\\
&\qquad - (I-\Pi_l)Q_{l+1}\{\Pi_l + I - \Pi_l\}Q_{l+1}P_l\Pi_l D^-(D\Pi_l x)'\\
&= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)'D\Pi_l x\\
&\qquad - (I-\Pi_l)(Q_l + Q_{l+1}\{\Pi_l + I - \Pi_l\}Q_{l+1}P_l)\Pi_l D^-(D\Pi_l x)'
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{E}_l &= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)'D\Pi_l x\\
&\qquad \underbrace{-(I-\Pi_l)(Q_l + Q_{l+1}\{I-\Pi_l\}Q_{l+1}P_l)\Pi_l D^-}_{\mathcal{U}_l}(D\Pi_l x)'\\
&\qquad - (I-\Pi_l)Q_{l+1}\underbrace{\Pi_l Q_{l+1}P_l\Pi_l D^-}_{\Pi_l Q_{l+1}}(D\Pi_l x)'\\
&= (I-\Pi_l)(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)'D\Pi_l x + \mathcal{U}_l(D\Pi_l x)'\\
&\quad - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)' + (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}D^-)'D\Pi_l x\\
&= (I-\Pi_l)\big\{(P_l - Q_{l+1}P_l)D^-(D\Pi_l D^-)' + Q_{l+1}D^-(D\underbrace{\Pi_l Q_{l+1}}_{\Pi_l - \Pi_{l+1}}D^-)'\big\}D\Pi_l x\\
&\qquad + \mathcal{U}_l(D\Pi_l x)' - (I-\Pi_l)\big\{Q_{l+1}D^-(D\Pi_l Q_{l+1}D^-)'\big\}\\
&= \mathcal{U}_l(D\Pi_l x)' - (I-\Pi_l)\big\{Q_{l+1}D^-(D\Pi_l Q_{l+1}D^-)'\big\}\\
&\quad + \underbrace{(I-\Pi_l)\big\{(P_l - Q_{l+1}P_l + Q_{l+1})D^-(D\Pi_l D^-)' - Q_{l+1}D^-(D\Pi_{l+1}D^-)'\big\}D\Pi_l}_{\mathcal{V}_l D}x.
\end{aligned}$$

In consequence, the representation (10.18) is nothing else than

$$\begin{aligned}
G_\kappa D^-(D\Pi_\kappa x)' + B_\kappa x + G_\kappa \sum_{l=0}^{\kappa-1}\{&Q_l x - (I-\Pi_l)Q_{l+1}D^-(D\Pi_l Q_{l+1}x)'\\
&+ \mathcal{V}_l D\Pi_l x + \mathcal{U}_l(D\Pi_l x)'\} = q,
\end{aligned}$$

which completes the proof. □

Throughout the rest of this section the DAE (10.12) is supposed to be tractable with index μ, and $Q_0, \ldots, Q_{\mu-1}$ denote admissible projector functions. We make use of the rearranged version (10.17) of (10.12).

$$G_{\mu-1}D^-(D\Pi_{\mu-1}x)' + B_{\mu-1}x \tag{10.19}$$
$$+G_{\mu-1}\sum_{\ell=0}^{\mu-2}\{Q_\ell x - (I-\Pi_\ell)Q_{\ell+1}D^-(D\Pi_\ell Q_{\ell+1}x)' + \mathcal{V}_\ell D\Pi_\ell x + \mathcal{U}_\ell(D\Pi_\ell x)'\} = q,$$

and the coefficients $\mathcal{V}_\ell, \mathcal{U}_\ell$ are from Proposition 10.3. By expressing

$$G_{\mu-1}Q_\ell = G_\mu Q_\ell,\ G_{\mu-1}\mathcal{V}_\ell = G_\mu \mathcal{V}_\ell,\ G_{\mu-1}\mathcal{U}_\ell = G_\mu \mathcal{U}_\ell,\ \ell = 0,\dots,\mu-2,$$

$$B_{\mu-1} = B_{\mu-1}P_{\mu-1} + B_{\mu-1}Q_{\mu-1} = B_{\mu-1}D^-D\Pi_{\mu-1} + B_{\mu-1}Q_{\mu-1},$$

formula (10.19) becomes

$$G_\mu\Big\{P_{\mu-1}D^-(D\Pi_{\mu-1}x)' + Q_{\mu-1}x \tag{10.20}$$
$$+\sum_{\ell=0}^{\mu-2}\big\{Q_\ell x - (I-\Pi_\ell)Q_{\ell+1}D^-(D\Pi_\ell Q_{\ell+1}x)' + \mathcal{V}_\ell D\Pi_\ell x + \mathcal{U}_\ell(D\Pi_\ell x)'\big\}\Big\}$$
$$+B_{\mu-1}D^-D\Pi_{\mu-1}x = q.$$

According to the definition of the tractability index μ, the matrix function G_μ has constant rank. We find a continuous generalized inverse G_μ^-, and a projector function $\mathcal{W}_\mu = I - G_\mu G_\mu^-$ along $\operatorname{im} G_\mu$. Notice that there is no need for the resulting projector function $G_\mu^- G_\mu$ to be also admissible. The projector functions $G_\mu G_\mu^-$ and $\mathcal{W}_\mu$ split the DAE (10.20) into two parts. Multiplication by $\mathcal{W}_\mu$ leads to equation (10.22) below. Multiplication by $G_\mu G_\mu^-$ yields

$$G_\mu\Big\{P_{\mu-1}D^-(D\Pi_{\mu-1}x)' + Q_{\mu-1}x$$
$$+\sum_{\ell=0}^{\mu-1}\{Q_\ell x - (I-\Pi_\ell)Q_{\ell+1}D^-(D\Pi_\ell Q_{\ell+1}x)' + \mathcal{V}_\ell D\Pi_\ell x + \mathcal{U}_\ell(D\Pi_\ell x)'\}$$
$$+G_\mu^- B_{\mu-1}D^-D\Pi_{\mu-1}x - G_\mu^- q\Big\} = 0.$$

This equation, $G_\mu\{\dots\} = 0$, may be rewritten as $\{\dots\} =: y$, where y denotes an arbitrary continuous function such that $G_\mu y = 0$. Altogether this leads to the system

$$P_{\mu-1}D^-(D\Pi_{\mu-1}x)' + Q_{\mu-1}x + \sum_{\ell=0}^{\mu-2}\{Q_\ell x - (I-\Pi_\ell)Q_{\ell+1}D^-(D\Pi_\ell Q_{\ell+1}x)' \tag{10.21}$$
$$+\mathcal{V}_\ell D\Pi_\ell x + \mathcal{U}_\ell(D\Pi_\ell x)'\} + y = G_\mu^-(q - B_{\mu-1}D^-D\Pi_{\mu-1}x),$$
$$\mathcal{W}_\mu B_{\mu-1}D^-D\Pi_{\mu-1}x = \mathcal{W}_\mu q, \tag{10.22}$$

where y can be chosen arbitrarily such that $G_\mu y = 0$. The relation

$$\ker G_\mu = (I - G_{\mu-1}^- B_{\mu-1}Q_{\mu-1})(N_{\mu-1} \cap S_{\mu-1}) \tag{10.23}$$

might be helpful. The undetermined part of y is actually $Q_{\mu-1}y \in N_{\mu-1} \cap S_{\mu-1}$.

Multiplication of (10.21) by projector functions uncovers some further structures. In particular, multiplication by $\Pi_{\mu-1}$ yields

$$\Pi_{\mu-1}D^-(D\Pi_{\mu-1}x)' + \Pi_{\mu-1}y = \Pi_{\mu-1}G_\mu^-(q - B_{\mu-1}D^-D\Pi_{\mu-1}x),$$

hence we recognize an inherent explicit regular ODE with respect to $D\Pi_{\mu-1}x$, namely

$$(D\Pi_{\mu-1}x)' - (D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}x + D\Pi_{\mu-1}y \\ + D\Pi_{\mu-1}G_\mu^-B_{\mu-1}D^-D\Pi_{\mu-1}x = D\Pi_{\mu-1}G_\mu^-q.$$

It is worth mentioning again that, in contrast to regular DAEs, the properties of the flow of this ODE may depend on the choice of the admissible projector functions, as it is the case for example (10.3).
Multiplying (10.21) by $\Pi_{\mu-2}Q_{\mu-1}$ gives

$$\Pi_{\mu-2}Q_{\mu-1}x + \Pi_{\mu-2}Q_{\mu-1}y + \Pi_{\mu-2}Q_{\mu-1}G_\mu^-B_{\mu-1}D^-D\Pi_{\mu-1}x = \Pi_{\mu-2}Q_{\mu-1}G_\mu^-q.$$

Apart from the terms including y, these two formulas are the counterparts of the corresponding ones in Section 2.6 for regular DAEs. However, the further equations that will be derived from (10.21) by multiplication with the additional projector functions are much more expensive to elaborate. We restrict ourselves to several case studies. For the case of index 0 we refer once again to Subsection 2.4.1 and the illustrative Example 2.24 therein.

10.2.2.1 G_μ has full column rank

This case can happen only if $k \geq m$, and $r_\mu = m$ holds true. Since G_μ is injective, due to Proposition 2.23, all intersections $(N_0 + \cdots + N_{i-1}) \cap N_i$, $i = 1, \ldots, \mu-1$, are trivial, the components $\mathcal{U}_0, \ldots, \mathcal{U}_{\mu-2}$ vanish, and $\mathcal{V}_\ell$ simplifies to $\mathcal{V}_\ell = V_\ell$, $\ell = 0, \ldots, \mu-2$. Moreover, $G_\mu y = 0$ implies $y = 0$.

The resulting special equation (10.21) reads

$$P_{\mu-1}D^-(D\Pi_{\mu-1}x)' + Q_{\mu-1}x \\ + \sum_{\ell=0}^{\mu-2}\left\{Q_\ell x - (I-\Pi_\ell)Q_{\ell+1}D^-(D\Pi_\ell Q_{\ell+1}x)' + V_\ell D\Pi_\ell x\right\} \qquad (10.24) \\ + G_\mu^-B_{\mu-1}D^-D\Pi_{\mu-1}x = G_\mu^-q.$$

For $k = m$, that is, for regular DAEs with tractability index μ, this formula coincides in essence with formula (2.48) (several terms are arranged in a different way).

Applying the decoupling procedure from Section 2.6, we can prove (10.24) to represent a regular index-μ DAE. Completed by an initial condition

$$D(t_0)\Pi_{\mu-1}(t_0)x(t_0) = z_0 \in \operatorname{im} D(t_0)\Pi_{\mu-1}(t_0), \tag{10.25}$$

equation (10.24) is uniquely solvable for x. This suggests the option of considering equation (10.24) to determine the solution x, and to treat the additional equation (10.22) as the resulting consistency condition.

Example 10.4 (Index-1 DAE). Set $m = 2$, $k = 3$, $n = 1$, and write the system

$$\begin{aligned} x_1' + x_2 &= q_1, \\ x_2 &= q_2, \\ x_2 &= q_3, \end{aligned} \tag{10.26}$$

as DAE (10.12) such that

$$A = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad D = [1\ 0], \quad G_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

$$G_1 = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{bmatrix}, \quad G_1^- = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

G_1 already has maximal possible rank, $r_1 = 2$, and hence this DAE is tractable with index 1. The consistency equation $\mathcal{W}_1(B\Pi_0 x - q) = 0$ means here $q_2 = q_3$. Equation (10.24) has the form

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} x_1' + \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} q_1 - q_2 \\ q_2 \end{bmatrix},$$

which is a regular index-1 DAE. □

Example 10.5 (Index-1 DAE). Set $m = 2$, $k = 3$, $n = 1$ and put the system

$$\begin{aligned} x_1' + x_2 &= q_1, \\ x_1 &= q_2, \\ x_2 &= q_3, \end{aligned} \tag{10.27}$$

into the form (10.12). This leads to

$$A = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad D = [1\ 0], \quad G_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$

$$G_1 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad G_1^- = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 0 & 1 \end{bmatrix}.$$

G_1 has maximal rank, $r_1 = 2$, this DAE is tractable with index 1. The condition $\mathcal{W}_1(B_0\Pi_0 x - q) = 0$ now means $x_1 = q_2$, and equation (10.24) specializes to

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} x_1' + \begin{bmatrix} 0 \\ x_2 \end{bmatrix} = \begin{bmatrix} q_1 - q_3 \\ q_3 \end{bmatrix},$$

which is a regular index-1 DAE. □

Example 10.6 ($\mu = 0$). Set $k = 5$, $m = 4$, $n = 4$, and put the DAE

$$\begin{aligned} x_1' &= q_1, \\ x_2' + x_1 &= q_2, \\ x_3' + x_2 &= q_3, \\ x_4' + x_3 &= q_4, \\ x_4 &= q_5, \end{aligned} \tag{10.28}$$

into the form (10.12). This yields

$$G_0 = \begin{bmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \\ 0&0&0&0 \end{bmatrix}, \quad B = \begin{bmatrix} 0&0&0&0 \\ 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{bmatrix}, \quad \mathcal{W}_0 = \begin{bmatrix} 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&0 \\ 0&0&0&0&1 \end{bmatrix},$$

and $\mu = 0$. This DAE is interpreted as an explicit ODE for the components x_1, x_2, x_3, x_4 and the consistency condition $x_4 = q_5$. □

Example 10.7 ($\mu = 2$). The DAE

$$\begin{aligned} x_2' + x_1 &= q_1, \\ x_3' + x_2 &= q_2, \\ x_3 &= q_3, \\ x_3' &= q_3', \end{aligned} \tag{10.29}$$

results from the index-3 system

$$\begin{aligned} x_2' + x_1 &= q_1, \\ x_3' + x_2 &= q_2, \\ x_3 &= q_3, \end{aligned} \tag{10.30}$$

by adding the differentiated version of the derivative-free equation. We may write (10.29) in the form (2.1) with $k = 4$, $m = 3$, $n = 2$,

$$A = \begin{bmatrix} 1&0 \\ 0&1 \\ 0&0 \\ 0&1 \end{bmatrix}, \; D = \begin{bmatrix} 0&1&0 \\ 0&0&1 \end{bmatrix}, \; B = \begin{bmatrix} 1&0&0 \\ 0&1&0 \\ 0&0&1 \\ 0&0&0 \end{bmatrix}, \; D^- = \begin{bmatrix} 0&0 \\ 1&0 \\ 0&1 \end{bmatrix}.$$

Compute

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; G_1 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; G_2 = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

$r_0 = 2$, $r_1 = 2$, $r_2 = 3$. It follows that (10.29) has tractability index 2 while (10.30) has tractability index 3.
System (10.29) is overdetermined, and, in our view, the subsystem $\mathcal{W}_2 Bx = \mathcal{W}_2 q$ (cf. (10.22)), which means here in essence $x_3 = q_3$, is interpreted as a consistency condition. The main part (10.24) of the DAE reads

$$\begin{aligned} x_2' + x_1 &= q_1, \\ x_2 &= q_2 - q_3', \\ x_3' &= q_3', \end{aligned}$$

and this is obviously a regular index 2 DAE. □

The last example addresses an interesting general phenomenon: If one adds to a given DAE the differentiated version of a certain part of the derivative-free equations, then the tractability index reduces.
There are several possibilities to choose appropriate derivative-free equations to be differentiated. Here we concentrate on the part

$$\mathcal{W}_{\mu-1} Bx = \mathcal{W}_{\mu-1} q,$$

supposing the original DAE (2.1) to have tractability index $\mu \geq 2$.
Considering the inclusion $N_0 \subseteq S_1 \subseteq S_{\mu-1} = \ker \mathcal{W}_{\mu-1} B$ we can write this derivative-free part as

$$\mathcal{W}_{\mu-1} BD^- Dx = \mathcal{W}_{\mu-1} q,$$

and differentiation yields

$$\mathcal{W}_{\mu-1} BD^- (Dx)' + (\mathcal{W}_{\mu-1} BD^-)' Dx = (\mathcal{W}_{\mu-1} q)'. \tag{10.31}$$

The enlarged DAE (10.12), (10.31) is now

$$\underbrace{\begin{bmatrix} A \\ \mathcal{W}_{\mu-1} BD^- \end{bmatrix}}_{=:\tilde{A}} (Dx)' + \underbrace{\begin{bmatrix} B \\ (\mathcal{W}_{\mu-1} BD^-)' D \end{bmatrix}}_{=:\tilde{B}} x = \begin{bmatrix} q \\ (\mathcal{W}_{\mu-1} q)' \end{bmatrix}, \tag{10.32}$$

with $k + m =: \tilde{k}$ equations. The DAE (10.32) inherits the properly stated leading term from (10.12) because of $\ker \tilde{A} = \ker A$.
The next proposition says that the tractability index of (10.32) is less by 1 than that of (10.12).

Proposition 10.8. *If the DAE (10.12) has tractability index $\mu \geq 2$ and characteristic values $r_0 \leq \cdots \leq r_{\mu-1} < r_\mu = m$, then the DAE (10.32) has tractability index $\tilde{\mu} = \mu - 1$, and characteristic values $\tilde{r}_i = r_i$, $i = 0, \ldots, \tilde{\mu} - 1$, $\tilde{r}_{\tilde{\mu}} = \tilde{r}_{\mu-1} = m$.*

Proof. We have $N_0 \subseteq \ker \mathcal{W}_{\mu-1}B = S_{\mu-1}$,

$$\tilde{G}_0 = \tilde{A}D = \begin{bmatrix} AD \\ \mathcal{W}_{\mu-1}BD^-D \end{bmatrix} = \begin{bmatrix} G_0 \\ \mathcal{W}_{\mu-1}B \end{bmatrix},\ \tilde{r}_0 = r_0.$$

Set $\tilde{Q}_0 = Q_0$ and form $\tilde{G}_1 = \tilde{G}_0 + \tilde{B}\tilde{Q}_0 = \begin{bmatrix} G_1 \\ \mathcal{W}_{\mu-1}B \end{bmatrix}$.
If $\mu = 2$, then $\ker \tilde{G}_1 = \ker G_1 \cap \ker \mathcal{W}_1 B = N_1 \cap S_1 = \{0\}$. Then, $\tilde{r}_1 = m$, $\tilde{r}_0 < \tilde{r}_1$, and hence the new DAE (10.32) has tractability index 1, and we are ready.
If $\mu \geq 3$ then $\ker \tilde{G}_1 = \ker G_1 \cap \ker \mathcal{W}_{\mu-1}B = N_1$, since $N_1 \subseteq S_2 \subseteq S_{\mu-1} = \ker \mathcal{W}_{\mu-1}B$. Moreover, $\tilde{r}_1 = r_1$.
Set $\tilde{Q}_1 = Q_1$ and form

$$\tilde{B}_1 = \begin{bmatrix} B_1 \\ (\mathcal{W}_{\mu-1}BD^-)'D - \mathcal{W}_{\mu-1}BD^-(D\Pi_1 D^-)'D \end{bmatrix} = \begin{bmatrix} B_1 \\ (\mathcal{W}_{\mu-1}BD^-)'D\Pi_1 \end{bmatrix},$$
$$\tilde{G}_2 = \begin{bmatrix} G_1 + B_1Q_1 \\ \mathcal{W}_{\mu-1}B \end{bmatrix} = \begin{bmatrix} G_2 \\ \mathcal{W}_{\mu-1}B \end{bmatrix},\ \tilde{N}_2 = N_2 \cap S_{\mu-1}.$$

If $\mu = 3$, then $\tilde{N}_2 = N_2 \cap S_2 = \{0\}$, and $\tilde{r}_2 = m$, i.e., $\tilde{G}_2$ is injective, and the DAE (10.32) has tractability index 2.
For $\mu > 3$, as long as $j \leq \mu - 2$, it follows that

$$\tilde{G}_j = \begin{bmatrix} G_j \\ \mathcal{W}_{\mu-1}B \end{bmatrix},\ \tilde{N}_j = N_j \cap S_{\mu-1} = N_j,\ \tilde{Q}_j = Q_j,\ \tilde{r}_j = r_j,$$
$$\tilde{B}_j = \begin{bmatrix} B_j \\ (\mathcal{W}_{\mu-1}BD^-)'D\Pi_{j-1} - \mathcal{W}_{\mu-1}BD^-(D\Pi_j D^-)'D\Pi_{j-1} \end{bmatrix}$$
$$= \begin{bmatrix} B_j \\ (\mathcal{W}_{\mu-1}BD^-)'D\Pi_j \end{bmatrix}.$$

Finally,

$$\tilde{G}_{\mu-1} = \begin{bmatrix} G_{\mu-1} \\ \mathcal{W}_{\mu-1}B \end{bmatrix},\ \tilde{N}_{\mu-1} = N_{\mu-1} \cap S_{\mu-1} = \{0\},\ \tilde{r}_{\mu-1} = m,$$

that is, $\tilde{G}_{\mu-1}$ is injective, and the DAE (10.32) has tractability index $\tilde{\mu} = \mu - 1$. □

We mention that $\tilde{\mathcal{W}}_{\tilde{\mu}} = \begin{bmatrix} \mathcal{W}_{\mu-1} & \\ & I - \mathcal{W}_{\mu-1} \end{bmatrix}$ is a projector function with $\ker \tilde{\mathcal{W}}_{\tilde{\mu}} = \operatorname{im} \tilde{G}_{\tilde{\mu}}$, and now the equation $\mathcal{W}_{\mu-1}Bx = \mathcal{W}_{\mu-1}q$ is interpreted as the consistency condition, whereas its differentiated version is included into the main part (10.24), as in Example 10.7.

10.2.2.2 Tractability index 1, G_1 has a nontrivial nullspace

The decomposed system (10.21), (10.22) has the form

$$D^-(Dx)' + Q_0x + y + G_1^-B_0D^-Dx = G_1^-q \tag{10.33}$$
$$\mathcal{W}_1B_0D^-Dx = \mathcal{W}_1q, \tag{10.34}$$

with $G_1y = 0$, i.e., $y = (I - G_0^-B_0Q_0)Q_0y$, $Q_0y \in N_0 \cap S_0$. The inherent explicit ODE is here

$$(Dx)' - R'Dx + Dy + DG_1^-BD^-Dx = DG_1^-q, \tag{10.35}$$

and multiplication of (10.33) by Q_0 gives

$$Q_0x + Q_0y + Q_0G_1^-B_0D^-Dx = Q_0G_1^-q. \tag{10.36}$$

For each arbitrarily fixed continuous $Q_0y \in N_0 \cap S_0$, equation (10.33) represents a regular index-1 DAE.

We consider (10.34) as a consistency condition. If $\operatorname{im} G_1 = \mathbb{R}^k$, $m \geq k$, are true, i.e., if G_1 has full row rank, then this condition disappears.
A regular index-1 DAE is solvable for each arbitrary continuous excitation q. The same holds true for general linear DAEs with tractability index 1 and full row-rank matrix function G_1.

Proposition 10.9. *Let the DAE (10.12) have continuous coefficients, a properly stated leading term, and the constants r, $\bar{k}$, ρ from (10.14); further $\bar{k} = k < m$. If the DAE has tractability index 1, and G_1 has full row rank, then the IVP*

$$A(Dx)' + Bx = q, \quad D(t_0)x(t_0) = z_0$$

is solvable for each arbitrary continuous excitation q and initial data $z_0 \in \operatorname{im} D(t_0)$.

Proof. There is no consistency condition (10.34) in this case. Put $y = 0$ in the inherent ODE (10.35) and in expression (10.36). In this way, taking any solution of the explicit ODE (10.35), satisfying the initial condition, and computing then Q_0x from (10.36), one obtains with $x := D^-Dx + Q_0x$ a solution of the DAE. □

Example 10.10 (Strangeness-reduced form). Set $m = m_1 + m_2 + m_3$, $k = k_1 + k_2 + k_3$, $n = m_1$, $m_1 = k_1$, $m_2 = k_2$, $k_3 \geq 0$, $m_3 \geq 0$, and consider the DAE (10.12) with the coefficients

$$A = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} I & 0 & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} I \\ 0 \\ 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 & B_{13} \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

which has the detailed form

$$\begin{aligned} x_1' + B_{13}x_3 &= q_1, \\ x_2 &= q_2, \\ 0 &= q_3. \end{aligned}$$

This special DAE plays its role in the strangeness index framework (e.g., [130]). Derive

$$G_0 = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \quad G_1 = \begin{bmatrix} I & 0 & B_{13} \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathcal{W}_1 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{bmatrix},$$

and $r_0 = m_1$, $r_1 = m_1 + m_2$ and $\operatorname{im} G_1 = \operatorname{im}[AD\ B] = \mathbb{R}^{m_1} \times \mathbb{R}^{m_1} \times \{0\}$. Therefore, G_1 has maximal possible rank, and hence the problem is tractable with index 1. The consistency condition (10.34) means simply $0 = q_3$, if $k_3 > 0$. It disappears for $k_3 = 0$.

Moreover, here we have $N_0 = \{z \in \mathbb{R}^m : z_1 = 0\}$, $S_0 = \{z \in \mathbb{R}^m : z_2 = 0\}$, $N_0 \cap S_0 = \{z \in \mathbb{R}^m : z_1 = 0, z_2 = 0\}$. $G_1 y = 0$ means $y_1 + B_{13} y_3 = 0$, $y_2 = 0$. The free component $Q_0 y \in N_0 \cap S_0$ is actually y_3 (if $m_3 > 0$), so that $y_1 = -B_{13} y_3$ follows.

It results that

$$G_1^- = \begin{bmatrix} I & & \\ & I & \\ & & 0 \end{bmatrix}, \quad G_1^- B_0 D^- = 0,$$

and equation (10.33) reads in detail

$$\begin{aligned} x_1' - B_{13} y_3 &= q_1, \\ x_2 &= q_2, \\ x_3 + y_3 &= 0. \end{aligned}$$

For each given function y_3, this is obviously a regular index-1 DAE. □

The characteristic values r_i as well as the tractability index are invariant under regular scalings and transformations of the unknown function (cf. Section 2.3). We derive a result on the special structure of an index-1 DAE via transformations.

Proposition 10.11. *Let $m > k$, and let the DAE (10.12) have tractability index 1. Then there are nonsingular matrix functions $L \in \mathcal{C}(J, L(\mathbb{R}^k))$, $L^* = L^{-1}$, $K \in \mathcal{C}(J, L(\mathbb{R}^m))$, $K^* = K^{-1}$, such that the premultiplication by L and the transformation of the unknown function $x = K\bar{x}$, $\bar{x} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} \begin{matrix} \} r_1 \\ \} m - r_1 \end{matrix}$, lead to the equivalent DAE*

$$\bar{A}_1 (\bar{D}_1 \bar{x}_1)' + \bar{B}_{11} \bar{x}_1 + \bar{B}_{12} \bar{x}_2 = \bar{q}_1, \tag{10.37}$$

$$\bar{B}_{21} \bar{x}_1 \qquad\qquad = \bar{q}_2, \tag{10.38}$$

with

$$LA = \begin{bmatrix} \bar{A}_1 \\ 0 \end{bmatrix}, \quad DK = \begin{bmatrix} \bar{D}_1 & 0 \end{bmatrix}, \quad LBK = \begin{bmatrix} \bar{B}_{11} & \bar{B}_{12} \\ \bar{B}_{21} & 0 \end{bmatrix}, \quad Lq = \begin{bmatrix} \bar{q}_1 \\ \bar{q}_2 \end{bmatrix} \begin{matrix} \} r_1 \\ \} k - r_1 \end{matrix},$$

and equation (10.37) is a regular DAE with tractability index 1 with respect to $\bar{x}_1$. If $r_1 = k$, i.e., if G_1 has full row rank, then the second equation (10.38) disappears. In general, it holds that $\ker \bar{B}_{21} \supseteq \ker \bar{D}_1$.

Proof. We choose Q_0, $\mathcal{W}_0$ to be the orthogonal projectors onto N_0 and $\operatorname{im} G_0$, and consider the matrix function

$$\mathcal{G}_1 = G_0 + \mathcal{W}_0 B Q_0,$$

which has constant rank r_1. Compute L so that

$$L\mathcal{G}_1 = \begin{bmatrix} \check{\mathcal{G}}_1 \\ 0 \end{bmatrix} \begin{matrix} \} r_1 \\ \} k - r_1 \end{matrix}, \quad \operatorname{rank} \check{\mathcal{G}}_1 = r_1.$$

Then we provide a K to obtain

$$\check{\mathcal{G}}_1 K = [\underbrace{\mathcal{S}}_{r_1} \quad \underbrace{0}_{m-r_1}], \qquad \mathcal{S} \text{ nonsingular.}$$

This yields

$$L(G_0 + \mathcal{W}_0 B Q_0)K = \begin{bmatrix} \mathcal{S} & 0 \\ 0 & 0 \end{bmatrix}, \quad L(G_0 + \mathcal{W}_0 B Q_0)K \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} = 0,$$

and further $G_0 K \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} = 0$, $\mathcal{W}_0 B Q_0 K \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} = 0$, $P_0 K \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} = 0$, $DK \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix} = 0$.
In particular, $\bar{D} := DK = [\bar{D}_1 \;\; 0]$ must be true, and $\operatorname{im} \bar{D}_1 = \operatorname{im} D$. Denoting $\tilde{P}_0 := \bar{D}_1^+ \bar{D}_1$, $\tilde{Q}_0 := I - \tilde{P}_0 \in \mathcal{C}(\mathcal{I}, L(\mathbb{R}^{r_1}))$ we find $\bar{Q}_0 = K^* Q_0 K = \begin{bmatrix} \tilde{Q}_0 & 0 \\ 0 & I \end{bmatrix}$ to be the orthogonal projector onto $\ker \bar{D} = K^* \ker D$.
Next we scale the DAE (10.12) by L and transform $x = K\bar{x}$. Because of $\operatorname{im} A \subseteq \operatorname{im} \mathcal{G}_1$, we must have

$$\bar{A} := LA = \begin{bmatrix} A_1 \\ 0 \end{bmatrix} \begin{matrix} \} r_1 \\ \} k - r_1 \end{matrix}, \quad \ker \bar{A} = \ker A = \ker A_1.$$

From $\operatorname{im} B Q_0 \subseteq \operatorname{im} G_1 = \operatorname{im} \mathcal{G}_1$ we derive, with $\bar{B} := LBK = \underbrace{\begin{bmatrix} \bar{B}_{11} & \bar{B}_{12} \\ \bar{B}_{21} & \bar{B}_{22} \end{bmatrix}}_{r_1 \quad m-r_1} \begin{matrix} \} r_1 \\ \} k - r_1 \end{matrix}$, that $\operatorname{im} \bar{B} \bar{Q}_0 \subseteq \operatorname{im} L\mathcal{G}_1$, hence $\bar{B}\bar{Q}_0$ has the form $\begin{bmatrix} * & * \\ 0 & 0 \end{bmatrix}$, and $\bar{B}_{21}\tilde{Q}_0 = 0$, $\ker \bar{D}_1 \subseteq \ker \bar{B}_{21}$, $\bar{B}_{22} = 0$ must hold.
It remains to show that (10.37) has regular index 1 as a DAE for x_1 in $\mathbb{R}^{r_1}$. Obviously, this DAE for x_1 has a properly stated leading term, too. If we succeed in showing $\bar{A}_1 \bar{D}_1 + \tilde{\mathcal{W}}_0 \bar{B}_{11} \tilde{Q}_0$ to be nonsingular, where $\tilde{\mathcal{W}}_0 := I - \bar{A}_1 \bar{A}_1^+$, we are done. Notice that $\tilde{\mathcal{W}}_0 := L\mathcal{W}_0 L^{-1}$ is the orthoprojector onto $\operatorname{im} \bar{G}_0^\perp = \operatorname{im} \bar{A}^\perp$. Because of $\bar{A} = \begin{bmatrix} A_1 \\ 0 \end{bmatrix}$, we have $\tilde{\mathcal{W}}_0 = \begin{bmatrix} \tilde{\mathcal{W}}_0 & 0 \\ 0 & I \end{bmatrix}$. Derive

$$\begin{aligned}\bar{A}_1\bar{D}_1+\tilde{\mathcal{W}}_0\bar{B}_{11}\tilde{Q}_0 &= \begin{bmatrix} I & 0\end{bmatrix} LADK \begin{bmatrix} I \\ 0\end{bmatrix} + \begin{bmatrix} I & 0\end{bmatrix} \bar{\mathcal{W}}_0 LBK\bar{Q}_0 \begin{bmatrix} I \\ 0\end{bmatrix} \\ &= \begin{bmatrix} I & 0\end{bmatrix} L(AD+\mathcal{W}_0BQ_0)K \begin{bmatrix} I \\ 0\end{bmatrix} \\ &= \begin{bmatrix} I & 0\end{bmatrix} \begin{bmatrix} \mathcal{S} & 0 \\ 0 & 0\end{bmatrix} \begin{bmatrix} I \\ 0\end{bmatrix} = S,\end{aligned}$$

and $\mathcal{S}$ is nonsingular. □

10.2.2.3 Tractability index 2, G_2 has a nontrivial nullspace

The decomposed system (10.21), (10.22) now reads

$$\begin{aligned}P_1D^-(D\Pi_1x)'+Q_1x+Q_0x-Q_0Q_1D^-(D\Pi_0Q_1x)'+\mathcal{V}_0Dx+\mathcal{U}_0(Dx)' \\ +G_2^-B_1D^-D\Pi_1x+y=G_2^-q, \qquad (10.39)\\ \mathcal{W}_2B_1D^-D\Pi_1x=\mathcal{W}_2q, \qquad (10.40)\end{aligned}$$

with coefficients (cf. Proposition 2.23)

$$\begin{aligned}\mathcal{U}_0 &= -Q_0\{Q_0+Q_1Q_0Q_1P_0\}\Pi_0D^- = -Q_0Q_1Q_0Q_1D^-,\\ \mathcal{V}_0 &= Q_0\{(P_0+Q_1Q_0)D^-R'-Q_1D^-(D\Pi_1D^-)'\}DD^- = -Q_0Q_1D^-(D\Pi_1D^-)'DD^-\end{aligned}$$

and an arbitrary continuous function y such that

$$G_2y=0. \qquad (10.41)$$

We multiply (10.39) by $D\Pi_1$, Q_1 and Q_0P_1, and obtain the system

$$(D\Pi_1x)'-(D\Pi_1D^-)'D\Pi_1x+D\Pi_1G_2^-B_1D^-D\Pi_1x+D\Pi_1y=D\Pi_1G_2^-q, \qquad (10.42)$$

$$\begin{aligned}Q_1x+Q_1Q_0x-Q_1Q_0Q_1D^-(D\Pi_0Q_1x)'+Q_1\mathcal{V}_0Dx+Q_1\mathcal{U}_0(Dx)' \qquad (10.43)\\ +Q_1G_2^-B_1D^-D\Pi_1x+Q_1y=Q_1G_2^-q,\end{aligned}$$

$$\begin{aligned}Q_0P_1Q_0x+Q_0P_1D^-(D\Pi_1x)'-Q_0P_1Q_0Q_1D^-(D\Pi_0Q_1x)'+Q_0P_1\mathcal{V}_0Dx \qquad (10.44)\\ +Q_0P_1\mathcal{U}_0(Dx)'+Q_0P_1G_2^-B_1D^-D\Pi_1x+Q_0P_1y=Q_0P_1G_2^-q,\end{aligned}$$

which is a decomposed version of (10.39) due to $\Pi_0+Q_0P_1+Q_1=I$, $\Pi_0=D^-D\Pi_0$. Multiplying equation (10.43) by Π_0 and taking into account the property $\Pi_0Q_1Q_0=0$ we derive

$$\Pi_0Q_1x+\Pi_0Q_1G_2^-B_1D^-D\Pi_1x+\Pi_0Q_1y=\Pi_0Q_1G_2^-q. \qquad (10.45)$$

Now it is evident that, for given y, and the initial condition

$$D(t_0)\Pi_1(t_0)x(t_0) = z_0 \in \operatorname{im} D(t_0)\Pi_1(t_0), \tag{10.46}$$

there is exactly one solution of the explicit ODE (10.42), that is, the solution component $\Pi_0 x = D^- D\Pi_0 x$ of the IVP for the DAE is uniquely determined. Having $D\Pi_1 x$, we obtain the next component $\Pi_0 Q_1 x$ from (10.45), and thus $Dx = D\Pi_1 x + D\Pi_0 Q_1 x$. Then, formula (10.44) provides an expression for $Q_0 P_1 Q_0 x$ in terms of the previous ones. Finally, multiplying (10.43) by Q_0 we find an expression $Q_0 Q_1 x + Q_0 Q_1 Q_0 x = \mathcal{E}$ with $\mathcal{E}$ depending on the already given terms y, $D\Pi_0 Q_1 x$, $D\Pi_1 x$, and Dx. In turn, this yields an expression for $Q_0 Q_1 Q_0 x$, and then for $Q_0 x = Q_0 Q_1 Q_0 x + Q_0 P_1 Q_0$. In summary, for each function y that satisfies condition (10.41), the system (10.42)–(10.44), completed by the initial condition (10.46), determines a unique solution $x = D^- D\Pi_1 x + \Pi_0 Q_1 x + Q_0 x$ of the DAE.

With regard of the discussion above (cf. (10.23)) the actual arbitrary part of y is $Q_1 y \in N_1 \cap S_1$.

We mention that, for solvability, the component $D\Pi_0 Q_1 x$ must be continuously differentiable. Equation (10.45) shows the terms being responsible for that. For instance, if $\Pi_0 Q_1 G_2^- B_1 D^-$ is a continuously differentiable matrix function, then the difference $D\Pi_0 Q_1 (G_2^- q - y)$ must be continuously differentiable.

Example 10.12 (G_2 has full row rank). Set $k = 3$, $m = 4$, $n = 2$, and consider the DAE (10.12) given by the coefficients

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

which means in detail

$$\begin{aligned} x_1' + x_1 &= q_1, \\ x_2' + x_3 + x_4 &= q_2, \\ x_2 &= q_3. \end{aligned} \tag{10.47}$$

We provide the sequence

$$G_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad Q_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$G_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \qquad Q_1 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 \end{bmatrix}, \qquad B_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

$$G_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \qquad G_2^- = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & -1 \\ 0 & 0 & 0 \end{bmatrix}, \qquad \mathcal{W}_2 = 0.$$

Hence the projector Q_1 satisfies the admissibility condition $X_1 \subset \ker Q_1$ with $X_1 := \{z \in \mathbb{R}^4 : z_1 = 0,\ z_2 = 0,\ z_3 = 0\}$ and $N_0 = (N_0 \cap N_1) \oplus X_1$. G_2 has maximal rank, $r_2 = k = 3$, thus the DAE is tractable with index 2. The consistency condition (10.40) disappears. Compute further $\mathcal{V}_l = 0$ and $\mathcal{U}_l = 0$, so that equation (10.39) simplifies to

$$P_1 D^-(D\Pi_1 x)' + Q_1 x + Q_0 x - Q_0 Q_1 D^-(D\Pi_0 Q_1 x)' + G_2^- B_1 D^- D\Pi_1 x + y = G_2^- q,$$

with

$$P_1 D^- = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad Q_0 Q_1 D^- = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & -1 \end{bmatrix}, \quad G_2^- B_1 D^- D\Pi_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Taking into account that $G_2 y = 0$ is equivalent to $y_1 = 0,\ y_2 = 0,\ y_4 = -y_3,$ we find equation (10.39) to be in detail:

$$\begin{aligned} x_1' + x_1 &= q_1, \\ x_2 &= q_3, \\ 2x_3 + y_3 &= q_2 - q_3, \\ x_4 - x_3 - x_2 + x_2' - y_3 &= 0. \end{aligned}$$

For each function y_3, this is a regular DAE with tractability index 2. Its solutions are the solutions of the original DAE. □

10.3 Underdetermined nonlinear DAEs

The discussion in Sections 3.1, 3.2 and 3.4 applies to general systems (3.1) comprising k equations and m unknown functions. In particular, we can construct admissible matrix function sequences in the same way as we used to do for regular DAEs. Lemma 3.27 offers the possibility to make use of linearizations.

In *nonregular* DAEs, we have to expect many more difficulties than in regular DAEs. Already for nonregular linear DAEs, there is much space left for different interpretations.

Though it seems to be straightforward now to generalize the tractability index for general nonlinear DAEs, we do not go this way. So far we have no idea how one could benefit from such a definition. We restrict our interest to underdetermined DAEs having tractability index 0 or 1, since this kind of system plays a certain role in optimal control problems with DAE constraints.

Consider DAEs of the form

$$f((D(t)x(t))', x(t), t) = 0, \tag{10.48}$$

consisting of k equations for $m > k$ unknown functions, and which satisfies the basic Assumption 3.16. We continue to use the denotations introduced in Sections 3.1 and 3.2.

Definition 10.13. Let the DAE (10.48) satisfy the basic Assumption 3.16 and $\operatorname{rank}[f_y D \; f_x] = k < m$.

(1) The DAE (10.48) is said to be *underdetermined with tractability index 0*, if the matrix function G_0 has maximal rank, that is $r_0 = k$.
(2) The DAE (10.48) is said to be *underdetermined with tractability index 1*, if $r_0 < k$ and the matrix function G_1 has maximal rank $r_1 = k$.

Example 10.14 (Strangeness-free reduced DAE). In [130] the strangeness-free reduced DAE

$$x_1'(t) + \mathcal{L}(x_1(t), x_2(t), x_3(t), t) = 0, \tag{10.49}$$
$$x_2(t) + \mathcal{R}(x_1(t), x_3(t), t) = 0, \tag{10.50}$$

plays its role. By means of

$$A = \begin{bmatrix} I \\ 0 \end{bmatrix}, \; D = \begin{bmatrix} I & 0 & 0 \end{bmatrix}, \; G_0 = \begin{bmatrix} I & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \; Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \; G_1 = \begin{bmatrix} I & \star & \star \\ 0 & I & \star \end{bmatrix}$$

we check this DAE to be underdetermined with tractability index 1. □

If the matrix function $G_0 = AD$ has full row rank k, then A is invertible, and D has full row rank k, too. Then we find a continuous orthogonal matrix function $K = [\underbrace{K_1}_{k} \; \underbrace{K_2}_{m-k}]$, such that

$$D(t)K(t) = [D(t)K_1(t) \;\; D(t)K_2(t)] = [\tilde{D}(t) \;\; 0], \quad t \in \mathcal{I}_f.$$

Transforming the unknown function $x = K_1\tilde{x} + K_2\tilde{u}$ within the DAE (10.48) yields

$$f((\tilde{D}(t)\tilde{x}(t))', K_1(t)\tilde{x}(t) + K_2(t)\tilde{u}(t), t) = 0. \tag{10.51}$$

We regard this equation as a DAE with respect to the unknown function $\tilde{x}$, and the function $\tilde{u}$ as an arbitrary continuous control. In order to meet the basic Assumption 3.16, for obtaining a continuously differentiable matrix function $\tilde{D}$ in (10.51), we suppose K_1 to be sufficiently smooth, which is, in turn, ensured by a sufficiently smooth coefficient AD. Since the matrix function $A\tilde{D}$ is nonsingular, our equation (10.51) is an implicit regular ODE, and a regular DAE with tractability index 0.
One could get along with transformations K being just continuous by adapting the theory for the special case of equations (10.48) with just continuous D, and $f_y D$ having full row rank. Notice at this point, that the linear theory in Chapter 2 is made with continuous coefficient D.

An analogous partition of the unknown function into two parts, one of which can be considered as an arbitrary control, is provided for linear DAEs that are tractable with index 1 by Proposition 10.11. For nonlinear DAEs, we obtain a local version of this property below.

Let the DAE (10.48) be underdetermined with index 1. Owing to Lemma 3.27, each linearization of this DAE along a function $x_* \in \mathcal{C}^2_*(\mathcal{D}_f \times \mathcal{I}_f)$ inherits the property of beeing underdetermined with tractability index 1. We fix such a function x_* and the corresponding linearization

$$A_*(t)((D(t)x(t))' + B_*(t)x(t) = q(t), \quad t \in \mathcal{I}_*. \tag{10.52}$$

Let Q_0 denote the orthoprojector function onto $\ker D$, and $\mathcal{W}_{*0}$ denote the orthoprojector function onto $\operatorname{im} G_{*0}^{\perp}$, with $G_{*0} = A_*D$. The asterisk indicates the possible dependence upon x_*. The matrix functions $G_{*1} = G_{*0} + B_*Q_0$ and $\mathcal{G}_{*1} = G_{*0} + \mathcal{W}_{*0}B_*Q_0 = G_{*1}(I - G_{*0}^{-}B_*Q_0)$ both have full row rank k. Therefore, there is a matrix function $K_* \in \mathcal{C}(\mathcal{I}_*, L(\mathbb{R}^m))$ that is pointwise orthogonal such that

$$\mathcal{G}_{*1}K_* = [\underbrace{\mathcal{S}_*}_{k} \ \underbrace{0}_{m-k}], \quad \mathcal{S}_* \text{ nonsingular.}$$

With the partition $K_* = [\underbrace{K_{*1}}_{k} \ \underbrace{K_{*2}}_{m-k}]$, this leads to the relations

$$\mathcal{G}_{*1}K_{*1} = \mathcal{S}_*, \quad \mathcal{G}_{*1}K_{*2} = 0, \quad DK_{*2} = 0.$$

Denote by $\tilde{Q}_{*0}$ the orthoprojector function onto $\ker DK_{*1}$. Then we have

$$\begin{bmatrix} \tilde{Q}_{*0} & 0 \\ 0 & I \end{bmatrix} = K_*^{-1}Q_0K_*, \quad \begin{bmatrix} \tilde{Q}_{*0} \\ 0 \end{bmatrix} = K_*^{-1}Q_0K_{*1}, \quad K_{*1}\tilde{Q}_{*0} = Q_0K_{*1}.$$

Next we apply the transformation $x = K_{*1}\tilde{x} + K_{*2}\tilde{u}$ arising from the linearized DAE to its nonlinear origin (10.48):

$$f((D(t)K_{*1}(t)\tilde{x}(t))', K_{*1}(t)\tilde{x}(t) + K_{*2}(t)\tilde{u}(t), t) = 0. \tag{10.53}$$

Denote $\begin{bmatrix} \tilde{x}_* \\ \tilde{u}_* \end{bmatrix} := K_*^{-1}x_*$.

Assumption 3.16 requires DK_{*1} to be continuously differentiable. Here we suppose this to be given. This can be ensured by a sufficiently smooth f. On the other hand, a more specific theory would get along with K_* being just continuous.
Equation (10.53) regarded as a DAE for $\tilde{x}$ fulfills the basic Assumption 3.16. We indicate the terms associated with this DAE by a tilde. With

$$\begin{aligned}
\tilde{f}(y,\tilde{x},\tilde{u},t) &:= f(y,K_{*1}(t)\tilde{x}+K_{*2}(t)\tilde{u},t),\\
\tilde{D}(t) &:= D(t)K_{*1}(t),\\
\tilde{A}(\tilde{x}^1,\tilde{x},t,\tilde{u}) &:= f_y(\tilde{D}(t)\tilde{x}^1+\tilde{D}'(t)\tilde{x},K_{*1}(t)\tilde{x}+K_{*2}(t)\tilde{u},t),\\
\tilde{B}(\tilde{x}^1,\tilde{x},t,\tilde{u}) &:= f_x(\tilde{D}(t)\tilde{x}^1+\tilde{D}'(t)\tilde{x},K_{*1}(t)\tilde{x}+K_{*2}(t)\tilde{u},t)K_{*1}(t),
\end{aligned}$$

it follows that

$$\begin{aligned}
\tilde{A}_*(t) &:= \tilde{A}(\tilde{x}_*'(t),\tilde{x}_*(t),t,\tilde{u}_*(t)) = A_*(t),\\
\tilde{B}_*(t) &:= \tilde{B}(\tilde{x}_*'(t),\tilde{x}_*(t),t,\tilde{u}_*(t)) = B_*(t)K_{*1}(t),\\
\tilde{\mathcal{W}}_{*0}(t) &= \mathcal{W}_{*0}(t),
\end{aligned}$$

and further, for the matrix function $\tilde{\mathcal{G}}_{*1} := \tilde{A}DK_{*1}+\tilde{\mathcal{W}}_0\tilde{B}\tilde{Q}_{*0}$,

$$\begin{aligned}
\tilde{\mathcal{G}}_{*1}(\tilde{x}_*'(t),\tilde{x}_*(t),t,\tilde{u}_*(t)) &= \tilde{A}_*(t)D(t)K_{*1}(t)+\tilde{\mathcal{W}}_{*0}(t)\tilde{B}_*(t)\tilde{Q}_{*0}(t)\\
&= A_*(t)D(t)K_{*1}(t)+\mathcal{W}_{*0}(t)B_*(t)K_{*1}(t)\tilde{Q}_{*0}(t)\\
&= A_*(t)D(t)K_{*1}(t)+\mathcal{W}_{*0}(t)B_*(t)Q_0(t)K_{*1}(t)\\
&= \big(A_*(t)D(t)+\mathcal{W}_{*0}(t)B_*(t)Q_0(t)\big)K_{*1}(t)\\
&= \mathcal{S}_*.
\end{aligned}$$

Since $\mathcal{S}_*$ is nonsingular, the matrix function $\tilde{\mathcal{G}}_{*1}$, and at the same time the matrix function $\tilde{G}_{*1}$, remain nonsingular in a neighborhood of the graph of $(\tilde{x}_*,\tilde{u}_*)$. This means that the DAE (10.53) is regular with tractability index 1 there. We summarize what we have shown:

Proposition 10.15. *If the DAE (10.48), with sufficiently smooth data f, is underdetermined with tractability index 1, then it can be transformed locally by a linear transformation into a regular index 1 DAE (10.53) for one component $\tilde{x}$ whereas the other part $\tilde{u}$ can be regarded as a kind of control function.*

We mention that, in our construction, $\ker \mathcal{G}_{*1} = N_0 \cap S_{*0}$ is a particular subspace associated with the linearized DAE (10.51), and the orthoprojector function onto this subspace reads $K_*\mathrm{diag}\,(0,I)K_*^{-1}$.
If the intersection subspace $N_0 \cap S_0$ associated with the nonlinear DAE does not vary with the arguments x^1,x, then the corresponding orthoprojector function $K_*\mathrm{diag}\,(0,I)K_*^{-1}$ is actually independent of the reference function x_*.

10.4 Notes and references

(1) The material of this chapter is to a large extent new. Proposition 10.11 is a modified version of [36, Proposition 3.2] and Proposition 10.15 is a slight generalization of [36, Proposition 5.1] where quasi-linear DAEs are considered.

A different version of a *singular tractability index* for the characterization of not necessarily regular DAEs is proposed in [62]. There, the level at which the admissible matrix function sequence becomes stationary is assigned to the tractability index. The paper deals then with index-1 DAEs.

(2) At least in the constant coefficient case, in admissible matrix (function) sequences, the G_i themselves as well as the subspaces $N_0 + \cdots + N_j$ become stationary. It is open how one could benefit from this property.

(3) We do not quote the various literature dealing with general DAEs and treating them by reduction into special forms. We refer to [130, 20] and the references therein for an overview. Applied to nonregular DAEs, the strangeness concept and the concept behind the tractability index follow widely different interpretations. They are so to say unrelated to each other.
Theorem 2.79, which relates the strangeness characteristics and the tractability characteristics for regular linear DAEs to each other, cannot be extended for nonregular DAEs. Whereas, for regular DAEs, the tractability index equals 1+ strangeness index, in the case of nonregular DAEs, the strangeness index can be equal to the tractability index but can also be arbitrarily higher.

(4) Proposition 10.11 precisely reflects the inherent structure of the system (10.33), (10.34) which has been obtained by the projector based decoupling.
The statement of Proposition 10.11 itself is a well-known fact in the context of control selection and feedback regularization (see the discussion, e.g., in [36]).

Chapter 11
Minimization with constraints described by DAEs

This chapter collects results obtained by means of the projector based approach to DAEs, which are relevant in view of optimization. We do not at all undertake to offer an overview concerning the large field of control and optimization with DAE constraints. We do not touch the huge arsenal of direct minimization methods.
We address the basic topics of adjoint and self-adjoint DAEs in Section 11.1 and provide extremal conditions in Sections 11.2 and 11.3. Section 11.4 is devoted to linear-quadratic optimal control problems (LQP) including also a generalization of the Riccati feedback solution. In each part, we direct particular attention to the properties of the resulting optimality DAE. If one intends to apply indirect optimization, that is, to solve the optimality DAE, then one should take great care to ensure appropriate properties, such as regularity with index 1, in advance by utilizing the scope of modeling. By providing criteria in terms of the original problem data we intend to assist specialists in modeling.
We direct the reader attention to several different denotations used in the sections of this chapter. In each case, the relevant denotations and basic assumptions are given at the beginning of the section.

11.1 Adjoint and self-adjoint DAEs

Adjoint and self-adjoint linear equations play an important role in various mathematical fields, in particular in the theory of ODEs. Usually in this context, one includes the investigation of the complex-valued case, and we do so, too. In the present section, $\mathbb{K}$ stands for the real numbers $\mathbb{R}$ or the complex numbers $\mathbb{C}$. We denote the inner product of the vector space $\mathbb{K}^s$ by $\langle .,. \rangle$. We consider linear DAEs

$$A(t)(D(t)x(t))' + B(t)x(t) = 0, \quad t \in \mathcal{I}, \tag{11.1}$$

with coefficients $A(t) \in L(\mathbb{K}^n, \mathbb{K}^k)$, $D(t) \in L(\mathbb{K}^m, \mathbb{K}^n)$, and $B(t) \in L(\mathbb{K}^m, \mathbb{K}^k)$ being continuous on the given interval $\mathcal{I}$. The DAE comprises k equations, whereas the

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_11,

unknown function x has m components.
A *solution* of this DAE is a function x which belongs to the function space

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

and which satisfies the DAE pointwise on $\mathcal{I}$.

For each pair of functions $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{K}^m)$, $y \in \mathcal{C}_{A^*}^1(\mathcal{I},\mathbb{K}^k)$, it results that

$$\begin{aligned}
&\big\langle A(t)(D(t)x(t))' + B(t)x(t), y(t)\big\rangle \\
&\quad = \big\langle (D(t)x(t))', A(t)^*y(t)\big\rangle + \big\langle x(t), B(t)^*y(t)\big\rangle \\
&\quad = \big(\big\langle D(t)x(t), A(t)^*y(t)\big\rangle\big)' - \big\langle D(t)x(t), (A(t)^*y(t))'\big\rangle + \big\langle x(t), B(t)^*y(t)\big\rangle \\
&\quad = \big(\big\langle D(t)x(t), A(t)^*y(t)\big\rangle\big)' + \big\langle x(t), -D(t)^*(A(t)^*y(t))' + B(t)^*y(t)\big\rangle,\ t \in \mathcal{I}.
\end{aligned}$$

If x is a solution of the DAE (11.1) and y is a solution of the DAE

$$-D(t)^*(A(t)^*y(t))' + B(t)^*y(t) = 0, \quad t \in \mathcal{I}, \tag{11.2}$$

then it follows that $(\langle D(t)x(t), A(t)^*y(t)\rangle)'$ vanishes identically, such that

$$\langle D(t)x(t), A(t)^*y(t)\rangle = \text{ constant }, \quad t \in \mathcal{I}. \tag{11.3}$$

In the particular case if $n = m = k$, $A = I$, $D = I$, if equation (11.1) is actually an explicit regular ODE, equation (11.2) is the adjoint ODE to (11.1), and the relation (11.3) is known as the Lagrange identity. We adopt these designations also in the general case.

Definition 11.1. The DAEs (11.1) and (11.2) are said to be *adjoint* to each other, and the identity (11.3) is called their *Lagrange identity*.

In minimization problems with DAE constraints, the DAEs usually have less equations than unknowns, that is $k < m$. Then, the adjoint DAE has more equations than unknowns, which makes the solvability problem highly nontrivial.

Definition 11.2. The DAE (11.1) is said to be *self-adjoint*, if $k = m$, $\mathcal{C}_D^1(\mathcal{I},\mathbb{K}^m) = \mathcal{C}_{A^*}^1(\mathcal{I},\mathbb{K}^k)$, and

$$A(Dx)' + Bx = -D^*(A^*x)' + B^*x \quad \text{for all } x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{K}^m).$$

Example 11.3 (Self-adjoint DAE). The DAE

$$\mathrm{i}\,(D(t)^*(D(t)x(t)))' + B(t)x(t) = 0,$$

with $B(t) = B(t)^*$, which is introduced and discussed in [1], is obviously self-adjoint.

The following assertion generalizes this example and provides a wider criterion of self-adjointness, which is of particular interest for minimization problems.

Proposition 11.4. *If $m = k$, J is a constant matrix such that $J^* = -J$ and $J^2 = -I$, $A(t) = D(t)^* J$, $B(t) = B(t)^*$ for $t \in \mathcal{I}$, then the DAE (11.1) is self-adjoint.*

Proof. Because of $m = k$, $A^* = J^* D$ and $D = JA^*$ the function spaces $\mathcal{C}_D^1(\mathcal{I}, \mathbb{K}^m)$ and $\mathcal{C}_{A^*}^1(\mathcal{I}, \mathbb{K}^k)$ coincide. Moreover, for each $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{K}^m)$ it holds that

$$A(Dx)' + Bx = D^* J(JA^* x)' + B^* x = -D^*(A^* x)' + B^* x.$$

□

As in Chapters 2 and 10 we intend to make use of proper formulations of the leading term of the DAEs. Though the material of these chapters is described for real valued functions, it applies analogously also in the complex valued case. For easier reading, here we again consider certain aspects separately. We emphasize the consistency with Chapters 2 and 10.

Definition 11.5. The DAE (11.1) has a *properly stated leading term*, if the transversality condition

$$\ker A(t) \oplus \operatorname{im} D(t) = \mathbb{K}^n, \quad t \in \mathcal{I}, \tag{11.4}$$

holds and the projector valued function $R : \mathcal{I} \to L(\mathbb{K}^n)$ uniquely defined by $\operatorname{im} R = \operatorname{im} D$, $\ker R = \ker A$ is continuously differentiable. The projector function R is named the *border projector* of the leading term of the DAE.
The DAE (11.1) has a *full rank proper leading term*, if

$$\ker A(t) = \{0\}, \quad \operatorname{im} D(t) = \mathbb{K}^n, \quad t \in \mathcal{I}. \tag{11.5}$$

The full rank proper leading term is associated with the trivial border projector $R = I$.

Proposition 11.6. *If the DAE (11.1) has a properly stated leading term, then its adjoint equation (11.3) also has a properly stated leading term.*
If the DAE (11.1) has a full rank proper leading term, then its adjoint equation (11.3) has also a full rank proper leading term.

Proof. The decomposition (11.4) can be written as $(\operatorname{im} A(t)^*)^\perp \oplus (\ker D(t)^*)^\perp = \mathbb{K}^n$, and it follows that $\ker D(t)^* \oplus \operatorname{im} A(t)^* = \mathbb{K}^n$. Furthermore, $R(t)^*$ is the projector onto $\operatorname{im} A(t)^*$ along $\ker D(t)^*$. The projector valued function R^* inherits the continuous differentiability from R. □

Admissible matrix function sequences can be built in the same way as given for the real valued case in Chapter 2. Also, as it is done there, one assigns characteristic values, regularity and the tractability index. The decouplings and the IERODEs keep their meaning. In particular, the so-called IERODE of a regular DAE is uniquely determined in the scope of fine decouplings. The IERODE of a regular index-1 DAE (11.1) is uniquely determined to have the form

$$u' = R'u - DG_1^{-1} B D^- u, \tag{11.6}$$

with $u = Dx$, see Section 2.4.

The next assertion characterizes an important class of DAE whose IERODEs show Hamiltonian structure.

Theorem 11.7. *Let the DAE (11.1) have the particular form described in Proposition 11.4 and be regular with tractability index 1. Additionally, let its leading term be full rank proper. Then, the IERODE of the DAE (11.1) applies to the component $u = Dx$ and has the form*

$$u'(t) = J^* E(t) u(t), \quad \text{with} \quad E(t) = E(t)^*. \tag{11.7}$$

Proof. This assertion is verified for $\mathbb{K} = \mathbb{R}$ in [8, Theorem 4.4], however, all arguments apply also in the complex valued case. □

The full rank proper leading term can be achieved by appropriate refactorizations of the leading term. The following example demonstrates that unless there is a full rank leading term, the Hamiltonian property (11.7) can get lost. Therefore, it is quite reasonable to ensure full rank proper leading terms.

Example 11.8 (Non-Hamiltonian IERODE). The DAE

$$\underbrace{\begin{bmatrix} 0&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&0\\ 0&0&t&1&0\\ 0&0&0&0&0 \end{bmatrix}}_{A(t)} \left(\underbrace{\begin{bmatrix} 0&0&0&t&0\\ 0&0&0&1&0\\ 0&0&0&0&0\\ 0&-1&0&0&0 \end{bmatrix}}_{D(t)} x(t) \right)' + \underbrace{\begin{bmatrix} 0&0&0&0&-1\\ 0&0&0&-1&0\\ 0&0&-1&0&0\\ 0&-1&0&-2&0\\ -1&0&0&0&-1 \end{bmatrix}}_{B(t)} x(t) = 0$$

is self-adjoint, it meets the conditions $A = D^* J$ and $B = B^*$ for

$$J = \begin{bmatrix} 0&0&1&0\\ 0&0&0&1\\ -1&0&0&0\\ 0&-1&0&0 \end{bmatrix}. \quad \text{We choose} \quad D(t)^- = \begin{bmatrix} 0&0&0&0\\ 0&0&-t&-1\\ 0&0&0&0\\ 0&1&0&0\\ 0&0&0&0 \end{bmatrix}$$

and we compute

$$R(t) = D(t)D(t)^- = \begin{bmatrix} 0&t&0&0\\ 0&1&0&0\\ 0&0&0&0\\ 0&0&t&1 \end{bmatrix}, \quad P_0(t) = D(t)^- D(t) = \begin{bmatrix} 0&0&0&0&0\\ 0&1&0&0&0\\ 0&0&0&0&0\\ 0&0&0&1&0\\ 0&0&0&0&0 \end{bmatrix}.$$

Regarding the relations $\operatorname{im} R = \operatorname{im} D$ and $\ker R = \ker A$ we see that the DAE has a properly stated leading term. However, $A(t)$ and $D(t)$ fail to have full rank, and the two-dimensional subspaces $\operatorname{im} D(t)$ and $\ker A(t)$ vary in $\mathbb{R}^4$ with t. Compute further

$$G_1(t) = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 \end{bmatrix}, \quad D(t)G_1(t)^{-1}B(t)D(t)^- = \begin{bmatrix} 0 & -t & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -2 & t & 1 \end{bmatrix},$$

so that the DAE is seen to be regular with tractability index 1. Its IERODE $u' - R'u + DG_1^{-1}BD^-u = 0$ reads in detail

$$u'(t) = \underbrace{\begin{bmatrix} 0 & t+1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 2 & 1-t & -1 \end{bmatrix}}_{M(t)} u(t).$$

The resulting

$$E(t) := JM(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 2 & 1-t & -1 \\ 0 & -1-t & 0 & 0 \\ 0 & -1 & 0 & 0 \end{bmatrix},$$

fails to be symmetric. The lack of symmetry is caused by the time-varying border projector $R(t)$. In this context, also a properly stated leading term with nontrivial border projector appears to be somewhat unqualified.

A corresponding refactorization of the leading term (see Section 2.3) improves the situation. Such a refactorization does not change the characteristic values including the tractability index of a regular DAE. Choose

$$H = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad H^- = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

such that H^- is a reflexive generalized inverse of H and the relation $RHH^-R = R$ is fulfilled. The refactorized DAE has the coefficients $\bar{A} = AH$, $\bar{D} = H^-D$ and $\bar{B} = B - ARH(H^-R)'D$, which means here

$$\underbrace{\begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}}_{\bar{A}(t)} \Big(\underbrace{\begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 0 & 0 \end{bmatrix}}_{\bar{D}(t)} x(t) \Big)' + \underbrace{\begin{bmatrix} 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & -2 & 0 \\ -1 & 0 & 0 & 0 & -1 \end{bmatrix}}_{\bar{B}(t)=B(t)} x(t) = 0. \tag{11.8}$$

The new DAE (11.8) has a full rank proper leading term. With

$$\bar{J} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}$$

it holds that $\bar{D}^* \bar{J} = \bar{A}$. The new IERODE applies to $\bar{u} = \bar{D}x$. It reads

$$\bar{u}'(t) = \underbrace{\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}}_{\bar{M}(t)} \bar{u}(t),$$

and the resulting

$$\bar{E}(t) := \bar{J}\bar{M}(t) = \begin{bmatrix} 2 & -1 \\ -1 & 0 \end{bmatrix},$$

is symmetric according to the theory. □

We finish this section by turning briefly to regular DAEs (11.1). We conjecture that regular DAEs and their adjoints share their characteristic values, in particular their tractability index. To date, the conjecture is approved for index 1 and index 2 cases. In any case, a regular DAE (11.1) and its adjoint (11.2) share the first level characteristic value $r_0 = \operatorname{rank} G_0 = \operatorname{rank} AD = \operatorname{rank} D^* A^*$.

Theorem 11.9. *The DAE (11.1) is regular with tractability index 1 and the characteristic values $r_0 < r_1 = m$, only if the adjoint DAE (11.2) is so.*
The DAE (11.1) is regular with tractability index 2 and the characteristic values $r_0 \le r_1 < r_2 = m$, only if the adjoint DAE (11.2) is so.

Proof. Let the DAE (11.1) be regular with tractability index 1. Denote by Q_0 and W_0 the orthoprojector functions onto $\ker G_0$, respectively $\operatorname{im} G_0^\perp = \ker G_0^*$. The matrix function $G_0 + W_0 B Q_0$ is nonsingular together with $G_1 = G_0 + BQ_0$. Further, also $-G_0 + W_0 B Q_0$, thus $(-G_0 + W_0 B Q_0)^* = -D^* A^* + Q_0 B^* W_0$ are nonsingular. This implies the invertibility of $-D^* A^* + B^* W_0$ which means that the adjoint DAE is regular with index 1.
The index-2 case is verified in [12, Theorem 5.1]. □

11.2 Extremal conditions and the optimality DAE

11.2.1 A necessary extremal condition and the optimality DAE

Consider the cost functional

$$J(x) = g(D(t_f)x(t_f)) + \int_{t_0}^{t_f} h(x(t),t)\mathrm{d}t \tag{11.9}$$

to be minimized on functions $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$, $\mathcal{I} = [t_0, t_f]$, subject to the constraints

$$f((D(t)x(t))',x(t),t) = 0, \quad t \in \mathcal{I}, \tag{11.10}$$

and

$$D(t_0)x(t_0) = z_0 \in \mathbb{R}^n. \tag{11.11}$$

For easier later use we collect the basic assumptions.

Assumption 11.10. *The function $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I} \to \mathbb{R}^k$ is continuous and has continuous partial derivatives f_y, f_x with respect to the first two variables $y \in \mathbb{R}^n$, $x \in \mathbb{R}^m$. $D : \mathcal{I} \to L(\mathbb{R}^m, \mathbb{R}^n)$ is continuous.*
The DAE (11.10) comprises $k < m$ equations. It has a full rank proper leading term, that is, $n \le m$, $n \le k$ and f_y has full column rank n, and D has full row rank n on their definition domains.
The functions h and g are continuously differentiable.

For a given function $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ we consider the linearization of the nonlinear DAE (11.10) along x_* to be the linear DAE

$$A_*(t)(D(t)x(t))' + B_*(t)x(t) = q(t),\ t \in \mathcal{I}, \tag{11.12}$$

with continuous coefficients

$$A_*(t) := f_y((D(t)x_*(t))',x_*(t),t), \quad B_*(t) := f_x((D(t)x_*(t))',x_*(t),t), \quad t \in \mathcal{I}.$$

The linear DAE (11.12) inherits the full rank proper leading term from (11.10). The following definition adapts Definition 10.13 to the present situation.

Definition 11.11. Let the DAE (11.10) satisfy Assumption 11.10. Let $\mathfrak{G} \subseteq \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I}$ be an open set.

(1) The DAE is said to be *underdetermined with tractability index 0* if $n = k$.
(2) The DAE is said to be *underdetermined with tractability index 1* on $\mathfrak{G}$ if $n < k$ and the rank condition

$$\operatorname{rank}\left[f_y(y,x,t)D(t) + f_x(y,x,t)(I - D(t)^+D(t))\right] = k, \quad (y,x,t) \in \mathfrak{G}, \tag{11.13}$$

is fulfilled.

It is evident that, in the index-1 case, the linearization (11.12) inherits the rank condition

$$\operatorname{rank}\left[A_*(t)D(t) + B_*(t)(I - D(t)^+D(t))\right] = k, \quad (y,x,t) \in \mathfrak{G}, \tag{11.14}$$

supposing the graph of x_* resides in $\mathfrak{G}$. This means that the linearization is also an underdetermined DAE with tractability index 1, if the nonlinear DAE is so in a neighborhood of the graph of x_*.

Theorem 11.12. *Let Assumption 11.10 be valid. Let $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ be a local solution of the optimization problem (11.9), (11.10), (11.11). Let the DAE (11.10) be*

underdetermined with tractability index 1 at least on an open set around the graph of x_*.
Then the terminal value problem

$$-D(t)^*(A_*(t)^*\lambda(t))' + B_*(t)^*\lambda(t) = h_x(x_*(t),t)^*, \quad t \in \mathcal{I}, \tag{11.15}$$

$$D(t_f)^*A_*(t_f)^*\lambda(t_f) = (g_\eta(D(t_f)x_*(t_f))D(t_f))^* \tag{11.16}$$

possesses a solution $\lambda_* \in \mathcal{C}^1_{A_*^*}(\mathcal{I},\mathbb{R}^k)$.

Proof. Owing to Proposition 10.9, the linear IVP

$$A_*(t)(D(t)x(t))' + B_*(t)x(t) = q(t),\ t \in \mathcal{I}, \quad D(t_0)x(t_0) = z_0,$$

is solvable for each arbitrary continuous q and $z_0 \in \mathbb{R}^n$. This means that the constraint operator $\mathcal{F} : \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m) \to \mathcal{C}(\mathcal{I},\mathbb{R}^m) \times \mathbb{R}^n$ defined by

$$(\mathcal{F}x)(t) := (f((D(t)x(t))',x(t),t),\ D(t_0)x(t_0)), \quad t \in \mathcal{I},$$

has a surjective derivative $\mathcal{F}_x(x_*)$ (cf. Section 3.9).
For the case of quasi-linear DAEs with $f(y,x,t) = A(x,t)y + b(x,t)$ the assertion is proved in [6, pages 121–139] by applying the famous Lyusternik theorem [151], then providing a representation of functionals on $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$, and further a representation of the Lagrange multiplier. The same arguments apply also in the slightly more general case discussed now. We emphasize the essential part of the surjectivity (closed range property) of $\mathcal{F}_x(x_*)$ in this context. □

Example 11.13 (A simple LQP). The linear-quadratic optimization problem given by the cost

$$J(x) = \frac{1}{2}x_1(T)^2 + \int_0^T x_2(t)^2 \mathrm{d}t$$

and the constraints

$$x_1'(t) - x_2(t) = 0, \quad x_1(0) = a \neq 0,$$

possesses the unique solution

$$x_*(t) = \begin{bmatrix} a - \frac{a}{T+2}\,t \\ -\frac{a}{T+2}, \end{bmatrix}$$

such that $J(x_*) = (\frac{a}{T+2})^2$. Theorem 11.12 applies. The resulting terminal value problem

$$-\begin{bmatrix}1\\0\end{bmatrix}\lambda'(t) + \begin{bmatrix}0\\-1\end{bmatrix}\lambda(t) = \begin{bmatrix}0\\2x_{*2}(t)\end{bmatrix}, \quad \begin{bmatrix}1\\0\end{bmatrix}\lambda(T) = \begin{bmatrix}1\\0\end{bmatrix}x_{*1}(T)$$

is uniquely solvable. □

The condition (11.14) is necessary for the existence of a solution of the terminal value problem (11.15), (11.16), as the following example demonstrates.

Example 11.14 (Backes' example). Consider the minimization problem (cf. [6, pages 50–52]) given by the cost functional

$$J(x) = \frac{1}{2}x_1(T)^2 + \frac{1}{2}\int_0^T (x_3(t)^2 + x_4(t)^2)\mathrm{d}t$$

and the constraints

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix} x(t) \right)' + \begin{bmatrix} 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 1 & 0 & 0 \end{bmatrix} x(t) = 0, \quad x_1(0) = a \neq 0,\ x_2(0) = 0.$$

This problem has the unique solution

$$x_*(t) = \begin{bmatrix} a - \frac{a}{T+2}\, t \\ 0 \\ -\frac{a}{T+2} \\ -\frac{a}{T+2}, \end{bmatrix}$$

which can easily be shown by reducing the problem to Example 11.13. However, in the setting given now the resulting terminal value problem

$$-\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \lambda(t) \right)' + \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & -1 & 0 \end{bmatrix} \lambda(t) = \begin{bmatrix} 0 \\ 0 \\ x_{*3}(t) \\ x_{*4}(t) \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \lambda(T) = \begin{bmatrix} x_{*1}(T) \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

has no solution. In particular, the terminal condition $\lambda_2(T) = 0$ contradicts the equation $\lambda_2 = x_{*3} \neq 0$. □

Corollary 11.15. *Let the Assumption 11.10 be fulfilled. If $x_* \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ is a local solution of the optimization problem (11.9), (11.10), (11.11), and the full rank condition (11.14) is satisfied, then the terminal value problem (11.15), (11.16) is solvable on $\mathcal{C}^1_{A_*^*}(\mathcal{I}, \mathbb{R}^k)$.*

Proof. The rank condition (11.14) implies the rank condition (11.13) to be given in a neighborhood $\mathfrak{G}$ of the graph of x_*. Then the assertion is a direct consequence of Theorem 11.12. □

Indirect optimization methods rely on the boundary value problem (BVP) for the composed so-called *optimality DAE*

$$f((D(t)x(t))',x(t),t) = 0, \tag{11.17}$$

$$-D(t)^*(f_y((D(t)x(t))',x(t),t)^*\lambda(t))' + f_x((D(t)x(t))',x(t),t)^*\lambda(t) = h_x(x(t),t)^*, \tag{11.18}$$

completed by the boundary conditions (11.11) and (11.16). Owing to Theorem 11.12 this BVP is solvable. By introducing the new function $y = (Dx)'$ and collecting the components λ, x, y in $\tilde{x}$, the DAE (11.17), (11.18) can be put into the more prevalent form

$$\tilde{f}((\tilde{d}(\tilde{x}(t),t))',\tilde{x}(t),t) = 0,$$

with properly involved derivative and nonlinear derivative term. This kind of equations is investigated in Chapter 3. Here we restrict our further interest to the easier quasi-linear case

$$f(y,x,t) = A(t)y + b(x,t), \tag{11.19}$$

which naturally comprises all semi-explicit systems.

In the case of the particular form (11.19), the optimality DAE simplifies to

$$A(t)(D(t)x(t))' + b(x(t),t) = 0, \tag{11.20}$$

$$-D(t)^*(A(t)^*\lambda(t))' + b_x(x(t),t)^*\lambda(t) = h_x(x(t),t)^*. \tag{11.21}$$

The optimality DAE combines $k+m$ equations for the same number of unknown functions. In view of a reliable practical treatment, when applying an indirect optimization method, it would be a great advantage to know whether the optimality DAE is regular with index 1. For this aim we consider the linearization of the DAE (11.20), (11.21) along (λ_*, x_*), namely

$$\begin{bmatrix} A(t) & 0 \\ 0 & D(t)^* \end{bmatrix} \left(\begin{bmatrix} 0 & D(t) \\ -A(t)^* & 0 \end{bmatrix} \begin{bmatrix} \lambda(t) \\ x(t) \end{bmatrix} \right)' + \begin{bmatrix} 0 & B_*(t) \\ B_*(t)^* & -H_*(t) \end{bmatrix} \begin{bmatrix} \lambda(t) \\ x(t) \end{bmatrix} = 0, \tag{11.22}$$

with the continuous matrix functions

$$H_*(t) := h_{xx}(x_*(t),t) - (b_x^*(x,t)\lambda_*(t))_x(x_*(t),t), \quad B_*(t) := b_x(x_*(t),t). \tag{11.23}$$

Theorem 11.16. *Let the Assumptions 11.10 be satisfied and let the DAE (11.10) have the special form given by (11.19). Let the functions b and h have the additional second continuous partial derivatives b_{xx}, h_{xx}. Set*

$$Q_0(t) = I - D(t)^+D(t), \quad W_0(t) = I - A(t)A(t)^+, \quad t \in \mathcal{I}.$$

Let $x_* \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ be a local solution of the optimization problem (11.9), (11.10), (11.11) and let the rank condition (11.14) be satisfied, that is, with the denotation (11.23)

$$\operatorname{rank}[A(t)D(t) + B_*(t)Q_0(t)] = k, \quad t \in \mathcal{I}. \tag{11.24}$$

If, additionally, λ_* denotes the resulting solution of the terminal value problem (11.15), (11.16), then

(1) *the optimality DAE (11.20), (11.21) is regular with index 1 in a neighborhood of the graph of (λ_*, x_*), exactly if*

$$\begin{aligned}(A(t)D(t)+W_0(t)B_*(t)Q_0(t))z &= 0,\\ H_*(t)Q_0(t)z \in \ker\,(A(t)D(t)+W_0(t)B_*(t)Q_0(t))^{\perp}&\\ \text{imply}\quad z=0,\quad \text{for all}\quad t\in\mathcal{I};&\end{aligned}\tag{11.25}$$

(2) *the linearized DAE (11.22) is self-adjoint and its inherent regular ODE has Hamiltonian structure such that*

$$\Theta' = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} \mathcal{E}\,\Theta, \quad \Theta := \begin{bmatrix} Dx \\ -A^*\lambda \end{bmatrix}, \tag{11.26}$$

with a symmetric continuous matrix function $\mathcal{E}$ of size $2n\times 2n$, supposing also condition (11.25) is given.

Furthermore, if $Q_0(t)H_*(t)Q_0(t)$ is semidefinite for all $t\in\mathcal{I}$, then condition (11.25) simplifies to the full rank condition

$$\operatorname{rank}\begin{bmatrix} A(t)D(t)+W_0(t)B_*(t)Q_0(t) \\ Q_0(t)H_*(t)Q_0(t) \end{bmatrix} = m, \quad t\in\mathcal{I}. \tag{11.27}$$

Proof. (1) Set $G=AD$. The proper formulation of the leading term yields $\ker G = \ker D$ and $\operatorname{im} G = \operatorname{im} A$, therefore $Q_0 = I - D^+D = I - G^+G$, $W_0 = I - GG^+ = I - AA^+$. Introduce the $(m+k)\times(k+m)$ matrix function

$$\hat{G}_{*1} = \begin{bmatrix} 0 & G+B_*Q_0 \\ -G^*+B_*^*W_0 & -H_*Q_0 \end{bmatrix}.$$

The optimality DAE (11.20), (11.21) is regular with index 1 around the graph of (x_*,λ_*), exactly if $\hat{G}_{*1}$ is nonsingular on $\mathcal{I}$. Compute the relations

$$\begin{aligned} G+B_*Q_0 &= (G+W_0B_*Q_0)(I+G^+B_*Q_0),\\ -G^*+B_*^*W_0 &= (-G^*+Q_0B_*^*W_0)(I-G^{+*}B_*^*W_0),\\ \operatorname{im}(G+W_0B_*Q_0) &= \operatorname{im} G \oplus \operatorname{im} W_0B_*Q_0,\\ \operatorname{im}(-G^*+Q_0B_*^*W_0) &= \operatorname{im} G^* \oplus \operatorname{im} Q_0B_*^*W_0 = \operatorname{im}(G^*+Q_0B_*^*W_0). \end{aligned}$$

From condition (11.24) it now follows that $\operatorname{rank}(-G^*+B_*^*W_0) = \operatorname{rank}(G+B_*Q_0) = k$, and hence $\ker(-G^*+B_*^*W_0) = \{0\}$.
The matrix function $\hat{G}_{*1}$ is nonsingular if, for $v\in\mathbb{R}^m$ and $w\in\mathbb{R}^k$, the system

$$(G+B_*Q_0)v = 0, \tag{11.28}$$

$$-H_*Q_0v + (-G^*+B_*^*W_0)w = 0 \tag{11.29}$$

has only the trivial solution. Since $-G^* + B_*^* W_0$ has full column rank and

$$\operatorname{im}(-G^* + B_*^* W_0) = \ker(G + W_0 B_* Q_0)^\perp,$$

equation (11.29) is equivalent to

$$-H_* Q_0 v \in \ker(G + W_0 B_* Q_0)^\perp, \quad w = (-G^* + B_*^* W_0)^+ H_* Q_0 v.$$

Introduce $\tilde{v} = (I + G^+ b_x Q_0)v$ so that $Q_0 v = Q_0 \tilde{v}$. Now it is clear that $\hat{G}_{*1}$ is nonsingular exactly if

$$(G + W_0 B_* Q_0)\tilde{v} = 0, \tag{11.30}$$

$$-H_* Q_0 \tilde{v} \in \ker(G + W_0 B_* Q_0)^\perp \tag{11.31}$$

imply $\tilde{z}_1 = 0$. This proves (1).
(2) The linearized DAE (11.22) is self-adjoint. As a self-adjoint index-1 DAE it has Hamiltonian structure.
It remains to verify the last assertion concerning condition (11.27). Equation (11.30) decomposes to $G\tilde{v} = 0$ and $W_0 b_x Q_0 \tilde{v} = 0$, thus $\tilde{v} = Q_0 \tilde{v}$ and $\tilde{v} \in \ker W_0 B_*$. Moreover, regarding condition (11.31), $\tilde{v}$ and $H_* \tilde{v}$ are orthogonal, $0 = \langle H_* \tilde{v}, \tilde{v} \rangle = \langle Q_0 H_* Q_0 \tilde{v}, \tilde{v} \rangle$. Since $Q_0 H_* Q_0$ is symmetric and semidefinite, it follows that $Q_0 H_* Q_0 v = 0$, and we are done. □

Example 11.17 (Driving a point to a circle). [6, p.144–146] Minimize the cost

$$J(x) = \frac{1}{2} \int_0^{t_f} (x_3(t)^2 + (x_4(t) - R^2)^2) \mathrm{d}t$$

subject to the constraint

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

with constants $r > 0, R > 0$. If $x_3(t)$ vanishes identically, the remaining IVP has a unique solution. Then the point $(x_1(t), x_2(t))$ orbits the origin with radius r and $x_4(t) = r$. By optimizing in view of the cost, the point $(x_1(t), x_2(t))$ becomes driven to the circle of radius R, with low cost of $x_3(t)$.
The resulting optimality DAE is everywhere regular with index 1. Namely, we have $m = 4, n = 2, k = 3$, and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad b(x,t) = \begin{bmatrix} x_2 \\ -x_1 - x_3 \\ -x_1^2 - x_2^2 + x_4 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

and condition (11.27) is satisfied, since

$$AD + b_x(x,t)(I - D^+D) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

has full row rank. The adjoint system comprises four equations and three unknown functions $\lambda_1, \lambda_2, \lambda_3$. The resulting optimality DAE

$$\begin{aligned} x_1'(t) + x_2(t) &= 0, \\ x_2'(t) - x_1(t) - x_3(t) &= 0, \\ -x_1(t)^2 - x_2(t)^2 + x_4(t) &= 0, \\ -\lambda_1'(t) - \lambda_2(t) - 2x_1(t)\lambda_3(t) &= 0, \\ -\lambda_2'(t) + \lambda_1(t) - 2x_2(t)\lambda_3(t) &= 0, \\ -\lambda_2(t) &= x_3(t), \\ \lambda_3(t) &= x_4(t) - R^2, \end{aligned}$$

has dimension 7 and is everywhere regular with index 1. Here we have

$$H_*(t) = \begin{bmatrix} 2\lambda_*(t) & 0 & 0 & 0 \\ 0 & 2\lambda_*(t) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad Q_0(t)H_*(t)Q_0(t) = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

so that condition (11.27) applies. □

The next example emphasizes that the only property that matters for the necessary extremal condition is the surjectivity of the respective operator, that is, the full rank condition (11.25). However, for obtaining a regular index-1 optimality DAE (11.20), (11.21), additionally, the cost must be somehow consistent with the DAE describing the constraint.

Example 11.18 (Consistency of cost and DAE constraint). Minimize the cost

$$J(x) = \frac{1}{2}\gamma x_2(t_f)^2 + \frac{1}{2}\int_0^{t_f} (\,\alpha\,(x_1(t) + x_3(t))^2 + \beta\, x_2(t)^2\,)\mathrm{d}t$$

subject to the constraint

$$\begin{aligned} x_2'(t) + x_2(t) + x_3(t) &= 0, \qquad & x_2(0) &= 1, \\ x_2'(t) + \sin t &= 0, \end{aligned}$$

with constants $\alpha, \beta, \gamma \geq 0$, $\alpha^2 + \beta^2 + \gamma^2 > 0$. The optimal solution is

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

We have

$$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad b(x,t) = \begin{bmatrix} x_2 + x_3 \\ \sin t \end{bmatrix}, \quad D = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix}, \quad AD + b_x(x,t)Q_0 = \begin{bmatrix} 0 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

so that condition (11.14) is satisfied for all α, β, γ. The optimality DAE reads

$$\begin{aligned} x_2'(t) + x_2(t) + x_3(t) &= 0, \\ x_2'(t) + \sin t &= 0, \\ -\alpha(x_1(t) + x_3(t)) &= 0, \\ -(\lambda_1(t) + \lambda_2(t))' + \lambda_1(t) - \beta x_2(t) &= 0, \\ \lambda_1(t) - \alpha(x_1(t) + x_3(t)) &= 0. \end{aligned}$$

This square DAE of dimension 5 is regular with index 1 exactly if α does not vanish. This condition reflects condition (11.27). Namely, we have

$$H_*(t) = h_{xx}(x_*(t),t) = \begin{bmatrix} \alpha & 0 & \alpha \\ 0 & \beta & 0 \\ \alpha & 0 & \alpha \end{bmatrix}, \quad Q_0(t)H_*(t)Q_0(t) = \begin{bmatrix} \alpha & 0 & \alpha \\ 0 & 0 & 0 \\ \alpha & 0 & \alpha \end{bmatrix},$$

$$G + W_0 B_* Q_0 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 & \frac{1}{2} \\ 0 & 1 & -\frac{1}{2} \end{bmatrix}.$$

In contrast, the optimality DAE fails to be regular for $\alpha = 0$.

In this example, the constraint DAE (11.10) consists of two equations for three unknown functions. If one regards x_3 as a control, the resulting controlled DAE (with respect to x_1, x_2) fails to be regular. In contrast, regarding x_1 as control, the resulting controlled DAE (with respect to x_2, x_3) is regular with index 1. This emphasizes that, in the context of minimization with DAE constraints the only property that matters for the extremal condition is the surjectivity of the operator representing the linearization (11.12). Moreover, for obtaining a regular index-1 optimality DAE (11.20), (11.21), the cost functional must be somehow consistent with the DAE describing the constraint. □

For BVPs in regular index-1 DAEs approved numerical solution methods are available such that indirect optimization can be expected to work well in practice. However, one should take great care to ensure the conditions are responsible for the nice properties. Otherwise one can go into trouble. Though the optimization problem is uniquely solvable and the associated optimality BVP is also solvable, if the BVP solution (x_*, λ_*) does not stay in a index-1 regularity region of the optimality DAE, then it can be hopeless to solve the optimality DAE numerically.

Example 11.19 (The optimality DAE has several regularity regions). Minimize the cost

$$J(x) = \frac{1}{2} \int_0^{2\pi} \big((x_1(t) - \sin t)^2 + (x_2(t) - \cos t)^2 + \gamma x_3(t)^2 + x_4(t)^2 \big) \mathrm{d}t$$

subject to the constraint

$$\begin{aligned} x_1'(t) - x_2(t) + x_3(t) &= 0, & x_1(0) &= 0, \\ x_2'(t) + x_1(t) &= 0, & x_2(0) &= 1, \\ x_1(t)^3 + \alpha(x_1(t))x_3(t) - (\sin t)^3 - x_4(t) &= 0, \end{aligned}$$

with a constant $\gamma \geq 0$ and the real function α given by

$$\alpha(s) := \begin{cases} s^3 \text{ if } s > 0 \\ 0 \text{ if } s \leq 0. \end{cases}$$

The minimization problem has the unique optimal solution

$$x_{*1}(t) = \sin t,\ x_{*2}(t) = \cos t,\ x_{*3}(t) = 0,\ x_{*4}(t) = 0.$$

We have $m = 4$, $k = 3$, $n = 2$, and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad b(x,t) = \begin{bmatrix} -x_2 + x_3 \\ x_1 \\ x_1^3 + \alpha(x_1)x_3 - (\sin t)^3 - x_4 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix},$$

such that the matrix function

$$AD + b_x(x,t)(I - D^+D) = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \alpha(x_1) & -1 \end{bmatrix}$$

results, which has full row rank independently of the behavior of $\alpha(x_1)$. Therefore, the associated terminal value problem and hence the optimality BVP are solvable. The optimality DAE reads in detail

$$\begin{aligned} x_1'(t) - x_2(t) + x_3(t) &= 0, \\ x_2'(t) + x_1(t) &= 0, \\ x_1(t)^3 + \alpha(x_1(t))x_3(t) - (\sin t)^3 - x_4(t) &= 0, \\ -\lambda_1'(t) + \lambda_2(t) + (3x_1(t)^2 + \alpha'(x_1(t))x_3(t))\lambda_3(t) &= x_1(t) - \sin t, \\ -\lambda_2'(t) - \lambda_1(t) &= x_2(t) - \cos t, \\ \lambda_1(t) + \alpha(x_1(t))\lambda_3(t) &= \gamma x_3(t), \\ -\lambda_3(t) &= x_4(t). \end{aligned}$$

It holds that

$$AD + W_0 B_* Q_0 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \alpha(x_1) & -1 \end{bmatrix},$$

$$H_*(t) = \begin{bmatrix} 1 - 6x_{*1}(t)\lambda_{*3}(t) & 0 & -\alpha'(x_{*1}(t))\lambda_{*3}(t) & 0 \\ 0 & 1 & 0 & 0 \\ -\alpha'(x_{*1}(t))\lambda_{*3}(t) & 0 & \gamma & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$Q_0 H_*(t) Q_0 = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \gamma & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Condition (11.27) requires $\gamma + (\alpha(x_1))^2 \neq 0$. Therefore, the optimality DAE is globally regular with index 1 in case of $\gamma > 0$.
In contrast, if $\gamma = 0$, then merely the set

$$\mathcal{G}_1 = \{(z,t) \in \mathbb{R}^7 \times (t_0, t_f) : z_1 > 0\}$$

is a regularity region with index $\mu = 1$, but there are two further regularity regions

$$\mathcal{G}_2 = \left\{(z,t) \in \mathbb{R}^7 \times (t_0, t_f) : z_1 < 0, \frac{9z_1^4 + 1}{6z_1} > z_7\right\},$$

$$\mathcal{G}_3 = \left\{(z,t) \in \mathbb{R}^7 \times (t_0, t_f) : z_1 < 0, \frac{9z_1^4 + 1}{6z_1} < z_7\right\}$$

with tractability index 3. Unfortunately, the optimal solution does not remain in the index-1 region but shuttles between $\mathcal{G}_1$ and $\mathcal{G}_3$. This is accompanied by a discontinuous neighboring flow and causes numerical difficulties. In practice, these difficulties are reflected also in case of small γ. □

Altogether, when intending to apply an indirect optimization method, it seems to be a good idea to make use of the modeling latitude and to reach an *optimality DAE which is regular with index-1* or, at least, to reach the situation that the expected solution stays in an index-1 regularity region. Only in this way can the indirect optimization work safely.

11.2.2 A particular sufficient extremal condition

Consider the quadratic cost functional

$$J(x) = \frac{1}{2}\langle VD(t_f)x(t_f), D(t_f)x(t_f)\rangle + \frac{1}{2}\int_{t_0}^{t_f} \langle W(t)x(t), x(t)\rangle \mathrm{d}t \tag{11.32}$$

to be minimized on functions $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$, subject to the constraints

$$A(t)(D(t)x(t))' + B(t)x(t) - q(t) = 0, \quad t \in \mathcal{I} = [t_0, t_f], \tag{11.33}$$

and

$$D(t_0)x(t_0) = z_0 \in \mathbb{R}^n. \tag{11.34}$$

The matrices V and $W(t)$, $t \in \mathcal{I}$, are supposed to be symmetric, positive semidefinite. Let the pair $(x_*, \lambda_*) \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \times \mathcal{C}_{A^*}^1(\mathcal{I},\mathbb{R}^k)$ be a solution of the BVP

$$A(t)(D(t)x(t))' + B(t)x(t) - q(t) = 0, \tag{11.35}$$

$$-D(t)^*(A(t)^*\lambda(t))' + B(t)^*\lambda(t) = W(t)x(t), \tag{11.36}$$

$$D(t_0)x(t_0) = z_0, \tag{11.37}$$

$$D(t_f)^*A(t_f)^*\lambda(t_f) = D(t_f)^*VD(t_f)x(t_f). \tag{11.38}$$

Then, for any $x \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m)$ and $\Delta x := x - x_*$, it holds that

$$\begin{aligned} &J(x) - J(x_*) \\ &\quad = \frac{1}{2}\int_{t_0}^{t_f} \langle W(t)\Delta x(t), \Delta x(t)\rangle \mathrm{d}t + \frac{1}{2}\langle VD(t_f)\Delta x(t_f), D(t_f)\Delta x(t_f)\rangle + \mathfrak{E} + \mathfrak{F}, \end{aligned} \tag{11.39}$$

with

$$\mathfrak{E} := \int_{t_0}^{t_f} \langle W(t)\Delta x(t), x_*(t)\rangle \mathrm{d}t, \quad \mathfrak{F} := \langle VD(t_f)\Delta x(t_f), D(t_f)x_*(t_f)\rangle.$$

The expression $\mathfrak{E} + \mathfrak{F}$ vanishes. To verify this, we derive

$$\begin{aligned} \mathfrak{E} &= \int_{t_0}^{t_f} \langle \Delta x(t), W(t)x_*(t)\rangle \mathrm{d}t \\ &= \int_{t_0}^{t_f} \langle \Delta x(t), -D(t)^*(A(t)^*\lambda_*(t))' + B(t)^*\lambda_*(t)\rangle \mathrm{d}t \\ &= \int_{t_0}^{t_f} \{-\langle D(t)\Delta x(t), (A(t)^*\lambda_*(t))'\rangle + \langle B(t)\Delta x(t), \lambda_*(t)\rangle\} \mathrm{d}t \\ &= \int_{t_0}^{t_f} \{-\langle D(t)\Delta x(t), (A(t)^*\lambda_*(t))'\rangle - \langle (D(t)\Delta x(t))', A(t)^*\lambda_*(t)\rangle\} \mathrm{d}t \\ &= -\int_{t_0}^{t_f} (\langle D(t)\Delta x(t), A(t)^*\lambda_*(t)\rangle)' \mathrm{d}t \\ &= -\langle D(t_f)\Delta x(t_f), A(t_f)^*\lambda_*(t_f)\rangle, \end{aligned}$$

further

$$\begin{aligned} \mathfrak{E} + \mathfrak{F} &= -\langle D(t_f)\Delta x(t_f), A(t_f)^*\lambda_*(t_f)\rangle + \langle VD(t_f)\Delta x(t_f), D(t_f)x_*(t_f)\rangle \\ &= \langle \Delta x(t_f), D(t_f)^*\{-A(t_f)^*\lambda_*(t_f) + VD(t_f)x_*(t_f)\}\rangle = 0. \end{aligned}$$

Proposition 11.20. *Let all coefficients in the minimization problem (11.32)–(11.34) be continuous. Let V and W be symmetric, positive semidefinite.*
Then, if the optimality BVP (11.35)–(11.38) possesses a solution $(x_*, \lambda_*) \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \times \mathcal{C}_{A^*}^1(\mathcal{I}, \mathbb{R}^k)$, the component x_* is a solution of the minimization problem. If W is even positive definite, then x_* is the only solution of the minimization problem.

Proof. Relation (11.39) yields the inequality $J(x) \geq J(x_*)$ for all $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$. If there are two minimizers x_* and $\bar{x}_*$, $J(\bar{x}_*) = J(x_*)$, with $\Delta_* := \bar{x}_* - x_*$, it results that $\langle W(t)\Delta_*(t), \Delta_*(t)\rangle = 0$ for all $t \in \mathcal{I}$. Owing to the positive definiteness of $W(t)$ it follows that $\Delta_*(t)$ vanishes identically. □

In Proposition 11.20 no restrictions concerning the constraint DAE and the optimality DAE are required. Neither properly stated leading terms nor index conditions are prescribed.

The DAE (11.35), (11.36) looks like the optimality DAE (11.20), (11.21) specified for the minimization problem considered now. In contrast to (11.20), (11.21), here we dispense with the requirements concerning the full rank proper leading term and condition (11.14).

11.3 Specification for controlled DAEs

In this section we specify results from Section 11.2 for easier application to the important case of constraints described by *controlled* DAEs. Here the DAE and the cost functional depend on a pair of functions, the *state* $x \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m)$ and the *control* $u \in \mathcal{C}(\mathcal{I}, \mathbb{R}^l)$. Now the DAE comprises m equations so that, for each fixed control, a square m-dimensional DAE results. Consider the cost functional

$$J(x,u) = g(D(t_f)x(t_f)) + \int_{t_0}^{t_f} h(x(t), u(t), t)\mathrm{d}t \tag{11.40}$$

to be minimized on pairs $(x,u) \in \mathcal{C}_D^1(\mathcal{I}, \mathbb{R}^m) \times \mathcal{C}(\mathcal{I}, \mathbb{R}^l)$, subject to the constraints

$$f((D(t)x(t))', x(t), u(t), t) = 0,\ t \in \mathcal{I}, \tag{11.41}$$

and

$$D(t_0)x(t_0) = z_0 \in \mathbb{R}^n. \tag{11.42}$$

We suppose an analogous setting as in Assumption 11.10, but with different denotations.

Assumption 11.21. *The function $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^l \times \mathcal{I} \to \mathbb{R}^m$ is continuous and has continuous partial derivatives f_y, f_x, f_u with respect to the first three variables $y \in \mathbb{R}^n$, $x \in \mathbb{R}^m$, $u \times \mathbb{R}^l$. The matrix function $D : \mathcal{I} \to L(\mathbb{R}^m, \mathbb{R}^n)$ is continuous.*

The DAE (11.10) comprises m equations. It has a full rank proper leading term, that is, $n \le m$, and f_y has full column rank n, and D has full row rank n on their definition domains.
The functions h and g are continuously differentiable.

Denote

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

such that now the linearization along (x_*, u_*) reads

$$A_*(t)(D(t)x(t))' + B_*(t)x(t) + C_*(t)u(t) = q(t), \quad t \in \mathcal{I}. \tag{11.43}$$

Our present problem constitutes a special case with partitioned variables of the general optimization problem considered in the previous sections. The following necessary extremal condition is a straightforward consequence of Theorem 11.12.

Theorem 11.22. *Let the Assumption 11.21 be given. If the optimization problem (11.40), (11.41), (11.42) has the local solution $(x_*, u_*) \in \mathcal{C}^1_D(\mathcal{I}, \mathbb{R}^m) \times \mathcal{C}(\mathcal{I}, \mathbb{R}^l)$ and if the full rank condition*

$$\operatorname{rank}\left[A_*(t)D(t) + B_*(t)(I - D(t)^+D(t)),\ C_*(t)\right] = m, \quad t \in \mathcal{I}, \tag{11.44}$$

is valid, then the terminal value problem

$$-D(t)^*(A_*(t)^*\lambda(t))' + B_*(t)^*\lambda(t) = h_x(x_*(t), u_*(t), t)^*, \tag{11.45}$$

$$C_*(t)^*\lambda(t) = h_u(x_*(t), u_*(t), t)^*, \quad t \in \mathcal{I} \tag{11.46}$$

$$D(t_f)^*A_*(t_f)^*\lambda(t_f) = D(t_f)^*(g_\eta(D(t_f)x_*(t_f)))^* \tag{11.47}$$

possesses a solution $\lambda_* \in \mathcal{C}^1_{A_*^*}(\mathcal{I}, \mathbb{R}^m)$.

If the controlled DAE is regular with index ≤ 1, then $A_*D + B_*(I - D^+D)$ is necessarily nonsingular such that condition (11.44) is valid independently of what C_* looks like. However, in all other cases, condition (11.44) entails structural requirements concerning C_*.

On the other hand, no regularity and index conditions for the given controlled DAE are required. For instance, in Example 11.17, one might consider x_3 or x_4 to serve as the control. In the first case, the resulting controlled DAE is regular with index 1, and in the second case it is regular with index 2 on the two regularity regions given by $x_2 > 0$ and $x_2 < 0$. Both versions result in the same regular index-1 optimality DAE.

Owing to Theorem 11.22, the BVP composed from the IVP (11.41), (11.42) and the terminal value problem (11.45)–(11.47) is solvable. Indirect optimization relies

on this BVP. Then, for practical reasons, the question arises whether the associated *optimality DAE* is regular with index 1. We answer for the quasi-linear case.

$$f(y,x,u,t) = A(t)y + b(x,u,t), \tag{11.48}$$

so that the optimality DAE simplifies to

$$A(t)(D(t)x(t))' + b(x(t),u(t),t) = 0, \tag{11.49}$$
$$-D(t)^*(A(t)^*\lambda(t))' + b_x(x(t),u(t),t)^*\lambda(t) = h_x(x(t),u(t),t)^*, \tag{11.50}$$
$$b_u(x(t),u(t),t)^*\lambda(t) = h_u(x(t),u(t),t)^*. \tag{11.51}$$

The optimality DAE (11.49)–(11.51) has the linearization (the argument t is dropped)

$$\begin{bmatrix} A & 0 \\ 0 & D^* \\ 0 & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & D & 0 \\ -A^* & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} \right)' + \begin{bmatrix} 0 & B_* & C_* \\ B_*^* & -\mathcal{W}_* & -\mathcal{S}_* \\ C_*^* & -\mathcal{S}_*^* & -\mathcal{R}_* \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix} = 0, \tag{11.52}$$

with continuous matrix functions

$$\begin{aligned} \mathcal{W}_*(t) &:= h_{xx}(x_*(t),u_*(t),t) - (b_x(x,u,t)^*\lambda_*(t))_x(x_*(t),u_*(t),t), \\ \mathcal{S}_*(t) &:= h_{xu}(x_*(t),u_*(t),t)^* - (b_x(x,u,t)^*\lambda_*(t))_u(x_*(t),u_*(t),t), \\ \mathcal{R}_*(t) &:= h_{uu}(x_*(t),u_*(t),t) - (b_u(x,u,t)^*\lambda_*(t))_u(x_*(t),u_*(t),t), \\ B_*(t) &:= b_x(x_*(t),u_*(t),t), \quad C_*(t) = b_u(x_*(t),u_*(t),t), \quad t \in \mathcal{I}. \end{aligned}$$

Theorem 11.23. *Let Assumption 11.21 be valid, let the DAE in (11.41) have the special form given by (11.48), and let the functions b and h have the necessary additional second continuous partial derivatives.*
Denote $Q_0(t) = I_m - D(t)^+D(t)$, $W_0(t) = I_m - A(t)A(t)^+$, $t \in \mathcal{I}$.
Let $(x_*,u_*) \in \mathcal{C}_D^1(\mathcal{I},\mathbb{R}^m) \times \mathcal{C}(\mathcal{I},\mathbb{R}^l)$ be a local solution of the optimization problem (11.40), (11.41), (11.42) and let the full rank condition (11.44) be satisfied. Denote by λ_* the solution of the terminal value problem (11.45)–(11.47).

(1) *Then the optimality DAE (11.49)–(11.51) is regular with index 1 in a neighborhood of the graph of (λ_*,x_*,u_*), exactly if*

$$\begin{aligned} &[A(t)D(t) + W_0(t)B_*(t)Q_0(t),\ W_0(t)C_*(t)]\,z = 0, \\ &\begin{bmatrix} \mathcal{W}_*(t)Q_0(t) & \mathcal{S}_*(t) \\ \mathcal{S}_*(t)^*Q_0(t) & \mathcal{R}_*(t) \end{bmatrix} z \in \ker\,[A(t)D(t) + W_0(t)B_*(t)Q_0(t),\ W_0(t)C_*(t)]^\perp \\ &\qquad\qquad \text{imply} \quad z = 0, \quad \text{for all} \quad t \in \mathcal{I}. \end{aligned} \tag{11.53}$$

(2) *If condition (11.53) is valid, then the linearized DAE (11.52) is self-adjoint and its inherent regular ODE has Hamiltonian structure such that*

$$\Theta' = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} \mathcal{E}\,\Theta, \quad \Theta := \begin{bmatrix} Dx \\ -A^*\lambda \end{bmatrix}, \tag{11.54}$$

with a symmetric continuous matrix function $\mathcal{E}$ of size $2n \times 2n$.

(3) *If the matrix*

$$\begin{bmatrix} Q_0(t)\mathcal{W}_*(t)Q_0(t) & Q_0(t)S_*(t) \\ \mathcal{S}_*(t)^*Q_0(t) & \mathcal{R}_*(t) \end{bmatrix}$$

is semidefinite for all $t \in \mathcal{I}$, then condition (11.53) simplifies to the full rank condition

$$\operatorname{rank} \begin{bmatrix} A(t)D(t) + W_0(t)B_*(t)Q_0(t) & W_0(t)C_*(t) \\ Q_0(t)\mathcal{W}_*(t)Q_0(t) & Q_0(t)\mathcal{S}_*(t) \\ \mathcal{S}_*(t)^*Q_0(t) & \mathcal{R}_*(t) \end{bmatrix} = m + l, \quad t \in \mathcal{I}. \tag{11.55}$$

Proof. These assertions are special cases of the corresponding assertions in Theorem 11.16. □

11.4 Linear-quadratic optimal control and Riccati feedback solution

We deal with the quadratic cost functional

$$J(u,x) := \frac{1}{2}\langle x(t_f), Vx(t_f)\rangle + \frac{1}{2}\int_0^{t_f} \{\langle x(t), \mathcal{W}(t)x(t)\rangle + 2\langle x(t), \mathcal{S}(t)u(t)\rangle + \langle u(t), \mathcal{R}(t)u(t)\rangle\}\,\mathrm{d}t \tag{11.56}$$

to be minimized on pairs $(u,x) \in \mathcal{C}(\mathcal{I},\mathbb{R}^l) \times \mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$, satisfying the IVP

$$A(t)(D(t)x(t))' + B(t)(t)x(t) + C(t)u(t) = 0, \quad t \in \mathcal{I} = [0,t_f], \tag{11.57}$$

$$A(0)D(0)x(0) = z_0. \tag{11.58}$$

Equation (11.57) comprises k equations and $m+l$ unknowns. We agree upon the following basic assumptions.

Assumption 11.24. *The cost matrices $\mathcal{W}(t) \in L(\mathbb{R}^m,\mathbb{R}^m)$, $\mathcal{R}(t) \in L(\mathbb{R}^l,\mathbb{R}^l)$, $\mathcal{S}(t) \in L(\mathbb{R}^l,\mathbb{R}^m)$, as well as the coefficients $A(t) \in L(\mathbb{R}^n,\mathbb{R}^k)$, $D(t) \in L(\mathbb{R}^m,\mathbb{R}^n)$, $B(t) \in L(\mathbb{R}^m,\mathbb{R}^k)$, $C(t) \in L(\mathbb{R}^l,\mathbb{R}^k)$ depend continuously on $t \in \mathcal{I}$.*
The DAE (11.68) has a properly stated leading term with the border projector R.
The cost matrices satisfy the standard assumptions: $V \in L(\mathbb{R}^m,\mathbb{R}^m)$ is symmetric, positive semidefinite, and it holds that $\ker V = \ker D(t_f)$. $\mathcal{W}(t)$ and $\mathcal{R}(t)$, are symmetric $\mathcal{R}(t)$ is positive definite, and $\begin{bmatrix} \mathcal{W}(t) & \mathcal{S}(t) \\ \mathcal{S}(t)^* & \mathcal{R}(t) \end{bmatrix}$ is positive semidefinite for all $t \in \mathcal{I}$.

We use the symbols $\mathcal{C}$, $\mathcal{C}^1_D$, and $\mathcal{C}^1_{A^*}$ for the function spaces $\mathcal{C}(\mathcal{I},\mathbb{R}^l)$, $\mathcal{C}^1_D(\mathcal{I},\mathbb{R}^m)$, and $\mathcal{C}^1_{A^*}(\mathcal{I},\mathbb{R}^k)$, respectively.
A pair $(u,x)\in\mathcal{C}\times\mathcal{C}^1_D$ that satisfies the IVP (11.57), (11.58) is said to be *admissible*.

We introduce the projector valued functions Q,Q_*,P,P_* by

$$\begin{aligned} Q(t),P(t)\in L(\mathbb{R}^m,\mathbb{R}^m),\ Q(t)&=Q(t)^*,\\ \operatorname{im}Q(t)=\ker A(t)D(t),\ P(t)&=I-Q(t),\\ Q_*(t),P_*(t)\in L(\mathbb{R}^k,\mathbb{R}^k),\ Q_*(t)&=Q_*(t)^*,\\ \operatorname{im}Q_*(t)=\ker D(t)^*A(t)^*,\ P_*(t)&=I-Q_*(t),\quad t\in\mathcal{I}. \end{aligned}$$

The projector functions Q,P,Q_*, and P_* are continuous owing to the properly stated leading term. Note that now Q and Q_* stand for Q_0 and W_0 used in previous sections.

For brevity and greater transparency we almost always drop the argument t. Then the relations are meant pointwise for $t\in\mathcal{I}$.

Having the projectors R,P, and P_*, we introduce the generalized inverses D^- of D and A^{*-} of A^* by

$$\begin{aligned} D^-DD^- &= D^-, & DD^-D &= D, & DD^- &= R, & D^-D &= P,\\ A^{*-}A^*A^{*-} &= A^{*-}, & A^*A^{*-}A^* &= A^*, & A^*A^{*-} &= R^*, & A^{*-}A^* &= P_*. \end{aligned} \tag{11.59}$$

The generalized inverses D^- and A^{*-} are uniquely determined by (11.59). They are continuous on $\mathcal{I}$. It holds further that

$$D^-R=D^-,\ A=AR,\ A^*=R^*A^*,\ D^{-*}=R^*D^{-*}. \tag{11.60}$$

11.4.1 Sufficient and necessary extremal conditions

The so-called *optimality DAE* (cf. (11.52))

$$\begin{bmatrix} A & 0\\ 0 & D^*\\ 0 & 0 \end{bmatrix} \Big(\begin{bmatrix} 0 & D & 0\\ -A^* & 0 & 0 \end{bmatrix} \begin{bmatrix} \lambda\\ x\\ u \end{bmatrix} \Big)' + \begin{bmatrix} 0 & B & C\\ B^* & -\mathcal{W} & -\mathcal{S}\\ C^* & -\mathcal{S}^* & -\mathcal{R} \end{bmatrix} \begin{bmatrix} \lambda\\ x\\ u \end{bmatrix} = 0 \tag{11.61}$$

is closely related to the minimization problem (11.67), (11.68), (11.69). Complete this DAE by the boundary conditions

$$A(0)D(0)x(0)=z_0,\quad D(t_f)^*A(t_f)^*\lambda(t_f)=Vx(t_f). \tag{11.62}$$

The DAE (11.61) is self-adjoint. It comprises $k+m+l$ equations and the same number of unknowns.
Since $\mathcal{R}$ is invertible, the equation $C^*\lambda-\mathcal{S}^*x-\mathcal{R}u=0$ can be solved for

$$u = \mathcal{R}^{-1}(C^*\lambda - \mathcal{S}^* x), \tag{11.63}$$

and then u can be eliminated. The resulting DAE

$$\begin{bmatrix} A & 0 \\ 0 & D^* \end{bmatrix} \frac{\mathrm{d}}{\mathrm{d}t} \left(\begin{bmatrix} 0 & D \\ -A^* & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ x \end{bmatrix} \right) + \begin{bmatrix} C\mathcal{R}^{-1}C^* & B - C\mathcal{R}^{-1}\mathcal{S}^* \\ B^* - \mathcal{S}\mathcal{R}^{-1}C^* & -\mathcal{W} + \mathcal{S}\mathcal{R}^{-1}\mathcal{S}^* \end{bmatrix} \begin{bmatrix} \lambda \\ x \end{bmatrix} = 0 \tag{11.64}$$

is also self-adjoint and has dimension $k+m$. If (11.57) is actually an explicit ODE, that is, $A = D = I$, then system (11.64) is also an explicit ODE which is called a *Hamiltonian system* associated to the minimization problem . Then this explicit system shows a Hamiltonian flow. We adopt this name for the general system (11.64) which is a DAE. At least, if A and D have full column rank, respectively full row rank, and the DAE is regular with index 1, then the IERODE of this ODE shows a Hamiltonian flow.

Theorem 11.25. *Let the Assumption 11.24 be valid.*

(1) *If the triple $(\lambda_*, x_*, u_*) \in \mathcal{C}^1_{A_*} \times \mathcal{C}^1_D \times \mathcal{C}$ is a solution of the BVP (11.61), (11.62), then $(x_*, u_*) \in \mathcal{C}^1_D \times \mathcal{C}$ is a solution of the minimization problem (11.56), (11.57), (11.58).*

(2) *Conversely, if $(x_*, u_*) \in \mathcal{C}^1_D \times \mathcal{C}$ is an optimal pair of the minimization problem (11.56), (11.57), (11.58) and the condition*

$$\operatorname{im}\left[A(t)D(t) + B(t)Q(t) \quad C(t)\right] = \mathbb{R}^k, \quad t \in \mathcal{I}, \tag{11.65}$$

is satisfied, then there exists a $\lambda_* \in \mathcal{C}^1_{A_*}$ such that the triple (λ_*, x_*, u_*) is a solution of the BVP (11.61), (11.62).

(3) *The DAE (11.61) is regular with tractability index 1, exactly if the condition (11.65) is satisfied and, additionally,*

$$\operatorname{im} \begin{bmatrix} D(t)^* A(t)^* + B(t)^* Q_*(t) & \mathcal{W}(t)Q(t) & \mathcal{S}(t) \\ C(t)^* Q_*(t) & \mathcal{S}(t)^* Q(t) & \mathcal{R}(t) \end{bmatrix} = \mathbb{R}^m \times \mathbb{R}^l, \quad t \in \mathcal{I}. \tag{11.66}$$

Proof. (1) is immediately verified along the lines of Proposition 11.20.
(2) is already given for $k = m$ and a DAE (11.57) with full rank proper leading term by Theorem 11.22. The general assertion is proved in [6, Theorem 3.22].
(3) is equivalent to Theorem 3.3 in [8]. □

11.4.2 Riccati feedback solution

Feedback solutions via Riccati differential equations are a known and proven tool for solving linear-quadratic optimal control problems given by the cost functional

$$J(u,x) := \frac{1}{2}\langle x(t_f), Vx(t_f)\rangle + \frac{1}{2}\int_0^{t_f} \left\langle \begin{bmatrix} x(t) \\ u(t) \end{bmatrix}, \begin{bmatrix} \mathcal{W}(t) & \mathcal{S}(t) \\ \mathcal{S}(t)^* & \mathcal{R}(t) \end{bmatrix} \begin{bmatrix} x(t) \\ u(t) \end{bmatrix} \right\rangle \mathrm{d}t \tag{11.67}$$

and the side conditions

$$x'(t) + B(t)x(t) + C(t)u(t) = 0, \quad t \in [0, t_f], \tag{11.68}$$

$$x(0) = z_0. \tag{11.69}$$

Solving the terminal value problem for the Riccati matrix differential equation

$$Y' = YB + B^*Y + (\mathcal{S} - YC)\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y) - \mathcal{W}, \tag{11.70}$$

$$Y(t_f) = V, \tag{11.71}$$

whose solution Y is symmetric, the minimization problem is traced back to the linear IVP

$$x' = -Bx + C\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)x, \quad x(0) = z_0. \tag{11.72}$$

The solution x_* of the IVP (11.72) together with $u_* := -\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)x_*$ solve the minimization problem (11.67), (11.68), (11.69). The minimal cost is then $J(u_*, x_*) = \langle z_0, Y(0)^* z_0 \rangle$.

If the explicit ordinary differential equation in (11.68) is replaced by the DAE (11.57) the situation is much more challenging. Different kinds of generalizations of the Ricatti approach can be imagined. Here, we work with the differential equation

$$D^*(A^*YD^-)'D = Y^*B + B^*Y + (\mathcal{S} - Y^*C)\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y) - \mathcal{W} \tag{11.73}$$

and the terminal value condition

$$A(t_f)^*Y(t_f)D(t_f)^- = D(t_f)^{-*}VD(t_f)^-, \tag{11.74}$$

which generalize the terminal value problem (11.70), (11.71). However, whereas the standard Riccati ODE (11.70) applies to a square matrix function Y, we are now looking for a rectangular one with k rows and m columns. The standard Riccati ODE is an explicit matrix differential equation, with continuously differentiable solutions. In contrast, equation (11.73) rather looks like a DAE so that we adopt the name *Riccati DAE*. The solution Y of the Riccati DAE is expected to be continuous with a continuously differentiable component A^*YD^-.

We prepare a symmetry property to be used later on.

Lemma 11.26. *If $Y : [0, t_f] \to L(\mathbb{R}^m, \mathbb{R}^k)$ is continuous with a continuously differentiable part A^*YD^-, and if it satisfies the terminal value problem (11.73), (11.74), then the symmetry relation*

$$A^*YD^- = D^{-*}Y^*A \tag{11.75}$$

becomes true.
If, additionally, Y satisfies the condition $A^*YQ = 0$, then it follows that

$$D^*A^*Y = Y^*AD. \tag{11.76}$$

Proof. Multiplying (11.73) by D^{-*} from the left, and by D^{-} from the right, leads to

$$R^*(A^*YD^-)'R = D^{-*}\{Y^*B + B^*Y + (\mathcal{S} - Y^*C)\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y) - \mathcal{W}\}D^- =: \mathfrak{A},$$

where $\mathfrak{A} = \mathfrak{A}^*$, and, further (cf. (11.60)),

$$(A^*YD^-)' = \mathfrak{A} + R^{*'}A^*YD^- + A^*YD^-R'.$$

It becomes clear that $U := A^*YD^-$ satisfies the ODE $U' = \mathfrak{A} + R^{*'}U + UR'$ as well as the condition $U(t_f) = D(t_f)^{-*}VD(t_f)^-$. Obviously, U^* is a further solution of the same final value problem; i.e., $U = U^*$ must be true.
Finally, condition $A^*YQ = 0$ yields $A^*YP = A^*Y$ and $Y^*A = Y^*AP$, thus $0 = D^*\{A^*YD^- - D^{-*}Y^*A\}D = D^*A^*YP - PY^*AB = D^*A^*Y - Y^*AB$. □

In turn, relation (11.76) implies $A^*YQ = 0$.

Theorem 11.27. *Let Assumption 11.24 be valid. Let Y be a solution of the terminal value problem (11.73), (11.74), and let the condition $A^*YQ = 0$ be fulfilled. Let $x_* \in \mathcal{C}_D^1$ be a solution of the IVP*

$$A(Dx)' = -Bx + C\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)x, \quad A(0)D(0)x(0) = z_0, \tag{11.77}$$

and

$$u_* := -\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)x_*. \tag{11.78}$$

Then, for each admissible pair $(u,x) \in \mathcal{C} \times \mathcal{C}_D^1$ it holds that

$$J(u,x) \geq J(u_*,x_*) = \frac{1}{2}\langle z_0, A(0)^{*-}D(0)^{-*}Y(0)^*z_0\rangle;$$

i.e., (u_*,x_*) is an optimal pair and (11.78) describes the optimal feedback.

Proof. It holds that $A^*Y = A^*YP = A^*YD^-D$, and that $D^{-*}Y^*AD = A^*Y$. Given an admissible pair (u,x), we derive

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle Dx, A^*Yx\rangle &= \langle (Dx)', A^*Yx\rangle + \langle Dx, (A^*YD^-Dx)'\rangle \\
&= \langle (Dx)', A^*Yx\rangle + \langle Dx, (A^*YD^-)'Dx\rangle + \langle Dx, A^*YD^-(Dx)'\rangle \\
&= \langle (Dx)', A^*Yx\rangle + \langle Dx, (A^*YD^-)'Dx\rangle + \langle A^*Yx, (Dx)'\rangle \\
&= 2\langle (Dx)', A^*Yx\rangle + \langle x, D^*(A^*YD^-)'Dx\rangle \\
&= 2\langle A(Dx)', Yx\rangle + \langle x, D^*(A^*YD^-)'Dx\rangle.
\end{aligned}$$

Taking into account (11.57) and (11.73) we obtain the expression

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\langle Dx, A^*Yx\rangle = &-\{\langle \mathcal{W}x, x\rangle + 2\langle \mathcal{S}u, x\rangle + \langle \mathcal{R}u, u\rangle\} \\
&+ \langle \mathcal{R}(u + \mathcal{R}^{-1}(\mathcal{S}^*x - C^*Yx)), u + \mathcal{R}^{-1}(\mathcal{S}^*x - C^*Yx)\rangle.
\end{aligned}$$

By this we find

$$J(u,x) = \frac{1}{2}\langle x(t_f), Vx(t_f)\rangle - \frac{1}{2}\int_0^{t_f} \frac{\mathrm{d}}{\mathrm{d}t}\langle D(t)x(t), A(t)^*Y(t)x(t)\rangle \mathrm{d}t + \mathfrak{B}(u,x),$$

$$\mathfrak{B}(u,x) = \frac{1}{2}\int_0^{t_f} \langle \mathcal{R}(t)(u(t) + \mathcal{R}(t)^{-1}(\mathcal{S}(t)^* - C(t)^*Y(t))x(t)), u(t) + \mathcal{R}(t)^{-1}(\mathcal{S}(t)^* - C(t)^*Y(t))x(t)\rangle \mathrm{d}t.$$

From the positive definiteness of $\mathcal{R}(t)$ it follows that $\mathfrak{B}(u,x) \geq 0$. Notice that $\mathfrak{B}(u_*,x_*) = 0$. Compute further

$$J(u,x) = \frac{1}{2}\langle x(t_f), Vx(t_f)\rangle - \frac{1}{2}\langle D(t_f)x(t_f), A(t_f)^*Y(t_f)x(t_f)\rangle + \frac{1}{2}\langle D(0)x(0), A(0)^*Y(0)x(0)\rangle + \mathfrak{B}(u,x).$$

Using the conditions (11.58) and (11.74), as well as the relations $V = VP(t_f)$ and $A^*Y = A^*YD^-D$, and (11.76), we arrive at

$$J(u,x) = \frac{1}{2}\langle z_0, A(0)^{*-}D(0)^{-*}Y(0)^*z_0\rangle + \mathfrak{B}(u,x).$$

Since the first term on the right-hand side is independent of the admissible pair (u,x), we conclude that

$$J(u,x) \geq \frac{1}{2}\langle z_0, A(0)^{*-}D(0)^{-*}Y(0)^*z_0\rangle = J(u_*,x_*).$$

□

The crucial question is now whether the terminal value problem for the Riccati DAE is solvable and the solution has the property $A_*YQ = 0$. We turn to the terminal value problem

$$D^*(A^*YD^-)'D = Y^*B + B^*Y + (\mathcal{S} - Y^*C)\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y) - \mathcal{W}, \tag{11.79}$$

$$P_*YQ = 0, \tag{11.80}$$

$$A(t_f)^*Y(t_f)D(t_f)^- = \widetilde{V} := D(t_f)^{-*}VD(t_f)^-. \tag{11.81}$$

Each solution Y can be decomposed by means of the projector functions as

$$\begin{aligned} Y &= P_*YP + Q_*YP + Q_*YQ \\ &= A^{*-}A^*YD^-D + Q_*YP + Q_*YQ. \end{aligned}$$

Also the DAE itself decouples by means of the projector functions. More precisely, multiplying (11.79) by Q from the left and right, then by Q from the left and P from the right, and also by D^{-*} from the left and D^- from the right, we obtain the system

$$0 = (YQ)^*BQ + QB^*YQ + (QS - (YQ)^*C)\mathcal{R}^{-1}(S^*Q - C^*YQ) - Q\mathcal{W}Q, \quad (11.82)$$

$$0 = (YQ)^*BP + QB^*YP + (QS - (YQ)^*C)\mathcal{R}^{-1}(S^*P - C^*YP) - Q\mathcal{W}P, \quad (11.83)$$

$$R^*(A^*YD^-)'R = (YD^-)^*BD^- + D^{-*}B^*YD^- + (D^{-*}S - (YD^-)^*C)\mathcal{R}^{-1}(S^*D^- - C^*YD^-) - D^{-*}\mathcal{W}D^-. \quad (11.84)$$

Since multiplication of (11.79) by P from the left and Q from the right yields again (11.83), we know (11.79) to be equivalent to (11.82)–(11.84). Obviously, the component $Z = Q_*YQ = YQ$ satisfies (cf. (11.82)) the algebraic Riccati equation

$$0 = Z^*Q_*BQ + QB^*Q_*Z + (QS - Z^*Q_*C)\mathcal{R}^{-1}(S^*Q - C^*Q_*Z) - Q\mathcal{W}Q \quad (11.85)$$

and the trivial conditions $P_*Z = 0,\ ZP = 0$. Denote

$$M := -QB^* + (QS - Z^*Q_*C)\mathcal{R}^{-1}C^*,\ \ M = QM. \quad (11.86)$$

Advancing the structural decoupling of the Riccati system one finds (see [135]) the components

$$U := A^*YD^- \in \mathcal{C}^1,\ \ \mathcal{V} := Q_*YP,\ \ Z := Q_*YQ = YQ \in \mathcal{C} \quad (11.87)$$

satisfy a standard Riccati differential equation, a linear equation, and an algebraic Riccati equation, respectively. The structural decoupling is the background of the following solvability result.

Theorem 11.28. *Let the Assumption 11.24 be valid. Let the algebraic Riccati system*

$$0 = Z^*Q_*BQ + QB^*Q_*Z + (QS - Z^*Q_*C)\mathcal{R}^{-1}(S^*Q - C^*Q_*Z) - Q\mathcal{W}Q$$
$$P_*Z = 0,$$
$$ZP = 0$$

have a continuous solution Z that satisfies the conditions

$$\operatorname{im} Z = \operatorname{im} Q_*, \qquad \ker Z = \ker Q, \quad (11.88)$$

$$\operatorname{im} MQ_* = \operatorname{im} Q, \qquad \ker MQ_* = \ker Q_*. \quad (11.89)$$

Then, the terminal value problem for the Riccati DAE (11.79)–(11.81) has a continuous solution Y whose component A^*YB^- is continuously differentiable and symmetric. Additionally, it holds that $A^*YQ = 0$.

Proof. See [135, pages 1284–1289]. □

To confirm the existence of an optimal control u_* with the minimal cost $J(u_*, x_*)$ from Theorem 11.59, in addition to the existence of a Riccati DAE solution Y, one necessarily needs to confirm the existence of a solution of the resulting closed loop DAE, that is (cf. (11.77)),

$$A(Dx)' = -Bx + C\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)x, \tag{11.90}$$

which satisfies the initial condition

$$A(0)D(0)x(0) = z_0, \tag{11.91}$$

where $z_0 \in \operatorname{im}(A(0)D(0))$ is fixed but chosen arbitrarily.

Clearly, if A and D are nonsingular, then the IVP (11.90), (11.91) always has a uniquely determined solution for each arbitrary z_0. In the case of singular A and D the situation is different, and so for time-invariant descriptor systems (see, e.g., [16]) one takes care to obtain a closed loop system that has no so-called *impulsive behavior* for any z_0. Within the scope of DAE theory, this means that one should have closed loop systems (11.90) that are regular with tractability index 1. Next we provide conditions ensuring the index-1 property for the DAE (11.90).

Theorem 11.29. *Let the conditions of Theorem 11.28 be given and let Y be a solution of the terminal value problem for the Riccati DAE (11.79)–(11.81).*

(1) *If $m = k$, then the DAE (11.90) is regular with tractability index 1, and there is exactly one solution $x_* \in \mathcal{C}_D^1$ of the IVP (11.90), (11.91).*
(2) *If $m > k$, then there are solutions $x_* \in \mathcal{C}_D^1$ of the IVP (11.90), (11.91).*

Proof. (1) The IVP solvability is a consequence of the index-1 property. We prove the DAE (11.90) to be regular with index 1. By Theorem 11.7, a DAE is regular with index 1, exactly if its adjoint is so. The adjoint equation to (11.90) reads

$$-D^*(A^*y)' = -B^*y + (-Y^*C + \mathcal{S})\mathcal{R}^{-1}C^*y. \tag{11.92}$$

The DAE (11.92) is regular with index 1 if the subspaces $\ker D^*A^* = \operatorname{im} Q_*$ and $\ker Q\{-B^* + (-Y^*C + \mathcal{S})\mathcal{R}^{-1}C^*\} =: S_*$ intersect trivially. Because of $QY^* = (YQ)^* = (ZQ)^* = QZ^*$ we have $S_* = \ker\{-QB^* + (-QZ^*C + Q\mathcal{S})\mathcal{R}^{-1}C^*\} = \ker M$. This means that the DAE (2.3) is regular with index 1 if $\ker M$ and $\operatorname{im} Q_*$ intersect trivially, but this is in turn a consequence of condition (11.89).
(2) Compute $G_1 := AD - \{-B + C\mathcal{R}^{-1}(\mathcal{S}^* - C^*Y)\}Q$ and ask whether this matrix function has full row rank k. This is in fact the case if $Q_*\{-BQ + C\mathcal{R}^{-1}(\mathcal{S}^*Q - C^*ZQ)\} = Q_*M^* = (MQ_*)^*$ has the same range as Q_*, i.e., if $\operatorname{im}(MQ_*)^* = \operatorname{im} Q_*$. However, this is ensured by (11.89). Then, solvability is ensured by Proposition 10.9. □

The following example demonstrates the appropriate solvability behavior of the introduced Riccati DAE.

Example 11.30 (Case study from [128]). We deal with the very special case of $k = m = 2$, $n = 1$, $l = 1$, $t_f = 1$,

$$J(u,x) = \frac{1}{2}\int_0^1 (\alpha x_1(t)^2 + \beta x_2(t)^2 + u(t)^2)\mathrm{d}t, \tag{11.93}$$

where $\alpha \geq 0$, $\beta \geq 0$, $\mathcal{W} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$, $R = 1$, $V = 0$, $S = 0$, and the DAE describing the side condition is

$$\begin{aligned} x_1'(t) &= c_{12}(t)x_2(t), \\ 0 &= c_{21}(t)x_1(t) + c_{22}(t)x_2(t) + u(t), \end{aligned} \tag{11.94}$$

i.e.,

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; D = [1 \;\; 0], \; D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \; R = 1, \; C = \begin{bmatrix} 0 \\ -1 \end{bmatrix}, \; B = -\begin{bmatrix} 0 & c_{12} \\ c_{21} & c_{22} \end{bmatrix}.$$

The initial condition for (11.94) reads

$$x_1(0) = z_0. \tag{11.95}$$

We have taken this problem from [128] and discuss the same three cases as considered there. In the first two cases, optimal controls exist, and we obtain them via our Riccati DAE system, whereas the Riccati DAE system used in [128] has no solutions. In the third case, if $z_0 \neq 0$, there is no optimal control, which is reflected by the failure of conditions (11.88) and (11.89). Notice that just in this case the Riccati DAE in [128] may have solutions.

To be more precise we consider the Riccati DAE system (11.79)–(11.81) for the 2×2 matrix function $Y = \begin{bmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{bmatrix}$. We describe (11.79) by means of the following three equations (cf. (11.82), (11.83), (11.84)), taking into account that we have here $Q = Q_* = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$, $P = P_* = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, and dropping the equations "$0 = 0$",

$$0 = -\beta - (Y_{12}c_{12} + Y_{22}c_{22}) - (c_{12}Y_{12} + c_{22}Y_{22}) + Y_{22}^2, \tag{11.96}$$

$$0 = -c_{21}Y_{22} - (c_{12}Y_{11} + c_{22}Y_{21}) + Y_{22}Y_{21}, \tag{11.97}$$

$$Y_{11}' = -\alpha - c_{21}Y_{21} - c_{21}Y_{21} + Y_{21}^2. \tag{11.98}$$

The terminal value condition (11.81) is

$$Y_{11}(1) = 0, \tag{11.99}$$

and condition (11.80) here means that

$$Y_{12} = 0. \tag{11.100}$$

Applying (11.100), (11.96) simplifies to

$$0 = -\beta + (Y_{22} - c_{22})^2 - c_{22}^2. \tag{11.101}$$

This algebraic equation has the solutions

$$Y_{22} = c_{22} \pm \sqrt{\beta + c_{22}^2}, \tag{11.102}$$

and the resulting matrix functions $Z = Q_* Y Q$ and MQ_* (cf. (11.86)) are

$$Z = \begin{bmatrix} 0 & 0 \\ 0 & Y_{22} \end{bmatrix}, \quad MQ_* = \begin{bmatrix} 0 & 0 \\ 0 & c_{22} - Y_{22} \end{bmatrix}.$$

In this case the conditions (11.88) and (11.89) are equivalent to the conditions that

$$Y_{22}(t) \quad \text{has no zeros} \tag{11.103}$$

and

$$Y_{22}(t) - c_{22}(t) = \pm\sqrt{\beta + c_{22}(t)^2} \quad \text{has no zeros, respectively.} \tag{11.104}$$

Case (1) c_{12} and c_{21} vanish identically, and $\beta > 0$.

Here, both Y_{22} and $Y_{22} - c_{22}$ do not have zeros; i.e., the conditions (11.88) and (11.89) are fulfilled. Equation (11.97) is simply $0 = (Y_{22} - c_{22})Y_{21}$, which leads to $Y_{21} = 0$. Equation (11.98) yields $Y_{11}' = -\alpha$. Hence, in this case

$$Y(t) = \begin{bmatrix} -\alpha(t-1) & 0 \\ 0 & c_{22}(t) \pm \sqrt{\beta + c_{22}(t)^2} \end{bmatrix}$$

solves the system. The feedback optimal control is given by

$$u = -(c_{22} \pm \sqrt{\beta + c_{22}^2})x_2.$$

The optimal trajectory, i.e., the solution of the IVP (11.90), (11.91) (cf. (11.77)) is $x_*(t) \equiv \binom{z_0}{0}$, the optimal control is $u_* = 0$, and the optimal cost is $J(u_*, x_*) = \frac{1}{2}\alpha z_0^2$.

Case (2) c_{22} vanishes identically, c_{12} and c_{21} have no zeros, and $\beta > 0$. Again, both $Y_{22} = \pm\sqrt{\beta}$ and $Y_{22} - c_{22} = Y_{22}$ have no zeros, and the conditions (11.88) and (11.89) are fulfilled. This time, (11.97) leads to

$$Y_{21} = c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}. \tag{11.105}$$

From (11.98) and (11.105) we derive the ODE

$$Y_{11}' = -\alpha - 2c_{21}\left(c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}\right) + \left(c_{21} \pm \frac{1}{\sqrt{\beta}} c_{12} Y_{11}\right)^2.$$

For example, for $c_{12} = c_{21} = 1$, $z_0 = 1$, the result is that

$$Y_{11}' = -(\alpha+1) + \frac{1}{\sqrt{\beta}} Y_{11}^2,$$
$$Y_{11}(t) = \beta\gamma \frac{1-e^{2\gamma(t-1)}}{1+e^{2\gamma(t-1)}} \quad \text{with } \gamma = \sqrt{\frac{1+\alpha}{\beta}}.$$

Then, $u = (\mp\frac{1}{\sqrt{\beta}} Y_{11} - 1)x_1 \mp \sqrt{\beta} x_2$ is an optimal feedback control. The DAE (11.90) is of the form

$$x_1' = x_2, \;\; 0 = \frac{1}{\sqrt{\beta}} Y_{11} x_1 + \sqrt{\beta} x_2,$$

and the optimal pair (u_*, x_*) consists of

$$u_*(t) = -x_{*1}(t), \;\; x_{*1}(t) = \frac{e^{\gamma t} + e^{\gamma(2-t)}}{1+e^{2\gamma}}, \;\; x_{*2}(t) = -\frac{1}{\beta} Y_{11}(t) x_{*1}(t).$$

The minimal costs are

$$J(u_*, x_*) = \frac{\beta\gamma}{2} \cdot \frac{1-e^{-2\gamma}}{1+e^{-2\gamma}}.$$

Case (3) $\beta = 0$, c_{22} vanishes identically, and c_{12}, c_{21} have no zeros. Here, (11.101) implies $Y_{22} = 0$, and hence, $Z = 0$, $MQ_* = 0$, and the conditions (11.88) and (11.89) fail to be valid. Equation (11.97) simplifies to $c_{12} Y_{11} = 0$, and hence $Y_{11} = 0$ must be true. By (11.98) we find $Y_{21} = c_{21} \pm \sqrt{\alpha + c_{21}^2}$. Therefore, the matrix function

$$Y = \begin{bmatrix} 0 & 0 \\ c_{21} \pm \sqrt{\alpha + c_{21}^2} & 0 \end{bmatrix}$$

solves system (11.79)–(11.81); however, the conditions (11.88) and (11.89) do not hold. The resulting closed loop DAE (11.90) is now $x_1' = c_{12} x_2$, $0 = \sqrt{\alpha + c_{21}^2}\, x_1$, and it has only the trivial solution. Consequently, for $z_0 \neq 0$, there is no solution of the IVP (11.90), (11.91). If $z_0 = 0$, then the trivial pair $u_* = 0$, $x_* = 0$ is optimal in accordance with Theorem 11.27. If $z_0 \neq 0$, then the linear-quadratic optimal control problem has no solution at all.

The Hamiltonian system (11.64) corresponding to our special problem (11.93)–(11.95) is the following:

$$\begin{aligned} x_1' &= c_{12} x_2, \\ 0 &= c_{21} x_1 + c_{22} x_2 - \lambda_2, \\ -\lambda_1' &= \alpha x_1 + c_{21} \lambda_2, \\ 0 &= \beta x_2 + c_{12} \lambda_1 + c_{22} \lambda_2. \end{aligned} \tag{11.106}$$

For this system, the initial and terminal conditions

$$x_1(0) = z_0, \;\; \lambda_1(1) = 0 \tag{11.107}$$

have to be taken into account. This linear DAE with respect to x, λ is regular with index 1 exactly if $\beta + c_{22}^2 \neq 0$. This index-1 condition is valid in Cases (1) and (2). In Case (3), the BVP (11.106), (11.107) has no solution for $z_0 \neq 0$. For $z_0 = 0$ it has the trivial solution. It may be checked that this DAE has index 2.
Let us add that, for the solvability of the corresponding Riccati DAE treated in [128], it is necessary to require $\beta = 0$, i.e., unfortunately, this Riccati DAE is no longer solvable in the unproblematic cases (1) and (2). In Case (3), the terminal value problem may or may not have solutions. From this point of view, those Riccati DAEs in [128] seem not to be appropriate tools for constructing optimal feedback solutions. □

11.5 Notes and references

(1) This chapter collects and slightly modifies results from [12, 8, 6, 135, 36]. An informative overview concerning extremal conditions is given in [6]. We refer to [8, 6, 172, 36] for further index relations and for discussions concerning the consistency with well-known facts in the context of linear-quadratic optimal control problems. Furthermore, different proposals to generalize the Riccati feedback are reported in [135].

(2) We have shown that optimal feedback controls of linear-quadratic optimal control problems with constraints described by general linear DAEs with variable coefficients can be obtained by suitably formulating a Riccati DAE system, similarly to the classical example in which the constraints are described by explicit ODEs. Compared to earlier results and some less suitable Riccati DAEs, one can now do without previous restrictive assumptions.
Furthermore, we would like to stress that it is not necessary and probably not even reasonable to transform the DAE describing the constraints (the descriptor system) or the DAE describing the Hamiltonian system with great expense into a special canonical form.

(3) One could surmise that the DAE

$$-E(t)^* y'(t) + F(t)^* y(t) = 0$$

is adjoint to the standard form DAE

$$E(t)x'(t) + F(t)x(t) = 0,$$

however, this is only true for constant matrix functions E. In general, the adjoint equation of this standard form DAE looks like

$$-\left(E(t)^* y(t)\right)' + F(t)^* y(t) = 0. \tag{11.108}$$

If E is continuously differentiable, this can be written as

$$-E(t)^* y'(t) + (F(t)^* - E'(t)^*) y(t) = 0.$$

Whereas the standard form DAE and its adjoint look quite quite different, the DAE (11.1) and its adjoint are formally of the same type, and show nice symmetries. At this point it should be mentioned that the motivation for introducing DAEs with properly stated leading terms in [11] was inspired by these symmetries.

(4) If the full rank condition (11.14) is not given in an optimization problem, then it might be a good idea to reformulate or reduce the problem so that the reduced DAE meets the condition. A different way consists in exploiting given structural peculiarities with the aim of obtaining surjectivity of the operator $\mathcal{F}_x(x_*)$ in specially adjusted function spaces, for instance, in the case of controlled Hessenberg size-2 DAEs, cf. [92, 91, 35]. Note that different function spaces may lead to different representations of the Lagrange multiplier, and hence yield another terminal value problem than (11.15), (11.16).

(5) We stress that our optimality criteria are clearly represented *algebraic conditions in terms of the original optimization problem*. In contrast, in [131] an analogous optimization problem with DAE constraint

$$\mathfrak{f}(x'(t), x(t), u(t), t) = 0, \quad t \in \mathcal{I},$$

is treated by transforming this equation first into the so-called reduced form

$$x_1'(t) - \mathcal{L}(x_1(t), x_2(t), u(t), t), \quad x_2(t) = \mathcal{R}(x_1(t), u(t), t), \tag{11.109}$$

and not till then formulating an extremal condition and the optimality DAE in terms of (11.109). This pre-handling is based on demanding assumptions (e.g., [130, Hypothesis 1]) and it needs considerable effort. In turn, the reduced system (11.109) represents a special case of a semi-explicit controlled regular index-1 DAE, such that condition (11.44) is given. The optimality DAE for the optimization problem with constraint DAE (11.109) is then the corresponding special case of the DAE (11.49)–(11.51).

(6) For linear-quadratic optimal control problems for descriptor systems

$$Ex' = Cx + Du, \tag{11.110}$$

with E being a singular constant square matrix, in the famous paper [16] it was first noted that the Riccati modification

$$E^* Y' E = -E^* Y C - C^* Y E + (\mathcal{S} + E^* Y D)\mathcal{R}^{-1}(\mathcal{S}^* + D^* Y E) - \mathcal{W}, \tag{11.111}$$

which is considered to be *obvious*, leads to unacceptable solvability conditions. Consequently, more specific Riccati approaches that skillfully make use of the inherent structures find favor with [16]. Starting from a singular value decomposition $UEV = \operatorname{diag}(\Sigma, 0)$ and certain rank conditions, lower dimensional Riccati equations of the form $\Sigma Y' \Sigma = \cdots$ are introduced. From the point of view of DAE theory the

rank conditions used in [16] imply that the related Hamilton–Lagrange system is a regular DAE with tractability index 1 (cf. [8]).
For example, in [134] (in a more general Hilbert space setting, with $\mathcal{S} = 0$) a different ansatz was followed with Riccati equations of the form

$$E^*Y' = -Y^*C - C^*Y + (\mathcal{S} + Y^*D)\mathcal{R}^{-1}(\mathcal{S}^* + D^*Y) - \mathcal{W}. \tag{11.112}$$

The solutions of the terminal value problem for (11.112) with the condition

$$E^*Y(T) = V \tag{11.113}$$

have the symmetry property $E^*Y = Y^*E$. Like (11.111), also (11.112) is primarily a matrix DAE, however, (11.112) is approved to be better solvable.
In [128] the Riccati DAE

$$(E^*YE)' = -E^*YC - C^*YE + (\mathcal{S} + E^*YD)\mathcal{R}^{-1}(\mathcal{S}^* + D^*YE) - \mathcal{W}, \tag{11.114}$$

which generalizes (11.111) for time-dependent coefficients E is introduced. However, this equation is as unsuitable as its time-invariant version (11.111), and the authors have to admit that, *unfortunately, this approach can be used only in very special cases since, for $E(t)$ singular, the solutions of* (11.114) *and the Euler–Lagrange equation are not related via $u = -\mathcal{R}^{-1}(\mathcal{S} + D^*YE)x$, as in the case of nonsingular $E(t)$.*

(7) In the theory of explicit ODEs, the fundamental solution matrix $X(t)$ of the ODE

$$x'(t) + B(t)x(t) = 0$$

and the fundamental solution matrix $Y(t)$ of its adjoint ODE

$$-y'(t) + B^*(t)y(t) = 0,$$

both normalized at t_0, $X(t_0) = I, Y(t_0) = I$, are related to each other by the identity

$$Y(t)^* = X(t)^{-1}. \tag{11.115}$$

Since fundamental solution matrices of regular DAEs are singular, an appropriate generalization of (11.115) should involve special generalized inverses. For regular index-1 and index-2 DAEs with properly stated leading term, such a generalization is derived in [11]. For standard form DAEs and their adjoints, various results are obtained, e.g., in [7].

(8) For an overview of the state of the art on the general field of optimization with DAE constraints we refer to [20].

(9) In [65], [66] optimal control problems with ODE and inequality restrictions are considered. The necessary solvability conditions lead to DAEs whose properties are investigated in detail using the tractability concept.

Chapter 12
Abstract differential-algebraic equations

This chapter is devoted to abstract differential-algebraic equations (ADAEs). We consider differential-algebraic equations operating in Hilbert spaces. Such a framework aims to provide a systematic approach for coupled systems of partial differential equations (PDEs) and differential-algebraic equations (DAEs).

We introduce abstract differential-algebraic equations as

$$A(t)\frac{\mathrm{d}}{\mathrm{d}t}d(x(t),t)+b(x(t),t)=0 \qquad \text{for } t\in\mathcal{I} \tag{12.1}$$

with $\mathcal{I}\subseteq\mathbb{R}$ being a bounded time interval. This equation is to be understood as an operator equation with operators $A(t)$, $d(\cdot,t)$ and $b(\cdot,t)$ acting in real Hilbert spaces. More precisely, let X, Y, Y_W, Z be Hilbert spaces and

$$A\ :\ Y_W\to Z,\quad d(\cdot,t)\ :\ X\to Y,\quad b(\cdot,t)\ :\ X\to Z. \tag{12.2}$$

In [166, 140], the case $Y_W=Y$ has been analyzed, which is adequate for coupled systems containing classical PDE formulations. In order to cover weak PDE formulations, it is useful to consider the case $Y_W=Y^*$.

Systems with A and d being invertible have already been studied in [79]. However, the classical formulation as well as the generalized formulation of the coupled system lead to abstract differential-algebraic systems with operators A or d that are not invertible. Such systems are also called degenerate differential equations (see e.g. [133, 74]).

In Section 12.1, we introduce an index concept for abstract differential-algebraic systems that orients towards the time domain behavior of a differential-algebraic system. One can compare it with the time index introduced in [149] and the modal index introduced in [47] for linear partial differential algebraic equation systems (PDAEs) with constant coefficients. By a few case studies in Section 12.2, we demonstrate the potential of the general approach for ADAEs for solving coupled systems with different types of differential and/or integral equations.

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*,
Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5_12,

A general theory for the existence and uniqueness of solutions for abstract differential-algebraic equations does not exist so far. Since already all types of PDEs can be considered as an ADAE, we cannot expect to find an approach treating all ADAEs in one framework. We need a classification of ADAEs reflecting the classification of PDEs as well as the index classification for DAEs. We do not have an answer providing a general classification. But we provide a starting point for such a classification in Section 12.3. We consider a class of linear ADAEs which covers parabolic PDEs and index-1 DAEs as well as couplings thereof. We treat this ADAE class by a Galerkin approach yielding an existence and uniqueness result for ADAE solutions as well as an error estimation for perturbed systems. In contrast to Galerkin methods for parabolic differential equations, the choice of the basis functions is more restricted since they have to satisfy certain constraints as we know from the theory of DAEs.

12.1 Index considerations for ADAEs

From the finite-dimensional case we know that the sensitivity of solutions of DAEs with respect to perturbations depends on their index, cf. Example 1.5. Since we are interested in the transient behavior of solutions of the coupled system, we follow the concept in [140]. We consider systems of the form

$$A(t)\frac{\mathrm{d}}{\mathrm{d}t}d(x(t),t)+b(x(t),t)=0 \qquad \text{for } t\in\mathcal{I} \tag{12.3}$$

with operators $A(t)$, $d(\cdot,t)$ and $b(\cdot,t)$ acting in Hilbert spaces X, Y and Z as follows:

$$A(t):\ Y\to Z, \qquad b(\cdot,t):\ X\to Z, \qquad d(\cdot,t):\ X\to Y.$$

We assume the existence of the Fréchet derivatives of the operators $b(\cdot,t)$ and $d(\cdot,t)$. More precisely, we assume the existence of linear, continuous operators $B(x,t)$ and $D(x,t)$ satisfying

$$\begin{aligned} b(x+h,t)-b(x,t)-B(x,t)h &= o(\|h\|), \quad h\to 0,\\ d(x+h,t)-d(x,t)-D(x,t)h &= o(\|h\|), \quad h\to 0, \end{aligned}$$

for all h in some neighborhood of zero in X, all $x\in X$ and all $t\in\mathcal{I}$. Furthermore, we assume that

$$\operatorname{im}D(x,t) \text{ and } \ker D(x,t) \text{ do not depend on } x \text{ and } t. \tag{12.4}$$

Since $D(x,t)$ is bounded, we find $N_0 := \ker D(x,t)$ to be a closed subspace in X. Finally, A and D are assumed to be well-matched in the sense that

$$\ker A(t)\oplus\operatorname{im}D(x,t)=Y \tag{12.5}$$

forms a topological direct sum for all $x \in X$ and $t \in \mathcal{I}$.

We introduce $G_0(x,t) := A(t)D(x,t)$ for all $x \in X$ and $t \in \mathcal{I}$. Since we are interested in abstract differential-algebraic systems containing equations without time derivatives, we assume that $\dim(N_0) > 0$.

Since A and D are assumed to be well-matched, the relations

$$\operatorname{im} G_0(x,t) = \operatorname{im} A(t) \quad \text{and} \quad \ker G_0(x,t) = \ker D(x,t) \tag{12.6}$$

are fulfilled for all $x \in X$ and $t \in \mathcal{I}$. Indeed, there is a constant, bounded projection operator $R : Y \to Y$ satisfying

$$\operatorname{im} R = \operatorname{im} D(x,t) \quad \text{and} \quad \ker R = \ker A(t)$$

for all $x \in X$ and $t \in \mathcal{I}$ because $\ker A(t)$ and $\operatorname{im} D(x,t)$ form a topological direct sum. Consequently,

$$\operatorname{im} G_0(x,t) = \operatorname{im} A(t)R = \operatorname{im} A(t)$$

and

$$\ker G_0(x,t) = \ker RD(x,t) = \ker D(x,t).$$

Considering the formulation of the DAE (12.3), it is natural that solutions belong to the set

$$C_d^1(\mathcal{I},X) := \{x \in C(\mathcal{I},X) : d(x(\cdot),\cdot) \in C^1(\mathcal{I},Y)\}.$$

For a linearization

$$A(t)\frac{\mathrm{d}}{\mathrm{d}t}(D_*(t)x(t)) + B_*(t)x(t) = q(t)$$

of the DAE (12.3) at a given function $x_* \in C_d^1(\mathcal{I},X)$ with $D_*(t) := D(x_*(t),t)$ and $B_*(t) := B(x_*(t),t)$ for all $t \in \mathcal{I}$, the natural solution space is given by

$$C_{D_*}^1(\mathcal{I},X) := \{x \in C(\mathcal{I},X) : D_*(\cdot)x(\cdot) \in C^1(\mathcal{I},Y)\}.$$

It is an analogon of the solution space for DAEs in finite dimensions (cf. Theorem 3.55). The next proposition shows that both solution sets coincide.

Proposition 12.1. *Let $D(x,t)$ depend continuously differentiably on x, t and satisfy (12.4). Additionally, assume the partial derivative $d_t(x,t)$ to exist and to be continuous. Moreover it is assumed that $d(x,t) \subseteq \operatorname{im} R$. Then, we have $C_d^1(\mathcal{I},X) = C_{D_*}^1(\mathcal{I},X)$.*

Proof. Since N_0 is a closed subspace of X, we find a projection operator $P_0 : X \to X$ with

$$\operatorname{im} P_0 = N_0^\perp \quad \text{and} \quad \ker P_0 = N_0.$$

Using the mean value theorem in Banach spaces, we get

$$d(x,t) - d(P_0x,t) = \int_0^1 D(sx + (1-s)P_0x,t)(I-P_0)x\,\mathrm{d}s = 0 \tag{12.7}$$

for all $x \in X$ and $t \in \mathcal{I}$. Additionally, $D(x,t)$ acts bijectively from $N_0^\perp$ to $\operatorname{im} R$ for all $x \in X$ and $t \in \mathcal{I}$. Applying the implicit function theorem to

$$F : Y \times N_0^\perp \times \mathcal{I} \to \operatorname{im} R \quad \text{with} \quad F(v,w,t) := d(w,t) - Rv$$

at $v = d(x(t),t)$ and $w = P_0 x(t)$ for $x \in C_d^1(\mathcal{I},X)$, we obtain a continuously differentiable function $g(\cdot,t) : Y \to N_0^\perp$ satisfying $F(v,g(v,t),t) = 0$ for v in a neighborhood of $d(x(t),t)$. In particular, we have

$$P_0 x(t) = g(d(x(t),t),t).$$

Due to the smoothness assumptions, g is also continuously differentiable with respect to t. Consequently,

$$d(x(\cdot),\cdot) \in C^1(\mathcal{I},Y) \quad \Leftrightarrow \quad P_0 x(\cdot) \in C^1(\mathcal{I},N_0^\perp) \tag{12.8}$$

if we regard equation (12.7). Analogously, we obtain

$$D(x,t) = D(P_0 x,t)$$

for all $x \in X$ and $t \in \mathcal{I}$. This implies $D(x_*(\cdot),\cdot) \in C^1(\mathcal{I},L(X,Y))$ for $x_* \in C_d^1(\mathcal{I},X)$. Thus,

$$D(x_*(\cdot),\cdot)x(\cdot) \in C^1(\mathcal{I},Y) \quad \Leftrightarrow \quad P_0 x(\cdot) \in C^1(\mathcal{I},N_0^\perp)$$

since $\ker D(x_*(\cdot),\cdot) = \ker P_0$ and $D(x_*(\cdot),\cdot)$ acts bijectively from $\operatorname{im} P_0$ to $\operatorname{im} R$. Regarding the equivalence relation (12.8), the proposition is proved. □

As the tractability index for finite-dimensional differential-algebraic systems, the following index concept for ADAEs is based on linearizations.

Definition 12.2. The abstract differential-algebraic system (12.3) has *index 0* if $G_0 := G$ is injective and $\operatorname{cl}(\operatorname{im} G_0(x,t)) = Z$ for all $x \in X$ and $t \in \mathcal{I}$.

The abstract differential-algebraic system (12.3) has *index 1* if there is a projection operator $Q_0 : X \to X$ onto the constant space $\ker G_0(x,t)$ such that the operator

$$G_1(x,t) := G_0(x,t) + B(x,t)Q_0$$

is injective and $\operatorname{cl}(\operatorname{im} G_1(x,t)) = Z$ for all $x \in X$ and $t \in \mathcal{I}$.

The definition is independent of the choice of the projection operator Q_0. If $\tilde{Q}_0$ is another projection operator onto $\ker G_0(t)$, then

$$\tilde{G}_1(x,t) = G_0(x,t) + B(x,t)\tilde{Q}_0 = G_1(x,t) \cdot (I + Q_0\tilde{Q}_0 P_0)$$

holds for $P_0 := I - Q_0$ since

$$Q_0 = \tilde{Q}_0 Q_0 \quad \text{and} \quad \tilde{Q}_0 = Q_0 \tilde{Q}_0.$$

The operator $I + Q_0\tilde{Q}_0 P_0$ is continuous and injective. Its inverse operator is given by $I - Q_0\tilde{Q}_0 P_0$ and, thus, is also continuous. Consequently,

$$\ker \tilde{G}_1(x,t) = \ker G_1(x,t) \quad \text{and} \quad \operatorname{im} \tilde{G}_1(x,t) = \operatorname{im} G_1(x,t).$$

Remark 12.3. Since the definition is independent of the choice of the projector and N_0 is bounded, we can easily characterize index-1 DAEs by using the orthoprojection operator onto N_0. The abstract DAE (12.3) has index 1 if and only if Q_0^* is the orthogonal projection onto N_0 and

$$G_1(x,t) := G_0(x,t) + B(x,t)Q_0^*$$

is injective and $\operatorname{cl}(\operatorname{im} G_1(x,t)) = Z$ for all $x \in X$ and $t \in \mathcal{I}$.

This index definition characterizes the behavior of abstract differential-algebraic systems with respect to perturbations of the right-hand side. It should not be confused with the Fredholm index of operators.

A general index definition for nonlinear abstract differential-algebraic systems is still a challenge of research. Formally, one could extend the index definition from Chapter 3 to abstract systems. But an index definition should give us some information about the solution and perturbation behavior of abstract DAEs. So far, such a general characterization is known only for abstract DAEs with constant coefficients. In [193, 191] the following theorem has been shown.

Theorem 12.4. *Let X, Z be Hilbert spaces and let (A,B) be a regular operator pair, i.e., the generalized resolvent is nontrivial, $\rho(A,B) \neq \{0\}$. Let $A : X \to Z$ be bounded and $B : dom_B \to Z$ be densely defined. Moreover, let the projector sequence $Q_i \in L_b(X) \cap L_b(dom_B)$, $P_i = I - Q_i$ with*

$$\begin{aligned} G_0 &= A, \quad B_0 = B, \\ \operatorname{im} Q_i &= \ker G_i, \qquad \sum_{j=0}^{i-1} \ker G_j \subset \ker Q_i, \\ G_{i+1} &= G_i + B_i Q_i, \quad B_{i+1} = B_i P_i \end{aligned}$$

exist and be stagnant, i.e., there exists a $\mu \in \mathbb{N}$ such that $\ker G_\mu = \{0\}$. Further, let

$$\operatorname{im} A + B\left(dom_B \cap \sum_{k=0}^{\mu-1} \ker G_k \right)$$

be closed. Then, there exist Hilbert spaces X_1, X_2, X_3 and bounded mappings $W \in L_b(Z, X_1 \times X_2 \times X_3)$, $T \in L_b(X_1 \times X_2, X)$, where T is bijective and W is injective and has dense range such that

$$WAT = \begin{bmatrix} N & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix} : \quad X_1 \times X_2 \to X_1 \times X_2 \times X_3, \tag{12.9}$$

$$WBT = \begin{bmatrix} I & K \\ 0 & L \\ 0 & M \end{bmatrix} : \quad X_1 \times dom_K \cap dom_L \cap dom_M \to X_1 \times X_2 \times X_3. \tag{12.10}$$

In particular, $N \in L_b(X_1)$ is a nilpotent operator with nilpotency index μ.

Notice that the space of bounded linear operators from X to Z is denoted by $L_b(X,Z)$ and $L_b(X) := L_b(X,X)$.

The nilpotency index μ takes the role of the Kronecker index for matrix pairs (A,B). Regarding the index definition for DAEs with constant coefficients (see Chapter 1), Theorem 12.4 justifies the definition of the index of abstract DAEs

$$Ax'(t) + Bx(t) = q(t)$$

with constant operators A, B satisfying the properties supposed in Theorem 12.4 by the first index μ with $\ker G_\mu = 0$.

In contrast to the finite dimensional case, additional operators K and M appear in (12.10). The operator M has its interpretation as a boundary control term. The coupling operator K is not always removable. [193, 191] present examples of abstract DAEs that do not possess a decoupling form with $K = 0$. Sufficient criteria for disappearing operators K are given in [191].

The decoupling form (12.10) for abstract DAEs provides structural information about the solution behavior and makes it possible to formulate perturbation results as well as consistent initializations, see [191, 192].

12.2 ADAE examples

This section demonstrates various application types of differential equation systems that are covered by the abstract differential-algebraic systems introduced in this chapter. First, we see that this approach not only includes (finite-dimensional) differential-algebraic systems as considered in the chapters before but also partial differential equations and integral equations as well as couplings thereof.

Aiming at unique solutions we have to equip such problems with appropriate initial and boundary conditions. As soon as we consider coupled systems, it is no longer clear for which components we have to provide initial and boundary conditions for unique solvability. We see that the index concept presented in Section 12.1 supports us in posing appropriate initial and boundary conditions.

12.2.1 Classical partial differential equations

In this section we consider the roots of some classical partial differential equations as they appear in applications. They, naturally, represent abstract DAEs. Applying the extended index analysis of DAEs to such abstract DAEs gives us an understanding of abstract DAEs. In particular, it provides guidance for forming appropriate function spaces for the solution as well as advice for finding consistent initial and boundary conditions.

12.2.1.1 Wave equation

The flow in an ideal gas is determined by three laws (cf. e.g., [22]). As usually, we denote the velocity by v, the density by ρ, and the pressure by p.

1. Continuity equation
$$\frac{\partial \rho}{\partial t} = -\rho_0 \operatorname{div} v. \tag{12.11}$$
Due to the conservation of mass, the variation of the mass in a (sub)volume V is equal to the flow over the surface, i.e., $\int_{\partial V} \rho v \cdot \mathfrak{n} \, \mathrm{d}O$. The Gaussian integral theorem implies the above equation, where ρ is approximated by the fixed density ρ_0.
2. Newton's theorem
$$\rho_0 \frac{\partial v}{\partial t} = -\operatorname{grad} p. \tag{12.12}$$
The pressure gradient induces a force field causing the acceleration of particles.
3. State equation
$$p = c^2 \rho. \tag{12.13}$$
In ideal gases the pressure is proportional to the density under constant temperature.

From the three laws above the wave equation

$$\frac{\partial^2}{\partial t^2} p = c^2 \frac{\partial^2 \rho}{\partial t^2} = -c^2 \frac{\partial}{\partial t} \rho_0 \operatorname{div} v = c^2 \operatorname{div} \rho_0 \frac{\partial v}{\partial t} = c^2 \operatorname{div} \operatorname{grad} p = c^2 \Delta p$$

is deduced. Considering equations (12.11)–(12.13) to form a system of equations we know that this is an abstract DAE of index 1. Namely, choosing

$$x(t) = \begin{bmatrix} \rho(t,\cdot) \\ v(t,\cdot) \\ p(t,\cdot) \end{bmatrix}, \qquad A = \begin{bmatrix} 1 & 0 \\ 0 & \rho_0 I \\ 0 & 0 \end{bmatrix}, \qquad D = \begin{bmatrix} 1 & 0 & 0 \\ 0 & I & 0 \end{bmatrix},$$

$$D^- = \begin{bmatrix} 1 & 0 \\ 0 & I \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & \rho_0 \operatorname{div} & 0 \\ 0 & 0 & \operatorname{grad} \\ -c^2 & 0 & 1 \end{bmatrix}, \qquad Q_0 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$

we find, independently of the special function spaces,

$$G_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho_0 I & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad G_1 = G_0 + BQ_0 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \rho_0 I & \operatorname{grad} \\ 0 & 0 & 1 \end{bmatrix}.$$

Obviously, G_1 is always nonsingular and its inverse is given by

$$G_1^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{\rho_0} I & -\frac{1}{\rho_0}\operatorname{grad} \\ 0 & 0 & 1 \end{bmatrix}.$$

Corresponding to the decoupling procedure for the finite-dimensional case (see [11, 167]) we compute

$$DG_1^{-1}BD = \begin{bmatrix} 0 & \rho_0 \operatorname{div} \\ \frac{c^2}{\rho_0}\operatorname{grad} & 0 \end{bmatrix}, \quad Dx = \begin{bmatrix} \rho \\ v \end{bmatrix}.$$

Hence, not surprisingly, the inherent regular differential equation reads

$$\begin{aligned} \rho' + \rho_0 \operatorname{div} v &= 0, \\ v' + \frac{c^2}{\rho_0} \operatorname{grad} \rho &= 0 \end{aligned} \tag{12.14}$$

while

$$\begin{bmatrix} 0 \\ 0 \\ p \end{bmatrix} = Q_0 x = Q_0 G_1^{-1} B D^- D x = \begin{bmatrix} 0 \\ 0 \\ c^2 \rho \end{bmatrix}$$

represents the constraint. The initial condition for the resulting DAE (12.3) consists of the initial condition for the inherent regular equation (12.14). Boundary conditions for (12.14) should be incorporated by specifying the function space X and the definition domain $\mathcal{D}_B \subset X$.

12.2.1.2 Heat equation

Let $T(x,t)$ be the distribution of temperature in a body. Then the heat flow reads

$$F = -\kappa \operatorname{grad} T, \tag{12.15}$$

where the diffusion constant κ represents a material constant. Due to the conservation of energy, the change of energy in a volume element is composed of the heat flow over the surface and the applied heat Q. Now it follows that

$$\begin{aligned}\frac{\partial E}{\partial t} &= -\operatorname{div} F + Q \\ &= \operatorname{div} \kappa \operatorname{grad} T + Q \\ &= \kappa \Delta u + Q,\end{aligned} \tag{12.16}$$

if κ is assumed to be constant. With the specific heat $a = \partial E/\partial T$ (a further material constant) we finally obtain the heat equation

$$\frac{\partial T}{\partial t} = \frac{\kappa}{a}\Delta T + \frac{1}{a}Q. \tag{12.17}$$

On the other hand, we may consider the original equations

$$\begin{aligned}a\frac{\partial T}{\partial t} &= -\operatorname{div} F + Q, \\ F &= -\kappa \operatorname{grad} T\end{aligned} \tag{12.18}$$

to form an abstract index-1 DAE (12.3) for $x(t) = \begin{bmatrix} T(t,\cdot) \\ F(t,\cdot) \end{bmatrix}$ with

$$A = \begin{bmatrix} a \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix},$$
$$B = \begin{bmatrix} 0 & \operatorname{div} \\ \kappa \operatorname{grad} & 1 \end{bmatrix}, \quad G_1 = \begin{bmatrix} a & \operatorname{div} \\ 0 & 1 \end{bmatrix}.$$

Also in this case, G_1 is always nonsingular. Following again the decoupling procedure for the description of the inherent regular differential equation and the constraints, we obtain

$$G_1^{-1} = \begin{bmatrix} \frac{1}{a} & -\frac{1}{a}\operatorname{div} \\ 0 & 1 \end{bmatrix}, \quad DG_1^{-1}BD^- = -\frac{\kappa}{a}\operatorname{div}\operatorname{grad}, \quad Dx = T,$$

which implies the inherent regular differential equation to be

$$T' - \frac{\kappa}{a}\operatorname{div}\operatorname{grad} T = \frac{1}{a}Q$$

and

$$\begin{bmatrix} 0 \\ F \end{bmatrix} = Q_0 x = -Q_0 G_1^{-1} B D^- Dx + Q_0 G_1^{-1} \begin{bmatrix} Q \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ -\kappa \operatorname{grad} T \end{bmatrix}$$

to represent the constraint. Again, the function spaces and incorporated boundary conditions as well as an initial condition for the inherent regular differential equation should enable us to obtain unique solvability.

12.2.2 A semi-explicit system with parabolic and elliptic parts

Let $\Omega \subset \mathbb{R}^d$ and $\mathcal{I} \subset \mathbb{R}$ be regular open domains. Consider the system

$$\begin{aligned} u_t + b_1 \Delta u + b_2 \Delta v + c_1 u + c_2 v &= r, \\ b_3 \Delta u + b_4 \Delta v + c_3 u + c_4 v &= s \end{aligned} \tag{12.19}$$

with the Laplacian Δ, unknown functions $u = u(z,t)$ and $v = v(z,t)$ and given functions $r = r(z,t)$ and $s = s(z,t)$ defined on $\Omega \times \mathcal{I}$. We assume

$$b_1 b_4 - b_2 b_3 \neq 0. \tag{12.20}$$

Regarding the system (12.19) in variation formulation and describing it as an abstract DAE (12.1) we choose $X = Z = L_2(\Omega) \times L_2(\Omega)$, $Y = L_2(\Omega)$ and

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

with the identity operator in $L_2(\Omega)$ represented as 1 and the zero operator in $L_2(\Omega)$ represented as 0. Further, we have

$$b(x) = \begin{bmatrix} b_1 \Delta u + c_1 u + b_2 \Delta v + c_2 v \\ b_3 \Delta u + c_3 u + b_4 \Delta v + c_4 v \end{bmatrix}$$

and

$$B = b_x(x) = \begin{bmatrix} b_1 \Delta + c_1 & b_2 \Delta + c_2 \\ b_3 \Delta + c_3 & b_4 \Delta + c_4 \end{bmatrix}.$$

Of course, even in variation formulation, b is not defined for all $x \in X = L_2(\Omega) \times L_2(\Omega)$. But it is defined on $\mathcal{D}_B^{(1)} := H^1(\Omega) \times H^1(\Omega)$. The question we want to discuss here is a proper choice of boundary conditions on Ω such that the system (12.19) has a unique solution.

Let $\sigma(\Delta)$ denote the spectrum of the Laplacian. It naturally depends on the region Ω. Injectivity of the operator

$$G_1 = \begin{bmatrix} 1 & b_2 \Delta + c_2 \\ 0 & b_4 \Delta + c_4 \end{bmatrix}$$

requires the operator $b_4 \Delta + c_4$ to be injective. Assuming

$$b_4 \gamma + c_4 \neq 0 \ \text{ for all } \ \gamma \in \sigma(\Delta) \tag{12.21}$$

we find G_1 to be injective on $\mathcal{D}_B^{(1)}$. This means that the system (12.19) has index $\mu = 1$ and no boundary condition for v has to be posed for v if the condition (12.21) is satisfied. If $c_4 = 0$ then condition (12.21) is violated. But we still obtain G_1 to be injective if $b_4 \neq 0$ and if we choose $\mathcal{D}_B^{(2)} := H^1(\Omega) \times H_0^1(\Omega)$ as the definition

domain for b. This means that we have to suppose boundary conditions on Ω for v if $c_4 = 0$ and $b_4 \neq 0$.

In both cases, the corresponding inherent regular differential equation is related to the component $Dx = u$. In order to obtain unique solvability, boundary and initial conditions for u have to be added. Hence, we select $\mathcal{D}_B = H_0^1(\Omega) \times H^1(\Omega)$ if condition (12.21) is satisfied and $\mathcal{D}_B = H_0^1(\Omega) \times H_0^1(\Omega)$ if $b_4 \neq 0$ and $c_4 = 0$. The initial condition is stated to be

$$Dx(0) = u(0) \in D\mathcal{D}_B = H_0^1(\Omega).$$

Next, consider the case that $b_4 = 0$, $c_4 = 0$. This implies that $b_2 \neq 0$, $b_3 \neq 0$ because of assumption (12.20). Obviously,

$$G_1 = \begin{bmatrix} 1 & b_2\Delta + c_2 \\ 0 & 0 \end{bmatrix}$$

is no longer injective. We form the next subspace

$$\ker G_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}_B^{(1)} : u + (b_2\Delta + c_2)v = 0 \right\}.$$

Provided that

$$b_2\gamma + c_2 \neq 0 \ \text{ for all } \ \gamma \in \sigma(\Delta) \tag{12.22}$$

is valid, we get

$$\ker G_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}_B^{(1)} : v = -(b_2\Delta + c_2)^- u \right\}$$

with $(b_2\Delta + c_2)^-$ being a generalized inverse of the operator $b_2\Delta + c_2$. The operator

$$Q_1 = \begin{bmatrix} 1 & 0 \\ -(b_2\Delta + c_2)^- & 0 \end{bmatrix}$$

is a bounded idempotent map acting on $X = L_2(\Omega) \times L_2(\Omega)$,

$$\operatorname{im} Q_1 = N_1 = \operatorname{cl} \ker G_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in X : u \in L_2(\Omega),\ v = -(b_2\Delta + c_2)^- u \right\}.$$

Note that $Q_1 Q_0 = 0$ is satisfied. Compute

$$G_2 = G_1 + BP_0Q_1 = \begin{bmatrix} 1 + c_1 + b_1\Delta & c_2 + b_2\Delta \\ c_3 + b_3\Delta & 0 \end{bmatrix}$$

and turn to $\mathcal{D}_B = H_0^1(\Omega) \times H^1(\Omega)$. Assuming

$$c_3 + b_3\gamma \neq 0 \ \text{ for all } \ \gamma \in \sigma(\Delta) \tag{12.23}$$

to be also satisfied, G_2 is injective and densely solvable. Hence, if $b_4 = c_4 = 0$ and the conditions (12.22), (12.23) are satisfied, then the resulting abstract DAE has index $\mu = 2$. The corresponding inherent regular differential equation is trivially $0 = 0$, since the component $DP_1 x = 0$ disappears. Consequently, no initial condition on $\mathcal{I}$ is allowed.

Finally, we deal with the case $b_4 \neq 0$ and $c_4 \neq 0$, but there is a $\gamma_* \in \sigma(\Delta)$ such that

$$b_4\gamma_* + c_4 = 0.$$

Then, with $\mathcal{D}_B^{(1)}$, we have that

$$\ker G_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}_B^{(1)} : u + (b_2\Delta + c_2)v = 0, v = \Pi_* v \right\},$$

where Π_* denotes the spectral projection onto the eigenspace of Δ corresponding to the eigenvalue γ_*, thus

$$\Delta\Pi_* = \gamma_*\Pi_*, \quad \Pi_*\Delta(I - \Pi_*) = 0.$$

Now, if $b_2\gamma_* + c_2 = 0$, we find $N_1 = 0 \times \operatorname{im}\Pi_*$ and

$$Q_1 = \begin{bmatrix} 0 & 0 \\ 0 & \Pi_* \end{bmatrix}, \quad G_2 = G_1 + BP_0Q_1 = G_1, \quad N_2 = N_1,$$

and so on. There is no injective G_j in the resulting sequence. Therefore, system (12.19) is no longer a regular abstract DAE. The operator pair $\{AD, B\}$ behaves like a singular matrix pencil. Typical for this situation is the particular case of $b_1 = -1$, $b_2 = 1$, $b_3 = 0$, $b_4 = 1$, $c_1 = 0$, $c_2 = c_4 = -\gamma_*$, $c_3 = 0$, i.e., system (12.19) reads

$$\begin{aligned} u_t - \Delta u + (\Delta - \gamma_*)v &= r, \\ (\Delta - \gamma_*)v &= s, \end{aligned}$$

which is no longer solvable for all sufficiently smooth functions $s : \mathcal{I} \to L_2(\Omega)$ and which does not determine the component $\Pi_* v$. Notice that this case is forbidden in [149] by means of the condition

$$(b_1 + \gamma c_1)(b_4 + \gamma c_4) - (b_2 + \gamma c_2)(b_3 + \gamma c_3) \neq 0, \ \gamma \in \sigma(\Delta). \tag{12.24}$$

Next we assume that

$$b_4\gamma_* + c_4 = 0, \ \ b_2\gamma_* + c_2 =: \alpha \neq 0. \tag{12.25}$$

We derive

$$\ker G_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}_B^{(1)} : v = -\frac{1}{\alpha} u,\ u = \Pi_* u \right\},$$
$$N_1 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in X : u = \Pi_* u,\ v = -\frac{1}{\alpha} u \right\},$$
$$Q_1 = \begin{bmatrix} \Pi_* & 0 \\ -\frac{1}{\alpha}\Pi_* & 0 \end{bmatrix}, \quad Q_1 Q_0 = 0,$$

further

$$G_2 = \begin{bmatrix} 1 + (b_1\Delta + c_1)\Pi_* & b_2\Delta + c_2 \\ (b_3\Delta + c_3)\Pi_* & b_4\Delta + c_4 \end{bmatrix} = \begin{bmatrix} 1 + (b_1\gamma_* + c_1)\Pi_* & b_2\Delta + c_2 \\ (b_3\gamma_* + c_3)\Pi_* & b_4\Delta + c_4 \end{bmatrix}.$$

The homogeneous equation $G_2 \begin{bmatrix} u \\ v \end{bmatrix} = 0$ reads in detail

$$u + (b_1\gamma_* + c_1)\Pi_* u + (b_2\Delta + c_2)v = 0, \tag{12.26}$$

$$(b_4\Delta + c_4)v = -(b_3\gamma_* + c_3)\Pi_* u. \tag{12.27}$$

If $b_3\gamma_* + c_3 = 0$, then equations (12.26), (12.27) yield

$$v = \Pi_* v, \quad u = \Pi_* u, \quad v = -\frac{1}{\alpha}(1 + b_1\gamma_* + c_1)u,$$

but $\Pi_* u$ is not determined and G_2 is not injective,

$$\ker G_2 = \left\{ \begin{bmatrix} u \\ v \end{bmatrix} \in \mathcal{D}_B^{(1)} : u = \Pi_* u,\ v = -\frac{1}{\alpha}(1 + b_1\gamma_* + c_1)u \right\},$$

$$Q_2 = \begin{bmatrix} \Pi_* & 0 \\ -\frac{1}{\alpha}(1 + b_1\gamma_* + c_1)\Pi_* & 0 \end{bmatrix}, \quad Q_2 Q_1 = Q_2, \quad Q_2 Q_0 = 0,$$

$G_3 = G_2$, and so on. Again, we arrive at a singular problem. This situation, too, is excluded by condition (12.24), cf. [149].

Provided that

$$b_4\gamma_* + c_4 = 0, \quad b_2\gamma_* + c_2 =: \alpha \neq 0, \quad b_3\gamma_* + c_3 =: \beta \neq 0, \tag{12.28}$$

equation (12.27) leads to $\Pi_* u = 0$, $v = \Pi_* v$ since the Laplacian, as a symmetric operator on $H_0^1(\Omega)$, does not have generalized eigenfunctions. Then, from (12.26), we obtain that $u = \Pi_* u$ and $v = -\frac{1}{\alpha}(1 + b_1\gamma_* + c_1)u$, hence $u = 0$, $v = 0$, i.e., G_2 is injective. G_2 is defined on $\mathcal{D}_B^{(1)}$ and we have $\operatorname{im} G_2 = Y$,

$$G_2^{-1} = \begin{bmatrix} I - \Pi_* & \frac{1}{\beta}\Pi_* - (b_2\Delta + c_2)(b_4\Delta + c_4)_*^-(I - \Pi_*) \\ \frac{1}{\alpha}\Pi_* & (b_4\Delta + c_4)_*^-(I - \Pi_*) - \frac{1}{\alpha\beta}(1 + b_1\gamma_* + c_1)\Pi_* \end{bmatrix},$$

where $(b_4\Delta + c_4)_*^-$ denotes the solution operator for $(b_4\Delta + c_4)f = g$, $g \in \operatorname{im}(I - \Pi_*)$, $f \in H_0^1(\Omega)$, $\Pi_* f = 0$. Now we are able to formulate the corresponding inherent regular differential equation that is related to the component $DP_1 x = (I - \Pi_*)u$. We have $DP_1 D^- = I - \Pi_*$, $Q_1 G_2^{-1} B P_0 = Q_1$,

$$DP_1 G_2^{-1} B D^- = (I - \Pi_*)(b_1\Delta + c_1) - (I - \Pi_*)(b_2\Delta + c_2)(b_4\Delta + c_4)_*^-(I - \Pi_*)(b_3\Delta + c_3)$$

and

$$DP_1 G_2^{-1} \begin{bmatrix} r \\ s \end{bmatrix} = (I - \Pi_*)r - (I - \Pi_*)(b_2\Delta + c_2)(b_4\Delta + c_4)_*^-(I - \Pi_*)s.$$

Therefore, the inherent regular equation is

$$\begin{aligned}((I - \Pi_*)u)_t + \{(b_1\Delta + c_1) - (b_2\Delta + c_2)(b_4\Delta + c_4)_*^-(b_3\Delta + c_3)\}(I - \Pi_*)u \\ = (I - \Pi_*)r - (b_2\Delta + c_2)(b_4\Delta + c_4)_*^-(I - \Pi_*)s.\end{aligned}$$

This equation has to be completed by boundary conditions by choosing $\mathcal{D}_B^{(2)}$, but also by an initial condition $(I - \Pi_*)u(0) \in H_0^1(\Omega)$. The part to be differentiated is

$$DQ_1 \begin{bmatrix} u \\ v \end{bmatrix} = \Pi_* u = DQ_1 G_2^{-1} \begin{bmatrix} r \\ s \end{bmatrix} = \begin{bmatrix} 0 & \frac{1}{\beta}\Pi_* \end{bmatrix} \begin{bmatrix} r \\ s \end{bmatrix} = \frac{1}{\beta}\Pi_* s.$$

Finally, the solutions of (12.19) can be expressed as

$$\begin{aligned}\begin{bmatrix} u \\ v \end{bmatrix} &= \begin{bmatrix} 1 \\ 0 \end{bmatrix}((I - \Pi_*)u + \frac{1}{\beta}\Pi_* s) + \begin{bmatrix} 0 \\ v \end{bmatrix}, \\ \begin{bmatrix} 0 \\ v \end{bmatrix} &= Q_0 P_1 G_2^{-1} \begin{bmatrix} r \\ s \end{bmatrix} - Q_0 P_1 G_2^{-1} B D^-(I - \Pi_*)u + \begin{bmatrix} 0 \\ \frac{1}{\alpha\beta}(\Pi_* s)_t \end{bmatrix}.\end{aligned}$$

Consequently, provided condition (12.28) is satisfied, the abstract formulation of the system (12.19) has index $\mu = 2$.

12.2.3 A coupled system of a PDE and Fredholm integral equations

Given a linear Fredholm integral operator $F : L_2(\Omega)^s \to L_2(\Omega)^s$, $\|F\| < 1$, a linear differential operator

$$K : C_0^2(\Omega) \to L_2(\Omega), \quad Kw := -\Delta w + cw \text{ for } w \in C_0^2(\Omega), \quad c \geq 0,$$

with linear bounded coupling operators $L : L_2(\Omega)^s \to L_2(\Omega)$, $E : L_2(\Omega) \to L_2(\Omega)^s$, the system to be considered is ([26])

$$\begin{aligned} x_1'(t) + Kx_1(t) + Lx_2(t) &= q_1(t), \\ Ex_1(t) + (I-F)x_2(t) &= q_2(t), \quad t \in [0,1]. \end{aligned} \tag{12.29}$$

Using the corresponding matrix representations for A, D, B, with $X := L_2(\Omega) \times L_2(\Omega)^s$, $Y := X$, $Z := L_2(\Omega)$, we rewrite (12.29) in the form (12.3). Namely, we have

$$A = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad G_0 = AD = P_0 = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad D^- = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad R = 1$$

and $B = \begin{bmatrix} K & L \\ E & I-F \end{bmatrix}$. By construction B is defined on $\mathcal{D}_B := C_0^2(\Omega) \times L_2(\Omega)^s$. Clearly, it holds that $N_0 = 0 \times L_2(\Omega)^s$. Choosing

$$Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}, \quad W_0 = \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$$

we obtain

$$W_0 B = \begin{bmatrix} 0 & 0 \\ E & I-F \end{bmatrix}, \quad G_1 = \begin{bmatrix} 1 & L \\ 0 & I-F \end{bmatrix}$$

defined on X (as trivial extensions of bounded maps). G_1 is a bijection such that this abstract DAE has the index 1. This implies

$$G_1^{-1} = \begin{bmatrix} 1 & -L(I-F)^{-1} \\ 0 & (I-F)^{-1} \end{bmatrix}, \quad DG_1^{-1}BD^- = K - L(I-F)^{-1}E.$$

Each solution of the DAE is given by the expression

$$\begin{aligned} x(t) &= D^- u(t) + Q_0 x(t), \\ Q_0 x(t) &= Q_0 G_1^{-1} q(t) - Q_0 G_1^{-1} B D^- u(t), \end{aligned}$$

where $u(t)$ is a solution of the abstract regular differential equation

$$u'(t) + DG_1^{-1}BD^- u(t) = DG_1^{-1} q(t),$$

which corresponds to

$$x_1'(t) + (K - L(I-F)^{-1}E)x_1(t) = q_1(t) - L(I-F)^{-1}Eq_2(t). \tag{12.30}$$

Obviously, one has to state an appropriate initial condition for (12.30), i.e., $x_1(0) = x_1^0 \in C_0^2(\Omega)$. Better solvability will be obtained by defining B (respectively K) on $H_0^1(\Omega)$ instead of $C_0^2(\Omega)$ and using weak solutions.

12.2.4 A PDE and a DAE coupled by a restriction operator

Here we consider systems where PDE and DAE model equations are coupled via certain boundary conditions. Such systems often arise by multiphysical modeling, for instance, when coupling circuit and electromagnetic simulation. Consider the system

$$\tilde{u}_t(y,t) - \tilde{u}_{yy}(y,t) + c\tilde{u}(y,t) = f(y,t),\ y \in \Omega,\ t \geq 0,\ \tilde{u}|_{\partial\Omega} = 0, \quad (12.31)$$

$$\tilde{A}(t)(\tilde{D}(t)\tilde{x}(t))' + \tilde{B}(t)\tilde{x}(t) + r(t)\tilde{u}(\cdot,t) = g(t), \quad t \geq 0. \quad (12.32)$$

Assume the linear restriction map $r(t) : H^1(\Omega) \to \mathbb{R}^m$ (usually describing boundary conditions for $\tilde{u}$) to be bounded and to depend continuously on t. Rewrite the system (12.31)–(12.32) as an abstract DAE for $x(t) = \begin{bmatrix} \tilde{u}(\cdot,t) \\ \tilde{x}(t) \end{bmatrix}$ and choose $X = Y = L_2(\Omega) \times \mathbb{R}^m$, $Z = L_2(\Omega) \times \mathbb{R}^n$,

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A} \end{bmatrix}, \quad D = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{D} \end{bmatrix}, \quad AD = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{A}\tilde{D} \end{bmatrix}, \quad Q_0 = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{Q}_0 \end{bmatrix},$$

$$D^- = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{D}^- \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{R} \end{bmatrix}, \quad B = \begin{bmatrix} -\Delta + c & 0 \\ r & \tilde{B} \end{bmatrix},$$

where $\mathcal{D}_B := H_0^1(\Omega) \times \mathbb{R}^m$. $G_1 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{G}_1 \end{bmatrix}$ is defined on X, $N_1 = 0 \times \tilde{N}_1$, $\operatorname{im} G_1 = L_2(\Omega) \times \operatorname{im} \tilde{G}_1$, $S_1 = \{x \in X : rx_1 + \tilde{B}\tilde{P}_0 x_2 \in \operatorname{im} \tilde{G}_1\}$. Supposing that the operator $r(t)$ maps into $\operatorname{im} \tilde{G}_1(t)$ it holds that

$$S_1 = \{x \in X : \tilde{B}\tilde{P}_0 x_2 \in \operatorname{im} \tilde{G}_1\} = L_2(\Omega) \times \tilde{S}_1.$$

Then, $Q_1 = \begin{bmatrix} 0 & 0 \\ 0 & \tilde{Q}_1 \end{bmatrix}$ is the projector onto N_1 along S_1, $G_2 = \begin{bmatrix} 1 & 0 \\ 0 & \tilde{G}_2 \end{bmatrix}$.

In the general case the projector Q_1 onto N_1 along S_1 is more difficult to construct. Obviously, we have $N_1 \cap S_1 = 0 \times (\tilde{N}_1 \cap \tilde{S}_1)$. G_2 is a bijection if $\tilde{G}_2$ is so. Consequently, the coupled system (12.31)–(12.32) interpreted as an abstract DAE has the same index as the DAE (12.32).

12.3 Linear ADAEs with monotone operators

The goal is to obtain existence and uniqueness of solutions for abstract differential-algebraic equations of the form

$$A(Dx)'(t) + B(t)x(t) = q(t) \qquad \text{for almost all} \quad t \in \mathcal{I} \quad (12.33)$$

with $\mathcal{I} := (t_0, T)$ and with linear operators A, D and $B(t)$ acting in linear Hilbert spaces. They naturally arise by a generalized formulation of coupled systems of differential-algebraic equations and partial differential equations of elliptic and parabolic type. Here, $(Dx)'$ denotes the generalized derivative of Dx in the sense of distributions.

Obviously, if A and D represent identity mappings and B is an elliptic operator, then the system (12.33) represents a system of parabolic differential equations. Usual approaches to obtain existence results in the theory of parabolic differential equations are the theory of semigroups and the Galerkin method (see e.g., [217]). From the numerical point of view, the Galerkin method is preferable. First, it provides canonically a numerical method for solving the ADAE. Secondly, the theory of semigroups treats the abstract system as an evolution on a manifold. However, this manifold is unknown at the outset and must be calculated when solving the system numerically. But, already in the finite-dimensional case, a calculation of the manifold (for a description see, e.g., [11, 189]) would be connected with significant computational expense, in particular for systems of higher index. Furthermore, it would be necessary to investigate the influence of perturbations of the manifold onto solutions, since we cannot expect to calculate the manifold exactly.

We follow the ideas in [207] and consider the following approach as a starting point for a general treatment of ADAEs. Existence proofs for nonlinear differential equations are often based on the construction of suitable fixed point mappings using the existence of solutions for linear differential equation systems. Therefore, we consider unique solvability statements for linear ADAEs as a substantial basis for a general approach treating abstract differential-algebraic systems.

12.3.1 Basic functions and function spaces

We start with the following assumptions.

Assumption 12.5. *The spaces V, Y and H are real Hilbert spaces. $Y \subseteq H \subseteq Y^*$ is an evolution triple. The mapping*

$$D : V \to Y$$

is linear, continuous and surjective. The mapping

$$A : Y^* \to V^*$$

represents the dual mapping of D, which means

$$\langle Af, v\rangle_V = \langle f, Dv\rangle_Y \qquad \text{for all} \quad v \in V.$$

The mapping

$$B(t) : V \to V^*$$

is linear, uniformly bounded and uniformly strongly monotone for all $t \in \mathcal{I}$. More precisely, there are constants $c_1, c_2 > 0$ such that

$$\langle B(t)x, v\rangle_V \le c_1 \|x\|\,\|v\|, \quad \langle B(t)x, x\rangle_V \ge c_2 \|x\|^2$$

for all $x, v \in V$ and $t \in \mathcal{I}$. Furthermore we assume the mapping

$$t \mapsto \langle B(t)x, v\rangle_V$$

to be measurable for all $x, v \in V$.

The assumption also implies that the operator AJD is monotone with $J : Y \to Y^*$ being the embedding operator mapping elements y of Y to the functional $f_y \in Y^*$ satisfying $f_y(z) = \langle y, z\rangle_Y$ for all $z \in Y$ and being well defined by the Riesz representation theorem. It is a simple consequence of

$$\langle AJDv, v\rangle_V = \langle JDv, Dv\rangle_Y = (Dv, Dv)_H = \|Dv\|_H^2$$

for all $v \in V$. We consider finite time intervals $\mathcal{I} := (t_0, T)$ with $t_0 < T < \infty$. For evolution equations, the natural solution space is given by the Sobolev space

$$W_2^1(\mathcal{I}, V, H) = \{x \in L_2(\mathcal{I}, V) : \; x' \in L_2(\mathcal{I}, V^*)\}$$

with x' being a generalized derivative of x satisfying

$$\int_{t_0}^{T} \varphi'(t)x(t)\mathrm{d}t = -\int_{t_0}^{T} \varphi(t)x'(t)\mathrm{d}t \quad \text{for all } \varphi(t) \in C_0^\infty(t_0, T).$$

For linear ADAEs of the form (12.33), we have to modify it, since we need the generalized derivative of $(Dx)(t)$ which belongs to Y and not to V. Consequently, we define

$$W_{2,D}^1(\mathcal{I}, V, Y, H) := \{x \in L_2(\mathcal{I}, V) : \; (Dx)' \in L_2(\mathcal{I}, Y^*)\}$$

where $(Dx)'$ denotes the generalized derivative of Dx, which means

$$\int_{t_0}^{T} \varphi'(t)Dx(t)\mathrm{d}t = -\int_{t_0}^{T} \varphi(t)(Dx)'(t)\mathrm{d}t \quad \text{for all } \varphi(t) \in C_0^\infty(t_0, T).$$

Here, $Dx : \mathcal{I} \to Y$ is defined by $(Dx)(t) = Dx(t)$ for all $t \in \mathcal{I}$.

Proposition 12.6. *The space $W_{2,D}^1(\mathcal{I}, V, Y, H)$ forms a real Banach space with the norm*

$$\|x\|_{W_{2,D}^1} := \|x\|_{L_2(\mathcal{I},V)} + \|(Dx)'\|_{L_2(\mathcal{I},Y^*)}.$$

Proof. Obviously, $\|\cdot\|_{W_{2,D}^1}$ is a norm since $\|\cdot\|_{L_2(\mathcal{I},V)}$ and $\|\cdot\|_{L_2(\mathcal{I},Y^*)}$ are norms. It remains to show that $W_{2,D}^1(\mathcal{I}, V, Y, H)$ is complete. Let (x_n) be a Cauchy sequence in $W_{2,D}^1(\mathcal{I}, V, Y, H)$. Since $L_2(\mathcal{I}, V)$ and $L_2(\mathcal{I}, Y^*)$ are Banach spaces, we find $x \in$

$L_2(\mathcal{I},V)$ and $v \in L_2(\mathcal{I},Y^*)$ with

$$x_n \to x \qquad \textit{and} \qquad (Dx_n)' \to v.$$

Since $D:\ V \to Y$ is continuous, we have

$$Dx_n \to Dx \qquad \text{in} \quad L_2(\mathcal{I},Y).$$

The continuous embedding $Y \subseteq Y^*$ yields the continuous embedding $L_2(\mathcal{I},Y) \subseteq L_2(\mathcal{I},Y^*)$. Consequently,

$$Dx_n \to Dx \qquad \text{in} \quad L_2(\mathcal{I},Y^*).$$

Since $L_2(\mathcal{I},Y^*) \subseteq L_1(\mathcal{I},Y^*)$, we get

$$Dx_n \to Dx \qquad \text{and} \qquad (Dx_n)' \to v \qquad \text{in} \quad L_1(\mathcal{I},Y^*). \tag{12.34}$$

For $\varphi \in C_0^\infty(\mathcal{I})$, we have

$$\int_{t_0}^{T} \varphi' Dx_n \,\mathrm{d}t = -\int_{t_0}^{T} \varphi (Dx_n)' \,\mathrm{d}t.$$

(12.34) allows us to apply the limit $n \to \infty$ which yields

$$\int_{t_0}^{T} \varphi' Dx \,\mathrm{d}t = -\int_{t_0}^{T} \varphi v \,\mathrm{d}t.$$

But this means $v = (Dx)'$ and hence $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$. □

Proposition 12.7. *If $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$, then Dx belongs to the classical Sobolev space*

$$W_2^1(\mathcal{I},Y,H) = \{v \in L_2(\mathcal{I},Y):\ v' \in L_2(\mathcal{I},Y^*)\}.$$

Proof. Let x belong to $W^1_{2,D}(\mathcal{I},V,Y,H)$. This implies $x \in L_2(\mathcal{I},V)$. Since $D:\ V \to Y$ is continuous, Dx belongs to $L_2(\mathcal{I},Y)$ and the proposition is proved. □

The last proposition implies immediately two important properties of the function Dx if $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$. The first one is a simple conclusion of the continuous embedding

$$W_2^1(\mathcal{I},Y,H) \subseteq C(\mathcal{I},H).$$

Corollary 12.8. *If $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$, then there exists a uniquely determined continuous function $y : \mathcal{I} \to H$ which coincides almost everywhere on $\mathcal{I}$ with the function Dx. Furthermore,*

$$\max_{t_0 \le t \le T} \|y(t)\|_H \le \text{const} \|Dx\|_{W_2^1}.$$

As a consequence of the generalized integration by parts formula, we obtain the next corollary.

Corollary 12.9. *The integration by parts formula is satisfied for all $u,\ v \in W^1_{2,D}(\mathcal{I},V,Y,H)$ and arbitrary $s,\ t$ with $t_0 \le s \le t \le T$, which means*

$$\big(Du(t)|Dv(t)\big)_H - \big(Du(s)|Dv(s)\big)_H = \int_s^t \langle (Du)'(\tau), Dv(\tau)\rangle_Y + \langle (Dv)'(\tau), Du(\tau)\rangle_Y \, d\tau. \tag{12.35}$$

Here, the values of Du and Dv are the values of the continuous functions $z_u,\ z_v : \mathcal{I} \to H$ in the sense of Corollary 12.8.

All statements and arguments of this section remain true if D depends on t provided that $D(\cdot,t) : V \to Y$ is uniformly Lipschitz continuous for almost all $t \in \mathcal{I}$, i.e.,

$$\|D(u,t) - D(v,t)\|_Y \le c\|u-v\|_V \quad \forall u,v \in V, \quad \text{for almost all } t \in \mathcal{I}$$

with a constant $c > 0$ being independent of t.

12.3.2 Galerkin approach

As explained in the section before, the natural solution space for ADAEs of the form (12.33) is given by $W^1_{2,D}(\mathcal{I},V,Y,H)$. Therefore, we consider the initial value problem

$$A[Dx(t)]' + B(t)x(t) = q(t) \qquad \text{for almost all} \quad t \in \mathcal{I}, \tag{12.36}$$

$$Dx(t_0) = y_0 \in H \tag{12.37}$$

with $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$. The initial condition (12.37) is to be understood as follows. For each $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$ we can modify Du on a subset of $\mathcal{I}$ such that the mapping $Du : \mathcal{I} \to H$ is continuous. This continuous representative Du makes (12.37) meaningful.

Additionally, we need the following assumption for the choice of basis functions. It ensures that the basis functions in the spaces V and V are consistent with each other.

Assumption 12.10. *Let $y_0 \in H$ and $q \in L_2(\mathcal{I},Y^*)$ be given. Furthermore, let $\{w_1,w_2,\dots\}$ be a basis in V and $\{y_1,y_2,\dots\}$ be a basis in Y such that, for all $n \in \mathbb{N}$,*

$$\exists m_n \in \mathbb{N}: \quad \{Dw_1,\dots,Dw_n\} \subseteq span\{y_1,\dots,y_{m_n}\}.$$

Furthermore, we require (y_{n0}) to be a sequence from H with

$$y_{n0} \to y_0 \quad in\ H \quad as\ n \to \infty,$$

where

$$y_{n0} \in span\{Dw_1, \ldots, Dw_n\} \quad \textit{for all } n.$$

In order to formulate the Galerkin method, we set

$$x_n(t) = \sum_{i=1}^{n} c_{in}(t) w_i.$$

Additionally, we use the formulation

$$\langle A[Dx(t)]', v\rangle_V + \langle B(t)x(t), v\rangle_V = \langle q(t), v\rangle_V \qquad \forall v \in V \tag{12.38}$$

which is equivalent to (12.36) for almost all $t \in \mathcal{I}$. Then, we obtain the Galerkin equations if we replace x by x_n and v by w_i:

$$\langle A[Dx_n(t)]', w_i\rangle_V + \langle B(t)x_n(t), w_i\rangle_V = \langle q(t), w_i\rangle_V, \tag{12.39}$$

$$Dx_n(t_0) = y_{n0}, \tag{12.40}$$

for all $i = 1, \ldots, n$. By Assumption 12.5, equation (12.39) reads as

$$\langle [Dx_n(t)]', Dw_i\rangle_Y + \langle B(t)x_n(t), w_i\rangle_V = \langle q(t), w_i\rangle_V. \tag{12.41}$$

Regarding the continuous embedding $H \subseteq Y^*$, we may also write

$$\big([Dx_n(t)]' | Dw_i\big)_H + \langle B(t)x_n(t), w_i\rangle_V = \langle q(t), w_i\rangle_V.$$

Consequently, the Galerkin equations are given by

$$\left(\sum_{j=1}^{n} [c_{jn}(t) Dw_j]' | Dw_i\right)_H + \sum_{j=1}^{n} \langle B(t)w_j, w_i\rangle_V c_{jn}(t) = \langle q(t), w_i\rangle_V, \tag{12.42}$$

$$Dx_n(t_0) = y_{n0}, \tag{12.43}$$

for all $i = 1, \ldots, n$. If we take into account Assumption 12.10, then we find coefficients a_{ik} with $i = 1, \ldots, n$ and $k = 1, \ldots, m_n$ such that

$$Dw_i = \sum_{k=1}^{m_n} a_{ik} y_k \quad \forall i = 1, \ldots, n.$$

Note that the coefficients are simply given by $a_{ik} = (Dw_i | y_k)_H$ if the basis $\{y_1, y_2, \ldots\}$ is an orthonormal basis in Y.

Consequently, equation (12.42) is equivalent to

$$\sum_{k=1}^{m_n} \left(\sum_{j=1}^{n} [c_{jn}(t) Dw_j]' | a_{ik} y_k\right)_H + \sum_{j=1}^{n} \langle B(t)w_j, w_i\rangle_V c_{jn}(t) = \langle q(t), w_i\rangle_V$$

for all $i = 1, \ldots, n$. This can be rewritten as

$$\sum_{k=1}^{m_n} a_{ik} \frac{\mathrm{d}}{\mathrm{d}t} \left(\sum_{j=1}^{n} (Dw_j | y_k)_H c_{jn}(t) \right) + \sum_{j=1}^{n} \langle B(t) w_j, w_i \rangle_V c_{jn}(t) = \langle q(t), w_i \rangle_V.$$

Furthermore, equation (12.43) is equivalent to

$$\sum_{j=1}^{n} c_{jn}(t_0)(Dw_j | y_k)_H = \sum_{j=1}^{n} \alpha_{jn} (Dw_j | y_k)_H \qquad \forall k = 1, \ldots, m_n,$$

where $y_{n0} = \sum_{j=1}^{n} \alpha_{jn} Dw_j$. The existence of the coefficients α_{jn} and the second equivalence are ensured by Assumption 12.10. Hence, the Galerkin equations represent an initial value differential-algebraic equation

$$A_n (D_n c_n(t))' + B_n(t) c_n(t) = r(t) \tag{12.44}$$

$$D_n c_n(t_0) = D_n \alpha_n \tag{12.45}$$

for the coefficients $c_{nj}(t)$ if we introduce the vector function $c_n(\cdot)$ as

$$c_n(t) = (c_{nj}(t))_{j=1,\ldots,n} \qquad \text{for all} \quad t \in \mathcal{I}$$

and the vector $\alpha_n = (\alpha_{jn}(t))_{j=1,\ldots,n}$. The matrices A_n, D_n and $B_n(t)$ are given by

$$A_n = (a_{ik})_{\substack{i=1,\ldots,n \\ k=1,\ldots,m_n}}, \quad D_n = (d_{kj})_{\substack{k=1,\ldots,m_n \\ j=1,\ldots,n}}, \quad B_n(t) = (b_{ij}(t))_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

with

$$d_{kj} = (Dw_j | y_k)_H \qquad \text{and} \qquad b_{ij}(t) = \langle B(t) w_j, w_i \rangle_V$$

for $i, j = 1, \ldots, n$ and $k = 1, \ldots, m_n$. Finally, the vector function for the right-hand side reads as

$$r(t) = (r_j(t))_{j=1,\ldots,n}$$

with $r_j(t) = \langle q(t), w_j \rangle_V$ for $j = 1, \ldots, n$.

Proposition 12.11. *The differential-algebraic equation* (12.44) *arising from the Galerkin approach for the IVP* (12.36)–(12.37) *has a properly stated leading term, which means*

$$\ker A_n \oplus \operatorname{im} D_n = \mathbb{R}^{m_n}.$$

Proof. We shall show first that the intersection of the spaces $\ker A_n$ and $\operatorname{im} D_n$ is trivial. We assume $z \in \mathbb{R}^{m_n}$ to belong to this intersection, which means

$$\sum_{k=1}^{m_n} a_{ik} z_k = 0 \quad \text{for all} \quad i = 1, \ldots, n.$$

Furthermore there exist $p_j \in \mathbb{R}^n$ for $j = 1, \ldots, n$ such that

$$z_k = \sum_{j=1}^{n} (Dw_j | y_k)_H p_j \quad \text{for all} \quad k = 1, \ldots, m_n.$$

Multiplying the last equations by a_{ik} and summing over $k = 1, \dots, m_n$ we get

$$\sum_{j=1}^{n} (Dw_j | \sum_{k=1}^{m_n} a_{ik} y_k)_H p_j = 0 \quad \text{for all } i = 1, \dots, n.$$

This implies

$$(\sum_{j=1}^{n} p_j Dw_j | Dw_i)_H = 0 \quad \text{for all } \ i = 1, \dots, n,$$

because of the definition of the coefficients a_{ik}. Multiplying this by p_i and building the sum over $i = 1, \dots, n$ we obtain

$$\| \sum_{j=1}^{n} p_j Dw_j \|_H^2 = 0, \text{ which implies } \sum_{j=1}^{n} p_j Dw_j = 0.$$

Consequently,

$$z_k = \sum_{j=1}^{n} (p_j Dw_j | y_k)_H = 0 \quad \text{for all } \ k = 1, \dots, m_n.$$

It remains to show that the sum of the spaces $\ker A_n$ and $\operatorname{im} D_n$ spans the whole space $\mathbb{R}^{m_n}$. For this, it is enough to verify that

$$\dim \ker A_n \geq m_n - \dim \operatorname{im} D_n.$$

Let d be the dimension of $\operatorname{im} D_n$. If $d = m_n$ we are done. If $d < m_n$, then, without loss of generality, let the first d rows of D_n be linearly independent. Then, we find for all k with $d < k \leq m_n$ real values $\lambda_{k1}, \dots, \lambda_{kd}$ such that

$$d_{ki} = \sum_{j=1}^{k} \lambda_{kj} d_{ji} \qquad \forall\, i = 1, \dots, m_n.$$

Regarding the definition of d_{ki} we may rewrite this equation as

$$(Dw_i | y_k - \sum_{j=1}^{k} \lambda_{kj} y_j)_H = 0 \qquad \forall\, i = 1, \dots, m_n.$$

Using the definition of A_n we find

$$\sum_{l=1}^{m_n} a_{il} (y_l | y_k - \sum_{j=1}^{k} \lambda_{kj} y_j)_H = 0 \qquad \forall\, i = 1, \dots, m_n.$$

This implies $A_n z_k = 0$ for $z_k = (z_{kl})_{l=1,\dots,m_n}$ with

$$z_{kl} = (y_l|y_k - \sum_{j=1}^{k} \lambda_{kj} y_j)_Y \quad \forall\, l = 1,\dots,m_n.$$

Consequently, $\dim\ker A_n \geq m_n - d$ holds if $\{z_k; k = d+1,\dots,m_n\}$ is linearly independent. We assume linear dependence, which means, we find $\mu_{d+1},\dots,\mu_{m_n}$ such that

$$\sum_{k=d+1}^{m_n} \mu_k z_k = 0.$$

This implies

$$\sum_{k=d+1}^{m_n} \mu_k (y_l|y_k - \sum_{j=1}^{k} \lambda_{kj} y_j)_H = 0 \quad \forall\, l = 1,\dots,m_n.$$

Defining

$$\xi_{kj} := \begin{cases} -\lambda_{kj} & \text{if } 1 \leq j \leq d \\ 1 & \text{if } j = k \\ 0 & \text{else} \end{cases}$$

for all $k = d+1,\dots,m_n$, the last equation reads as

$$(y_l| \sum_{k=d+1}^{m_n} \mu_k \sum_{j=1}^{m_n} \xi_{kj} y_j)_H = 0 \quad \forall\, l = 1,\dots,m_n.$$

This yields

$$(\sum_{l=1}^{m_n} \sum_{k=d+1}^{m_n} \mu_k \xi_{kl} y_l| \sum_{j=1}^{m_n} \sum_{k=d+1}^{m_n} \mu_k \xi_{kj} y_j)_H = 0$$

and hence

$$\sum_{j=1}^{m_n} \sum_{k=d+1}^{m_n} \mu_k \xi_{kj} y_j = 0.$$

Since $\{y_1,\dots,y_{m_n}\}$ is linearly independent, we get

$$\sum_{k=d+1}^{m_n} \mu_k \xi_{kj} = 0 \quad \forall\, j = 1,\dots,m_n.$$

Considering the definition of ξ_{kj}, we obtain that $\mu_j = 0$ for $j = d+1,\dots,m_n$, which means $\{z_k; k = d+1,\dots,m_n\}$ is linearly independent. □

Proposition 12.12. *For the leading term matrix functions of the differential-algebraic equation* (12.44)*:*

(1) $(\mathrm{im}\, A_n)^\perp = \ker D_n$,
(2) $A_n D_n$ *is positive semidefinite,*
(3) $B_n(t)$ *is positive definite for all* $t \in \mathcal{I}$.

Proof. (1) Any vector $z \in \mathbb{R}^n$ belongs to $\ker D_n$ if and only if

$$\sum_{j=1}^{n}(Dw_j|y_l)_H z_j = 0 \qquad \forall\, l = 1,\dots,m_n$$

or, equivalently,

$$\sum_{j=1}^{n}(\sum_{k=1}^{m_n} a_{jk}y_k|z_l)_H z_j = 0 \qquad \forall\, l = 1,\dots,m_n.$$

This is equivalent to

$$\sum_{k=1}^{m_n}\sum_{j=1}^{n} z_j a_{jk} y_k = 0,$$

which means, $\sum_{j=1}^{n} z_j a_{jk} = 0$ for all $k = 1,\dots,m_n$, since $\{y_k; k = 1,\dots,m_n\}$ is linearly independent. But this means nothing else than $z \in (\ker A_n)^{\perp}$.

(2) For any $z \in \mathbb{R}^n$ we have

$$\begin{aligned} z^{\mathrm{T}}A_nD_nz &= \sum_{k=1}^{m_n}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ik}d_{kj}z_iz_j \;=\; \sum_{i=1}^{n}\sum_{j=1}^{n}(Dw_j|Dw_i)_H z_iz_j \\ &= (\sum_{j=1}^{n} z_jDw_j|\sum_{i=1}^{n} z_iDw_i)_H \;=\; \|\sum_{j=1}^{n} z_jDw_j\|_H^2 \geq 0. \end{aligned}$$

(3) Let $z \neq 0$ be any vector in $\mathbb{R}^n$. Then we get

$$z^{\mathrm{T}}B_n(t)z = \sum_{i=1}^{n}\sum_{j=1}^{n} z_i\langle B(t)w_j, w_i\rangle_V z_j = \langle B(t)(\sum_{j=1}^{n} z_jw_j), \sum_{i=1}^{n} z_iw_i\rangle_V \;> 0$$

since $B(t)$ is strongly monotone and $w_1,\dots,w_n$ are linearly independent. □

Proposition 12.13. *The differential-algebraic equation* (12.44) *arising from the Galerkin approach for the IVP* (12.36)–(12.37) *has at most index 1.*

Proof. If A_nD_n is singular, then let Q_n be any projector onto $\ker A_nD_n$. We shall show that the matrix

$$G_{1n} := A_nD_n + B_n(t)Q_n$$

is nonsingular. We assume z to belong to the nullspace of G_{1n}, i.e.,

$$A_nD_nz + B_n(t)Q_nz = 0. \tag{12.46}$$

Multiplying the equation by $(Q_nz)^{\mathrm{T}}$ we get

$$(Q_nz)^{\mathrm{T}}B_n(t)Q_nz = 0$$

since $\ker Q_n^T = \mathrm{im}A_n$ (see Proposition 12.12). The positive definiteness of B_n yields $Q_nz = 0$ and, regarding (12.46), we have $A_nD_nz = 0$. But the latter equation means nothing else than $z = Q_nz$ and, finally, $z = 0$. □

12.3.3 Solvability

In order to obtain unique solutions via the Galerkin method we shall need, additionally, the following assumption.

Assumption 12.14. *Let $Q: V \to V$ be a projection operator with $\operatorname{im} Q = \ker D$. The existence of such an operator is ensured by the continuity of D implying $\ker D$ to be closed in V. Then, we assume that the basis $\{w_1, w_2, \ldots\}$ of V is chosen such that there are index sets I_1, $I_2 \subset \mathbb{N}$ with*

$$\operatorname{span}\{w_i \in V \mid i \in I_1\} = \ker Q \text{ and } \operatorname{span}\{w_i \mid i \in I_2\} = \operatorname{im} Q = \ker D.$$

This assumption guarantees that the dynamic part of the solution will be approximated by linear combinations of the basis functions w_i for $i \in I_1$. Correspondingly, the nondynamic part of the solution will be approximated by linear combinations of the basis functions w_i for $i \in I_2$. In applications, it should not be a problem to fulfill Assumption 12.14.

In the proof of the existence and uniqueness of solutions, we will need some properties of the adjoint operator of the projection operator Q. Therefore, we summarize them in the following lemma.

Lemma 12.15. *Let V and Y be Banach spaces. Furthermore, let D be a linear, continuous, and surjective operator $D: V \to Y$.*

If $Q: V \to V$ is a projection operator onto $\ker D$, then the adjoint operator $Q^*: V^* \to V^*$ defined by

$$\langle Q^*\bar{v}, v\rangle_V = \langle \bar{v}, Qv\rangle_V \qquad \forall\, \bar{v} \in V^*,\ v \in V$$

is a projection operator along $\operatorname{im} D^*$ for the adjoint operator $D^*: Y^* \to V^*$ defined by

$$\langle D^*\bar{y}, v\rangle_V = \langle \bar{y}, Dv\rangle_Y \qquad \forall\, \bar{y} \in Y^*,\ v \in V.$$

Proof. (i) $Q^{*2} = Q^*$. For all $\bar{v} \in V^*$ and $v \in V$, we have

$$\langle Q^{*2}\bar{v}, v\rangle_V = \langle Q^*\bar{v}, Qv\rangle_V = \langle \bar{v}, Q^2 v\rangle_V = \langle \bar{v}, Qv\rangle_V = \langle Q^*\bar{v}, v\rangle_V.$$

(ii) Continuity. Let $v \neq 0$ belong to V and $\bar{v} \in V^*$. This implies

$$\left|\left\langle Q^*\bar{v}, \frac{v}{\|v\|}\right\rangle\right| = \left|\left\langle \bar{v}, \frac{Qv}{\|v\|}\right\rangle\right| = \left|\left\langle \bar{v}, \frac{Qv}{\|Qv\|}\right\rangle\right| \frac{\|Qv\|}{\|v\|} \leq \text{const}\|\bar{v}\|_{V^*}$$

since Q is continuous. But this means $\|Q^*\bar{v}\|_{V^*} \leq \text{const}\|\bar{v}\|_{V^*}$.

(iii) $\operatorname{im} D^* \subseteq \ker Q^*$. For all $\bar{y} \in Y^*$, we get

$$\langle Q^*(D^*\bar{y}), v\rangle_V = \langle D^*\bar{y}, Qv\rangle_V = \langle \bar{y}, D(Qv)\rangle_Y = 0.$$

(iv) $\ker Q^* \subseteq \operatorname{im} D^*$. Let $\bar{v} \in \ker Q^*$, i.e.,

$$0 = \langle Q^* \bar{v}, v \rangle_V = \langle \bar{v}, Qv \rangle_V \tag{12.47}$$

for all $v \in V$. Since D is surjective, we find, for all $y \in Y$, a $v \in V$ such that $y = Dv$. This allows us to define a functional $\bar{y} \in Y^*$ by

$$\langle \bar{y}, y \rangle_Y := \langle \bar{v}, v \rangle_V$$

for any $v \in V$ with $y = Dv$. The functional $\bar{y}$ is well defined since, for any v_1, $v_2 \in V$ with

$$Dv_1 = y = Dv_2,$$

it follows that $v_1 - v_2 \in \ker D = \operatorname{im} Q$ and, consequently,

$$\langle \bar{v}, v_1 \rangle_V = \langle \bar{v}, v_2 \rangle_V$$

if we regard (12.47). Finally, for all $v \in V$,

$$\langle D^* \bar{y}, v \rangle_V = \langle \bar{y}, Dv \rangle_Y = \langle \bar{v}, v \rangle_V,$$

which yields $\bar{v} = D^* \bar{y} \in \operatorname{im} D^*$. □

Note that the surjectivity of D is needed for the relation $\ker Q^* \subseteq \operatorname{im} D^*$ only.

Theorem 12.16. *Let the Assumptions 12.5, 12.10 and 12.14 be satisfied. Then, the ADAEs (12.36)–(12.37) have exactly one solution* $x \in W^1_{2,D}(\mathcal{I}, V, Y, H)$.

The following proof is oriented towards the existence proof for first-order linear evolution equations presented in [217]. The main differences are the following.

1. We are looking for solutions $x \in W^1_{2,D}(\mathcal{I}, V, Y, H)$ instead of $x \in W^1_2(\mathcal{I}, V, H)$.
2. The Galerkin equations represent a differential-algebraic equation instead of an explicit ordinary differential equation.
3. Appropriate initial conditions are given only for $Dx(t_0)$ instead of the whole of $x(t_0)$.
4. Assumption 12.14 is needed to ensure the existence of the generalized derivative $(Dx)'$.

Proof. For brevity we set $W = W^1_{2,D}(\mathcal{I}, V, Y, H)$.

Step 1: Uniqueness. We suppose x_1 and x_2 to be two solutions of the system (12.36)–(12.37). Then, the difference $x = x_1 - x_2$ satisfies the initial value problem

$$\begin{aligned} A[Dx(t)]' + B(t)x(t) &= 0 \qquad \text{for almost all} \quad t \in (t_0, T), \\ Dx(t_0) &= 0 \end{aligned}$$

with $x \in W$. This yields

$$\int_{t_0}^{T} \left\langle A[Dx(t)]', x(t) \right\rangle_V \mathrm{d}t + \int_{t_0}^{T} \left\langle B(t)x(t), x(t) \right\rangle_V \mathrm{d}t = 0.$$

Regarding Assumption 12.5, we have

$$\big\langle A(Dx)'(t), x(t)\big\rangle_V = \big((Dx)'(t)|Dx(t)\big)_H.$$

By the integration by parts formula (12.35) we get

$$\tfrac{1}{2}\|Dx(T)\|_H^2 - \tfrac{1}{2}\|Dx(t_0)\|_H^2 = -\int_{t_0}^{T} \big\langle Bx(t), x(t)\big\rangle_V \,\mathrm{d}t.$$

Since $B(t)$ is uniformly strongly monotone, there is a constant $c > 0$ such that

$$\tfrac{1}{2}\|Dx(T)\|_H^2 - \tfrac{1}{2}\|Dx(t_0)\|_H^2 \le -c\int_{t_0}^{T} \|x(t)\|_V^2 \,\mathrm{d}t.$$

The initial condition $Dx(t_0) = 0$ implies

$$\tfrac{1}{2}\|Dx(T)\|_H^2 + c\int_{t_0}^{T} \|x(t)\|_V^2 \,\mathrm{d}t \le 0,$$

and, consequently, $x(t) = 0$ for almost all $t \in (t_0, T)$.

Step 2: Existence proof via the Galerkin method.

(I) Solution of the Galerkin equations. The Galerkin equations (12.44)–(12.45) represent an initial value differential-algebraic equation with index 1 (see Proposition 12.13). Since $q \in L^2(\mathcal{I}, V^*)$, the right-hand side r of the Galerkin equations belongs to $L^2(\mathcal{I}, \mathbb{R}^n)$. Applying Proposition 2.83, the Galerkin equations have a unique solution in

$$H_D^1(\mathcal{I}, \mathbb{R}^n) = \{x \in L^2(\mathcal{I}, \mathbb{R}^n) : \ Dx \in H^1(\mathcal{I}, \mathbb{R}^{m_n})\}.$$

(II) A priori estimates for the Galerkin solution. Multiplying the Galerkin equations (12.42) by $c_{nj}(t)$ and summing over $j = 1, \dots, n$, we obtain

$$\big((Dx_n)'(t)|Dx_n(t)\big)_H + \big\langle B(t)x_n(t), x_n(t)\big\rangle_V = \big\langle q(t), x_n(t)\big\rangle_V.$$

Due to the product formula for real valued functions we get

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(Dx_n(t)|Dx_n(t)\big)_H = 2\,\big((Dx_n)'(t)|Dx_n(t)\big)_H.$$

This implies

$$\frac{\mathrm{d}}{\mathrm{d}t}\|Dx_n(t)\|_H^2 + 2\big\langle B(t)x_n(t), x_n(t)\big\rangle_V = 2\big\langle q(t), x_n(t)\big\rangle_V.$$

Integration over t yields

$$\begin{aligned}&\|Dx_n(T)\|_H^2 - \|Dx_n(t_0)\|_H^2 \\ &+ 2\int_{t_0}^{T} \big\langle B(t)x_n(t), x_n(t)\big\rangle_V \,\mathrm{d}t = 2\int_{t_0}^{T} \big\langle q(t), x_n(t)\big\rangle_V \,\mathrm{d}t.\end{aligned}$$

Since $B(t)$ is strongly monotone with a constant C_0 independent of t, we get

$$\|Dx_n(T)\|_H^2 + 2C_0 \int_{t_0}^{T} \|x_n(t)\|_V^2 \,\mathrm{d}t \;\le\; \|Dx_n(t_0)\|_H^2 + 2\int_{t_0}^{T} \langle q(t), x_n(t)\rangle_V \,\mathrm{d}t.$$

Using the classical inequality

$$2|xy| \le C_0^{-1}x^2 + C_0 y^2,$$

and the assumption that q belongs to $L_2(\mathcal{I}, V^*)$, we find

$$\|Dx_n(T)\|_H^2 + 2C_0 \int_{t_0}^{T} \|x_n(t)\|_V^2 \,\mathrm{d}t \le \|Dx_n(t_0)\|_H^2 + C_0^{-1}\int_{t_0}^{T} \|q\|_{V^*}^2 \,\mathrm{d}t + C_0 \int_{t_0}^{T} \|x_n(t)\|_V^2 \,\mathrm{d}t.$$

Consequently, there is a constant C such that

$$\int_{t_0}^{T} \|x_n(t)\|_V^2 \,\mathrm{d}t \le C\left(\|Dx_n(t_0)\|_H^2 + \int_{t_0}^{T} \|q\|_{V^*}^2 \,\mathrm{d}t\right). \tag{12.48}$$

(III) Weak convergence of the Galerkin method in $L_2(\mathcal{I}, V)$. Because of $Dx_n(t_0) = y_{n0} \to y_0$ in H as $n \to \infty$, the a priori estimate (12.48) yields the boundedness of the sequence (x_n) in the Hilbert space $L_2(\mathcal{I}, V)$. Therefore there is a weakly convergent subsequence $(x_{n'})$ with

$$x_{n'} \rightharpoonup x \quad \text{in } L_2(\mathcal{I}, V) \qquad \text{as} \quad n \to \infty. \tag{12.49}$$

The goal is now to show that x belongs to W and that x is a solution of the original equation (12.36)–(12.37). If this done, then we know because of uniqueness (see Step 1) that all weakly convergent subsequences $(x_{n'})$ have the same limit x and thus

$$x_n \rightharpoonup x \quad \text{in } L_2(\mathcal{I}, V) \qquad \text{as} \quad n \to \infty.$$

(III-1) We shall show the key equation

$$-\big(y_0|Dv\big)_H \varphi(t_0) - \int_{t_0}^{T} \big(Dx(t)|Dv\big)_H \varphi'(t)\,\mathrm{d}t + \int_{t_0}^{T} \big\langle B(t)x(t), v\big\rangle_V \varphi(t)\,\mathrm{d}t = \int_{t_0}^{T} \big\langle q(t), v\big\rangle_V \varphi(t)\,\mathrm{d}t \tag{12.50}$$

for all $v \in V$ and real functions

$$\varphi \in C^1(\mathcal{I}) \qquad \text{with} \quad \varphi(T) = 0. \tag{12.51}$$

Let φ be as in (12.51). We multiply the Galerkin equations (12.42) by φ and use the integration by parts formula (12.35) in order to get

$$\begin{aligned} &- \big(y_0|Dw_i\big)_H \varphi(t_0) - \int_{t_0}^{T} \big(Dx_n(t)|Dw_i\big)_H \varphi'(t)\,\mathrm{d}t \\ &\quad + \int_{t_0}^{T} \big\langle B(t)x_n(t), w_i\big\rangle_V \varphi(t)\,\mathrm{d}t = \int_{t_0}^{T} \big\langle q(t), w_i\big\rangle_V \varphi(t)\,\mathrm{d}t \end{aligned} \tag{12.52}$$

for all $i = 1,\dots,n$. In order to be able to apply the weak limit, we shall show that the integral terms on the left-hand side are linear continuous functionals on the space $L_2(\mathcal{I},V)$. Using the Hölder inequality and the continuity of D, we get

$$\begin{aligned} \left|\int_{t_0}^{T} \big(Dx_n(t)|Dw_i\big)_H \varphi'(t)\,\mathrm{d}t\right| &\le \int_{t_0}^{T} \|Dx_n(t)\|_H \|Dw_i\|_H |\varphi'(t)|\,\mathrm{d}t \\ &\le C_1 \|w_i\|_V \left(\int_{t_0}^{T} \|x_n(t)\|_V^2\,\mathrm{d}t\right)^{\frac{1}{2}} = C_1 \|w_i\|_V \|x_n\|_{L_2(\mathcal{I},V)} \end{aligned} \tag{12.53}$$

for all $i = 1,\dots,n$. Since $B(t)$ is bounded with a constant independent of t, we find

$$\begin{aligned} \left|\int_{t_0}^{T} \big\langle B(t)x_n(t), w_i\big\rangle_V \varphi(t)\,\mathrm{d}t\right| &\le C_2 \int_{t_0}^{T} \|x_n(t)\|_V \|w_i\|_V |\varphi(t)|\,\mathrm{d}t \\ &\le C_3 \|w_i\|_V \|x_n\|_{L_2(\mathcal{I},V)} \end{aligned} \tag{12.54}$$

for all $i = 1,\dots,n$. Applying now the weak limit (12.49) to equation (12.52), we obtain

$$\begin{aligned} &- \big(y_0|Dw_i\big)_H \varphi(t_0) - \int_{t_0}^{T} \big(Dx(t)|Dw_i\big)_H \varphi'(t)\,\mathrm{d}t \\ &\quad + \int_{t_0}^{T} \big\langle B(t)x(t), w_i\big\rangle_V \varphi(t)\,\mathrm{d}t = \int_{t_0}^{T} \big\langle q(t), w_i\big\rangle_V \varphi(t)\,\mathrm{d}t \end{aligned} \tag{12.55}$$

for all $i = 1,\dots,n$. By Assumption 12.10, there exists a sequence (v_n) with

$$v_n \to v \quad \text{in } V \qquad \text{as} \quad n \to \infty,$$

where each v_n is a finite linear combination of certain basis elements w_i. Regarding the continuity of D, the inequalities (12.53), (12.54), and $q \in L_2(\mathcal{I},V^*)$, we obtain that equation (12.55) is also satisfied if we replace w_i by v, which means the key equation (12.50) is satisfied.

(III-2) Proof that x belongs to $W^1_{2,D}(\mathcal{I},V,Y,H)$. Considering Assumption 12.14, the Galerkin equations (12.39) with basis elements with w_i for $i \in I_2$ imply that

$$\begin{aligned}\langle q(t) - B(t)x_n(t), w_i\rangle_V &= \langle A[Dx_n(t)]', w_i\rangle_V \\ &= \langle [Dx_n(t)]', Dw_i\rangle_Y \;=\; 0 \end{aligned} \tag{12.56}$$

for all $n = 1, 2, \ldots$ and almost all $t \in (t_0, T)$. Recall that $w_i \in \operatorname{im} Q = \ker D$ for $i \in I_2$. Due to (12.54), we may again apply the weak limit, which means

$$\int_{t_0}^{T} \langle q(t), w_i\rangle_V \varphi(t)\, \mathrm{d}t - \int_{t_0}^{T} \langle B(t)x(t), w_i\rangle_V \varphi(t)\, \mathrm{d}t = 0$$

for all $\varphi \in C_0^\infty(\mathcal{I})$. Applying the variational lemma, we get

$$\langle q(t), w_i\rangle_V - \langle B(t)x(t), w_i\rangle_V = 0$$

for almost all $t \in \mathcal{I}$. For any $v \in \operatorname{im} Q$, we find a sequence (v_n) with

$$v_n \to v \qquad \text{in} \quad V \qquad \text{as} \quad n \to \infty$$

where each v_n is a linear combination of the basis elements w_i with $i \in I_2$. Consequently,

$$\langle q(t), v\rangle_V - \langle B(t)x(t), v\rangle_V = 0 \tag{12.57}$$

for all $v \in \operatorname{im} Q$. This allows us to define a functional $\bar{y}(t) \in Y^*$ such that

$$\langle \bar{y}(t), y\rangle_Y = \langle q(t) - B(t)x(t), v\rangle_V \tag{12.58}$$

for almost all $t \in \mathcal{I}$ and for all $y \in Y$ with $y = Dv$ for some $v \in V$. Since D is surjective, the functional is defined for all $y \in Y$. Furthermore, $\bar{y}(t)$ is well defined since

$$\langle \bar{y}(t), Dv_1\rangle_Y = \langle \bar{y}(t), Dv_2\rangle_Y$$

for any v_1, $v_2 \in V$ with $Dv_1 = Dv_2$. This is a conclusion from the fact that $v_1 - v_2$ belongs to $\operatorname{im} Q$ and (12.57). We shall show that $\bar{y}$ belongs to $L_2(\mathcal{I}, Y^*)$. We have

$$\begin{aligned}\|\bar{y}(t)\|_{Y^*} = \sup_{\|z\|_Y \le 1} \left|\langle \bar{y}(t), y\rangle_Y\right| &= \sup_{\|Dv\|_Y \le 1,\ v \in \ker Q} \left|\langle q(t) - B(t)x(t), v\rangle_V\right| \\ &\le \sup_{\|Dv\|_Y \le 1,\ v \in \ker Q} \|q(t) - B(t)x(t)\|_{V^*} \|v\|_V.\end{aligned}$$

Note that D is bijective from $\ker Q$ to Y. Using the open mapping theorem and recalling that D is also linear and continuous, we find a constant $C \ge 0$ such that

$$\|v\|_V \le C\|Dv\|_Y \qquad \text{for all} \quad v \in \ker Q.$$

This implies

$$\|\bar{y}(t)\|_{Y^*} \le C\|q(t) - B(t)x(t)\|_{V^*} \tag{12.59}$$

for almost all $t \in \mathcal{I}$. Since $q \in L_2(\mathcal{I}, V^*)$, $x \in L_2(\mathcal{I}, V)$, and $B(t)$ is uniformly bounded, we obtain $\bar{z} \in L_2(\mathcal{I}, Y^*)$. Using the key equation (12.50)

and (12.58), we arrive at

$$-\int_{t_0}^{T} \big(Dx(t)|Dv\big)_H \varphi'(t) = \int_{t_0}^{T} \big\langle \bar{y}(t), Dv\big\rangle_V \varphi(t)\,\mathrm{d}t$$

for all $v \in V$ and $\varphi \in C_0^\infty(\mathcal{I})$. Since D is surjective, we have

$$-\int_{t_0}^{T} \langle Dx(t), y\rangle_Y \varphi'(t) = \int_{t_0}^{T} \langle \bar{y}(t), y\rangle_Y \varphi(t)\,\mathrm{d}t$$

for all $y \in Y$. This is equivalent to

$$\left\langle -\int_{t_0}^{T} Dx(t)\varphi'(t) - \int_{t_0}^{T} \bar{y}(t)\varphi(t)\,\mathrm{d}t, y \right\rangle_Y = 0 \qquad \forall y \in Y$$

since $\varphi' Dx$ and φz belong to $L_2(\mathcal{I}, Y^*)$ for all $\varphi \in C_0^\infty(\mathcal{I})$. But this means that

$$-\int_{t_0}^{T} Dx(t)\varphi'(t) = \int_{t_0}^{T} \bar{y}(t)\varphi(t)\,\mathrm{d}t,$$

and, finally, Dx has the generalized derivative $\bar{y} \in L_2(\mathcal{I}, Y^*)$. Hence, x belongs to $W^1_{2,D}(\mathcal{I}, V, Y, H)$.

(III-3) Proof that x fulfills (12.36). Since $\bar{y} = (Dx)'$, we have, by (12.58),

$$\langle (Dx)'(t), Dv\rangle_Y = \langle q(t) - B(t)x(t), v\rangle_V \tag{12.60}$$

for all $v \in V$ and almost all $t \in \mathcal{I}$. By Assumption 12.5, we get

$$\langle A(Dx)'(t), v\rangle_V = \langle q(t) - B(t)x(t), v\rangle_V.$$

But this means that (12.36) is satisfied.

(III-4) Proof that x fulfills (12.37). Since $x \in W^1_{2,D}(\mathcal{I}, V, Y, H)$, we can apply the integration by parts formula (12.35). This yields

$$(Dx(T), \varphi(T)Dv)_H - (Dx(t_0), \varphi(t_0)Dv)_H = \int_{t_0}^{T} \langle (Dx)'(t), \varphi(t)Dv\rangle_Y + (Dx(t)|\varphi'(t)Dv)_H\,\mathrm{d}t$$

for all $\varphi \in C^1\mathcal{I}$ and $v \in V$. In particular, if $\varphi(t_0) = 1$ and $\varphi(T) = 0$, then equation (12.50) along with (12.60) yields

$$(Dx(t_0) - y_0|Dv)_H = 0 \qquad \text{for all} \quad v \in V.$$

Since D is surjective and Y is dense in H, we get $Dx(t_0) = y_0$.

□

12.3.4 Continuous dependence on the data

Lemma 12.17. *[177] Let W, Y be Banach spaces. Let $Y \subseteq H \subseteq Y^*$ be an evolution triple. Let $D : W \to Y$ be linear, continuous and bijective. Define $G : W \to W^*$ as follows:*

$$\langle G\tilde{w}, w\rangle_W := (D\tilde{w}|Dw)_H \qquad \text{for all} \quad \tilde{w},\ w \in W.$$

Then G is linear and continuous and $cl(\operatorname{im} G) = W^*$.

Proof. The linearity of G is clear and

$$\langle G\tilde{w}, w\rangle_W \le \|D\tilde{w}\|_H \|Dw\|_H \le \tilde{c}\|\tilde{w}\|_W \|w\|_W$$

because the embedding $Y \subseteq H$ is continuous and D is continuous. We have to show that for all $\bar{w} \in W^*$ there is a sequence $(\bar{w}_n) \subseteq \operatorname{im} G$ such that

$$\|\bar{w} - \bar{w}_n\|_{W^*} \to 0 \qquad \text{as} \quad n \to \infty.$$

Since D is bijective the adjoint operator $D^* : Y^* \to W^*$ is linear, continuous and bijective. Let $\bar{w} \in W^*$ be arbitrary then there exists a $\bar{y} \in Y^*$ such that $D^*\bar{y} = \bar{w}$. Since $H^* \subseteq Y^*$ dense there is a sequence $(\bar{u}_n) \subseteq H^*$ such that

$$\forall \varepsilon > 0 \exists N \in \mathbb{N} \forall n \ge N : \|\bar{y} - \bar{u}_n\|_{Y^*} \le \frac{\varepsilon}{2}.$$

By the Riesz representation theorem there is a unique $u_n \in H$ such that $\langle \bar{u}_n, u\rangle_H = (u_n|u)_H$ for all $u \in H$. Since also $Y \subseteq H$ dense there exists $y_n \in Y$ such that

$$\|y_n - u_n\|_H \le \frac{\varepsilon}{2c}$$

where $c > 0$ is the constant from the continuous embedding $Y \subseteq H$. We define

$$\langle \bar{y}_n, y\rangle_Y := (y_n|y)_H \qquad \text{for all} \quad y \in Y.$$

Clearly $\bar{y}_n \in Y^*$ because the embedding $Y \subseteq H$ is continuous and we have for $y \in Y$

$$\begin{aligned}\langle \bar{y} - \bar{y}_n, y\rangle_Y &= \langle \bar{y} - \bar{u}_n, y\rangle_Y + \langle \bar{u}_n, y\rangle_Y - (y_n|y)_H \\ &= \langle \bar{y} - \bar{u}_n, y\rangle_Y + (u_n - y_n|y)_H \\ &\le \|\bar{y} - \bar{u}_n\|_{Y^*}\|y\|_Y + c\|u_n - y_n\|_H\|y\|_Y \le \varepsilon\|y\|_Y.\end{aligned}$$

We conclude that $\|\bar{y} - \bar{y}_n\|_{Y^*} \to 0$ as $n \to \infty$. Since D is bijective there is a unique $w_n \in W$ such that $Dw_n = y_n$. We define

$$\langle \bar{w}_n, w\rangle_W := \langle \bar{y}_n, Dw\rangle_Y \qquad \text{for all} \quad w \in W.$$

Clearly $\bar{w}_n \in W^*$ and also $\bar{w}_n \in \operatorname{im} G$ because

$$\langle \bar{w}_n, w\rangle_W = \langle \bar{y}_n, Dw\rangle_Y = (Dw_n|Dw)_H = \langle Gw_n, w\rangle_W.$$

Finally we see

$$\begin{aligned}\langle \bar{w} - \bar{w}_n, w\rangle_W &= \langle \bar{y}, Dw\rangle_Y - \langle \bar{y}_n, Dw\rangle_Y \\ &\le \tilde{c}\|\bar{y} - \bar{y}_n\|_{Y^*}\|w\|_W\end{aligned}$$

for a $\tilde{c} > 0$ because D is continuous. This completes the proof.

Theorem 12.18. *[177] Let the Assumptions 12.5, 12.10 and 12.14 be satisfied. Then, the ADAE (12.36)–(12.37) has at most index 1, which means that the map $G_1(t) : V \to V^*$ defined as*

$$\langle G_1(t)x, v\rangle_V := \langle A(Dx), v\rangle_V + \langle B(t)Qx, v\rangle_V \qquad \text{for all} \quad v \in V$$

is injective and densely solvable for all $x \in V$. The system has index 0 if and only if D is injective. Here, Dx is considered as the unique element of Y^* satisfying

$$\langle Dx, y\rangle_Y = (Dx|y)_H \qquad \forall y \in Y.$$

Proof. Step 1. Injectivity of $G_1(t)$. Let x belong to the nullspace of $G_1(t)$. This implies

$$\langle A(Dx), v\rangle_V + \langle B(t)Qx, v\rangle_V = 0 \qquad \text{for all} \quad v \in V.$$

Due to Assumption 12.5 we have

$$(Dx|Dv)_H + \langle B(t)Qx, v\rangle_V = 0 \qquad \text{for all} \quad v \in V. \tag{12.61}$$

In particular, for $v = Qx$, we get

$$\langle B(t)Qx, Qx\rangle_V = 0.$$

Since $B(t)$ is strictly monotone, we have $Qx = 0$. This yields $Dx = 0$ if we use $v := x$ in (12.61), i.e. $x \in \operatorname{im} Q$. Consequently, $x = Qx = 0$.

Step 2. Fix $t \in [t_0, T]$. We have to show that $\operatorname{cl}(\operatorname{im} G_1(t)) = V^*$, i.e. that for all $\bar{v} \in V^*$ there is a sequence $(\bar{v}_n) \subseteq \operatorname{im} G_1(t)$ such that

$$\|\bar{v} - \bar{v}_n\|_{V^*} \to 0 \qquad \text{as} \quad n \to \infty.$$

Let $\bar{v} \in V^*$. $U := \operatorname{im} Q$ is a closed subspace of V because Q is a projection operator. Hence U is a reflexive Banach space with the norm $\|\cdot\|_V$. We define the operator $\tilde{B}(t) : U \to U^*$ as follows

$$\langle \tilde{B}(t)\tilde{u}, u\rangle_U := \langle B(t)Q\tilde{u}, u\rangle_V \qquad \text{for all} \quad \tilde{u}, u \in U.$$

Surely $\tilde{B}(t)$ is well defined. It is coercive and bounded because $B(t)$ is and the norm on U is $\|\cdot\|_V$. Define furthermore

$$\langle \bar{u}, u\rangle_U := \langle \bar{v}, Qu\rangle_V$$

and clearly $\bar{u} \in U^*$ because $\bar{v} \in V^*$. With the Browder–Minty theorem, cf. [218], we can uniquely solve the equation

$$\langle \tilde{B}(t)\tilde{u}, u\rangle_U = \langle \bar{v}, Qu\rangle_V \tag{12.62}$$

on U with $\tilde{u} \in U$. Let $P := I - Q$ be the complementary projection operator of Q and define $W := \operatorname{im} P$. W is a Banach space with the norm $\|\cdot\|_V$. Furthermore we introduce

$$\langle \bar{w}, w\rangle_W := \langle \bar{v} - B(t)Q\tilde{u}, Pw\rangle_V \qquad \text{for all} \quad w \in W. \tag{12.63}$$

Clearly $\bar{w} \in W^*$ because $\bar{v} \in V^*$ and $B(t)$ is bounded. The map

$$D|_W : W \to Y$$

is bijective because D is surjective and $\ker D = \operatorname{im} Q = \ker P$. Setting $G : W \to W^*$ as

$$\langle G\tilde{w}, w\rangle_W := (D|_W \tilde{w} | D|_W w)_H = (D\tilde{w}|Dw)_H \qquad \text{for all} \quad \tilde{w}, w \in W$$

we can apply Lemma 12.17. So there is a sequence $(\bar{w}_n) \subseteq \operatorname{im} G$ such that

$$\|\bar{w} - \bar{w}_n\|_{W^*} \to 0 \qquad \text{as} \quad n \to \infty.$$

$\bar{w}_n \in \operatorname{im} G$ means that there exists a $\tilde{w}_n \in W$ such that

$$(D\tilde{w}_n|Dw)_H = \langle G\tilde{w}_n, w\rangle_W = \langle \bar{w}_n, w\rangle_W \qquad \text{for all} \quad w \in W. \tag{12.64}$$

We define now $\tilde{v}_n := P\tilde{w}_n + Q\tilde{u} \in V$ and set

$$\langle \bar{v}_n, v\rangle_V := \langle G_1(t)\tilde{v}_n, v\rangle_V \qquad \text{for all} \quad v \in V.$$

By definition $\bar{v}_n \in \operatorname{im} G_1(t)$ and we observe for all $v \in V$:

$$\begin{aligned}
\langle \bar{v} - \bar{v}_n, v\rangle_V &= \langle \bar{v}, v\rangle_V - \langle G_1(t)\tilde{v}_n, v\rangle_V \\
&= \langle \bar{v}, v\rangle_V - (D\tilde{w}_n|Dv)_H - \langle B(t)Q\tilde{u}, v\rangle_V \\
&= \langle \bar{v}, Pv\rangle_V - \langle B(t)Q\tilde{u}, Pv\rangle_V - (D\tilde{w}_n|Dv)_H + \langle \bar{v} - \langle B(t)Q\tilde{u}, Qv\rangle_V \\
&= \langle \bar{w} - \bar{w}_n, Pv\rangle_W \\
&\le \tilde{c}\|\bar{w} - \bar{w}_n\|_{W^*}\|v\|_V.
\end{aligned}$$

Here we used (12.62), (12.63) and (12.64). So we obtain $\|\bar{v} - \bar{v}_n\|_{V^*} \to 0$ as $n \to \infty$.

Step 3. If D is injective, then Q is the zero mapping and $G_0(t) = G$ from Lemma 12.17. Hence $G_0(t)$ is then densely solvable and $G_0(t)$ is injective because of step 1. Conversely, injectivity of $G_0(t)$ implies injectivity of D. Consequently, (12.36)–(12.37) has index 0 if and only if D is injective. □

From the theory of DAEs we know that index-1 systems have solutions which depend continuously on the data. The following theorem shows that this is also the case for the solution of the index-1 ADAE (12.36)–(12.37).

Theorem 12.19. *Let the Assumptions 12.5, 12.10 and 12.14 be satisfied. Furthermore, let $x \in W^1_{2,D}(\mathcal{I},V,Y,H)$ be the unique solution of the ADAE (12.36)–(12.37). Then, the map*

$$(y_0, q) \mapsto x$$

is linear and continuous from $Y \times L_2(\mathcal{I},V^*)$ to $W^1_{2,D}(\mathcal{I},V,Y,H)$, i.e., there is a constant $C > 0$ such that

$$\|x\|_{W^1_{2,D}} \le C(\|y_0\|_H + \|q\|_{L_2(\mathcal{I},V^*)}),$$

for all $y_0 \in H$ and $q \in L_2(\mathcal{I},V^*)$.

Proof. In the proof of Theorem 12.16 we see that

$$x_n \rightharpoonup x \qquad \text{in} \quad L_2(\mathcal{I},V) \qquad \text{as} \quad n \to \infty.$$

From the Banach–Steinhaus theorem it follows that

$$\|x\|_{L_2(\mathcal{I},V)} \le \varliminf_{n\to\infty} \|x_n\|_{L_2(\mathcal{I},V)}.$$

Using the a priori estimate (12.48), the continuity of D and Assumption 12.10, we find a constant $C_1 \ge 0$ such that

$$\|x\|_{L_2(\mathcal{I},V)} \le C_1 \left(\|y_0\|_H + \|q\|_{L_2(\mathcal{I},V_*)}\right). \tag{12.65}$$

Using inequality (12.59), we obtain

$$\|(Dx)'\|_{L_2(\mathcal{I},Y^*)} \le C_2 \left(\|x\|_{L_2(\mathcal{I},V)} + \|q\|_{L_2(\mathcal{I},V^*)}\right)$$

if we recall that $\bar{y} = (Dx)'$ in the proof of Theorem 12.16. Together with (12.65), this implies the assertion with the constant $C = C_1C_2 + C_1 + C_2 > 0$. □

12.3.5 Strong convergence of the Galerkin method

Theorem 12.20. *Let the Assumptions 12.5, 12.10 and 12.14 be satisfied. Then, for all $n = 1,2,\ldots$, the Galerkin equations (12.44)–(12.45) have exactly one solution*

$$x_n \in W^1_{2,D}(\mathcal{I},V,Y,H).$$

The sequence (x_n) converges as $n \to \infty$ to the solution x of (12.36)–(12.37) in the following sense:

$$x_n \to x \quad \text{in } L_2(\mathcal{I},V) \qquad \text{and} \qquad \max_{t_0 \le t \le T} \|Dx_n(t) - Dx(t)\|_H \to 0.$$

Proof. In the proof of Theorem 12.16, the Galerkin equations (12.44)–(12.45) are shown to have a unique solution $x_n \in W^1_{2,D}(\mathcal{I},V,Y,H)$. Furthermore, it was shown that the unique solution x of (12.36)–(12.37) belongs to $W^1_{2,D}(\mathcal{I},V,Y,H)$.

Step 1. We shall show that $\max_{t_0\le t\le T} \|Dx_n(t)-Dx(t)\|_H \to 0$. We refer here to [175] where this is proven in a more general setting.

Step 2. Strong convergence of (x_n) in $L_2(\mathcal{I},V)$. From Theorem 12.16 we know that

$$x_n \rightharpoonup x \qquad \text{in} \quad L_2(\mathcal{I},V) \qquad \text{as} \quad n\to\infty.$$

Since B is linear and uniformly continuous, we get

$$Bx_n \rightharpoonup Bx \qquad \text{in} \quad L_2(\mathcal{I},V^*) \qquad \text{as} \quad n\to\infty$$

and, hence,

$$\int_{t_0}^{T} \langle Bx_n,x\rangle_V \to \int_{t_0}^{T} \langle Bx,x\rangle_V \qquad \text{as} \quad n\to\infty. \tag{12.66}$$

Furthermore, the assumption $q\in L_2(\mathcal{I},V^*)$ yields

$$\int_{t_0}^{T} \langle q,x_n\rangle_V \to \int_{t_0}^{T} \langle q,x\rangle_V \qquad \text{as} \quad n\to\infty. \tag{12.67}$$

From the integration by parts formula (12.35), we have

$$\tfrac{1}{2}\|(Dx(T)-Dx_n(T)\|_H^2 - \tfrac{1}{2}\|(Dx(t_0)-Dx_n(t_0)\|_H^2 = \int_{t_0}^{T} \langle (Dx)'(t)-(Dx_n)'(t), Dx(t)-Dx_n(t)\rangle_Y \,\mathrm{d}t$$

and

$$\big(Dx_n(T)|Dx(T)\big)_H - \big(Dx_n(t_0)|Dx(t_0)\big)_H = \int_{t_0}^{T} \langle (Dx_n)'(t),Dx(t)\rangle_Y + \langle (Dx)'(t),Dx_n(t)\rangle_Y \,\mathrm{d}t.$$

Using again (12.58) for the generalized derivative $(Dx)'=\bar{y}$, we obtain

$$\tfrac{1}{2}\|(Dx(T)-Dx_n(T)\|_H^2 - \tfrac{1}{2}\|(Dx(t_0)-Dx_n(t_0)\|_H^2 = \int_{t_0}^{T} \langle q(t)-B(t)x(t),x(t)-x_n(t)\rangle_V - \langle (Dx_n)'(t),Dx(t)-Dx_n(t)\rangle_Y \,\mathrm{d}t \tag{12.68}$$

as well as

$$\big(Dx_n(T)|Dx(T)\big)_H - \big(Dx_n(t_0)|Dx(t_0)\big)_H = \int_{t_0}^{T} \langle (Dx_n)'(t),Dx(t)\rangle_Y + \langle (q(t)-B(t)x(t),x_n(t)\rangle_V \,\mathrm{d}t. \tag{12.69}$$

From the Galerkin equations (12.41) we get

$$\langle (Dx_n)'(t), Dx_n(t)\rangle_Y + \langle B(t)x_n(t), x_n(t)\rangle_V = \langle q(t), x_n(t)\rangle_V. \tag{12.70}$$

The strong monotonicity of $B(t)$ implies

$$C\,\|x - x_n\|^2_{L_2(\mathcal{I},V)} \le \int_{t_0}^{T} \langle B(t)x(t) - B(t)x_n(t), x(t) - x_n(t)\rangle_V \mathrm{d}t,$$

where C is a positive constant. Applying (12.68), we obtain

$$\begin{aligned} C\,\|x - x_n\|^2_{L_2(\mathcal{I},V)} &\le \int_{t_0}^{T} \langle q(t) - B(t)x_n(t), x(t) - x_n(t)\rangle_V \mathrm{d}t \\ &- \int_{t_0}^{T} \langle (Dx_n)'(t), Dx(t) - Dx_n(t)\rangle_Y \mathrm{d}t + \tfrac{1}{2}\|Dx(t_0) - Dx_n(t_0)\|^2_H. \end{aligned}$$

Regarding (12.70), this yields

$$\begin{aligned} C\,\|x - x_n\|^2_{L_2(\mathcal{I},V)} &\le \int_{t_0}^{T} \langle q(t) - B(t)x_n(t), x(t) - x_n(t)\rangle_V \mathrm{d}t \\ &- \int_{t_0}^{T} \langle (Dx_n)'(t), Dx(t)\rangle_Y \mathrm{d}t + \tfrac{1}{2}\|Dx(t_0) - Dx_n(t_0)\|^2_H. \end{aligned}$$

Using (12.69), we get

$$\begin{aligned} C\,\|x - x_n\|^2_{L_2(\mathcal{I},V)} &\le \int_{t_0}^{T} \langle q(t) - B(t)x_n(t), x(t)\rangle_V \mathrm{d}t \\ &+ \int_{t_0}^{T} \langle q(t) - B(t)x(t), x_n(t)\rangle_V \mathrm{d}t - (Dx_n(T), Dx(T))_H \\ &+ (Dx_n(t_0), Dx(t_0))_H + \tfrac{1}{2}\|Dx(t_0) - Dx_n(t_0)\|^2_H. \end{aligned} \tag{12.71}$$

If we apply (12.66) and (12.67), we see that the right-hand side of inequality (12.71) converges to

$$2\int_{t_0}^{T} \langle q(t) - B(t)x(t), x(t)\rangle_V \mathrm{d}t - (Dx(T), Dx(T))_H + (Dx(t_0), Dx(t_0))_H \tag{12.72}$$

as $n \to \infty$. Note that we have already proved in step 1 that

$$\|Dx(t_0) - Dx_n(t_0)\|_H \to 0 \qquad \text{as} \quad n \to \infty.$$

Applying once more the integration by parts formula (12.35) and regarding (12.58) for the generalized derivative $(Dx)' = \bar{y}$, we get

$$(Dx(T),Dx(T))_H-(Dx(t_0),Dx(t_0))_H \;=\; 2\int_{t_0}^{T}\langle (Dx)'(t),Dx(t)\rangle_Y \mathrm{d}t \;=\; 2\int_{t_0}^{T}\langle q(t)-B(t)x(t),x(t)\rangle_V \mathrm{d}t. \tag{12.73}$$

Summarizing (12.71)–(12.73), we obtain

$$\|x-x_n\|^2_{L_2(\mathcal{I},V)}\to 0 \qquad \text{as} \quad n\to\infty$$

which implies the assertion. □

12.4 Notes and references

(1) In [176], the results for linear abstract differential systems

$$A(Dx)'(t)+B(t)x(t)=q(t)$$

satisfying Assumption 12.5 presented in Section 12.3 have been extended to abstract differential-algebraic equations with

$$A=(TD)^*$$

with $T:Y\to Y$ being linear, continuous and bijective. The unique extension operator $T_H:H\to H$, $T_H|_V=T$ is self-adjoint on H and there is a linear, continuous operator $S:H\to H$ with $T_H=S^*S$.

(2) Approaches to extend the perturbation index concept to partial differential algebraic equations have been presented in [47], [149] and [190]. The various examples discussed there show that one has to characterize the solution behavior not only with respect to time-varying perturbations but also with respect to space-varying perturbations.

(3) In [74], systems of the form (12.1) with linear operators have been studied. More precisely, initial value systems of the form

$$A\frac{\mathrm{d}}{\mathrm{d}t}Dx+Bx=Ag(t) \qquad \text{for } t\in(0,T], \tag{12.74}$$

$$Dx(0)=v_0\in \operatorname{im} D, \tag{12.75}$$

with $X=Y=Z$ and $A=\mathcal{I}$ are treated by a semigroup approach. $\mathcal{I}$ denotes the identity operator. The system (12.74)–(12.75) is reduced to multivalued differential equations of the form

$$\frac{\mathrm{d}v}{\mathrm{d}t}\in Av+f(t) \qquad \text{for } t\in(0,T], \tag{12.76}$$

$$v(0)=v_0, \tag{12.77}$$

where $A := -BD^{-1}$. Here, the operator D^{-1} is defined as the (multivalued) function satisfying

$$D^{-1}v = \{x \in \mathcal{G}_D : \ Dx = v\} \qquad \text{for all } v \in \operatorname{im} D$$

with $\mathcal{G}_D$ being the definition domain of D. The existence and uniqueness of classical solutions of (12.74)–(12.75) satisfying

$$Dx \in C^1([0,T];X) \quad \text{and} \quad Bx \in C([0,T];X)$$

is shown provided that $g \in C^1([0,T];X)$, $Bx(0) \in \operatorname{im} A$,

$$\operatorname{Re}(-Bx|Dx)_X \le \beta \|Dx\|_X^2, \quad \text{for all } x \in \mathcal{G}_B \subseteq \mathcal{G}_D \tag{12.78}$$

as well as the operator

$$\lambda_0 AD + B : \ \mathcal{G}_B \to X \quad \text{is bijective for some } \lambda_0 > \beta. \tag{12.79}$$

A similar result is obtained for systems of the form (12.74)–(12.75) with $A = D^*$ or $D = I$.

This approach via multivalued differential equations concentrates on the dynamic part of the system. It is limited to systems with special constraints. So, for instance, condition (12.78) implies

$$\operatorname{Re}(Bw, Dv) = 0 \qquad \text{for all } w \in \ker D \cap \mathcal{G}_B, \ v \in \mathcal{G}_D.$$

Together with condition (12.79), we obtain, for the finite-dimensional case, a DAE of index 1 with the constraint

$$(BQ)^* Bx = (BQ)^* g(t)$$

where Q denotes a projector onto $\ker D$. This is obvious since $(BQ)^* D = 0$ if we use the Euclidean norm.

(4) The semigroup approach has been extended to linear systems with time-dependent operators in [74] supposing that $A = I$ or $D = I$. There, it is assumed that the operator $D(t)(\lambda D(t) + B(t))^{-1}$ or $(\lambda D(t) + B(t))^{-1} D(t)$, respectively, is bounded in a certain way for λ from a specific region in $\mathbb{C}$. However, it is usually quite difficult to verify this condition for coupled systems in practice. Additionally, having DAEs in mind, we think that we should not concentrate on the pencil operators $\lambda D(t) + B(t)$ in the nonstationary case. Even in the finite-dimensional case, it is not reasonable to demand nonsingularity of $\lambda D(t) + B(t)$ for existence and uniqueness of solutions (see, e.g., [25, 96], also Example 2.4). Regarding the trivial example

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

we see that $\lambda D(t)+B(t)$ is nonsingular for all λ but a solution exists only if $f_1(t)+t\,f_2(t)=0$. On the other hand, considering the example

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

the matrix pencil $\lambda D(t)+B(t)$ is singular of all λ but there is a unique solution satisfying

$$x_1(t) = -t, \qquad x_2(t) = 1.$$

Hence, investigating time-dependent differential-algebraic equations, one has turned away from matrix pencils.

(5) Using an operational method developed in [184], more general linear systems of the form (12.74)–(12.75) with either $A=I$ or $D=I$ are treated in [74]. But it is restricted to systems with constant injective operators B having a bounded inverse.

(6) Nonlinear abstract systems of the form

$$\frac{\mathrm{d}}{\mathrm{d}t}(Dx)+Bx = F(t,Kx), \quad t\in[0,T]$$

where D, B, K are linear closed operators from a complex Banach space X into a Banach space Y, have been investigated in [71, 72, 13, 73]. The theory developed there is based mainly on properties of the operator

$$T := D(\lambda D+B)^{-1},$$

λ being a regular point of the operator pencil $\lambda D+B$. Most results are presented for problems with the resolvent operator $(T-\xi I)^{-1}$ having a simple pole at $\xi=0$. In the finite-dimensional case, such problems are DAEs of index 1. In [73] also problems are investigated where $(T-\xi I)^{-1}$ has a pole of multiple order at $\xi=0$. Considering again the finite-dimensional case, these are problems of higher index. Existence and uniqueness of solutions of such problems are obtained by a study of the transformed problem

$$\frac{\mathrm{d}}{\mathrm{d}t}(Tv)+v = f(t,Nv), \quad t\in[0,T]$$

with $N=\mathcal{K}(\lambda D+B)^{-1}$, $f(t,w)=e^{-\lambda t}F(t,e^{\lambda t}w)$ and

$$v(t) = e^{-\lambda t}(\lambda D+B)x(t).$$

Besides certain smoothness conditions and consistent initial conditions, the nonlinear function f has to fulfill a structural condition of the form

$$\pi_k f(t,Nv) = \pi_k f(t,N\sum_{j=k}^{m-1}\Pi_j v), \quad k=m-1,\dots,1$$

for certain projectors Π_j satisfying $\Pi_k v = \Pi_k(Pv)\varphi_k$ if m is the order of the pole of $(T-\xi I)^{-1}$ in $\xi = 0$ and $\ker T^m$ is spanned by $\{\varphi_k = T^{m-1-k}\varphi_n\}_{k=0}^{m-1}$. Although the assumptions are shown to be satisfied for a sample circuit with an LI-cutset in [73], we do not know the network topological conditions for general networks that guarantee all assumptions. In particular, the determination of the order of the pole of the operator $(T-\xi I)^{-1}$ in $\xi = 0$ often becomes a problem for coupled systems.

Nevertheless, the order of the pole plays a significant role for the characterization of the systems. Indeed, in the linear, finite-dimensional case, the pole order equals the index of the DAE.

Appendix A
Linear algebra – basics

In this appendix we collect and complete well-known facts concerning projectors and subspaces of $\mathbb{R}^m$ (Section A.1), and generalized inverses (Section A.2). Section A.3 provides material on matrix and projector valued functions with proofs, since these proofs are not easily available. In Section A.4 we introduce $\mathcal{C}^k$-subspaces of $\mathbb{R}^m$ via $\mathcal{C}^k$-projector functions. We show $\mathcal{C}^k$-subspaces to be those which have local $\mathcal{C}^k$ bases.

A.1 Projectors and subspaces

We collect some basic and useful properties of projectors and subspaces.

Definition A.1. (1) A linear mapping $Q \in L(\mathbb{R}^m)$ is called a projector if $Q^2 = Q$.
(2) A projector $Q \in L(\mathbb{R}^m)$ is called a projector onto $S \subseteq \mathbb{R}^m$ if $\operatorname{im} Q = S$.
(3) A projector $Q \in L(\mathbb{R}^m)$ is called a projector along $S \subseteq \mathbb{R}^m$ if $\ker Q = S$.
(4) A projector $Q \in L(\mathbb{R}^m)$ is called an orthogonal projector if $Q = Q^*$.

Example A.2. The m-dimensional matrix $Q = \begin{bmatrix} 1 & 0 & \dots & 0 \\ * & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ * & 0 & \dots & 0 \end{bmatrix}$ with arbitrary entries for $*$ becomes a projector onto the one-dimensional subspace spanned by the first column of Q along the $(m-1)$-dimensional subspace $\left\{ v : v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{bmatrix}, v_1 = 0 \right\}$. □

Lemma A.3. *Let P and $\bar{P}$ be projectors, and $Q := I - P$, $\bar{Q} := I - \bar{P}$ the complementary projectors. Then the following properties hold:*

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5,

(1) $z \in \operatorname{im} Q \Leftrightarrow z = Qz$.
(2) *If Q and $\bar{Q}$ project onto the same subspace S, then $\bar{Q} = Q\bar{Q}$ and $Q = \bar{Q}Q$ are valid.*
(3) *If P and $\bar{P}$ project along the same subspace S, then $\bar{P} = \bar{P}P$ and $P = P\bar{P}$ are true.*
(4) *Q projects onto S iff $P := I - Q$ projects along S.*
(5) *Each matrix of the form $I + PZQ$, with arbitrary matrix Z, is nonsingular and its inverse is $I - PZQ$.*
(6) *Each projector P is diagonalizable. Its eigenvalues are 0 and 1. The multiplicity of the eigenvalue 1 is $r = \operatorname{rank} P$.*

Proof. (1) $z = Qy \quad \rightarrow \quad Qz = Q^2 y = Qy = z$.
(2) $\bar{Q}z \in \operatorname{im}\bar{Q} = S = \operatorname{im} Q$, also $\bar{Q}z = Q\bar{Q}z \; \forall z$.
(3) $\bar{P}P = (I - \bar{Q})(I - Q) = I - \bar{Q} - Q + \bar{Q}Q = I - \bar{Q} = \bar{P}$.
(4) $P^2 = P \Leftrightarrow (I - Q)^2 = I - Q \Leftrightarrow -Q + Q^2 = 0 \Leftrightarrow Q^2 = Q$ and $z \in \ker P \Leftrightarrow Pz = 0 \Leftrightarrow z = Qz \Leftrightarrow z \in \operatorname{im} Q$.
(5) Multiplying $(I + PZQ)z = 0$ by $Q \Rightarrow Qz = 0$. Now with $(I + PZQ)z = 0$ follows $z = 0$.
$(I + PZQ)(I - PZQ) = I - PZQ + PZQ = I$.
(6) Let $\bar{P}_1$ be a matrix of the r linearly independent columns of P and $\bar{Q}_2$ a matrix of the $m - r$ linearly independent columns of $I - P$. Then by construction $P\left[\bar{P}_1 \; \bar{Q}_2\right] = \left[\bar{P}_1 \; \bar{Q}_2\right] \begin{bmatrix} I & \\ & 0 \end{bmatrix}$. Because of the nonsingularity of $\left[\bar{P}_1 \; \bar{Q}_2\right]$ we have the structure $P = \left[\bar{P}_1 \; \bar{Q}_2\right] \begin{bmatrix} I & \\ & 0 \end{bmatrix} \left[\bar{P}_1 \; \bar{Q}_2\right]^{-1}$. The columns of $\bar{P}_1$, respectively $\bar{Q}_2$ are the eigenvectors to the eigenvalues 1, respectively 0. □

Lemma A.4. *Let $A \in L(\mathbb{R}^n, \mathbb{R}^k)$, $D \in L(\mathbb{R}^m, \mathbb{R}^n)$ be given, and $r := \operatorname{rank}(AD)$. Then the following two implications are valid:*

(1) $\ker A \cap \operatorname{im} D = 0$, $\operatorname{im}(AD) = \operatorname{im} A \Rightarrow \ker A \oplus \operatorname{im} D = \mathbb{R}^n$.
(2) $\ker A \oplus \operatorname{im} D = \mathbb{R}^n \Rightarrow$

- $\ker A \cap \operatorname{im} D = \{0\}$,
- $\operatorname{im} AD = \operatorname{im} A$,
- $\ker AD = \ker D$,
- $\operatorname{rank} A = \operatorname{rank} D = r$.

Proof. (1) Because of $\operatorname{im}(AD) = \operatorname{im} A$, the matrix A has $\operatorname{rank} r$ and $\ker A$ has dimension $n - r$. Moreover, $\operatorname{rank} D \geq r$ must be true. The direct sum $\ker A \oplus \operatorname{im} D$ is well-defined, and it has dimension $n - r + \operatorname{rank} D \leq n$. This means that D has $\operatorname{rank} r$. We are done with (1).

(2) The first relation is an inherent property of the direct sum. Let $R \in L(\mathbb{R}^n)$ denote the projector onto $\operatorname{im} D$ along $\ker A$. By means of suitable generalized inverses D^- and A^- of D and A we may write (Appendix A.2) $R = A^- A = DD^-$, $D = RD$, $A = AR$. This leads to

$$\begin{aligned}\operatorname{im} AD &\subseteq \operatorname{im} A = \operatorname{im} ADD^- \subseteq \operatorname{im} AD,\\ \ker AD &\subseteq \ker A^- AD = \ker D \subseteq \ker AD.\end{aligned}$$

The remaining rank property now follows from (1). □

Lemma A.5. *[94, Ch. 12.4.2]*
Given matrices G, Π, $\mathcal{N}$, $\mathcal{W}$ of suitable sizes such that

$$\begin{aligned}\ker G &= \operatorname{im} \mathcal{N},\\ \ker \Pi\mathcal{N} &= \operatorname{im} \mathcal{W},\end{aligned}$$

then it holds that

$$\ker G \cap \ker \Pi = \ker \mathcal{N}\mathcal{W}.$$

Proof. For $x \in \ker G \cap \ker \Pi$ we find $x = \mathcal{N}y, \Pi x = 0$, further $\Pi\mathcal{N}y = 0$, and hence $y = \mathcal{W}z$, $x = \mathcal{N}\mathcal{W}z \in \operatorname{im}\mathcal{N}\mathcal{W}$.
Conversely, each $x = \mathcal{N}\mathcal{W}z$ obviously belongs to $\ker G$, and $\Pi x = \Pi\mathcal{N}\mathcal{W}z = 0$. □

Lemma A.6. *$N, M \subseteq \mathbb{R}^m$ subspaces $\Rightarrow (N+M)^\perp = N^\perp \cap M^\perp$.*

Proof.

$$\begin{aligned}(N+M)^\perp &= \{z \in \mathbb{R}^m : \forall w \in N+M : \langle z, w\rangle = 0\}\\ &= \{z \in \mathbb{R}^m : \forall w_N \in N, \forall w_M \in M : \langle z, w_N + w_M\rangle = 0\}\\ &= \{z \in \mathbb{R}^m : \forall w_N \in N, \forall w_M \in M : \langle z, w_N\rangle = 0,\ \langle z, w_M\rangle = 0\}\\ &= N^\perp \cap M^\perp.\end{aligned}$$

□

Lemma A.7. (1) *Given two subspaces $N, X \subseteq \mathbb{R}^m$, $N \cap X = \{0\}$, then $\dim N + \dim X \leq m$, and there is a projector $Q \in L(\mathbb{R}^m)$ such that $\operatorname{im} Q = N$, $\ker Q \supseteq X$.*

(2) *Given two subspaces $S, N \subseteq \mathbb{R}^m$. If the decomposition*

$$\mathbb{R}^m = S \oplus N$$

holds true, i.e., S and N are transversal , then there is a uniquely determined projector $P \in L(\mathbb{R}^m)$ such that $\operatorname{im} P = S$, $\ker P = N$.

(3) *An orthoprojector P projects onto $S := \operatorname{im} P$ along $S^\perp = \ker P$.*

(4) *Given the subspaces $K, N \subseteq \mathbb{R}^m$, $\widehat{N} := N \cap K$, if a further subspace $X \subseteq \mathbb{R}^m$ is a complement of $\widehat{N}$ in K, which means $K = \widehat{N} \oplus X$, then there is a projector $Q \in L(\mathbb{R}^m)$ onto N such that*

$$X \subseteq \ker Q. \tag{A.1}$$

Let d_K, d_N, u denote the dimensions of the subspaces $K, N, \widehat{N}$, respectively. Then

$$d_K + d_N \leq m + u \tag{A.2}$$

holds.

(5) *If the subspace K in (4) is the nullspace of a certain projector $\Pi \in L(\mathbb{R}^m)$, that is $K = \ker \Pi = \operatorname{im}(I - \Pi)$, then*

$$\Pi Q(I - \Pi) = 0 \tag{A.3}$$

becomes true.

(6) *Given the two projectors $\Pi, Q \in L(\mathbb{R}^m)$, further $P := I - Q$, $N := \operatorname{im} Q$, $K := \ker \Pi$, then, supposing (A.3) is valid, the products ΠP, ΠQ, $P\Pi P$, $P(I - \Pi)$, $Q(I - \Pi)$ are projectors, too. The relation*

$$\ker \Pi P = \ker P\Pi P = N + K \tag{A.4}$$

holds true, and the subspace $X := \operatorname{im} P(I - \Pi)$ is the complement of $\widehat{N} := N \cap K$ in K, such that $K = \widehat{N} \oplus X$.
Moreover, the decomposition

$$\mathbb{R}^m = (N + K) \oplus \operatorname{im} P\Pi P = N \oplus \underbrace{X \oplus \operatorname{im} P\Pi P}_{\operatorname{im} P}$$

is valid.

(7) *If the projectors Π, Q in (6) are such that $\Pi^* = \Pi$, $(\Pi P)^* = \Pi P$, $(P(I - \Pi))^* = P(I - \Pi)$ and $Q\Pi P = 0$, then it follows that*

$$X = K \cap \widehat{N}^{\perp},\ \operatorname{im} P = X \oplus (N + K)^{\perp}.$$

Proof. (1) Let $x_1, \ldots, x_r \in \mathbb{R}^m$ and $n_1, \ldots, n_t \in \mathbb{R}^m$ be bases of X and N. Because of $X \cap N = \{0\}$ the matrix

$$F := [x_1 \ldots x_r n_1 \ldots n_t]$$

has full column rank and $r + t = \dim X + \dim N \leq m$. The matrix F^*F is invertible, and

$$Q := F \begin{bmatrix} 0 & \\ & I \end{bmatrix}_{\substack{r \\ t}} (F^*F)^{-1} F^*$$

is a projector we looked for. Namely,

$$Q^2 = F \begin{bmatrix} 0 & \\ & I \end{bmatrix} (F^*F)^{-1} F^* F \begin{bmatrix} 0 & \\ & I \end{bmatrix} (F^*F)^{-1} F^* = Q, \quad \operatorname{im} Q = \operatorname{im} F \begin{bmatrix} 0 & \\ & I \end{bmatrix} = N,$$

and $z \in X$ implies that it has to have the structure $z = F \begin{bmatrix} \alpha \\ 0 \end{bmatrix} \begin{matrix} \}r \\ \}t \end{matrix}$, which leads to $Qz = 0$.

(2) For transversal subspaces S and N we apply Assertion (1) with $t = m - r$, i.e., F is square. We have to show that P is unique. Supposing that there are two projectors $P, \bar{P}$ such that $\ker P = \ker \bar{P} = N$, $\operatorname{im} P = \operatorname{im} \bar{P} = S$, we immediately have $P = (\bar{P} + \bar{Q})P = \bar{P}P + \bar{Q}P = \bar{P}P = \bar{P}$.

(3) Let $S := \operatorname{im} P$ and $N := \ker P$. We choose a $v \in N$ and $y \in S$. Lemma A.3 (1) implies $y = Py$, therefore $\langle v, y \rangle = \langle v, Py \rangle = \langle P^* v, y \rangle$. With the symmetry of P we obtain $\langle P^* v, y \rangle = \langle Pv, y \rangle = 0$, i.e., $N = S^\perp$.

(4) X has dimension $d_K - u$. Since the sum space $K + N = X \oplus N \subseteq \mathbb{R}^m$ may have at most dimension m, it results that $\dim(K+N) = \dim X + \dim N = d_K - u + d_N \leq m$, and assertion (1) provides Q.

(5) Take an arbitrary $z \in \operatorname{im}(I - \Pi) = K$ and decompose $z = z_{\widehat{N}} + z_X$. It follows that $\Pi Q z = \Pi Q z_{\widehat{N}} + \Pi \underbrace{Q z_X}_{=0} = \Pi z_{\widehat{N}} = 0$, and hence (A.3) is true.

(6) (A.3) means $\Pi Q = \Pi Q \Pi$ and hence

$$
\begin{aligned}
\Pi Q \Pi Q &= \Pi Q Q = \Pi Q, \\
\Pi P \Pi P &= \Pi (I - Q) \Pi P = \Pi P - \Pi \underbrace{Q \Pi P}_{=0} = \Pi P, \\
(P \Pi P)^2 &= P \Pi P \Pi P = P \Pi P, \\
(P(I - \Pi))^2 &= P(I - \Pi)(I - Q)(I - \Pi) = P(I - \Pi) - P(I - \Pi) Q (I - \Pi) \\
&= P(I - \Pi) + P \underbrace{\Pi Q (I - \Pi)}_{=0}, \\
(Q(I - \Pi))^2 &= Q(I - \Pi) - Q \Pi Q (I - \Pi) = Q(I - \Pi).
\end{aligned}
$$

The representation $I - \Pi = Q(I - \Pi) + P(I - \Pi)$ corresponds to the decomposition $K = \widehat{N} \oplus X$.

Next we verify (A.4). The inclusion $\ker \Pi P \subseteq \ker P \Pi P$ is trivial. On the other hand, $P \Pi P z = 0$ implies $\Pi P \Pi P z = 0$ and hence $\Pi P z = 0$, and it follows that $\ker \Pi P = \ker P \Pi P$. Now it is evident that $K + N \subseteq \ker \Pi P$. Finally, $\Pi P z = 0$ implies $Pz \in K, z = Qz + Pz \in N + K$.

(7) From $Q \Pi P = 0$ and the symmetry of ΠP we know that $P \Pi P = \Pi P$, $\operatorname{im} P \Pi P = (N+K)^\perp$, $\operatorname{im} P = X \oplus (N+K)^\perp$. Next using Lemma A.6, compute $\widehat{N}^\perp = N^\perp + K^\perp$, and further

$$
\begin{aligned}
K \cap \widehat{N}^\perp &= K \cap (N^\perp + K^\perp) \\
&= \{ z \in \mathbb{R}^m : \Pi z = 0, z = z_{N^\perp} + z_{K^\perp}, z_{N^\perp} \in N^\perp, z_{K^\perp} \in K^\perp \} \\
&= \{ z \in \mathbb{R}^m : z = (I - \Pi) z_{N^\perp}, z_{N^\perp} \in N^\perp \} = (I - \Pi) N^\perp \\
&= \operatorname{im}(I - \Pi) P^* = \operatorname{im}(P(I - \Pi))^* = \operatorname{im} P(I - \Pi) = X.
\end{aligned}
$$

□

Lemma A.8. *Let $D \in L(\mathbb{R}^m, \mathbb{R}^n)$ be given, and let $M \subseteq \mathbb{R}^m$ be a subspace. Let $D^+ \in L(\mathbb{R}^n, \mathbb{R}^m)$ be the Moore–Penrose inverse of D. Then,*

(1) $\ker D^* = \operatorname{im} D^\perp$, $\operatorname{im} D = \ker D^{*\perp}$, $\ker D = \ker D^{+*}$, $\operatorname{im} D = \operatorname{im} D^{+*}$.
(2) $\ker D \subseteq M \Rightarrow (DM)^\perp = (\operatorname{im} D)^\perp \oplus D^{+*}M^\perp$.
(3) $\ker D \subseteq M \Rightarrow M^\perp = D^*(DM)^\perp$.

Proof. (1) The first two identities are shown in [15] (Theorem 1, p.12).
If $z \in \ker D = \operatorname{im} I - D^+D$ with Lemma A.3(1) it is valid that $z = (I - D^+D)z$ or $D^+Dz = 0$. With (A.11) it holds that $0 = D^+Dz = (D^+D)^*z = D^*D^{+*}z \Leftrightarrow D^{+*}z = 0$ because of (A.8) for D^* and we have that $z \in \ker D^{+*}$. We prove $\operatorname{im} D = \operatorname{im} D^{+*}$ analogously.
(2) Let $T \in L(\mathbb{R}^m)$ be the orthoprojector onto M, i.e., $\operatorname{im} T = M$, $\ker T = M^\perp$, $T^* = T$.
$\Rightarrow DM = \operatorname{im} DT$,

$$\begin{aligned}(DM)^\perp &= (\operatorname{im} DT)^\perp = \ker (DT)^* = \ker TD^* = \{z \in \mathbb{R}^n : D^*z \in M^\perp\} \\ &= \underbrace{\ker D^*}_{=\operatorname{im} D^\perp} \oplus \{v \in \operatorname{im} D : D^*v \in M^\perp\}.\end{aligned}$$

It remains to show that

$$\{v \in \operatorname{im} D : D^*v \in M^\perp\} = D^{+*}M^\perp.$$

From $v \in \operatorname{im} D = \operatorname{im} DD^+$ we get with Lemma A.3(1) $v = DD^+v = (DD^+)^*v = D^{+*}D^*v$. Because of $D^*v \in M^\perp$ it holds that $v \in D^{+*}M^\perp$. Conversely with Lemma A.3(4), $u \in D^{+*}M^\perp = \operatorname{im} D^{+*}(I - T)$ implies $u \in \operatorname{im} D^{+*} = \operatorname{im} D$, and $\exists w : u = D^{+*}(I - T)w$, $D^*u = D^*D^{+*}(I - T)w = D^+D(I - T)w$. Since $\operatorname{im}(I - T) = M^\perp \subseteq \ker D^\perp = \ker D^+D^\perp = \operatorname{im}(D^+D)^* = \operatorname{im} D^+D$, it holds that $D^+D(I - T) = I - T$, hence $D^*u = (I - T)w \in M^\perp$.
(3) This is a consequence of (2), because of

$$D^*(DM)^\perp = D^*[(\operatorname{im} D)^\perp \oplus D^{+*}M^\perp] = D^*D^{+*}M^\perp = D^+DM^\perp = M^\perp.$$

□

Lemma A.9. *[96, Appendix A, Theorem 13]*
Let $A, B \in L(\mathbb{R}^m)$, $\operatorname{rank} A = r < m$, $N := \ker A$, and $S := \{z \in \mathbb{R}^m : Bz \in \operatorname{im} A\}$. The following statements are equivalent:

(1) *Multiplication by a nonsingular $E \in L(\mathbb{R}^m)$ such that*

$$EA = \begin{bmatrix} \bar{A}_1 \\ 0 \end{bmatrix}, \quad EB = \begin{bmatrix} \bar{B}_1 \\ \bar{B}_2 \end{bmatrix}, \quad \operatorname{rank} \bar{A}_1 = r,$$

yields a nonsingular $\begin{bmatrix} \bar{A}_1 \\ \bar{B}_2 \end{bmatrix}$.

(2) $N \cap S = \{0\}$.
(3) *$A + BQ$ is nonsingular for each projector Q onto N.*
(4) $N \oplus S = \mathbb{R}^m$.
(5) *The pair $\{A, B\}$ is regular with Kronecker index 1.*
(6) *The pair $\{A, B + AW\}$ is regular with Kronecker index 1 for each arbitrary $W \in L(\mathbb{R}^m)$.*

Proof. (1) $\Rightarrow$ (2): With $\bar{N} := \ker \bar{A}_1 = \ker EA = \ker A = N$,

$$\bar{S} := \ker \bar{B}_2 = \{z \in \mathbb{R}^m : EBz \in \operatorname{im} EB\} = S,$$

we have

$$0 = \ker \begin{bmatrix} \bar{A}_1 \\ \bar{B}_2 \end{bmatrix} = \bar{N} \cap \bar{S} = N \cap S.$$

(2) $\Rightarrow$ (3): $(A + BQ)z = 0$ implies $BQz = -Az$, that is $Qz \in N \cap S$, thus $Qz = 0$, $Az = 0$, therefore $z = Qz = 0$.

(3) $\Rightarrow$ (4): Fix any projector $Q \in L(\mathbb{R}^m)$ onto N and introduce $Q_* := Q(A + BQ)^{-1}B$. We show Q_* to be a projector with $\operatorname{im} Q_* = N$, $\ker Q_* = S$ so that the assertion follows. Compute

$$Q_*Q = Q(A + BQ)^{-1}BQ = Q(A + BQ)^{-1}(A + BQ)Q = Q,$$

hence $Q_*^2 = Q_*$, $\operatorname{im} Q_* = N$. Further, $Q_*z = 0$ implies $(A + BQ)^{-1}Bz = (I - Q)(A + BQ)^{-1}Bz$, thus

$$Bz = (A + BQ)(I - Q)(A + BQ)^{-1}Bz = A(A + BQ)^{-1}Bz,$$

that is, $z \in S$. Conversely, $z \in S$ leads to $Bz = Aw$ and

$$Q_*z = Q(A + BQ)^{-1}Bz = Q(A + BQ)^{-1}Aw = Q(A + BQ)^{-1}(A + BQ)(I - Q)w = 0.$$

This proves the relation $\ker Q_* = S$.

(4) $\Rightarrow$ (5): Let Q_* denote the projector onto N along S, $P_* := I - Q_*$. Since $N \cap S = 0$ we know already that $G_* := A + BQ_*$ is nonsingular as well as the representation $Q_* = Q_*G_*^{-1}B$. It holds that

$$\begin{aligned} G_*^{-1}A &= G_*^{-1}(A + BQ_*)P_* = P_*, \\ G_*^{-1}B &= G_*^{-1}BQ_* + G_*^{-1}BP_* = G_*^{-1}(A + BQ_*)Q_* + G_*^{-1}BP_* = Q_* + G_*^{-1}BP_*. \end{aligned}$$

Consider the equation $(\lambda A + B)z = 0$, or the equivalent one $(\lambda G_*^{-1}A + G_*^{-1}B)z = 0$, i.e.,

$$(\lambda P_* + G_*^{-1}BP_* + Q_*)z = 0. \tag{A.5}$$

Multiplying (A.5) by Q_* and taking into account that $Q_*G_*^{-1}BP_* = Q_*P_* = 0$ we find $Q_*z = 0$, $z = P_*z$. Now (A.5) becomes

$$(\lambda I + G_*^{-1}B)z = 0.$$

If λ does not belong to the spectrum of the matrix $-G_*^{-1}B$, then it follows that $z=0$. This means that $\lambda A+B$ is nonsingular except for a finite number of values λ, hence the pair $\{A,B\}$ is regular.
Transform $\{A,B\}$ into Weierstraß–Kronecker canonical form (cf. Section 1.1):

$$\bar{A} := EAF = \begin{bmatrix} I & 0 \\ 0 & J \end{bmatrix}, \quad \bar{B} := EBF = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}, \quad J^{\mu} = 0,\ J^{\mu-1} \neq 0.$$

We derive further

$$\bar{N} := \ker \bar{A} = F^{-1}\ker A, \quad \bar{S} := \{z \in \mathbb{R}^m : \bar{B}z \in \operatorname{im}\bar{A}\} = F^{-1}S,$$
$$\bar{N} \cap \bar{S} = F^{-1}(N \cap S) = \{0\}, \text{ and}$$
$$\bar{N} \cap \bar{S} = \left\{ \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} \in \mathbb{R}^m : z_1 = 0,\ Jz_2 = 0,\ z_2 \in \operatorname{im} J \right\}.$$

Now it follows that $J=0$ must be true since otherwise $\bar{N}\cap\bar{S}$ would be nontrivial.

(5) $\Rightarrow$ (1): This follows from $\bar{A} = EAF = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$, $\bar{B} = EBF = \begin{bmatrix} W & 0 \\ 0 & I \end{bmatrix}$, $\bar{N}\cap\bar{S}=0$ and $\bar{N}\cap\bar{S} = F^{-1}(N\cap S) = \{0\}$.

(6) $\Rightarrow$ (5) is trivial.

(2) $\Rightarrow$ (6): Set $\tilde{B} := B + AW, \tilde{S} := \{z \in \mathbb{R}^m : \tilde{B}z \in \operatorname{im}A\} = S$. Because of $\tilde{S}\cap N = S\cap N = \{0\}$, and the equivalence of assertion (2) and (5), which is proved already, the pair $\{A,\tilde{B}\}$ is regular with Kronecker index 1. □

Lemma A.10. *Let $A,B \in L(\mathbb{R}^m)$ be given, A singular, $N := \ker A$, $S := \{z \in \mathbb{R}^m : Bz \in \operatorname{im}A\}$, and $N \oplus S = \mathbb{R}^m$. Then the projector Q onto N along S satisfies the relation*

$$Q = Q(A+BQ)^{-1}B. \tag{A.6}$$

Proof. First we notice that Q is uniquely determined. $A+BQ$ is nonsingular due to Lemma A.9. The arguments used in that lemma apply to show $Q(A+BQ)^{-1}B$ to be the projector onto N along S so that (A.6) becomes valid. □

For any matrix $A \in L(\mathbb{R}^m)$ there exists an integer k such that

$$\mathbb{R}^m = \operatorname{im}A^0 \supset \operatorname{im}A \supset \cdots \supset \operatorname{im}A^k = \operatorname{im}A^{k+1} = \cdots,$$
$$\{0\} = \ker A^0 \subset \ker A \subset \cdots \subset \ker A^k = \ker A^{k+1} = \cdots,$$

and $\operatorname{im}A^k \oplus \ker A^k = \mathbb{R}^m$. This integer $k \in \mathbb{N}\cup\{0\}$ is said to be the *index of* A, and we write $k = \operatorname{ind}A$.

Lemma A.11. *[96, Appendix A, Theorem 4]*
Let $A \in L(\mathbb{R}^m)$ be given, $k = \operatorname{ind}A$, $r = \operatorname{rank}A^k$, and let $s_1,\ldots,s_r \in \mathbb{R}^m$ and $s_{r+1},\ldots,s_m \in \mathbb{R}^m$ be bases of $\operatorname{im}A^k$ and $\ker A^k$, respectively. Then, for $S = [s_1 \ldots s_m]$ the product $S^{-1}AS$ has the special structure

$$S^{-1}AS = \begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}$$

where $M \in L(\mathbb{R}^r)$ is nonsingular and $N \in L(\mathbb{R}^{m-r})$ is nilpotent, $N^k = 0$, $N^{k-1} \neq 0$.

Proof. For $i \leq r$, it holds that $As_i \in A\,\mathrm{im}A^k = \mathrm{im}A^{k+1} = \mathrm{im}A^k$, therefore $As_i = \sum_{j=1}^{r} s_j m_{ji}$. For $i \geq r+1$, it holds that $As_i \in \ker A^{k+1} = \ker A^k$, thus $As_i = \sum_{j=r+1}^{m} s_j n_{ji}$. This yields the representations $A[s_1 \ldots s_r] = [s_1 \ldots s_r]M$ with $M = (m_{ij})_{i,j=1}^{r}$, and $A[s_{r+1} \ldots s_m] = [s_{r+1} \ldots s_m]N$, with $N = (n_{ij})_{i,j=r+1}^{m}$. The block M is nonsingular. Namely, for a $z \in \mathbb{R}^r$ with $Mz = 0$, we have $A[s_1 \ldots s_r]z = 0$, that is,

$$\sum_{j=1}^{r} z_j s_j \in \mathrm{im}A^k \cap \ker A \subseteq \mathrm{im}A^k \cap \ker A^k = \{0\},$$

which shows the matrix M to be nonsingular. It remains to verify the nilpotency of N. We have $AS = S\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}$, hence $A^\ell S = S\begin{bmatrix} M^\ell & 0 \\ 0 & N^\ell \end{bmatrix}$. From $A^k s_i = 0$, $i \geq r+1$ it follows that $N^k = 0$ must be valid. It remains to prove the fact that $N^{k-1} \neq 0$. Since $\ker A^{k-1}$ is a proper subspace of $\ker A^k$ there is an index $i_* \geq r+1$ such that the basis element $s_{i_*} \in \ker A^k$ does not belong to $\ker A^{k-1}$. Then, $S\begin{bmatrix} M^{k-1} & 0 \\ 0 & N^{k-1} \end{bmatrix} e_{i_*} = A^{k-1} s_{i_*} \neq 0$, that is, $N^{k-1} \neq 0$. □

A.2 Generalized inverses

In [15] we find a detailed collection of properties of generalized inverses for theory and application. We report here the definitions and relations of generalized inverses we need for our considerations.

Definition A.12. For a matrix $Z \in L(\mathbb{R}^n, \mathbb{R}^m)$, we call the matrix $Z^- \in L(\mathbb{R}^m, \mathbb{R}^n)$ a *reflexive generalized inverse*, if it fulfills

$$ZZ^-Z = Z \quad \text{and} \tag{A.7}$$

$$Z^-ZZ^- = Z^-. \tag{A.8}$$

Z^- is called a $\{1,2\}$-inverse of Z in [15].

The products $ZZ^- \in L(\mathbb{R}^m)$ and $Z^-Z \in L(\mathbb{R}^n)$ are projectors (cf. Appendix A.1). We have $(ZZ^-)^2 = ZZ^-ZZ^- = ZZ^-$ and $(Z^-Z)^2 = Z^-ZZ^-Z = Z^-Z$. We know that the rank of a product of matrices does not exceed the rank of any factor. Let Z have $\mathrm{rank}\, r_z$. From (A.7) we obtain $\mathrm{rank}\, r_z \leq \mathrm{rank}\, r_{z^-}$ and from (A.8) the opposite, i.e., that both Z and Z^- and also the projectors ZZ^- and Z^-Z have the same rank.

Let $R \in L(\mathbb{R}^n)$ be any projector onto $\mathrm{im}Z$ and $P \in L(\mathbb{R}^m)$ any projector along $\ker Z$.

Lemma A.13. *With (A.7), (A.8) and the conditions*

$$Z^- Z = P \quad \text{and} \tag{A.9}$$
$$ZZ^- = R \tag{A.10}$$

the reflexive inverse Z^- is uniquely determined.

Proof. Let Y be a further matrix fulfilling (A.7), (A.8), (A.9) and (A.10). Then

$$\begin{aligned} Y &\overset{(A.8)}{=} YZY \overset{(A.7)}{=} YZZ^-ZY \overset{(A.10)}{=} YRZY \\ &\overset{(A.10)}{=} YR \overset{(A.10)}{=} YZZ^- \overset{(A.9)}{=} PZ^- \overset{(A.8)}{=} Z^-. \end{aligned}$$

□

If we choose for the projectors P and R the orthogonal projectors the conditions (A.9) and (A.10) could be replaced by

$$Z^- Z = (Z^- Z)^*, \tag{A.11}$$
$$ZZ^- = (ZZ^-)^*. \tag{A.12}$$

The resulting generalized inverse is called the Moore–Penrose inverse and denoted by Z^+.

To represent the generalized reflexive inverse Z^- we want to use a decomposition of

$$Z = U \begin{bmatrix} S \\ 0 \end{bmatrix} V^{-1}$$

with nonsingular matrices U, V and S. Such a decomposition is, e.g., available using an SVD or a Householder decomposition of Z.
A generalized reflexive inverse is given by

$$Z^- = V \begin{bmatrix} S^{-1} & M_2 \\ M_1 & M_1 S M_2 \end{bmatrix} U^{-1} \tag{A.13}$$

with M_1 and M_2 being matrices of free parameters that fulfill

$$P = Z^- Z = V \begin{bmatrix} I & 0 \\ M_1 S & 0 \end{bmatrix} V^{-1}$$

and

$$R = ZZ^- = U \begin{bmatrix} I & SM_2 \\ 0 & 0 \end{bmatrix} U^{-1}$$

(cf. also [219]). There are two ways of looking at the parameter matrices M_1 and M_2. We can compute an arbitrary Z^- with fixed M_1 and M_2. Then also the projectors P and R are fixed by these parameter matrices. Or we provide the projectors P and R, then M_1 and M_2 are given and Z^- is fixed, too.

A.3 Parameter-dependent matrices and projectors

For any two continuously differentiable matrix functions of appropriate size $F : \mathcal{I} \to L(\mathbb{R}^m.\mathbb{R}^k)$ and $G : \mathcal{I} \to L(\mathbb{R}^l.\mathbb{R}^m)$, $\mathcal{I} \subseteq \mathbb{R}$, an interval, the product $FG : \mathcal{I} \to L(\mathbb{R}^l.\mathbb{R}^k)$ is defined pointwise by $(FG)(t) := F(t)G(t)$, $t \in \mathcal{I}$, and the product rule applies to the derivatives, i.e.,

$$(FG)'(t) = F'(t)G(t) + F(t)G'(t).$$

In particular, this is valid for projector valued functions.
Let $P \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^m))$ be a projector valued function and $Q = I - P$ the complementary one. The following three simple rules are useful in computations:

(1) $Q + P = I$, and hence $Q' = -P'$.
(2) $QP = PQ = 0$, and hence $Q'P = -QP'$, $P'Q = -PQ'$.
(3) $PP'P = -PQ'P = PQP' = 0$ and, analogously, $QQ'Q = 0$.

Lemma A.14. (1) *If the matrix function $P \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^m))$ is projector valued, that is, $P(t)^2 = P(t)$, $t \in \mathcal{I}$, then it has constant rank r, and there are r linearly independent functions $\eta_1, \ldots, \eta_r \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$ such that* $\operatorname{im} P(t) = \operatorname{span}\{\eta_1(t), \ldots, \eta_r(t)\}$, $t \in \mathcal{I}$.

(2) *If a time-dependent subspace $L(t) \subseteq \mathbb{R}^m$, $t \in \mathcal{I}$, with constant dimension r is spanned by functions $\eta_1, \ldots, \eta_r \in \mathcal{C}^1(\mathcal{I}, \mathbb{R}^m)$, which means $L(t) = \operatorname{span}\{\eta_1(t), \ldots, \eta_r(t)\}$, $t \in \mathcal{I}$, then the orthoprojector function onto this subspace is continuously differentiable.*

(3) *Let the matrix function $A \in \mathcal{C}^k(\mathcal{I}, L(\mathbb{R}^m))$ have constant rank r. Then, there is a matrix function $M \in \mathcal{C}^k(\mathcal{I}, L(\mathbb{R}^m))$ that is pointwise nonsingular such that $A(t)M(t) = [\underbrace{\tilde{A}(t)}_{r}\ 0]$, $\operatorname{rank}\tilde{A}(t) = r$ for all $t \in \mathcal{I}$.*

Proof. (1) Denote $Q = I - P$, and let r be the maximal rank of $P(t)$ for $t \in \mathcal{I}$. We fix a value $\bar{t} \in \mathcal{I}$ such that $\operatorname{rank} P(\bar{t}) = r$. Let $\bar{\eta}_1, \ldots, \bar{\eta}_r$ be a basis of $\operatorname{im} P(\bar{t})$.
For $i = 1, \ldots, r$, the ordinary IVP

$$\eta'(t) = P'(t)\eta(t),\ t \in \mathcal{I}, \quad \eta(\bar{t}) = \bar{\eta}_i,$$

is uniquely solvable. The IVP solutions $\eta_1, \ldots, \eta_r$ remain linearly independent on the entire interval $\mathcal{I}$ since they are so at $\bar{t}$.
Moreover, the function values of these functions remain in $\operatorname{im} P$, that is, $\eta_i(t) = P(t)\eta_i(t)$. Namely, multiplying the identity $\eta_i = P'\eta_i$ by Q gives $(Q\eta_i)' = -Q'Q\eta_i$, and because of $Q(\bar{t})\eta_i(\bar{t}) = Q(\bar{t})\bar{\eta}_i = 0$, the function $Q\eta_i$ must vanish identically.
It follows that $\operatorname{span}\{\eta_1(t), \ldots, \eta_r(t)\} \subseteq \operatorname{im} P(t)$ for all $t \in \mathcal{I}$, and $r \le \operatorname{rank} P(t)$, and hence $r = \operatorname{rank} P(t)$ and $\operatorname{span}\{\eta_1(t), \ldots, \eta_r(t)\} = \operatorname{im} P(t)$.
(2) The matrix function $\Gamma := [\eta_1\ \ \eta_r]$, the columns of which are the given functions $\eta_1, \ldots, \eta_r$, is continuously differentiable and injective, and $\Gamma^*\Gamma$ is invertible. Then $P := \Gamma(\Gamma^*\Gamma)^{-1}\Gamma^*$ is continuously differentiable. The value $P(t)$ is an or-

thoprojector, further $\operatorname{im} P \subseteq \operatorname{im} \Gamma$ by construction, and $P\Gamma = \Gamma$, in consequence $\operatorname{im} P = \operatorname{im} \Gamma = L$.
(3) See [61]. □

For matrix functions depending on several variables we define products pointwise, too. More precisely, for $F : \Omega \to L(\mathbb{R}^m.\mathbb{R}^k)$ and $G : \Omega \to L(\mathbb{R}^l.\mathbb{R}^m)$, $\Omega \subseteq \mathbb{R}^p$, the product $FG : \Omega \to L(\mathbb{R}^l.\mathbb{R}^k)$ is defined pointwise by $(FG)(x) := F(x)G(x)$, $x \in \Omega$.
We speak of a projector function $P : \Omega \to L(\mathbb{R}^l)$, if for all $x \in \Omega$, $P(x)^2 = P(x)$ holds true, and of an orthoprojector function, if, additionally, $P(x)^* = P(x)$. Saying that P is a projector function onto the subspace L we mean that P and L have a common definition domain, say Ω, and $\operatorname{im} P(x) = L(x)$, $x \in \Omega$.

Lemma A.15. *Given a matrix function $A \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m, \mathbb{R}^n))$, $k \in \mathbb{N} \cup \{0\}$, $\Omega \subseteq \mathbb{R}^p$ open, that has constant rank r, then*

(1) *The orthoprojector function onto* $\operatorname{im} A$ *is k times continuously differentiable.*
(2) *The orthoprojector function onto* $\ker A$ *is also k times continuously differentiable.*

Proof. (1) Let $\bar{x} \in \Omega$ be fixed, and $\bar{z}_1, \ldots, \bar{z}_r$ be an orthonormal basis of $\operatorname{im} A(\bar{x})^\perp$. Denote $\bar{u}_i := A(\bar{x})\bar{z}_i$, $i = 1, \ldots, r$. By construction, $\bar{u}_1, \ldots, \bar{u}_r$ are linearly independent.
We form $u_i(x) := A(x)\bar{z}_i$ for $i = 1, \ldots, r$, and then the matrix $U(x) := [u_1(x) \;\; u_r(x)]$, $x \in \Omega$. The matrix $U(\bar{x})$ has full column rank r. Therefore, there is a neighborhood $\mathcal{N}_{\bar{x}}$ of $\bar{x}$ such that $U(x)$ has full column rank r on $\mathcal{N}_{\bar{x}}$. The Gram–Schmidt orthogonalization yields the factorization

$$U(x) = Q(x)R(x), \quad Q(x) \in L(\mathbb{R}^r, \mathbb{R}^n), \quad Q(x)^*Q(x) = I_r, \quad x \in \mathcal{N}_{\bar{x}},$$

with $R(x)$ being upper triangular and nonsingular. It follows that $\operatorname{im} U(x) = \operatorname{im} Q(x)$ is true for $x \in \mathcal{N}_{\bar{x}}$.
Further, $U = A[\bar{z}_1 \;\; \bar{z}_r]$ shows that U is k times continuously differentiable together with A. By construction, Q is as smooth as U. Finally, the matrix function $R_A := Q(Q^*Q)^{-1}Q^*$ is k times continuously differentiable, and it is an orthoprojector function, $\operatorname{im} R_A = \operatorname{im} Q = \operatorname{im} U = \operatorname{im} A$.
(2) This assertion is a consequence of (1). Considering the well-known relation $\ker A^\perp = \operatorname{im} A^*$ we apply (1) and find the orthoprojector function P_A onto $\ker A^\perp$ along $\ker A$ to be k times continuously differentiable, and $I - P_A$ has this property, too. □

Remark A.16. By Lemma A.14 the orthogonal projector function $P \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^m))$, $\mathcal{I} \subseteq \mathbb{R}$ an interval, generates globally on $\mathcal{I}$ defined bases $\eta_1, \ldots, \eta_r \in \mathcal{C}^1(\mathcal{I}, L(\mathbb{R}^m))$, $r = \operatorname{rank} P(t)$, $\operatorname{im} P(t) = \operatorname{im}[\eta_1(t), \ldots, \eta_r(t)]$, $t \in \mathcal{I}$.
In the higher dimensional case, if $P \in \mathcal{C}^1(\Omega, L(\mathbb{R}^m))$, $\Omega \subseteq \mathbb{R}^p$ open, $p > 1$, the situation is different. By Lemma A.20, item (8), there are local bases. However, in general, global bases do not necessarily exist.
For instance, the orthoprojector function onto the nullspace of the matrix function $M(x) = [x_1, x_2, x_3]$, $x \in \mathbb{R}^3 \setminus \{0\}$, reads

$$P(x) = \frac{1}{x_1^2+x_2^2+x_3^2} \begin{bmatrix} x_2^2+x_3^2 & -x_1x_2 & -x_1x_3 \\ -x_1x_2 & x_1^2+x_3^2 & -x_2x_3 \\ -x_1x_3 & -x_2x_3 & x_1^2+x_2^2 \end{bmatrix}.$$

This projector function is obviously continuously differentiable. On the other hand, the nullspace $\ker M(x) = \{z \in \mathbb{R}^3 : x_1z_1 + x_2z_2 + x_3z_3 = 0\}$ allows only locally different descriptions by bases, e.g.,

$$\ker M(x) = \operatorname{im} \begin{bmatrix} -\frac{x_2}{x_1} & -\frac{x_3}{x_1} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \text{if } x_1 \neq 0,$$

$$\ker M(x) = \operatorname{im} \begin{bmatrix} 1 & 0 \\ 0 & -\frac{x_3}{x_2} \\ 0 & 1 \end{bmatrix} \qquad \text{if } x_1 = 0,\ x_2 \neq 0,$$

$$\ker M(x) = \operatorname{im} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \text{if } x_1 = 0,\ x_2 = 0,\ x_3 \neq 0.$$

Proposition A.17. *For $k \in \mathbb{N} \cup \{0\}$, let the matrix function $D \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m, \mathbb{R}^n))$ have constant rank on the open set $\Omega \subseteq \mathbb{R}^p$.*

(1) *Then the Moore–Penrose generalized inverse D^+ of D is as smooth as D.*
(2) *Let $R \in \mathcal{C}^k(\Omega, L(\mathbb{R}^n))$ be a projector function onto $\operatorname{im} D$, and $P \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ be a projector function such that $\ker P = \ker D$. Then the four conditions*

$$DD^-D = D, \quad D^-DD^- = D, \quad D^-D = P, \quad DD^- = R,$$

determine uniquely a function D^- that is pointwise a generalized inverse of D, and D^- is k times continuously differentiable.

Proof. The first assertion is well-known, and can be found, e.g., in [49].
The second assertion follows from the first one. We simply show the matrix function $D^- := PD^+R$ to be the required one. By Lemma A.13, the four conditions define pointwise a unique generalized inverse. Taking into account that $\operatorname{im} D = \operatorname{im} R = \operatorname{im} DD^+$ and $\ker D = \ker D^+D = \ker P$ we derive

$$\begin{aligned} D(PD^+R)D &= DD^+R = R, \\ (PD^+R)D(PD^+R) &= PD^+DD^+R = (PD^+R), \\ (PD^+R)D &= PD^+D = P, \\ D(PD^+R) &= DD^+R = R, \end{aligned}$$

so that the four conditions are fulfilled. Obviously, the product PD^+R inherits the smoothness of its factors. □

For what concerns the derivatives, the situation is more difficult, if several variables are involved. We use the symbols $F_x(x,t)$, $F_t(x,t)$ for the partial derivatives and partial Jacobian matrices of the function $F \in \mathcal{C}^1(\Omega \times \mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^k))$ with respect to $x \in \mathbb{R}^p$ and $t \in \mathbb{R}$, taken at the point $(x,t) \in \Omega \times \mathcal{I}$.
For the two functions $F \in \mathcal{C}^1(\Omega \times \mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^k))$ and $G \in \mathcal{C}^1(\Omega \times \mathcal{I}, L(\mathbb{R}^l, \mathbb{R}^m))$, the product $FG \in \mathcal{C}^1(\Omega \times \mathcal{I}, L(\mathbb{R}^l, \mathbb{R}^k))$ is defined pointwise. We have

$$(FG)_x(x,t)z = [F_x(x,t)z]G(x,t) + F(x,t)G_x(x,t)z \quad \text{for all } z \in \mathbb{R}^p.$$

Besides the partial derivatives we apply the *total derivative in jet variables*. For the function $F \in \mathcal{C}^1(\Omega \times \mathcal{I}, L(\mathbb{R}^m, \mathbb{R}^k))$, $\Omega \times \mathcal{I} \subseteq \mathbb{R}^p \times \mathbb{R}$, the function $F' \in \mathcal{C}(\Omega \times \mathcal{I} \times \mathbb{R}^p, L(\mathbb{R}^m, \mathbb{R}^k))$ defined by

$$F'(x,t,x^1) := F_x(x,t)x^1 + F_t(x,t), \quad x \in \Omega,\ t \in \mathcal{I},\ x^1 \in \mathbb{R}^p,$$

is named the total derivative of F in jet variables. For the total derivative, the product rule

$$(FG)' = F'G + FG'$$

is easily checked to be valid.

Lemma A.18. *The total derivatives in jet variables P' and Q' of a continuously differentiable projector function P and its complementary one $Q = I - P$ satisfy the following relations:*

$$\begin{aligned} Q' &= -P', \\ Q'P &= -QP', \\ PP'P &= 0. \end{aligned}$$

Proof. The assertion follows from the identities $Q + P = I$ and $QP = 0$ by regarding the product rule. □

Notice that, for each given function $x_* \in \mathcal{C}^1(\mathcal{I}_*, \mathbb{R}^p)$, $\mathcal{I}_* \subseteq \mathcal{I}$, with values in Ω, the resulting superposition $F(x_*(t),t)$ is continuously differentiable with respect to t on $\mathcal{I}_*$, and it possesses the derivative

$$(F(x_*(t),t))' := (F(x_*(.),.))'(t) = F'(x_*(t),t,x_*'(t)).$$

A.4 Variable subspaces

Definition A.19. Let $\Omega \subseteq \mathbb{R}^P$ be open and connected, and $L(x) \subseteq \mathbb{R}^m$ be a subspace for each $x \in \Omega$. For $k \in \mathbb{N} \cup \{0\}$, L is said to be a *$\mathcal{C}^k$-subspace on Ω*, if there exists a projector function $R \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ which projects pointwise onto L, i.e., $R(x) = R(x)^2$, $\operatorname{im} R(x) = L(x)$, $x \in \Omega$. We write $\operatorname{im} R = L$.

It should be mentioned at this point that the notion of *smooth subspace* (smooth stands for $\mathcal{C}^1$) is applied in [96], Subsection 1.2.1, to subspaces depending on one real variable ($p = 1$) in the same way.

Lemma A.20. *Let $k \in \mathbb{N} \cup \{0\}$.*

(1) *A $\mathcal{C}^k$-subspace on an open connected Ω has constant dimension.*
(2) *The orthoprojector function onto a $\mathcal{C}^k$-subspace belongs to $\mathcal{C}^k$.*
(3) *If L is a $\mathcal{C}^k$-subspace, so is $L^\perp$.*
(4) *If L and N are $\mathcal{C}^k$-subspaces, and $L \cap N$ has constant dimension, then $L \cap N$ is a $\mathcal{C}^k$-subspace, too.*
(5) *If N and L are $\mathcal{C}^k$-subspaces, and $N \oplus L = \mathbb{R}^m$, then the projector onto N along L belongs to $\mathcal{C}^k$.*
(6) *If L and N are $\mathcal{C}^k$-subspaces, and $L \cap N$ has constant dimension, then there is a $\mathcal{C}^k$-subspace X such that $X \subseteq L$, and*

$$L = X \oplus (N \cap L),$$

as well as a projector $R \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ with $\operatorname{im} R = N$, $\ker R \supseteq X$.
(7) *If L and N are $\mathcal{C}^k$-subspaces, and $N \cap L = 0$, then $L \oplus N$ is a C^k-subspace, too.*
(8) *L is a $\mathcal{C}^k$-subspace on $\Omega \Leftrightarrow$ for each $\bar{x} \in \Omega$ there is a neighborhood $U_{\bar{x}} \subseteq \Omega$ and a local $\mathcal{C}^k$-basis $\eta_1, \ldots, \eta_{r(\bar{x})} \in \mathcal{C}^k(U_{\bar{x}}, \mathbb{R}^m)$ spanning L on $U_{\bar{x}}$, i.e.,*

$$\operatorname{span}\{\eta_1(x), \ldots, \eta_{r(\bar{x})}(x)\} = L(x), \quad x \in U_{\bar{x}}.$$

Proof. (1) Let $x_0 \in \Omega$, and let the columns of $\xi^0 := [\xi^0_1, \ldots, \xi^0_{r_{x_0}}]$ form a basis of $L(x_0)$, i.e., $L(x_0) = \operatorname{im} \xi^0$. $\xi(x) := R(x)\xi^0$ is a $\mathcal{C}^k$ matrix function, and since $\xi(x_0) = R(x_0)\xi^0 = \xi^0$ has full column rank r_{x_0}, there is a neighborhood $U_{x_0} \subset \Omega$ such that $\xi(x)$ has rank r_{x_0} for all $x \in U_{x_0}$. This means $\operatorname{im} \xi(x) \subseteq \operatorname{im} R(x)$,

$$\operatorname{rank} R(x) \geq \operatorname{rank} \xi(x) = r_{x_0}, \quad x \in U_{x_0}.$$

Denote by $r_{\min}$, $r_{\max}$ the minimal and maximal ranks of $R(x)$ on Ω, $0 \leq r_{\min} \leq r_{\max} \leq m$, and by $x_{\min}, x_{\max} \in \Omega$ points with $\operatorname{rank} R(x_{\min}) = r_{\min}$, $\operatorname{rank} R(x_{\max}) = r_{\max}$.
Since Ω is connected, there is a connecting curve of $x_{\min}$ and $x_{\max}$ belonging to Ω. We move on this curve from $x_{\max}$ to $x_{\min}$. If $r_{\min} < r_{\max}$, there must be a x_* on this curve with

$$r_* := \operatorname{rank} R(x_*) < r_{\max},$$

and in each arbitrary neighborhood of x_* there are points $\hat{x}$ with $\operatorname{rank} R(\hat{x}) = r_{\max}$.
At each $x \in \Omega$, as a projector, $R(x)$ has only the eigenvalues 1 and 0 (cf. Lemma A.3(6)). Hence, $R(x_*)$ has eigenvalue 1 with multiplicity r_*, and eigenvalue 0 with multiplicity $m - r_*$, $R(\hat{x})$ has eigenvalue 1 with multiplicity $r_{\max}$ and eigenvalue 0 with multiplicity $m - r_{\max}$.
Since eigenvalues depend continuously on the entries of a matrix, and the entries of

$R(x)$ are $\mathcal{C}^k$-functions in x, the existence of x_* contradicts the continuity of eigenvalues. Therefore, $r_{\min} = r_{\max}$ must be valid.

(2) If L is a $\mathcal{C}^k$-subspace, by definition, there is a projector $R \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ onto L, and the $\operatorname{rank} R(x)$ is constant on Ω. By Lemma A.15, the orthoprojector function onto $\operatorname{im} R = L$ is k times continuously differentiable.

(3) If L is a $\mathcal{C}^k$-subspace, the orthoprojector R onto L belongs to $\mathcal{C}^k$. Then, $I - R$ is a $\mathcal{C}^k$-projector onto $\operatorname{im}(I-R) = L^\perp$.

(4) Suppose L, N are $\mathcal{C}^k$-subspaces in $\mathbb{R}^m$, and R_L, R_N corresponding projectors onto L and N. Then $F := \begin{bmatrix} I - R_L \\ I - R_N \end{bmatrix}$ is a $\mathcal{C}^k$-function, and $\ker F = L \cap N$. Since $L \cap N$ has constant dimension, F has constant rank, and therefore F^+ and F^+F are $\mathcal{C}^k$-functions. F^+F is the orthoprojector onto $\ker F$, thus $\ker F = L \cap N$ is a $\mathcal{C}^k$-subspace.

(5) Let N, L be $\mathcal{C}^k$-subspaces, $N \oplus L = \mathbb{R}^m$. For each arbitrary $x \in \Omega$, $R(x)$ is uniquely determined by $\operatorname{im} R(x) = L(x)$, $\ker R(x) = N(x)$, $R(x)^2 = R(x)$. We have to make sure that R belongs to $\mathcal{C}^k$. To each fixed $x_0 \in \Omega$ we consider bases $\xi_1^0, \ldots, \xi_r^0$ of $L(x_0)$, and $\eta_1^0, \ldots, \eta_{m-r}^0$ of $N(x_0)$, and consider

$$\xi(x) := R_L(x)\xi^0, \quad \eta(x) := R_N(x)\eta^0, \quad x \in \Omega,$$

where

$$\xi^0 = [\xi_1^0, \ldots, \xi_r^0], \quad \eta^0 = [\eta_1^0, \ldots, \eta_{m-r}^0],$$

and R_L, R_N are $\mathcal{C}^k$-projectors according to the $\mathcal{C}^k$-subspaces L and N. There is a neighborhood $U_{x_0} \subset \Omega$ of x_0, such that the columns of $\xi(x)$ and $\eta(x)$, for $x \in U_{x_0}$, are bases of $L(x)$ and $N(x)$, and the matrix $F(x) := [\xi(x), \eta(x)]$ is nonsingular for $x \in U_{x_0}$. Define, for $x \in U_{x_0}$,

$$\tilde{R}(x) := F(x) \begin{bmatrix} I_r & \\ & 0 \end{bmatrix} F(x)^{-1},$$

such that

$$\tilde{R} \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m)), \quad \operatorname{im} \tilde{R}(x) = L(x), \quad \ker \tilde{R}(x) = N(x).$$

Since the projector corresponding to the decomposition $N(x) \oplus L(x) = \mathbb{R}^m$ is unique, we have $R(x) = \tilde{R}(x)$, $x \in U_{x_0}$, and hence R is $\mathcal{C}^k$ on U_{x_0}.

(6) Let L, N be $\mathcal{C}^k$-subspaces, $\dim(N \cap L) = \text{constant} =: u$. By (d), $N \cap L$ is a $\mathcal{C}^k$-subspace. We have $\mathbb{R}^m = (L \cap N) \oplus (L \cap N)^\perp$, $L = (L \cap N) \oplus (L \cap (L \cap N)^\perp)$, and $X := L \cap (L \cap N)^\perp$ is a $\mathcal{C}^k$-subspace, too. Further (cf. Lemma A.6), $(N+L)^\perp = N^\perp \cap L^\perp$ is also a $\mathcal{C}^k$-subspace. With $N + L = N \oplus X$ we find

$$\mathbb{R}^m = (N+L)^\perp \oplus (N+L) = (N+L)^\perp \oplus X \oplus N = S \oplus N, \quad S := (N+L)^\perp \oplus X.$$

Denote by $R^\perp$ and R_X the orthoprojectors onto the $\mathcal{C}^k$-subspaces $(N+L)^\perp$ and X. Due to $X \subseteq N + L$, $(N+L)^\perp \subseteq X^\perp$, hence $\operatorname{im} R_X \subseteq \ker R^\perp$, $\operatorname{im} R^\perp \subseteq \ker R_X$, it holds

that $R_X R^\perp = 0$, $R^\perp R_X = 0$, hence $R_S := R^\perp + R_X$ is a projector and belongs to $\mathcal{C}^k$, $\operatorname{im} R_S = \operatorname{im} R^\perp + \operatorname{im} R_X = S$. This makes it clear that S is also a $\mathcal{C}^k$-subspace.
Finally, due to $\mathbb{R}^m = S \oplus N$, there is a projector $R \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ with $\operatorname{im} R = N$, $\ker R = S \supset X$.

(7) By (6), due to $N \cap L = 0$, there are projectors $R_L, R_N \in \mathcal{C}^k(\Omega, L(\mathbb{R}^m))$ such that $\operatorname{im} R_L = L$, $N \subset \ker R_L$, $\operatorname{im} R_N = N$, $L \subset \ker R_N$, thus $R_L R_N = 0$, $R_N R_L = 0$, and $R := R_L + R_N$ is a $\mathcal{C}^k$-projector, too, and finally $\operatorname{im} R = \operatorname{im} R_L + \operatorname{im} R_N = L \oplus N$.

(8) If L is a $\mathcal{C}^k$-subspace then the orthogonal projector R on L along $L^\perp$ is $\mathcal{C}^k$. For each $x_0 \in \Omega$ and a basis $\xi_1^0, \dots, \xi_r^0$ of $L(x_0)$, the columns of $\xi(x) := R(x)\xi^0$, $\xi = [\xi_1^0, \dots, \xi_r^0]$, form a $\mathcal{C}^k$-basis of $L(x)$ locally on a neighborhood $U_{x_0} \subset \Omega$ of x_0. Conversely, if there is a local $\mathcal{C}^k$-basis on the neighborhood $U_{\bar{x}}$ of $\bar{x}$, then one can show that the orthoprojector onto $L(x)$, $x \in U_{\bar{x}}$, can be represented by means of this basis. That means, L is $\mathcal{C}^k$ on $U_{\bar{x}}$. □

Corollary A.21. *Any projector function being continuous on an open connected set has constant rank there.*

Proof. The continuous projector function, say $P : \Omega \to L(\mathbb{R}^p)$, defines the $\mathcal{C}$-space $\operatorname{im} P$. Owing to Lemma A.20 item (1), $\operatorname{im} P$ has constant dimension, and hence P has constant rank. □

Appendix B
Technical computations

B.1 Proof of Lemma 2.12

Lemma 2.12

If two projector function sequences $Q_0,\dots,Q_k$ and $\bar{Q}_0,\dots,\bar{Q}_k$ are both admissible, then the corresponding matrix functions and subspaces are related by the following properties:

(a) $\ker \bar{\Pi}_j = \bar{N}_0+\cdots+\bar{N}_j = N_0+\cdots+N_j = \ker \Pi_j,\ j=0,\dots,k,$

(b) $\bar{G}_j = G_j Z_j,$

$$\bar{B}_j = B_j - G_j Z_j \bar{D}^-(D\bar{\Pi}_j\bar{D}^-)'D\Pi_j + G_j\sum_{l=0}^{j-1} Q_l\mathfrak{A}_{jl},\ j=1,\dots,k,$$

with nonsingular matrix functions $Z_0,\dots,Z_{k+1}$ given by
$Z_0 := I,\ Z_{i+1} := Y_{i+1}Z_i,\ i=0,\dots,k,$

$$Y_1 := I + Q_0(\bar{Q}_0 - Q_0) = I + Q_0\bar{Q}_0P_0,$$
$$Y_{i+1} := I + Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i) + \sum_{l=0}^{i-1} Q_l\mathfrak{A}_{il}\bar{Q}_i,\ i=1,\dots,k,$$

and certain continuous coefficients $\mathfrak{A}_{il}$ that satisfy the condition $\mathfrak{A}_{il} = \mathfrak{A}_{il}\bar{\Pi}_{i-1}$,

(c) $Z_i(\bar{N}_i \cap (\bar{N}_0+\cdots+\bar{N}_{i-1})) = N_i \cap (N_0+\cdots+N_{i-1}),\quad i=1,\dots,k,$

(d) $\bar{G}_{k+1} = G_{k+1}Z_{k+1},\quad \bar{N}_0+\cdots+\bar{N}_{k+1} = N_0+\cdots+N_{k+1},$
$Z_{k+1}(\bar{N}_{k+1}\cap(\bar{N}_0+\cdots+\bar{N}_k)) = N_{k+1}\cap(N_0+\cdots+N_k).$

Proof. We have $G_0 = AD = \bar{G}_0,\ B_0 = B = \bar{B}_0,\ \ker P_0 = N_0 = \bar{N}_0 = \ker \bar{P}_0$, hence $P_0 = P_0\bar{P}_0,\ \bar{P}_0 = \bar{P}_0P_0$.
The generalized inverses D^- and $\bar{D}^-$ of D satisfy the properties $DD^- = D\bar{D}^- = R$, $D^-D = P_0,\ \bar{D}^-D = \bar{P}_0$, and therefore $\bar{D}^- = \bar{D}^-D\bar{D}^- = \bar{D}^-DD^- = \bar{P}_0D^-$, $D^- = P_0\bar{D}^-$.
Compare $G_1 = G_0 + B_0Q_0$ and

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5,

$$\begin{aligned}\bar{G}_1 &= \bar{G}_0 + \bar{B}_0\bar{Q}_0 = G_0 + B_0\bar{Q}_0 = G_0 + B_0Q_0\bar{Q}_0\\ &= (G_0 + B_0Q_0)(P_0 + \bar{Q}_0) = G_1Z_1,\end{aligned}$$

where $Z_1 := Y_1 := P_0 + \bar{Q}_0 = I + Q_0\bar{Q}_0P_0 = I + Q_0(\bar{Q}_0 - Q_0)$. Z_1 is invertible, and it has inverse $Z_1^{-1} = I - Q_0\bar{Q}_0P_0$.
The nullspaces N_1 and $\bar{N}_1$ are, due to $\bar{G}_1 = G_1Z_1$, related by $\bar{N}_1 = Z_1^{-1}N_1 \subseteq N_0 + N_1$. This implies $\bar{N}_0 + \bar{N}_1 = N_0 + (Z_1^{-1}N_1) \subseteq N_0 + N_1$. From $N_1 = Z_1\bar{N}_1 \subseteq N_0 + \bar{N}_1 = \bar{N}_0 + \bar{N}_1$, we obtain $\bar{N}_0 + \bar{N}_1 = N_0 + N_1$.
Since the projectors $\Pi_1 = P_0P_1$ and $\bar{\Pi}_1 = \bar{P}_0\bar{P}_1$ have the common nullspace $N_0 + N_1 = \bar{N}_0 + \bar{N}_1$, we may now derive

$$\begin{aligned}D\bar{P}_0\bar{P}_1\bar{D}^- &= D\bar{P}_0\bar{P}_1P_0P_1\bar{P}_0D^- = D\bar{P}_0\bar{P}_1P_0P_1D^- = D\bar{P}_0\bar{P}_1\bar{D}^-DP_0P_1D^-,\\ DP_0P_1D^- &= DP_0P_1D^-D\bar{P}_0\bar{P}_1\bar{D}^-.\end{aligned}$$

Next we compute

$$\begin{aligned}\bar{B}_1 &= \bar{B}_0\bar{P}_0 - \bar{G}_1\bar{D}^-(D\bar{P}_0\bar{P}_1\bar{D}^-)'D\bar{P}_0\\ &= B_0(P_0 + Q_0)\bar{P}_0 - G_1Z_1\bar{D}^-(D\bar{P}_0\bar{P}_1\bar{D}^-DP_0P_1D^-)'D\\ &= B_0P_0 + B_0Q_0\bar{P}_0 - G_1Z_1\bar{D}^-(D\bar{P}_0\bar{P}_1\bar{D}^-)'DP_0P_1 - G_1Z_1\bar{P}_0\bar{P}_1\bar{D}^-(DP_0P_1D^-)'D\\ &= B_1 + G_1D^-(DP_0P_1D^-)'D - G_1Z_1\bar{D}^-(D\bar{P}_0\bar{P}_1\bar{D}^-)'DP_0P_1\\ &\qquad - G_1Z_1\bar{P}_0\bar{P}_1D^-(DP_0P_1D^-)'D + B_0Q_0\bar{P}_0\\ &= B_1 - G_1Z_1\bar{D}^-(D\bar{P}_0\bar{P}_1\bar{D}^-)'DP_0P_1 + \mathfrak{B}_1\end{aligned}$$

with $\mathfrak{B}_1 := G_1Q_0\bar{P}_0 + G_1(I - Z_1\bar{\Pi}_1)D^-(D\Pi_1D^-)'D$.
The identity $0 = \bar{G}_1\bar{Q}_1 = G_1Z_1\bar{Q}_1 = G_1\bar{Q}_1 + G_1(Z_1 - I)\bar{Q}_1$ leads to $G_1\bar{Q}_1 = -G_1(Z_1 - I)\bar{Q}_1$ and further to

$$\begin{aligned}G_1(I - Z_1\bar{\Pi}_1) &= G_1(I - \bar{\Pi}_1 - (Z_1 - I)\bar{\Pi}_1) = G_1(\bar{Q}_1 + \bar{Q}_0\bar{P}_1 - Q_0\bar{Q}_0P_0\bar{\Pi}_1)\\ &= G_1(-Q_0\bar{Q}_0P_0\bar{Q}_1 + \bar{Q}_0\bar{P}_1 - Q_0\bar{Q}_0P_0\bar{P}_1) = G_1(-Q_0\bar{Q}_0P_0 + \bar{Q}_0\bar{P}_1)\\ &= G_1(-Q_0\bar{Q}_0 + Q_0 + Q_0\bar{Q}_0\bar{P}_1) = G_1(-Q_0\bar{Q}_0\bar{Q}_1 + Q_0).\end{aligned}$$

Inserting into the expression for $\mathfrak{B}_1$ yields $\mathfrak{B}_1 = G_1Q_0\bar{P}_0 - G_1Q_0\bar{Q}_0\bar{Q}_1D^-(D\Pi_1D^-)'D = G_1Q_0\mathfrak{A}_{10}$ with $\mathfrak{A}_{10} := \bar{P}_0 - \bar{Q}_0\bar{Q}_1D^-(D\Pi_1D^-)'D$ and $\mathfrak{A}_{10} = \mathfrak{A}_{10}\bar{P}_0$.
In order to verify assertions (a) and (b) by induction, we assume the relations

$$\begin{aligned}\bar{N}_0 + \cdots + \bar{N}_j &= N_0 + \cdots + N_j,\\ \bar{G}_j &= G_jZ_j,\\ \bar{B}_j &= B_j - G_jZ_j\bar{D}^-(D\bar{\Pi}_j\bar{D}^-)'D\Pi_j + G_j\sum_{l=0}^{j-1}Q_l\mathfrak{A}_{jl}\end{aligned}\qquad\text{(B.1)}$$

to be valid for $j = 1,\ldots,i$, $i < k$, with nonsingular Z_i as described above.
By construction, Z_i is of the form $Z_j = Y_jZ_{j-1} = Y_jY_{j-1}\cdots Y_1$. By carrying out the

multiplication and rearranging the terms we find the expression

$$Z_j - I = \sum_{l=0}^{j-1} Q_l \mathcal{C}_{jl} \tag{B.2}$$

with continuous coefficients $\mathcal{C}_{jl}$.
It holds that $Y_1 - I = Q_0 \bar{Q}_0 P_0$ and

$$Y_j - I = (Y_j - I)\Pi_{j-2},\ j = 2, \ldots, i, \tag{B.3}$$

such that $(Y_j - I)(Z_{j-1} - I) = 0$ must be true. From this it follows that $Y_j(Z_{j-1} - I) = Z_{j-1} - I$, and $Z_j = Y_j Z_{j-1} = Y_j + Y_j(Z_{j-1} - I) = Y_j + Z_{j-1} - I = Y_j - I + Z_{j-1}$, i.e.,

$$\begin{aligned} Z_j &= Y_j - I + \cdots + Y_1 - I + Z_0, \\ Z_j - Z_0 &= Z_j - I = \sum_{l=1}^{j} (Y_l - I). \end{aligned} \tag{B.4}$$

From (B.4) one can obtain special formulas for the coefficients $\mathcal{C}_{jl}$ in (B.2), but in our context there is no need for these special descriptions.
Now we compare $\bar{G}_{i+1}$ and G_{i+1}. We have

$$\bar{G}_{i+1} = \bar{G}_i + \bar{B}_i \bar{Q}_i = G_i Z_i + \bar{B}_i \bar{Q}_i.$$

Because of $\bar{B}_i = \bar{B}_i \bar{\Pi}_{i-1}$ we may write

$$\bar{B}_i \bar{Q}_i (Z_i - I) = \bar{B}_i \bar{\Pi}_{i-1} \bar{Q}_i (Z_i - I) = \bar{B}_i \bar{\Pi}_{i-1} \bar{Q}_i \bar{\Pi}_{i-1} (Z_i - I)$$

and using (B.2) and $Q_i = \bar{Q}_i Q_i$ we obtain $\bar{B}_i \bar{Q}_i (Z_i - I) = 0$, i.e., $\bar{B}_i \bar{Q}_i = \bar{B}_i \bar{Q}_i Z_i$. This yields

$$\bar{G}_{i+1} = (G_i + \bar{B}_i \bar{Q}_i) Z_i.$$

Derive further

$$\bar{G}_{i+1} Z_i^{-1} = G_i + \bar{B}_i \bar{Q}_i = G_{i+1} + (\bar{B}_i \bar{Q}_i - B_i Q_i)$$

and using (B.1) and $\bar{Q}_i = Q_i \bar{Q}_i$ we obtain

$$\begin{aligned} &= G_{i+1} + B_i(\bar{Q}_i - Q_i) + G_i \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i \\ &= G_{i+1} + B_i(\bar{\Pi}_{i-1} \bar{Q}_i - \Pi_{i-1} Q_i) + G_{i+1} \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i \\ &= G_{i+1} + B_i Q_i (\bar{\Pi}_{i-1} \bar{Q}_i - \Pi_{i-1} Q_i) + G_{i+1} \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il} \bar{Q}_i \\ &= G_{i+1} Y_{i+1}, \end{aligned}$$

and $\bar{G}_{i+1} = G_{i+1}Y_{i+1}Z_i = G_{i+1}Z_{i+1}$, that is, $\bar{G}_{i+1}$ and G_{i+1} are related as demanded. Next we show the invertibility of Y_{i+1} and compute the inverse. Consider the linear equation $Y_{i+1}z = w$, i.e.,

$$z + Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i)z + \sum_{l=0}^{i-1} Q_l \mathfrak{A}_{il}\bar{Q}_i z = w.$$

Because of (B.3) we immediately realize that

$$\Pi_i z = \Pi_i w, \quad z = w - (Y_{i+1} - I)\Pi_{i-1}z,$$

and

$$\Pi_{i-1}z + \Pi_{i-1}Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i)z = \Pi_{i-1}w.$$

Taking into account that

$$\begin{aligned}\Pi_{i-1}Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i) &= \Pi_{i-1}Q_i\bar{Q}_i - \Pi_{i-1}Q_i = -\Pi_{i-1}Q_i\bar{P}_i \\ &= -\Pi_{i-1}Q_i\bar{\Pi}_{i-1}\bar{P}_i = -\Pi_{i-1}Q_i\bar{P}_i\Pi_i\end{aligned}$$

we conclude

$$\Pi_{i-1}z = \Pi_{i-1}w - \Pi_{i-1}Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i)w$$

and

$$\begin{aligned}z &= w - (Y_{i+1} - I)(I - Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i))w, \\ Y_{i+1}^{-1} &= I - (Y_{i+1} - I)(I - Q_i(\bar{\Pi}_{i-1}\bar{Q}_i - \Pi_{i-1}Q_i)).\end{aligned}$$

The inverse $Z_{i+1}^{-1} = (Y_{i+1}\cdots Y_1)^{-1} = Y_1^{-1}\cdots Y_{i+1}^{-1}$ may be expressed as

$$Z_{i+1}^{-1} = I + \sum_{l=0}^{i} Q_l \mathfrak{E}_{i+1,l}$$

with certain continuous coefficients $\mathfrak{E}_{i+1,l}$. We have

$$\begin{aligned}\bar{N}_{i+1} = Z_{i+1}^{-1}N_{i+1} &\subseteq N_0 + \cdots + N_{i+1}, \\ \bar{N}_0 + \cdots + \bar{N}_{i+1} = N_0 + \cdots + N_i + \bar{N}_{i+1} &\subseteq N_0 + \cdots + N_{i+1}, \\ N_0 + \cdots + N_{i+1} = N_0 + \cdots + N_i + (Z_{i+1}\bar{N}_{i+1}) & \\ \subseteq N_0 + \cdots + N_i + \bar{N}_{i+1} &= \bar{N}_0 + \cdots + \bar{N}_{i+1},\end{aligned}$$

thus $\bar{N}_0 + \cdots + \bar{N}_{i+1} = N_0 + \cdots + N_{i+1}$. It follows that

$$D\bar{\Pi}_{i+1}\bar{D}^- = D\bar{\Pi}_{i+1}\bar{D}^- D\Pi_{i+1}D^-.$$

Now we consider the terms $\bar{B}_{i+1}$ and B_{i+1}. We have

$$\begin{aligned}
\bar{B}_{i+1} &= \bar{B}_i\bar{P}_i - \bar{G}_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\bar{\Pi}_i \\
&= \bar{B}_i\bar{P}_i - \bar{G}_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-D\Pi_{i+1}D^-)'D\bar{\Pi}_i \\
&= \bar{B}_i\bar{P}_i - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} - G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}\bar{D}^-(D\Pi_{i+1}D^-)'D\bar{\Pi}_i \\
&= \bar{B}_i\bar{P}_i - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} \\
&\quad -G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}\bar{D}^-\{(D\Pi_{i+1}D^-)'D\Pi_i - D\Pi_{i+1}D^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i\} \\
&= \bar{B}_i\bar{P}_i - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} \\
&\quad -G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i + G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i.
\end{aligned}$$

Taking into account the given result for $\bar{B}_i$ we obtain

$$\begin{aligned}
\bar{B}_{i+1} &= \{B_i - G_iZ_i\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i + G_i\sum_{l=0}^{i-1}Q_l\mathfrak{A}_{il}\}(P_i+Q_i)\bar{P}_i \\
&\quad - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} - G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i \\
&\quad + G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i \\
&= B_iP_i - G_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i + G_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i + B_iQ_i\bar{P}_i \\
&\quad - G_iZ_i\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i + G_i\sum_{l=0}^{i-1}Q_l\mathfrak{A}_{il}\bar{P}_i - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} \\
&\quad - G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}D^-(D\Pi_{i+1}D^-)'D\Pi_i + G_{i+1}Z_{i+1}\bar{\Pi}_{i+1}\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i,
\end{aligned}$$

hence

$$\bar{B}_{i+1} = B_{i+1} - G_{i+1}Z_{i+1}\bar{D}^-(D\bar{\Pi}_{i+1}\bar{D}^-)'D\Pi_{i+1} + \mathfrak{B}_{i+1}$$

with

$$\begin{aligned}
\mathfrak{B}_{i+1} &= B_iQ_i\bar{P}_i + G_i\sum_{l=0}^{i-1}Q_l\mathfrak{A}_{il}\bar{P}_i + G_{i+1}(I - Z_{i+1}\bar{\Pi}_{i+1})D^-(D\Pi_{i+1}D^-)'D\Pi_i \\
&\quad -G_{i+1}(P_iZ_i - Z_{i+1}\bar{\Pi}_{i+1})\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i.
\end{aligned}$$

It remains to show that $\mathfrak{B}_{i+1}$ can be expressed as $G_{i+1}\sum_{l=0}^{i}Q_l\mathfrak{A}_{i+1\,l}$. For this purpose we rewrite

$$\begin{aligned}
\mathfrak{B}_{i+1} &= G_{i+1}Q_i\bar{P}_i + G_{i+1}\sum_{l=0}^{i-1}Q_l\mathfrak{A}_{il}\bar{P}_i \\
&\quad +G_{i+1}(I - \bar{\Pi}_{i+1} - (Z_{i+1} - I)\bar{\Pi}_{i+1})D^-(D\Pi_{i+1}D^-)'D\Pi_i \\
&\quad -G_{i+1}(Z_i - I - Q_iZ_i + I - \bar{\Pi}_{i+1} - (Z_{i+1} - I)\bar{\Pi}_{i+1})\bar{D}^-(D\bar{\Pi}_iD^-)'D\Pi_i.
\end{aligned}$$

Take a closer look at the term $G_{i+1}(I - \bar{\Pi}_{i+1}) = G_{i+1}(\bar{Q}_{i+1} + (I - \bar{\Pi}_i)\bar{P}_{i+1})$. By means of the identity $0 = \bar{G}_{i+1}\bar{Q}_{i+1} = G_{i+1}Z_{i+1}\bar{Q}_{i+1} = G_{i+1}\bar{Q}_{i+1} +$

$G_{i+1}(Z_{i+1}-I)\bar{Q}_{i+1}$ we obtain the relation

$$G_{i+1}\bar{Q}_{i+1} = -G_{i+1}(Z_{i+1}-I)\bar{Q}_{i+1}$$

and hence

$$G_{i+1}(I-\bar{\Pi}_{i+1}) = G_{i+1}(-(Z_{i+1}-I)\bar{Q}_{i+1}+(I-\bar{\Pi}_i)\bar{P}_{i+1}).$$

This yields

$$\begin{aligned}\mathfrak{B}_{i+1} = {} & G_{i+1}Q_i\bar{P}_i + G_{i+1}\sum_{l=0}^{i-1} Q_l\mathfrak{A}_{il}\bar{P}_i \\ & + G_{i+1}\{-(Z_{i+1}-I)\bar{Q}_{i+1}+(I-\bar{\Pi}_i)\bar{P}_{i+1}) - (Z_{i+1}-I)\bar{\Pi}_{i+1}\}\times \\ & \times D^-(D\Pi_{i+1}D^-)'D\Pi_i - G_{i+1}\{Z_i - I - Q_iZ_i - (Z_{i+1}-I)\bar{Q}_{i+1} \\ & + (I-\bar{\Pi}_i)\bar{P}_{i+1} - (Z_{i+1}-I)\bar{\Pi}_{i+1}\}\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i.\end{aligned}$$

With

$$\begin{aligned}Z_{i+1}-I = \sum_{l=0}^{i} Q_l\mathfrak{C}_{i+1\,l},\ Z_i - I = \sum_{l=0}^{i-1} Q_l\mathfrak{C}_{il}, \\ I-\bar{\Pi}_i = (I-\Pi_i)(I-\bar{\Pi}_i) = Q_i + Q_{i-1}P_i + \cdots + Q_0P_1\cdots P_i)(I-\bar{\Pi}_i),\end{aligned}$$

by rearranging the terms we arrive at

$$\mathfrak{B}_{i+1} = G_{i+1}\sum_{l=0}^{i} Q_l\mathfrak{A}_{i+1\,l},$$

e.g., with

$$\begin{aligned}\mathfrak{A}_{i+1\,i} := {} & \bar{P}_i + \{-\mathfrak{C}_{i+1\,i}(\bar{Q}_{i+1}+\bar{\Pi}_{i+1}) + (I-\bar{\Pi}_i)\bar{P}_{i+1}\}D^-(D\Pi_{i+1}D^-)'D\Pi_i \\ & - \{-Z_i - \mathfrak{C}_{i+1\,i}(\bar{Q}_{i+1}+\bar{\Pi}_{i+1}) + (I-\bar{\Pi}_i)\bar{P}_{i+1}\}\bar{D}^-(D\bar{\Pi}_i\bar{D}^-)'D\Pi_i.\end{aligned}$$

It is evident that all coefficients have the required property $\mathfrak{A}_{i+1\,l} = \mathfrak{A}_{i+1\,l}\bar{\Pi}_i$.
Finally, we are done with assertions (a), (b). At the same time, we have proved the first two relations in (d).
Assertion (c) is a consequence of (a), (b) and the special form (B.2) of the nonsingular matrix function Z_i. Namely, we have $Z_i(N_0+\cdots+N_{i-1}) = N_0+\cdots+N_{i-1}$, $Z_i\bar{N}_i = N_i$, thus

$$\begin{aligned}Z_i(\bar{N}_i\cap(\bar{N}_0+\cdots+\bar{N}_{i-1})) = (Z_i\bar{N}_i)\cap(Z_i(\bar{N}_0+\cdots+N_{i-1})) = \\ N_i\cap(Z_i(N_0+\cdots+N_{i-1})) = N_i\cap(N_0+\cdots+N_{i-1}).\end{aligned}$$

The same arguments apply for obtaining the third relation in (d). □

B.2 Proof of Lemma 2.41

Lemma 2.41 *Let the DAE (2.44) with sufficiently smooth coefficients be regular with tractability index $\mu \geq 3$, and let $Q_0, \ldots, Q_{\mu-1}$ be admissible projector functions. Let $k \in \{1, \ldots, \mu-2\}$ be fixed, and let $\bar{Q}_k$ be an additional continuous projector function onto $N_k = \ker G_k$ such that $D\Pi_{k-1}\bar{Q}_k D^-$ is continuously differentiable and the inclusion $N_0 + \cdots + N_{k-1} \subseteq \ker \bar{Q}_k$ is valid. Then the following becomes true:*

(1) *The projector function sequence*

$$\begin{aligned} &\bar{Q}_0 := Q_0, \ldots, \bar{Q}_{k-1} := Q_{k-1}, \\ &\bar{Q}_k, \\ &\bar{Q}_{k+1} := Z_{k+1}^{-1} Q_{k+1} Z_{k+1}, \ldots, \bar{Q}_{\mu-1} := Z_{\mu-1}^{-1} Q_{\mu-1} Z_{\mu-1}, \end{aligned}$$

is also admissible with the continuous nonsingular matrix functions $Z_{k+1}, \ldots, Z_{\mu-1}$, determined below.

(2) *If, additionally, the projector functions $Q_0, \ldots, Q_{\mu-1}$ provide an advanced decoupling in the sense that the conditions (cf. Lemma 2.31)*

$$Q_{\mu-1*}\Pi_{\mu-1} = 0, \ldots, \; Q_{k+1*}\Pi_{\mu-1} = 0$$

are given, then also the relations

$$\bar{Q}_{\mu-1*}\bar{\Pi}_{\mu-1} = 0, \ldots, \; \bar{Q}_{k+1*}\bar{\Pi}_{\mu-1} = 0, \tag{B.5}$$

are valid, and further

$$\bar{Q}_{k*}\bar{\Pi}_{\mu-1} = (Q_{k*} - \bar{Q}_k)\Pi_{\mu-1}. \tag{B.6}$$

The matrix functions Z_i are consistent with those given in Lemma 2.12, however, for easier reading we do not access this general lemma in the proof below. In the special case given here, Lemma 2.12 yields simply $Z_0 = I, Y_1 = Z_1 = I, \ldots, Y_k = Z_k = I$, and further

$$\begin{aligned} &Y_{k+1} = I + Q_k(\bar{Q}_k - Q_k) + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{kl} \bar{Q}_k = (I + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{kl} Q_k)(I + Q_k(\bar{Q}_k - Q_k)), \\ &Z_{k+1} = Y_{k+1}, \\ &Y_j = I + \sum_{l=0}^{j-2} Q_l \mathfrak{A}_{j-1l} Q_{j-1}, \quad Z_j = Y_j Z_{j-1}, \quad j = k+2, \ldots, \mu. \end{aligned}$$

Besides the general property $\ker \bar{\Pi}_j = \ker \Pi_j$, $j = 0, \ldots, \mu-1$, which follows from Lemma 2.12, now it additionally holds that

$$\operatorname{im} \bar{Q}_k = \operatorname{im} Q_k, \quad \text{but} \quad \ker \bar{Q}_j = \ker Q_j, \quad j = k+1, \ldots, \mu-1.$$

Proof (of Lemma 2.41). **(1)** Put $\bar{Q}_i = Q_i$ for $i = 0, \dots, k-1$ such that $\bar{Q}_0, \dots, \bar{Q}_k$ are admissible by the assumptions and the following relations are valid:

$$\begin{aligned}
\Pi_k &= \Pi_k \bar{\Pi}_k,\ \bar{\Pi}_k = \bar{\Pi}_k \Pi_k,\\
\bar{Q}_k P_k &= \bar{Q}_k \Pi_k,\\
Q_k \bar{P}_k &= Q_k (I - \bar{Q}_k) = Q_k - \bar{Q}_k = \bar{Q}_k Q_k - \bar{Q}_k = -\bar{Q}_k P_k,\\
\bar{\Pi}_k &= \Pi_{k-1}(P_k + Q_k)\bar{P}_k = \Pi_k + \Pi_{k-1} Q_k \bar{P}_k = (I - \Pi_{k-1}\bar{Q}_k)\Pi_k.
\end{aligned}$$

We verify the assertion level by level by induction. Set $\bar{G}_i = G_i, Z_i = I, \bar{B}_i = B_i$, for $i = 0, \dots, k-1$, $\bar{G}_k = G_k$, $Z_k = I$, and derive

$$\begin{aligned}
\bar{B}_k &= B_{k-1}P_{k-1} - G_k D^-(D\bar{\Pi}_k D^-)' D\Pi_{k-1}\\
&= B_{k-1}P_{k-1} - G_k D^- \{D\bar{\Pi}_k D^-(D\Pi_k D^-)' + (D\bar{\Pi}_k D^-)' D\Pi_k D^-\} D\Pi_{k-1}\\
&= B_{k-1}P_{k-1} - G_k \bar{\Pi}_k D^-(D\Pi_k D^-)' D\Pi_{k-1} - G_k D^-(D\bar{\Pi}_k D^-)' D\Pi_k\\
&= B_k + G_k(I - \bar{\Pi}_k) D^-(D\Pi_k D^-)' D\Pi_{k-1} - G_k D^-(D\bar{\Pi}_k D^-)' D\Pi_k\\
&= B_k + G_k \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} - G_k D^-(D\bar{\Pi}_k D^-)' D\Pi_k,
\end{aligned}$$

where we have used $G_k \bar{Q}_k = 0$ and $I - \bar{\Pi}_k = \bar{Q}_k + Q_{k-1}\bar{P}_k + \cdots + Q_0 P_1 \cdots P_{k-1}\bar{P}_k$ and with coefficients

$$\mathfrak{A}_{k,l} = Q_l P_{l+1} \cdots P_{k-1} \bar{P}_k D^-(D\bar{\Pi}_k D^-)' D\Pi_{k-1}.$$

Next we compute

$$\begin{aligned}
\bar{G}_{k+1} &= G_k + \bar{B}_k \bar{Q}_k = G_k + B_k \bar{Q}_k + G_k \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} \bar{Q}_k\\
&= G_{k+1} + B_k(\bar{Q}_k - Q_k) + G_k \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} \bar{Q}_k = G_{k+1} Z_{k+1},
\end{aligned}$$

$$Z_{k+1} = I + Q_k(\bar{Q}_k - Q_k) + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} \bar{Q}_k = (I + \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} Q_k)(I + Q_k(\bar{Q}_k - Q_k)),$$

$$Z_{k+1}^{-1} = (I - Q_k(\bar{Q}_k - Q_k))(I - \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} Q_k) = I - Q_k(\bar{Q}_k - Q_k) - \sum_{l=0}^{k-1} Q_l \mathfrak{A}_{k,l} Q_k.$$

Put $\bar{Q}_{k+1} = Z_{k+1}^{-1} Q_{k+1} Z_{k+1} = Z_{k+1}^{-1} Q_{k+1}$ such that

$$\bar{Q}_{k+1} P_{k+1} = 0,\ \bar{Q}_{k+1} = \bar{Q}_{k+1}\Pi_{k-1},\ \Pi_k \bar{Q}_{k+1} = \Pi_k Q_{k+1},$$

$\bar\Pi_k\bar Q_{k+1} = \bar\Pi_k\Pi_k Q_{k+1}$ is continuous and $D\bar\Pi_k\bar Q_{k+1}D^- = D\bar\Pi_k D^- D\Pi_k Q_{k+1}D^-$ is continuously differentiable, and hence $\bar Q_0,\dots,\bar Q_k,\bar Q_{k+1}$ are admissible. It holds that

$$\Pi_{k+1} = \Pi_{k+1}\bar\Pi_{k+1}, \quad \bar\Pi_{k+1} = \bar\Pi_{k+1}\Pi_{k+1}, \quad \bar\Pi_{k+1} = (I - \Pi_{k-1}\bar Q_k)\Pi_{k+1}.$$

We obtain the expression

$$\bar B_{k+1} = B_{k+1} - \bar G_{k+1}D^-(D\bar\Pi_{k+1}D^-)'D\Pi_{k+1} + G_{k+1}\sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l},$$

with continuous coefficients $\mathfrak{A}_{k+1,l} = \mathfrak{A}_{k+1,l}\Pi_k = \mathfrak{A}_{k+1,l}\bar\Pi_k$, and then

$$\begin{aligned}\bar G_{k+2} &= \bar G_{k+1} + \bar B_{k+1}\bar Q_{k+1} = (G_{k+1} + \bar B_{k+1}Q_{k+1})Z_{k+1}\\ &= (G_{k+1} + B_{k+1}Q_{k+1} + G_{k+1}\sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l}Q_{k+1})Z_{k+1}\\ &= G_{k+2}(I + \sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l}Q_{k+1})Z_{k+1} =: G_{k+2}Z_{k+2},\end{aligned}$$

with the nonsingular matrix function

$$\begin{aligned}Z_{k+2} &= (I + \sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l}Q_{k+1})Z_{k+1}\\ &= I + Q_k(\bar Q_k - Q_k) + \sum_{l=0}^{k-1} Q_l\mathfrak{A}_{k,l}\bar Q_k + \sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l}Q_{k+1}\end{aligned}$$

such that

$$Z_{k+1}Z_{k+2}^{-1} = I - \sum_{l=0}^{k} Q_l\mathfrak{A}_{k+1,l}Q_{k+1}.$$

Letting $\bar Q_{k+2} = Z_{k+2}^{-1}Q_{k+2}Z_{k+2} = Z_{k+2}^{-1}Q_{k+2}$ we find

$$\begin{aligned}&Q_{k+2}\bar Q_{k+2} = Q_{k+2}, \quad \bar Q_{k+2}Q_{k+2} = \bar Q_{k+2}, \quad \bar Q_{k+2} = \bar Q_{k+2}\Pi_{k+1} = \bar Q_{k+2}\bar\Pi_{k+1},\\ &\bar\Pi_{k+1}\bar Q_{k+2} = \bar\Pi_{k+1}\Pi_{k+1}Q_{k+2}, \quad D\bar\Pi_{k+1}\bar Q_{k+2}D^- = D\bar\Pi_{k+1}D^-D\Pi_{k+1}Q_{k+2}D^-,\end{aligned}$$

so that $\bar Q_0,\dots,\bar Q_{k+2}$ are known to be admissible.
Further, we apply induction. For a certain $\kappa \ge k+2$, let, the projector functions $\bar Q_0,\dots,\bar Q_\kappa$ be already shown to be admissible and, for $i = k+2,\dots,\kappa$,

$$\bar{B}_{i-1} = B_{i-1} - \bar{G}_{i-1}D^-(D\bar{\Pi}_{i-1}D^-)'D\Pi_{i-1} + G_{i-1}\sum_{l=0}^{i-2} Q_l\mathfrak{A}_{i-1,l},$$

$$\mathfrak{A}_{i-1,l} = \mathfrak{A}_{i-1,l}\Pi_{i-2},$$

$$\bar{G}_i = G_iZ_i, \quad Z_i = (I + \sum_{l=0}^{i-2} Q_l\mathfrak{A}_{i-1,l}Q_{i-1})Z_{i-1},$$

$$\bar{Q}_i = Z_i^{-1}Q_iZ_i = Z_i^{-1}Q_i, \quad \bar{\Pi}_i = (I - \Pi_{k-1}\bar{Q}_k)\Pi_i.$$

Now we consider

$$\begin{aligned}\bar{B}_\kappa &= \bar{B}_{\kappa-1}\bar{P}_{\kappa-1} - \bar{G}_\kappa D^-(D\bar{\Pi}_\kappa D^-)'D\bar{\Pi}_{\kappa-1}\\ &= \bar{B}_{\kappa-1}P_{\kappa-1} - \bar{G}_\kappa D^-(D\bar{\Pi}_\kappa D^-)'D\Pi_\kappa - \bar{G}_\kappa\bar{\Pi}_\kappa D^-(D\Pi_\kappa D^-)'D\bar{\Pi}_{\kappa-1}\\ &= B_\kappa - \bar{G}_\kappa D^-(D\bar{\Pi}_\kappa D^-)'D\Pi_\kappa + G_\kappa\sum_{l=0}^{\kappa-2} Q_l\mathfrak{A}_{\kappa-1,l}P_{\kappa-1} + \mathfrak{C}_\kappa,\end{aligned}$$

with

$$\begin{aligned}\mathfrak{C}_\kappa &:= G_\kappa D^-(D\Pi_\kappa D^-)'D\Pi_{\kappa-1} - \bar{G}_\kappa\bar{\Pi}_\kappa D^-(D\Pi_\kappa D^-)'D\bar{\Pi}_{\kappa-1}\\ &\quad - \bar{G}_{\kappa-1}D^-(D\bar{\Pi}_{\kappa-1}D^-)'D\Pi_{\kappa-1}\\ &= G_\kappa D^-(D\Pi_\kappa D^-)'D\Pi_{\kappa-1} - \bar{G}_\kappa\bar{\Pi}_\kappa D^-\{(D\Pi_\kappa D^-)' - D\Pi_\kappa D^-(D\bar{\Pi}_{\kappa-1}D^-)'\}\times\\ &\quad \times D\Pi_{\kappa-1} - \bar{G}_{\kappa-1}D^-(D\bar{\Pi}_{\kappa-1}D^-)'D\Pi_{\kappa-1}\\ &= G_\kappa(I - Z_\kappa\bar{\Pi}_\kappa)D^-(D\Pi_\kappa D^-)'D\Pi_{\kappa-1}\\ &\quad - G_\kappa(P_{\kappa-1}Z_{\kappa-1} - Z_\kappa\bar{\Pi}_\kappa)D^-(D\bar{\Pi}_{\kappa-1}D^-)'D\Pi_{\kappa-1}.\end{aligned}$$

Regarding the relations $\Pi_\kappa Z_\kappa = \Pi_\kappa$ and $\Pi_\kappa Z_{\kappa-1} = \Pi_\kappa$ we observe that

$$\Pi_\kappa(I - Z_\kappa\bar{\Pi}_\kappa) = 0, \quad \Pi_\kappa(P_{\kappa-1}Z_{\kappa-1} - Z_\kappa\bar{\Pi}_\kappa) = 0.$$

The representation $I - \Pi_\kappa = Q_\kappa + Q_{\kappa-1}P_\kappa + \cdots + Q_0P_1\cdots P_\kappa$ admits of the expressions

$$I - Z_\kappa\bar{\Pi}_\kappa = \sum_{l=0}^{\kappa} Q_l\mathfrak{E}_{\kappa,l}, \quad P_{\kappa-1}Z_{\kappa-1} - Z_\kappa\bar{\Pi}_\kappa = \sum_{l=0}^{\kappa} Q_l\mathfrak{F}_{\kappa,l}.$$

Considering $G_\kappa Q_\kappa = 0$, this leads to the representations

$$\mathfrak{C}_\kappa = \sum_{l=0}^{\kappa-1} Q_l\{\mathfrak{E}_{\kappa,l}D^-(D\Pi_\kappa D^-)'D\Pi_{\kappa-1} - \mathfrak{F}_{\kappa,l}D^-(D\bar{\Pi}_{\kappa-1}D^-)'D\Pi_{\kappa-1}\},$$

and hence

$$\bar{B}_\kappa = B_\kappa - \bar{G}_\kappa D^-(D\bar{\Pi}_\kappa D^-)'D\Pi_\kappa + G_\kappa\sum_{l=0}^{\kappa-1} Q_l\mathfrak{A}_{\kappa,l},$$

with continuous coefficients

$$\mathfrak{A}_{\kappa,l} = \mathfrak{A}_{\kappa,l}\Pi_{\kappa-1},\ l = 0,\dots,\kappa-1.$$

It follows that

$$\begin{aligned}\bar{G}_{\kappa+1} &= \bar{G}_\kappa + \bar{B}_\kappa\bar{Q}_\kappa = G_\kappa Z_\kappa + \bar{B}_{\kappa+1}Z_\kappa^{-1}Q_\kappa Z_\kappa \\ &= \{G_\kappa + B_\kappa Q_\kappa + G_\kappa\sum_{l=0}^{\kappa-1} Q_l\mathfrak{A}_{\kappa,l}Q_\kappa\}Z_\kappa \\ &= G_{\kappa+1}\{I + \sum_{l=0}^{\kappa-1} Q_l\mathfrak{A}_{\kappa,l}Q_\kappa\}Z_\kappa =: G_{\kappa+1}Z_{\kappa+1}.\end{aligned}$$

Letting $\bar{Q}_{\kappa+1} = Z_{\kappa+1}^{-1}Q_{\kappa+1}Z_{\kappa+1} = Z_{\kappa+1}^{-1}Q_{\kappa+1}$ we find

$$\begin{aligned}\bar{Q}_{\kappa+1} &= \bar{Q}_{\kappa+1}\Pi_\kappa = \bar{Q}_{\kappa+1}\Pi_\kappa\bar{\Pi}_\kappa = \bar{Q}_{\kappa+1}\bar{\Pi}_\kappa, \\ \bar{\Pi}_\kappa\bar{Q}_{\kappa+1} &= \bar{\Pi}_\kappa\Pi_\kappa Q_{\kappa+1} \quad D\bar{\Pi}_\kappa\bar{Q}_{\kappa+1}D^- = D\bar{\Pi}_\kappa D^- D\Pi_\kappa Q_{\kappa+1}D^-,\end{aligned}$$

which shows the sequence $\bar{Q}_0,\dots,\bar{Q}_{\kappa+1}$ to be admissible and all required relations to be valid. We are done with Assertion (1).

(2) Owing to Lemma 2.31, the functions

$$\begin{aligned}Q_{\mu-1*} &= Q_{\mu-1}G_\mu^{-1}B_{\mu-1}, \\ Q_{i*} &= Q_iP_{i+1}\cdots P_{\mu-1}G_\mu^{-1}\underbrace{\{B_i + G_iD^-(D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\}}_{=:\mathfrak{B}_i}, \quad i = 1,\dots,\mu-2,\end{aligned}$$

are continuous projector-valued functions such that

$$\operatorname{im}Q_{i*} = \operatorname{im}Q_i = \ker G_i, \quad Q_{i*} = Q_{i*}\Pi_{i-1}, \quad i = 1,\dots,\mu-1.$$

Since $Q_0,\dots,Q_{\mu-1}$ are admissible, for $j = 1,\dots,\mu-2$, it holds that

$$\begin{aligned}Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_j &= Q_jP_{j+1}\cdots P_{\mu-1}P_{\mu-1}\cdots P_j = Q_jP_{j+1}\cdots P_{\mu-1}P_j \\ &= Q_jP_{j+1}\cdots P_{\mu-1} - Q_j = -Q_j(I - P_{j+1}\cdots P_{\mu-1}) \\ &= -Q_j\{Q_{j+1} + P_{j+1}Q_{j+2} + \cdots + P_{j+1}\cdots P_{\mu-2}Q_{\mu-1}\}.\end{aligned} \tag{B.7}$$

Property (B.7) immediately implies

$$Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_j = Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_j\Pi_j, \tag{B.8}$$

$$Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_j\Pi_{\mu-1} = 0, \tag{B.9}$$

$$Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_i = Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}G_j \quad \text{for} \quad i < j. \tag{B.10}$$

Analogous relations are valid also for the new sequence $\bar{Q}_0,\dots,\bar{Q}_{\mu-1}$, and, additionally,

$$\bar{Q}_j\bar{P}_{j+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_j = \bar{Q}_j\bar{P}_{j+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}G_j, \tag{B.11}$$

$$\bar{Q}_j\bar{P}_{j+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_j = \bar{Q}_j\bar{P}_{j+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_j\Pi_j. \tag{B.12}$$

Noting that $\bar{Q}_l = \bar{Q}_l Q_l$, $Q_l = Q_l\bar{Q}_l$ for $l \geq k+1$, we have further

$$\bar{Q}_j\bar{P}_{j+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_j\Pi_{\mu-1} = 0, \quad \text{for} \quad j \geq k. \tag{B.13}$$

Now, assume the projector function sequence $Q_0,\ldots,Q_{\mu-1}$ provides an already advanced decoupling such that

$$Q_{\mu-1*}\Pi_{\mu-1} = 0,\ldots,Q_{k+1*}\Pi_{\mu-1} = 0.$$

Recall that $k \leq \mu-2$. Taking into account the relation $Q_{\mu-1}G_\mu^{-1}G_{\mu-1} = Q_{\mu-1}P_{\mu-1} = 0$, we immediately conclude

$$\begin{aligned}\bar{Q}_{\mu-1*}\bar{\Pi}_{\mu-1} &= \bar{Q}_{\mu-1}\bar{G}_\mu^{-1}\bar{B}_{\mu-1}\bar{\Pi}_{\mu-1} = \bar{Q}_{\mu-1}\underbrace{Q_{\mu-1}Z_\mu^{-1}}_{=Q_{\mu-1}}G_\mu^{-1}\bar{B}_{\mu-1}\underbrace{\Pi_{\mu-2}\bar{\Pi}_{\mu-1}}_{=\Pi_{\mu-1}}\\ &= \bar{Q}_{\mu-1}Q_{\mu-1}G_\mu^{-1}B_{\mu-1}\Pi_{\mu-1} = \bar{Q}_{\mu-1}Q_{\mu-1*}\Pi_{\mu-1} = 0.\end{aligned}$$

Next, for $k \leq i \leq \mu-2$, we investigate the terms

$$\begin{aligned}\bar{Q}_{i*}\bar{\Pi}_{\mu-1} &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{\mathfrak{B}}_i\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1} + \mathfrak{D}_i,\end{aligned}$$

with $\mathfrak{D}_i := \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\{\bar{\mathfrak{B}}_i - \mathfrak{B}_i\}\bar{\Pi}_{\mu-1}$. First we show that $\mathfrak{D}_i = 0$ thanks to (B.11)–(B.13). Namely, we have by definition

$$\begin{aligned}\mathfrak{D}_i &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\{\bar{B}_i + \bar{G}_iD^-(D\bar{\Pi}_{\mu-1}D^-)'D\bar{\Pi}_{i-1} - B_i\\ &\quad - G_iD^-(D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\}\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\{-\bar{G}_iD^-(D\bar{\Pi}_iD^-)'D\Pi_i + G_i\sum_{l=0}^{i-1}Q_l\mathfrak{A}_{i,l}\\ &\quad + \bar{G}_iD^-(D\bar{\Pi}_{\mu-1}D^-)'D\bar{\Pi}_{i-1} - G_iD^-(D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\}\bar{\Pi}_{\mu-1},\end{aligned}$$

yielding

$$\begin{aligned}\mathfrak{D}_i &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_i\Pi_iD^-\{-(D\Pi_iD^- - D\Pi_{k-1}\bar{Q}_kD^-D\Pi_iD^-)'D\Pi_i\\ &\quad + (D\Pi_{\mu-1}D^- - D\Pi_{k-1}\bar{Q}_kD^-D\Pi_{\mu-1}D^-)'(D\Pi_{i-1}D^- - D\Pi_{k-1}\bar{Q}_kD^-D\Pi_{i-1})\\ &\quad - (D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\}\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_i\Pi_iD^-\{(D\Pi_{k-1}\bar{Q}_kD^-)'D\Pi_i + (D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\\ &\quad - (D\Pi_{\mu-1}D^-)'D\Pi_{k-1}\bar{Q}_kD^-D\Pi_{i-1} - (D\Pi_{k-1}\bar{Q}_kD^-)'D\Pi_{\mu-1}\\ &\quad - (D\Pi_{\mu-1}D^-)'D\Pi_{i-1}\}\bar{\Pi}_{\mu-1}.\end{aligned}$$

Due to $\Pi_i\bar{\Pi}_{\mu-1}=\Pi_{\mu-1}$ we arrive at

$$\begin{aligned}\mathfrak{D}_i &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_i\Pi_i D^-\{-(D\Pi_{\mu-1}D^-)'D\Pi_{k-1}\bar{Q}_kD^-D\Pi_{i-1}\}\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_i\Pi_i D^-D\Pi_{\mu-1}D^-(D\Pi_{k-1}\bar{Q}_kD^-)'D\Pi_{i-1}\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\bar{G}_i\Pi_{\mu-1}D^-(D\Pi_{k-1}\bar{Q}_kD^-)'D\Pi_{i-1}\bar{\Pi}_{\mu-1}=0,\end{aligned}$$

which proves the relation

$$\bar{Q}_{i*}\bar{\Pi}_{\mu-1}=\bar{Q}_i\bar{P}_{i+1}\cdots\bar{P}_{\mu-1}\bar{G}_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1} \tag{B.14}$$

for $k\le i\le\mu-2$. By means of the formula

$$Z_jZ_{j+1}^{-1}=I-\sum_{l=0}^{j-1}Q_l\mathfrak{A}_{j,l}Q_j$$

being available for $j=k+1,\ldots,\mu-1$, we rearrange the terms in (B.14) as

$$\begin{aligned}\bar{Q}_{i*}\bar{\Pi}_{\mu-1} &= \bar{Q}_iZ_{i+1}^{-1}P_{i+1}Z_{i+1}Z_{i+2}^{-1}P_{i+2}\cdots Z_{\mu-1}^{-1}P_{\mu-1}Z_{\mu-1}Z_\mu^{-1}G_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1}\\ &= \bar{Q}_iZ_{i+1}^{-1}P_{i+1}\cdots P_{\mu-1}G_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1}\\ &\quad+\sum_{j=i+1}^{\mu-2}\mathfrak{E}_{i,j}Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1}+\mathfrak{E}_{i,\mu-1}Q_{\mu-1}G_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1}.\end{aligned}$$

The very last term in this formula disappears because of

$$\begin{aligned}Q_{\mu-1}G_\mu^{-1}\mathfrak{B}_i\bar{\Pi}_{\mu-1} &= Q_{\mu-1}G_\mu^{-1}B_i\bar{\Pi}_{\mu-1}=Q_{\mu-1}G_\mu^{-1}B_{\mu-1}\bar{\Pi}_{\mu-1}\\ &= Q_{\mu-1*}(I-\Pi_{k-1}Q_k)\Pi_{\mu-1}=Q_{\mu-1*}\Pi_{\mu-1}=0.\end{aligned}$$

Next we prove the involved sum also vanishes. For this aim we consider the relation

$$(B_j-B_i)\Pi_{\mu-1}=-\sum_{l=i+1}^{j}G_lD^-(D\Pi_lD^-)'D\Pi_{\mu-1},\quad\text{for}\quad j\ge i+1. \tag{B.15}$$

We first assume $i>k$ leading to $\mathfrak{B}_i\bar{\Pi}_{\mu-1}=\mathfrak{B}_i\Pi_{i-1}\Pi_{\mu-1}=\mathfrak{B}_i\Pi_{\mu-1}$ and further

$$\begin{aligned}&Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}\mathfrak{B}_i\Pi_{\mu-1}\\ &\quad=\underbrace{Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}\mathfrak{B}_j\Pi_{\mu-1}}_{=Q_{j*}\Pi_{\mu-1}=0}+Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}(\mathfrak{B}_i-\mathfrak{B}_j)\Pi_{\mu-1}\\ &\quad=Q_jP_{j+1}\cdots P_{\mu-1}G_\mu^{-1}\{\sum_{l=i+1}^{j}G_lD^-(D\Pi_lD^-)'D\Pi_{\mu-1}\\ &\qquad\qquad+(G_j-G_i)(D\Pi_{\mu-1}D^-)'D\Pi_{\mu-1}\}.\end{aligned}$$

Applying once more the properties (B.8) and (B.10), we derive

$$
\begin{aligned}
&Q_j P_{j+1} \cdots P_{\mu-1} G_\mu^{-1} \mathfrak{B}_i \Pi_{\mu-1} \\
&\quad = Q_j P_{j+1} \cdots P_{\mu-1} G_\mu^{-1} \{ \sum_{l=i+1}^{j} G_l D^- (D\Pi_l D^-)' D\Pi_{\mu-1} \\
&\qquad\qquad + (G_j - G_i)(D\Pi_{\mu-1} D^-)' D\Pi_{\mu-1} \} \\
&\quad = Q_j P_{j+1} \cdots P_{\mu-1} G_\mu^{-1} G_j \Pi_j D^- \sum_{l=i+1}^{j} (D\Pi_l D^-)' D\Pi_{\mu-1} = 0.
\end{aligned}
$$

Now, for $i > k$, it results that

$$
\begin{aligned}
\bar{Q}_{i*} \Pi_{\mu-1} &= \bar{Q}_i Z_{i+1}^{-1} P_{i+1} \cdots P_{\mu-1} G_\mu^{-1} \mathfrak{B}_i \Pi_{\mu-1} = \bar{Q}_i Q_i P_{i+1} \cdots P_{\mu-1} G_\mu^{-1} \mathfrak{B}_i \Pi_{\mu-1} \\
&= \bar{Q}_i Q_{i*} \Pi_{\mu-1} = 0,
\end{aligned}
$$

which verifies property (B.5). By the same means one obtains

$$
\begin{aligned}
\bar{Q}_{k*} \Pi_{\mu-1} &= \underbrace{\bar{Q}_k Z_{k+1}^{-1}}_{=Q_k} P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} \mathfrak{B}_k \Pi_{\mu-1} = Q_k P_{k+1} \cdots P_{\mu-1} G_\mu^{-1} \mathfrak{B}_k \Pi_{\mu-1} \\
&= Q_{k*} \Pi_{\mu-1}.
\end{aligned}
$$

Finally, it remains to investigate the expression $\bar{Q}_{k*} \bar{\Pi}_{\mu-1}$. Since $\bar{Q}_{k*}$ also projects onto $\operatorname{im} \bar{Q}_k = \ker G_k$, it follows that $\bar{Q}_{k*} \bar{Q}_k = \bar{Q}_k$. This proves property (B.6), namely

$$
\begin{aligned}
\bar{Q}_{k*} \bar{\Pi}_{\mu-1} &= \bar{Q}_{k*} (I - \Pi_{k-1} \bar{Q}_k) \Pi_{\mu-1} = \bar{Q}_{k*} \Pi_{\mu-1} - \bar{Q}_{k*} \Pi_{k-1} \bar{Q}_k \Pi_{\mu-1} \\
&= Q_{k*} \Pi_{\mu-1} - \bar{Q}_k \Pi_{\mu-1} = (Q_{k*} - \bar{Q}_k) \Pi_{\mu-1}.
\end{aligned}
$$

□

B.3 Admissible projectors for $Nx' + x = r$

In this part, admissible projectors are generated for the DAE (B.16) with a nilpotent matrix function N typical for the normal form in the framework of strangeness index (cf. [130]). Our admissible projectors are given explicitly by formulas (B.26) below; they have upper block triangular form corresponding to the strict upper block triangular form of N.

Roughly speaking Lemma B.1 below is the technical key when proving that any DAE which has a well-defined regular strangeness index is at the same time regular in the tractability-index framework, and, in particular, the constant-rank requirements associated to the strangeness index are sufficient for the constant-rank conditions associated to the tractability index.

We deal with the special DAE

$$
Nx' + x = r, \tag{B.16}
$$

given by a matrix function $N \in C(\mathcal{I}, L(\mathbb{R}^m))$, $\mathcal{I} \subseteq \mathbb{R}$ an interval, that has strict upper block triangular structure uniform on $\mathcal{I}$

$$N = \begin{bmatrix} 0 & N_{12} & \dots & N_{1\mu} \\ & 0 & \ddots & \vdots \\ & & \ddots & \vdots \\ & & & 0 \; N_{\mu-1\mu} \\ & & & 0 \end{bmatrix} \begin{matrix} \} \ell_1 \\ \\ \\ \\ \} \ell_{\mu-1} \\ \} \ell_\mu \end{matrix},$$

$1 \le \ell_1 \le \cdots \le \ell_\mu$, $\ell_1 + \cdots + \ell_\mu = m$, $\mu \ge 2$. The blocks N_{ii+1}, $i = 1, \dots, \mu-1$, are supposed to have full row rank each, i.e.,

$$\operatorname{rank} N_{ii+1} = \ell_i, \quad i = 1, \dots, \mu-1. \tag{B.17}$$

This implies that all powers of N have constant rank, namely

$$\begin{aligned} \operatorname{rank} N &= \ell_1 + \cdots + \ell_{\mu-1}, \\ \operatorname{rank} N^k &= \ell_1 + \cdots + \ell_{\mu-k}, \quad k = 1, \dots, \mu-1, \\ \operatorname{rank} N^\mu &= 0. \end{aligned} \tag{B.18}$$

N is nilpotent with index μ, i.e., $N^{\mu-1} \neq 0$, $N^\mu = 0$. For $i = 1, \dots, \mu-1$, we introduce projectors $\mathcal{V}^{[1]}_{i+1,i+1} \in C(\mathcal{I}, L(\mathbb{R}^{\ell_{i+1}}))$ onto the continuous subspace $\ker N_{i,i+1}$, and $\mathcal{U}^{[1]}_{i+1,i+1} := I_{\ell_{i+1}} - \mathcal{V}^{[1]}_{i+1,i+1}$. $\mathcal{V}^{[1]}_{i+1,i+1}$ and $\mathcal{U}^{[1]}_{i+1,i+1}$ have constant rank $\ell_{i+1} - \ell_i$ and ℓ_i, respectively. Exploiting the structure of N we build a projector $\mathcal{V}^{[1]} \in C(\mathcal{I}, L(\mathbb{R}^m))$ onto the continuous subspace $\ker N$, which has a corresponding upper block triangular structure

$$\mathcal{V}^{[1]} = \begin{bmatrix} I & & & & \\ & \mathcal{V}^{[1]}_{22} & * & \dots & * \\ & & \ddots & \ddots & \vdots \\ & & & \ddots & * \\ & & & & \mathcal{V}^{[1]}_{\mu\mu} \end{bmatrix} \begin{matrix} \} \ell_1 \\ \\ \\ \} \ell_{\mu-1} \\ \} \ell_\mu \end{matrix}. \tag{B.19}$$

The entries indicated by "$*$" are uniquely determined by the entries of N and generalized inverses $N^-_{i,i+1}$ with

$$N^-_{i,i+1} N_{i,i+1} = \mathcal{V}^{[1]}_{i+1,i+1}, \quad N_{i,i+1} N^-_{i,i+1} = I_{\ell_i}, \quad i = 1, \dots, \mu-1.$$

In the following, we assume the nullspace $\ker N$ to be just a C^1 subspace, and the projector $\mathcal{V}^{[1]}$ to be continuously differentiable. Obviously, the property $N \in C^1(\mathcal{I}, L(\mathbb{R}^m))$ is sufficient for that but might be too generous. For this reason, we do not specify further smoothness conditions in terms of N but in terms of projectors

and subspaces.
Making use of $N = N\mathcal{U}^{[1]}$, $\mathcal{U}^{[1]} := I - \mathcal{V}^{[1]}$, we reformulate the DAE (B.16) as

$$N(\mathcal{U}^{[1]}x)' + (I - N\mathcal{U}^{[1]\prime})x = r. \tag{B.20}$$

The matrix function $N\mathcal{U}^{[1]\prime}$ is again strictly upper block triangular, and $I - N\mathcal{U}^{[1]\prime}$ is nonsingular, upper block triangular with identity diagonal blocks.

$$M_0 := (I - N\mathcal{U}^{[1]\prime})^{-1}N = \sum_{\ell=0}^{\mu-1} (N\mathcal{U}^{[1]\prime})^{\ell}N$$

has the same strict upper block triangular structure as N, the same nullspace, and entries $(M_0)_{i,i+1} = N_{i,i+1}$, $i = 1, \ldots, \mu-1$. Scaling equation (B.20) by $(I - N\mathcal{U}^{[1]\prime})^{-1}$ yields

$$M_0(\mathcal{U}^{[1]}x)' + x = q, \tag{B.21}$$

where $q := (I - N\mathcal{U}^{[1]\prime})r$. By construction, the DAE (B.21) has a properly stated leading term (cf. Definition 2.1). Written as a general linear DAE

$$A(Dx)' + Bx = q$$

with $A = M_0$, $D = \mathcal{U}^{[1]}$, $B = I$, we have $\ker A = \ker M_0 = \ker N = \ker \mathcal{U}^{[1]}$, $\operatorname{im} D = \operatorname{im}\mathcal{U}^{[1]}$, $R = \mathcal{U}^{[1]}$.
Next we choose $D^- = \mathcal{U}^{[1]}$, and, correspondingly $P_0 = \mathcal{U}^{[1]}$, $Q_0 = \mathcal{V}^{[1]}$. With these projectors, $\Pi_0 = P_0$, and $G_0 = AD = M_0\mathcal{U}^{[1]} = M_0$, $B_0 = I$, we form a matrix function sequence and admissible projectors $Q_0, \ldots, Q_\kappa$ for the DAE (B.21) as described in Section 2.2.2. In particular, we shall prove this DAE to be regular with tractability index μ.
The first matrix function (cf. Section 2.2.2) G_1 is

$$G_1 = M_0 + Q_0,$$

and $G_1 z = 0$, i.e., $(M_0 + Q_0)z = 0$, leads to $P_0M_0z = 0$, $Q_0z = -Q_0M_0P_0z$, $z = (I - Q_0M_0)P_0z$, $z \in \ker P_0M$. Because of $P_0M_0 = M_0^- M_0M_0$, $M_0^2 = M_0P_0M_0$ the nullspaces of P_0M_0 and M_0^2 coincide. The inclusion $\ker M_0 \subset \ker M_0^2 = \ker P_0M_0$ allows for the decomposition $\ker M_0^2 = \ker M_0 \oplus P_0 \ker M_0^2$. If $\mathcal{V}^{[2]}$ denotes a projector onto $\ker M_0^2$, $\mathcal{U}^{[2]} := I - \mathcal{V}^{[2]}$, then it follows that

$$\begin{aligned}
\operatorname{im}\mathcal{V}^{[2]} &= \operatorname{im}\mathcal{V}^{[1]} \oplus \operatorname{im}\mathcal{U}^{[1]}\mathcal{V}^{[2]},\\
\mathcal{V}^{[2]}\mathcal{V}^{[1]} &= \mathcal{V}^{[1]}, \quad (\mathcal{U}^{[1]}\mathcal{U}^{[2]})^2 = \mathcal{U}^{[1]}\mathcal{U}^{[2]},\\
(\Pi_0\mathcal{V}^{[2]})^2 &= \Pi_0\mathcal{V}^{[2]},\\
\operatorname{rank}\mathcal{U}^{[2]} &= \operatorname{rank} M_0^2 = \ell_1 + \cdots + \ell_{\mu-2},\\
\operatorname{rank}\mathcal{V}^{[2]} &= \ell_{\mu-1} + \ell_\mu,\\
\operatorname{rank}\Pi_0\mathcal{V}^{[2]} &= \operatorname{rank}\mathcal{V}^{[2]} - \operatorname{rank}\mathcal{V}^{[1]} = \ell_{\mu-1}.
\end{aligned}$$

The matrix function

$$Q_1 := (I - Q_0 M_0)\Pi_0 \mathcal{V}^{[2]} \tag{B.22}$$

has the properties

$$Q_1 Q_0 = (I - Q_0 M_0)\Pi_0 \mathcal{V}^{[2]} \mathcal{V}^{[1]} = (I - Q_0 M_0)\Pi_0 \mathcal{V}^{[1]} = (I - Q_0 M_0)\Pi_0 Q_0 = 0,$$

hence $Q_1 \cdot Q_1 = Q_1$, and

$$\begin{aligned} G_1 Q_1 &= (M_0 + Q_0)(I - Q_0 M_0)\Pi_0 \mathcal{V}^{[2]} = (M_0 - Q_0 M_0 + Q_0)\Pi_0 \mathcal{V}^{[2]} \\ &= P_0 M_0 \Pi_0 \mathcal{V}^{[2]} = P_0 M_0 \mathcal{V}^{[2]} = 0. \end{aligned}$$

It becomes clear that Q_1 is actually the required projector onto $\ker G_1$, if $\operatorname{rank} Q_1 = m - \operatorname{rank} G_1$. $I - Q_0 M_0$ is nonsingular, and Q_1 has the same rank as $\Pi_0 \mathcal{V}^{[2]}$, that is, $\operatorname{rank} Q_1 = \ell_{\mu-1}$. Proposition 2.5(3) allows for an easy rank determination of the matrix function G_1. With

$$\mathcal{W}_0 := \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & I \end{bmatrix} \begin{matrix} \\ \\ \\ \} \ \ell_\mu \end{matrix}$$

we find $\operatorname{im} G_1 = \operatorname{im} G_0 \oplus \operatorname{im} \mathcal{W}_0 B_0 Q_0 = \operatorname{im} M_0 \oplus \operatorname{im} \mathcal{W}_0 Q_0$, thus $r_1 = r_0 + \operatorname{rank} \mathcal{V}^{[1]}_{\mu\mu} = m - \ell_\mu + \ell_\mu - \ell_{\mu-1} = m - \ell_{\mu-1}$. It turns out that Q_0, Q_1 are admissible, supposing $\pi_1 = \mathcal{U}^{[1]}\mathcal{U}^{[2]}$ is continuously differentiable.
Next, due to the structure of M_0^2, the projector $\mathcal{V}^{[2]}$ can be chosen to be upper block triangular,

$$\mathcal{V}^{[2]} = \begin{bmatrix} I & & & \\ & I & & \\ & & * \cdots * & \\ & & \ddots \ \vdots & \\ & & & * \end{bmatrix}, \quad \mathcal{U}^{[2]} = I - \mathcal{V}^{[2]} = \begin{bmatrix} 0 & & & \\ & 0 & & \\ & & * \cdots * & \\ & & \ddots \ \vdots & \\ & & & * \end{bmatrix}.$$

The entries in the lower right corners play their role in rank calculations. They are

$$\mathcal{V}^{[2]}_{\mu\mu} = I - \mathcal{U}^{[2]}_{\mu\mu}, \quad \mathcal{U}^{[2]}_{\mu\mu} = (N_{\mu-2,\mu-1} N_{\mu-1,\mu})^- N_{\mu-2,\mu-1} N_{\mu-1,\mu}.$$

To realize this we just remember that the entry $(\mu-2, \mu)$ of M_0^2 is $[M_0^2]_{\mu-2,\mu} = N_{\mu-2,\mu-1} N_{\mu-1,\mu}$. Both $N_{\mu-2,\mu-1}$ and $N_{\mu-1,\mu}$ have full row rank $\ell_{\mu-2}$, respectively $\ell_{\mu-1}$. Therefore, the product $N_{\mu-2,\mu-1} N_{\mu-1,\mu}$ has full row rank equal to $\ell_{\mu-2}$. From this it follows that

$$\operatorname{rank} \mathcal{V}^{[2]}_{\mu\mu} = \dim \ker N_{\mu-2,\mu-1} N_{\mu-1,\mu} = \ell_\mu - \ell_{\mu-2}.$$

Taking into account the inclusion

$$\operatorname{im}\mathcal{V}^{[1]}_{\mu\mu} = \ker N_{\mu-1,\mu} \subseteq \ker N_{\mu-2,\mu-1}N_{\mu-1,\mu} = \operatorname{im}\mathcal{V}^{[2]}_{\mu\mu}$$

we find

$$\operatorname{rank}\mathcal{U}^{[1]}_{\mu\mu}\mathcal{V}^{[2]}_{\mu\mu} = \operatorname{rank}\mathcal{V}^{[2]}_{\mu\mu} - \operatorname{rank}\mathcal{V}^{[1]}_{\mu\mu} = \ell_{\mu-1} - \ell_{\mu-2}.$$

By Proposition 2.5(3), with the projector along $\operatorname{im} G_1$

$$\mathcal{W}_1 := \begin{bmatrix} 0 & & \\ & \ddots & \\ & & 0 \\ & & \mathcal{U}^{[1]}_{\mu\mu} \end{bmatrix}, \quad \mathcal{W}_1 = \mathcal{W}_0\mathcal{U}^{[1]},$$

we compute (before knowing G_2 in detail)

$$\operatorname{im} G_2 = \operatorname{im} G_1 \oplus \operatorname{im}\mathcal{W}_1Q_1, \quad \mathcal{W}_1Q_1 = \mathcal{W}_0\mathcal{U}^{[1]}\mathcal{V}^{[2]},$$
$$r_2 = r_1 + \operatorname{rank}\mathcal{W}_1Q_1 = r_1 + \operatorname{rank}\mathcal{U}^{[1]}_{\mu\mu}\mathcal{V}^{[2]}_{\mu\mu} = m - \ell_{\mu-1} + \ell_{\mu-1} - \ell_{\mu-2} = m - \ell_{\mu-2}.$$

We compute $G_2 = G_1 + (B_0\Pi_0 - G_1D^-(D\Pi_1D^-)'D\Pi_0)Q_1$ (cf. Section 2.2.2) itself as

$$\begin{aligned} G_2 &= M_0 + Q_0 + \Pi_0Q_1 - (M_0 + Q_0)P_0\Pi_1'\Pi_0Q_1 \\ &= M_0 + Q_0 + \Pi_0Q_1 - M_0F_1\Pi_0Q_1, \end{aligned}$$

where $F_1 := P_0\Pi_1'\Pi_0Q_1$ is upper block triangular as are all its factors. It follows that

$$G_2 = M_0 + Q_0 + (I - M_0F_1)P_0(I - \Pi_1),$$

and G_2 is upper block triangular. Due to the nonsingularity of $I - M_0F_1$, as well as the simple property $(I - M_0F_1)Q_0 = Q_0$, we may use the description

$$G_2 = (I - M_0F_1)^{-1}\{M_1 + I - \Pi_1\},$$

where $M_1 := (I - M_0F_1)^{-1}M_0$ again has the strict upper block triangular structure of N, and entries $[M_1]_{ii+1} = N_{ii+1}$, $i = 1, \ldots, \mu - 1$. From the representation

$$\begin{aligned} \Pi_1M_1 &= \Pi_1P_0M_1 = \Pi_1P_0(I + M_0F_1 + \cdots + (M_0F_0)^{\mu-1})M_0 \\ &= \Pi_1(I + M_0F_1 + \cdots + (M_0F_1)^{\mu-1})P_0M_0 \end{aligned}$$

we know the inclusion $\ker\Pi_0M_0 \subseteq \ker\Pi_1M_1$ to be valid. Furthermore, we have $\ker M_0^2M_1 = \ker\Pi_1M_1$ because of the representations $\ker\mathcal{U}^{[2]} = \ker M_0^2 = \ker P_0M_0$, $\Pi_1M_1 = P_0\mathcal{U}^{[2]}M_1 = P_0(M_0^2)^-M_0^2M_1$, and $M_0^2M_1 = M_0^2\mathcal{U}^{[2]}M_1 = M_0^2P_0\mathcal{U}^{[2]}M_1 = M_0^2\Pi_1M_1$.

The next lemma shows that we may proceed further in this way to construct admissible projectors for the DAE (B.21). We shall use certain auxiliary continuous matrix functions which are determined from level to level as

$$
\begin{aligned}
F_0 &:= 0,\\
F_i &:= F_{i-1} + \sum_{\ell=1}^{i} P_0\Pi_\ell'\Pi_{i-1}Q_i = \sum_{j=1}^{i}\sum_{\ell=1}^{j} P_0\Pi_\ell'\Pi_{j-1}Q_i, \quad i\ge 1, \qquad \text{(B.23)}\\
H_2 &:= H_1 := H_0 := 0,\\
H_i &:= H_{i-1} + \sum_{\ell=2}^{i-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{i-1}Q_i\\
&= \sum_{j=3}^{i}\sum_{\ell=2}^{j-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{j-1}Q_j, \quad i\ge 3. \qquad \text{(B.24)}
\end{aligned}
$$

These matrix functions inherit the upper block triangular structure. They disappear if the projectors $\Pi_1,\ldots,\Pi_i$ do not vary with time (what is given at least in the constant coefficient case).
It holds that $F_i = F_iP_0$, $H_i = H_iP_0$. The products F_iM_0 are strictly upper block triangular so that $I-M_0F_i$ is nonsingular, and

$$M_i := (I-M_0F_i)^{-1}M_0 \qquad \text{(B.25)}$$

again has strict upper block triangular structure. The entries $(j,j+1)$ of M_i coincide with those of N, i.e.,

$$[M_i]_{j,j+1} = N_{j,j+1}. \qquad \text{(B.26)}$$

If the projectors $\Pi_0,\ldots,\Pi_i$ are constant, then we simply have $M_i = M_0 = N$.

Lemma B.1. *Let N be sufficiently smooth so that the continuous projectors Π_i arising below are even continuously differentiable. Let $k\in\mathbb{N}$, $k\le\mu-1$, and let $Q_0 := \mathcal{V}^{[1]}$ be given by (B.19), and, for $i=1,\ldots,k$,*

$$Q_i := \Big(I - \sum_{j=0}^{i-1} Q_j(I-H_{i-1})^{-1}M_{i-1}\Big)\Pi_{i-1}\mathcal{V}^{[i+1]}, \qquad \text{(B.27)}$$

$\mathcal{V}^{[i+1]}\in C(\mathcal{I},L(\mathbb{R}^m))$ an upper block triangular projector onto $\ker M_0^2M_1\cdots M_{i-1}$, $\mathcal{U}^{[i+1]} := I-\mathcal{V}^{[i+1]}$. Then, the matrix functions $Q_0,\ldots,Q_k$ are admissible projectors for the DAE (B.21) on $\mathcal{I}$, and, for $i=1,\ldots,k$, it holds that

$$
\begin{aligned}
&\Pi_{i-1}Q_i = \Pi_{i-1}\mathcal{V}^{[i+1]}, \quad \Pi_i = \mathcal{U}^{[1]}\cdots\mathcal{U}^{[i+1]}, && \text{(B.28)}\\
&\ker\Pi_{i-1}M_{i-1}\subset\ker\Pi_iM_i, && \text{(B.29)}\\
&\ker\Pi_iM_i = \ker M_0^2M_1\ldots M_i, && \text{(B.30)}\\
&G_{i+1} = M_0+Q_0+(I-M_0F_i)(I-H_i)P_0(I-\Pi_i), && \text{(B.31)}\\
&r_{i+1} = \operatorname{rank}G_{i+1} = m-\ell_{\mu-i-1}, \quad \operatorname{im}G_{i+1} = \operatorname{im}G_i\oplus\operatorname{im}\mathcal{W}_0\Pi_{i-1}Q_i,
\end{aligned}
$$

and $I - H_i$ is nonsingular.

Before we turn to the proof of Lemma B.1 we realize that it provides admissible projectors $Q_0, \dots, Q_{\mu-1}$ and characteristics $r_0 = m - \ell_\mu, \dots, r_{\mu-1} = m - \ell_1 < m$. Because of the strict upper block triangular structure of $M_0, \dots, M_{\mu-2}$, the product $M_0^2 M_1 \cdots M_{\mu-2}$ disappears (as N^μ does). This leads to $\mathcal{V}^{[\mu]} = I$, $\mathcal{U}^{[\mu]} = 0$, thus $\Pi_{\mu-1} = 0$, and

$$\begin{aligned} G_\mu &= M_0 + Q_0 + (I - M_0 F_{\mu-1})(I - H_{\mu-1}) P_0 (I - \Pi_{\mu-1}) \\ &= M_0 + Q_0 + (I - M_0 F_{\mu-1})(I - H_{\mu-1}) P_0 \\ &= (I - M_0 F_{\mu-1})(I - H_{\mu-1})\{(I - H_{\mu-1})^{-1} M_{\mu-1} + I\}. \end{aligned}$$

The factors $I - M_0 F_{\mu-1}$ and $I - H_{\mu-1}$ are already known to be nonsingular. $(I - H_{\mu-1})^{-1} M_{\mu-1}$ inherits the strict upper block triangular structure from $M_{\mu-1}$, but then $I + (I - H_{\mu-1})^{-1} M_{\mu-1}$ is nonsingular, and so is G_μ. Hence we have proved an important consequence of Lemma B.1:

Proposition B.2. *Let N be sufficiently smooth to make the continuous projectors $\Pi_0, \dots, \Pi_{\mu-2}$ even continuously differentiable. Then the DAE (B.21) is on $\mathcal{I}$ regular with tractability index μ and characteristic values*

$$r_i = m - \ell_{\mu-i}, \quad i = 0, \dots, \mu - 1, \quad r_\mu = m.$$

It holds that $\Pi_{\mu-1} = 0$, and there is no inherent regular ODE within the DAE.

To prepare the proof of Lemma B.1 we give the following lemma

Lemma B.3. *Let $\mathcal{V}_i \in L(\mathbb{R}^m)$ be idempotent, $\mathcal{U}_i := I - \mathcal{V}_i$, $L_i := \operatorname{im} \mathcal{V}_i$, $\nu_i := \operatorname{rank} \mathcal{V}_i$, $i = 1, \dots, k$, and $L_i \subseteq L_{i+1}$, $i = 1, \dots, k-1$.*
Then the products $\mathcal{U}_1\mathcal{V}_2, \dots, \mathcal{U}_1 \cdots \mathcal{U}_{k-1}\mathcal{V}_k, \mathcal{U}_1\mathcal{U}_2, \dots, \mathcal{U}_1 \cdots \mathcal{U}_k$ are projectors, too, and it holds that

$$\begin{aligned} &\mathcal{U}_1 \cdots \mathcal{U}_i \mathcal{V}_{i+1} \mathcal{V}_j = 0, \quad 1 \le j \le i, \quad i = 1, \dots, k-1, \\ &\ker \mathcal{U}_1 \cdots \mathcal{U}_i = L_i, \quad i = 1, \dots, k, \\ &L_k = L_1 \oplus \mathcal{U}_1 L_2 \oplus \cdots \oplus \mathcal{U}_1 \cdots \mathcal{U}_{k-1} L_k, \\ &\dim \mathcal{U}_1 \cdots \mathcal{U}_{k-1} L_k = \nu_k - \nu_{k-1}. \end{aligned} \tag{B.32}$$

Proof. The inclusions $L_1 \subseteq L_2 \subseteq \cdots \subseteq L_{i+1}$ lead to $\mathcal{V}_{i+1}\mathcal{V}_j = \mathcal{V}_j$, for $j = 1, \dots, i$. Compute

$$\begin{aligned} \mathcal{U}_1\mathcal{V}_2\mathcal{U}_1\mathcal{V}_2 &= \mathcal{U}_1\mathcal{V}_2(I - \mathcal{V}_1)\mathcal{V}_2 = \mathcal{U}_1\mathcal{V}_2 - \mathcal{U}_1\mathcal{V}_1\mathcal{V}_2 = \mathcal{U}_1\mathcal{V}_2, \\ \mathcal{U}_1\mathcal{U}_2\mathcal{U}_1\mathcal{U}_2 &= \mathcal{U}_1(I - \mathcal{V}_2)(I - \mathcal{V}_1)\mathcal{U}_2 = \mathcal{U}_1(I - \mathcal{V}_1 - \mathcal{V}_2 + \mathcal{V}_1)\mathcal{U}_2 = \mathcal{U}_1\mathcal{U}_2. \end{aligned}$$

$L_2 = \operatorname{im} \mathcal{V}_2 \subseteq \ker \mathcal{U}_1\mathcal{U}_2$ holds trivially. $z \in \ker \mathcal{U}_1\mathcal{U}_2$ means $(I - \mathcal{V}_1)(I - \mathcal{V}_2)z = 0$, hence $z = \mathcal{V}_1 z + \mathcal{V}_2 z - \mathcal{V}_1\mathcal{V}_2 z \in L_2$, so that $\ker \mathcal{U}_1\mathcal{U}_2 = L_2$ is true.
By induction, if $\mathcal{U}_1 \cdots \mathcal{U}_{i-1}Q_i$, $\mathcal{U}_1 \cdots \mathcal{U}_i$ are projectors, $\ker \mathcal{U}_1 \cdots \mathcal{U}_i = L_i$, then these properties remain valid for $i+1$ instead of i. Namely,

$$\begin{aligned}\mathcal{U}_1\cdots\mathcal{U}_{i+1}\mathcal{U}_1\cdots\mathcal{U}_{i+1} &= \mathcal{U}_1\cdots\mathcal{U}_i(I-\mathcal{V}_{i+1})\mathcal{U}_1\cdots\mathcal{U}_{i+1}\\ &= \mathcal{U}_1\cdots\mathcal{U}_i\mathcal{U}_1\cdots\mathcal{U}_{i+1} = \mathcal{U}_1\cdots\mathcal{U}_{i+1},\end{aligned}$$

$$\begin{aligned}\mathcal{U}_1\cdots\mathcal{U}_i\mathcal{V}_{i+1}\mathcal{U}_1\cdots\mathcal{U}_i\mathcal{V}_{i+1} &= \mathcal{U}_1\cdots\mathcal{U}_i\mathcal{V}_{i+1},\\ L_{i+1} = \ker\mathcal{U}_{i+1} &\subseteq \ker\mathcal{U}_1\cdots\mathcal{U}_{i+1},\end{aligned}$$

and $z\in\ker\mathcal{U}_1\cdots\mathcal{U}_{i+1}$ implies $\mathcal{U}_{i+1}z\in\mathrm{im}\mathcal{U}_1\cdots\mathcal{U}_i = L_i$, $z-\mathcal{V}_{i+1}z\in L_i$, hence $z\in L_i+L_{i+1}=L_{i+1}$. Now we can decompose

$$\begin{aligned}L_2 &= L_1\oplus\mathcal{U}_1L_2,\\ L_3 &= L_1\oplus\mathcal{U}_1L_2\oplus\mathcal{U}_1\mathcal{U}_2L_3 = L_2\oplus\mathcal{U}_1\mathcal{U}_2L_3,\\ L_{i+1} &= \underbrace{L_1\oplus\mathcal{U}_1L_2\oplus\cdots\oplus}_{=L_i}\mathcal{U}_1\cdots\mathcal{U}_iL_{i+1} = L_i\oplus\mathcal{U}_1\cdots\mathcal{U}_iL_{i+1},\end{aligned}$$

and it follows that $\dim\mathcal{U}_1\cdots\mathcal{U}_iL_{i+1}=v_{i+1}-v_i$, $i=1,\ldots,k-1$. □

Proof (of Lemma B.1). We apply induction. For $k=1$ the assertion is already proved, and the corresponding projector Q_1 is given by (B.22).
Let the assertion be true up to level k. We are going to show its validity for level $k+1$. We stress once more that we are dealing with structured triangular matrices.
We already know that $Q_0,\ldots,Q_k$ are admissible, and, in particular, it holds that $Q_iQ_j=0$, for $0\le j<i\le k$. A closer look at the auxiliary matrix functions H_i (cf. (B.24)) shows that $H_iQ_1=0$, $H_iQ_2=0$, further $H_i\Pi_i=0$, and $\Pi_{i-2}H_i=0$.
Namely, $\Pi_1H_3=\Pi_1P_0(I-\Pi_1)\Pi_2'\Pi_1Q_2=0$, and $\Pi_{j-3}H_{j-1}=0$, for $j\le i$, implies $\Pi_{i-2}H_i=0$ (due to $\Pi_{i-2}H_\ell=0$, $\Pi_{i-2}P_0(I-\Pi_{\ell-1})=0$, $\ell=1,\ldots,i-1$).
The functions $F_1,\ldots,F_k$ (cf. (B.23)) are well-defined, and they have the properties

$$(F_k-F_j)\Pi_k=0,\quad (F_k-F_j)\Pi_j=F_k-F_j,\quad \text{for } j=1,\ldots,k. \tag{B.33}$$

It follows that, for $j=1,\ldots,k$,

$$(I-M_0F_k)^{-1}(I-M_0F_j)=I+(I-M_0F_k)^{-1}M_0(F_k-F_j)\Pi_j.$$

Next we verify the property

$$\Pi_{j-1}M_kQ_j=0,\quad j=0,\ldots,k. \tag{B.34}$$

From $G_jQ_j=0$, $j=0,\ldots,k$, we know

$$M_0Q_j+Q_0Q_j+(I-M_0F_{j-1})(I-H_{j-1})P_0(I-\Pi_{j-1})Q_j=0. \tag{B.35}$$

Multiplication by $(I-M_0F_k)^{-1}$ leads to

$$M_kQ_j+Q_0Q_j+\{I+(I-M_0F_k)^{-1}M_0(F_k-F_{j-1})\Pi_{j-1}\}(I-H_{j-1})P_0(I-\Pi_{j-1})Q_j=0,$$

and further, taking account of $\Pi_{j-1}H_{j-1}=0$, $\Pi_{j-1}P_0(I-\Pi_{j-1})=0$,

$$M_k Q_j + Q_0 Q_j + (I - H_{j-1}) P_0 (I - \Pi_{j-1}) Q_j = 0, \tag{B.36}$$

and hence $\Pi_{j-1} M_k Q_j = 0$, i.e., (B.34). Now it follows that $\Pi_k M_k Q_j = 0$, for $j = 0, \ldots, k$, hence

$$\Pi_k M_k = \Pi_k M_k \Pi_k, \tag{B.37}$$

a property that will appear to be very helpful.
Recall that we already have a nonsingular $I - H_k$, as well as

$$\begin{aligned} G_{k+1} &= M_0 + Q_0 + (I - M_0 F_k)(I - H_k) P_0 (I - \Pi_k) \\ &= (I - M_0 F_k)(I - H_k)\{(I - H_k)^{-1} M_k + I - \Pi_k\}, \end{aligned} \tag{B.38}$$

and G_{k+1} has rank $r_{k+1} = m - \ell_{\mu-k-1}$. We have to show the matrix function

$$Q_{k+1} := \Big(I - \sum_{j=0}^{k} Q_j (I - H_k)^{-1} M_k\Big) \Pi_k \mathcal{V}^{[k+2]}$$

to be a suitable projector. We check first whether $G_{k+1} Q_{k+1} = 0$ is satisfied. Derive (cf. (B.38))

$$\begin{aligned} G_{k+1} Q_{k+1} &= (I - M_0 F_k)\{M_k + (I - H_k)(I - \Pi_k)\}\Big(I - \sum_{j=0}^{k} Q_j (I - H_k)^{-1} M_k\Big) \Pi_k \mathcal{V}^{[k+2]} \\ &= (I - M_0 F_k)\Big\{M_k - \sum_{j=1}^{k} M_k Q_j (I - H_k)^{-1} M_k \\ &\qquad - (I - H_k) \sum_{j=0}^{k} Q_i (I - H_k)^{-1} M_k\Big\} \Pi_k \mathcal{V}^{[k+2]} \\ &= (I - M_0 F_k)\Big\{I - H_k - \sum_{j=1}^{k} M_k Q_j - (I - H_k) \sum_{j=0}^{k} Q_j\Big\} \times \\ &\qquad \times (I - H_k)^{-1} M_k \Pi_k \mathcal{V}^{[k+2]}. \end{aligned} \tag{B.39}$$

From (B.36) we obtain, for $j = 1, \ldots, k$,

$$\begin{aligned} M_k Q_j + (I - H_k) Q_j &= -Q_0 Q_j - (I - H_{j-1}) P_0 (I - \Pi_{j-1}) Q_j + (I - H_k) Q_j \\ &= P_0 Q_j - H_k Q_j - (I - H_{j-1})(I - \Pi_{j-1}) P_0 Q_j \\ &= P_0 Q_j - H_k Q_j - (I - H_{j-1}) P_0 Q_j + \Pi_{j-1} Q_j \\ &= -(H_k - H_{j-1}) P_0 Q_j + \Pi_{j-1} Q_j \end{aligned}$$

and, therefore,

$$\sum_{j=1}^{k}(M_kQ_j + (I-H_k)Q_j) = \sum_{j=1}^{k}\Pi_{j-1}Q_j - \sum_{j=1}^{k}(H_k - H_{j-1})P_0Q_j$$
$$= \sum_{j=1}^{k}\Pi_{j-1}Q_j - H_k.$$

The last relation becomes true because of $(H_k - H_0)Q_1 = 0$, $(H_k - H_1)Q_2 = 0$, and the construction of H_i (cf. (B.24)),

$$\sum_{j=1}^{k}(H_k - H_{j-1})P_0Q_j = \sum_{j=3}^{k}(H_k - H_{j-1})P_0Q_j$$
$$= \sum_{j=3}^{k}\Big[\sum_{\nu=j}^{k}\sum_{\ell=2}^{\nu-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{\nu-1}Q_\nu\Big]P_0Q_j$$
$$= \sum_{j=3}^{k}\sum_{\ell=2}^{j-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{j-1}Q_j = H_k.$$

Together with (B.39) this yields

$$G_{k+1}Q_{k+1} = (I-M_0F_k)\Big\{I - H_k - \Big(\sum_{j=1}^{k}\Pi_{j-1}Q_j - H_k\Big) - Q_0\Big\}(I-H_k)^{-1}M_k\Pi_k\mathcal{V}^{[k+2]}$$
$$= (I-M_0F_k)\Big\{I - Q_0 - \sum_{j=1}^{k}\Pi_{j-1}Q_j\Big\}(I-H_k)^{-1}M_k\Pi_k\mathcal{V}^{[k+2]}$$
$$= (I-M_0F_k)\Pi_k(I-H_k)^{-1}M_k\Pi_k\mathcal{V}^{[k+2]}. \tag{B.40}$$

For more specific information on $(I-H_k)^{-1}$ we consider the equation $(I-H_k)z = w$, i.e. (cf. (B.24))

$$(I-H_{k-1})z - \sum_{\ell=2}^{k-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{k-1}Q_kz = w. \tag{B.41}$$

Because of $\Pi_{k-1}H_{k-1} = 0$, $\Pi_{k-1}H_{\ell-1} = 0$, $\Pi_{k-2}P_0(I-\Pi_{\ell-1}) = 0$, multiplication of (B.41) by $\Pi_{k-1}Q_k = \Pi_{k-1}Q_k\Pi_{k-1}$ yields $\Pi_{k-1}Q_kz = \Pi_{k-1}Q_kw$, such that

$$z = (I-H_{k-1})^{-1}\Big\{w + \sum_{\ell=2}^{k-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{k-1}Q_kw\Big\}$$

results, and further,

$$(I-H_k)^{-1} = (I-H_{k-1})^{-1}\Big(I - \sum_{\ell=2}^{k-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{k-1}Q_k\Big)$$

$$= (I-H_3)^{-1}\Big(I+\sum_{\ell=2}^{3}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_3Q_4\Big)\times\cdots$$
$$\cdots\times\Big(I+\sum_{\ell=2}^{k-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{k-1}Q_k\Big)$$
$$= (I+P_0Q_1\Pi_2'\Pi_2Q_3)\times\cdots\times\Big(I+\sum_{\ell=2}^{k-1}(I-H_{\ell-1})P_0(I-\Pi_{\ell-1})\Pi_\ell'\Pi_{k-1}Q_k\Big).$$

This shows that $\Pi_k(I-H_k)^{-1}=\Pi_k$ holds true. On the other hand $F_k\Pi_k=0$ is also given, which leads to

$$G_{k+1}Q_{k+1}=\Pi_kM_k\Pi_k\mathcal{V}^{[k+2]}.$$

With the help of (B.37), and taking into account that $\ker\Pi_kM_k=\ker M_0^2M_1\cdots M_k$, we arrive at

$$G_{k+1}Q_{k+1}=\Pi_kM_k\mathcal{V}^{[k+2]}=0,$$

that is, the matrix function Q_{k+1} satisfies the condition $\operatorname{im}Q_{k+1}\subseteq\ker G_{k+1}$. The inclusions (cf. (B.29), (B.30))

$$\ker\Pi_{i-1}M_{i-1}=\ker M_0^2M_1\cdots M_{i-1}\subset\ker\Pi_iM_i=\ker M_0^2M_1\cdots M_i$$

are valid for $i=1,\ldots,k$. This leads to

$$\operatorname{im}\mathcal{V}^{[1]}\subset\operatorname{im}\mathcal{V}^{[2]}\subset\cdots\subset\mathcal{V}^{[\kappa+2]}$$

which allows an application of Lemma B.3. We make use of the structural properties

$$\operatorname{rank}M_0^2M_1\cdots M_i=\operatorname{rank}N^{i+2}=\ell_1+\cdots+\ell_{\mu-i-2},$$
$$\operatorname{rank}\mathcal{V}^{[i+2]}=m-(\ell_1+\cdots+\ell_{\mu-i-2})=\ell_{\mu-i-1}+\cdots+\ell_\mu,$$

so that Lemma B.3 yields

$$\operatorname{rank}\mathcal{U}^{[1]}\cdots\mathcal{U}^{[k+1]}\mathcal{V}^{[k+2]}=\operatorname{rank}\mathcal{V}^{[k+2]}-\operatorname{rank}\mathcal{V}^{[k+1]}=\ell_{\mu-k-1}.$$

Writing Q_{k+1} in the form

$$Q_{k+1}=\Big(I-\sum_{j=0}^{k}Q_j(I-H_k)^{-1}M_k\Pi_k\Big)\Pi_k\mathcal{V}^{[k+2]},$$

and realizing that the first factor is nonsingular, we conclude

$$\operatorname{rank}Q_{k+1}=\operatorname{rank}\Pi_k\mathcal{V}^{[k+2]}=\ell_{\mu-k-1}=m-\operatorname{rank}G_{k+1}.$$

Applying Lemma B.3 again we derive, for $j=0,\ldots,k$,

$$Q_{k+1}Q_j = \Big(I - \sum_{j=0}^{k} Q_j(I-H_k)^{-1}M_k\Big)\Pi_k\mathcal{V}^{[k+2]}Q_j,$$

$$\begin{aligned}\Pi_k\mathcal{V}^{[k+2]}Q_j &= \mathcal{U}^{[1]}\cdots\mathcal{U}^{[k+1]}\mathcal{V}^{[k+2]}Q_j\\ &= \mathcal{U}^{[1]}\cdots\mathcal{U}^{[k+1]}\mathcal{V}^{[k+2]}\mathcal{U}^{[1]}\cdots\mathcal{U}^{[j]}Q_j\\ &= \mathcal{U}^{[1]}\cdots\mathcal{U}^{[k+1]}\mathcal{V}^{[k+2]}\mathcal{U}^{[1]}\cdots\mathcal{U}^{[j]}\mathcal{V}^{[j+1]} = 0,\end{aligned}$$

such that $Q_{k+1}Q_j = 0$, $j = 0,\ldots,k$, and furthermore $Q_{k+1}Q_{k+1} = Q_{k+1}$. This completes the proof that Q_{k+1} is a suitable projector function, and that $Q_0,\ldots,Q_k,Q_{k+1}$ are admissible.
It remains to verify (B.29)–(B.31) for $i = k+1$, to consider the rank of G_{k+2} as well as to show the nonsingularity of $I - H_{k+1}$.
First we consider the rank of G_{k+2}. Following Proposition 2.5(3) it holds that

$$\operatorname{im} G_{k+2} = \operatorname{im} G_{k+1} \oplus \operatorname{im} \mathcal{W}_{k+1}\Pi_k Q_{k+1},$$

with a projector $\mathcal{W}_{k+1}$ such that $\ker \mathcal{W}_{k+1} = \operatorname{im} G_{k+1}$. Because of

$$\begin{aligned}\operatorname{im} G_{k+1} &= \operatorname{im} G_k \oplus \operatorname{im} \mathcal{W}_0\Pi_{k-1}Q_k\\ &= \operatorname{im} G_0 \oplus \operatorname{im} \mathcal{W}_0Q_0 \oplus \cdots \oplus \operatorname{im} \mathcal{W}_0\Pi_{k-1}Q_k\\ &= \operatorname{im} G_0 \oplus \operatorname{im} \mathcal{W}_0(Q_0 + \cdots + \Pi_{k-1}Q_k)\\ &= \operatorname{im} G_0 \oplus \operatorname{im} \mathcal{W}_0(I - \Pi_k)\end{aligned}$$

we may choose the projector

$$\mathcal{W}_{k+1} = \mathcal{W}_0\Pi_k = \mathcal{W}_0\Pi_k\mathcal{W}_0.$$

This leads to

$$\operatorname{im} G_{k+2} = \operatorname{im} G_{k+1} \oplus \operatorname{im} \mathcal{W}_0\Pi_k Q_{k+1},$$

as well as to

$$\begin{aligned}r_{k+2} &= r_{k+1} + \operatorname{rank} \mathcal{W}_0\Pi_k Q_{k+1} = r_{k+1} + \operatorname{rank}\,[\Pi_k Q_{k+1}]_{\mu\mu}\\ &= r_{k+1} + \operatorname{rank} \mathcal{U}_{\mu\mu}^{[1]}\cdots\mathcal{U}_{\mu\mu}^{[k+1]}\mathcal{V}_{\mu\mu}^{[k+2]} = m - \ell_{\mu-k-1} + (\ell_{\mu-k-1} - \ell_{\mu-k-2})\\ &= m - \ell_{\mu-k-2}.\end{aligned}$$

Therefore, to show that $\operatorname{rank} \mathcal{U}_{\mu\mu}^{[1]}\cdots\mathcal{U}_{\mu\mu}^{[k+1]}\mathcal{V}_{\mu\mu}^{[k+2]} = \ell_{\mu-k-1} - \ell_{\mu-k-2}$ we recall that

$$\begin{aligned}&\mathcal{V}_{\mu\mu}^{[1]} \text{ projects onto } \ker N_{\mu-1,\mu},\\ &\mathcal{V}_{\mu\mu}^{[2]} \text{ projects onto } \ker N_{\mu-2,\mu-1}N_{\mu-1,\mu},\\ &\cdots\\ &\mathcal{V}_{\mu\mu}^{[k+1]} \text{ projects onto } \ker N_{\mu-k-1,\mu-k}\cdots N_{\mu-1,\mu}\end{aligned}$$

and

$$\mathcal{V}_{\mu\mu}^{[k+2]} \text{ projects onto } \ker N_{\mu-k-2,\mu-k-1}\cdots N_{\mu-1,\mu},$$

and

$$\begin{aligned}&\operatorname{im}\mathcal{V}_{\mu\mu}^{[1]} \subset \operatorname{im}\mathcal{V}_{\mu\mu}^{[2]} \subset \cdots \subset \operatorname{im}\mathcal{V}_{\mu\mu}^{[k+2]},\\ &\operatorname{rank}\mathcal{V}_{\mu\mu}^{[i]} = \ell_\mu - \ell_{\mu-i}, \quad i = 1,\ldots,k+2.\end{aligned}$$

Here, Lemma B.3 applies again, and it follows that

$$\begin{aligned}&\operatorname{rank}\mathcal{U}_{\mu\mu}^{[1]}\cdots\mathcal{U}_{\mu\mu}^{[k+1]}\mathcal{V}_{\mu\mu}^{[k+2]} = \operatorname{rank}\mathcal{V}_{\mu\mu}^{[k+2]} - \operatorname{rank}\mathcal{V}_{\mu\mu}^{[k+1]}\\ &= \ell_\mu - \ell_{\mu-k-2} - (\ell_\mu - \ell_{\mu-k-1}) = \ell_{\mu-k-1} - \ell_{\mu-k-2}.\end{aligned}$$

So we are done with the range and rank of G_{k+2}.
In the next step we provide G_{k+2} itself (cf. Section 2.2.2). Compute

$$\begin{aligned}G_{k+2} &= G_{k+1} + \Pi_k Q_{k+1} - \sum_{j=1}^{k+1} G_j P_0 \Pi_j' \Pi_k Q_{k+1}\\ &= M_0 + Q_0 + (I - M_0F_k)(I - H_k)P_0(I - \Pi_k) + \Pi_k Q_{k+1} - M_0\Pi_1'\Pi_k Q_{k+1}\\ &\quad - \sum_{j=2}^{k+1}\{M_0 + (I - M_0F_{j-1})(I - H_{j-1})P_0(I - \Pi_{j-1})\}\Pi_j'\Pi_k Q_{k+1}\\ &= M_0 + Q_0 + (I - M_0F_k)P_0(I - \Pi_k) - (I - M_0F_k)H_k + \Pi_k Q_{k+1}\\ &\quad - \sum_{j=1}^{k+1} M_0\Pi_j'\Pi_k Q_{k+1} - \sum_{j=2}^{k+1}(I - M_0F_{j-1})(I - H_{j-1})P_0(I - \Pi_{j-1})\Pi_j'\Pi_k Q_{k+1}\end{aligned}$$

and rearrange (cf. (B.23), (B.24)) certain terms to

$$(I - M_0F_k)P_0(I - \Pi_k) + \Pi_k Q_{k+1} - M_0\sum_{j=1}^{k+1} P_0\Pi_j'\Pi_k Q_{k+1} = (I - M_0F_{k+1})P_0(I - \Pi_{k+1})$$

and

$$\begin{aligned}&(I - M_0F_k)H_k + \sum_{j=2}^{k}(I - M_0F_{j-1})(I - H_{j-1})P_0(I - \Pi_{j-1})\Pi_j'\Pi_k Q_{k+1}\\ &= (I - M_0F_k)\Big\{H_k + \sum_{j=2}^{k}(I - M_0F_k)^{-1}(I - M_0F_{j-1})(I - H_{j-1}) \times\\ &\qquad\qquad \times P_0(I - \Pi_{j-1})\Pi_j'\Pi_k Q_{k+1}\Big\}\\ &= (I - M_0F_k)\Big\{H_k + \sum_{j=2}^{k}(I - H_{j-1})P_0(I - \Pi_{j-1})\Pi_j'\Pi_k Q_{k+1}\Big\}\\ &= (I - M_0F_k)H_{k+1} = (I - M_0F_{k+1})(I - M_0F_{k+1})^{-1}(I - M_0F_k)H_{k+1}\end{aligned}$$

$$= (I - M_0F_{k+1})H_{k+1} = (I - M_0F_{k+1})H_{k+1}P_0(I - \Pi_{k+1}),$$

which leads to

$$\begin{aligned} G_{k+2} &= M_0 + Q_0 + (I - M_0F_{k+1})P_0(I - \Pi_{k+1}) - (I - M_0F_{k+1})H_{k+1}P_0(I - \Pi_{k+1}) \\ &= M_0 + Q_0 + (I - M_0F_{k+1})(I - H_{k+1})P_0(I - \Pi_{k+1}), \end{aligned}$$

and we are done with G_{k+2} (cf. (B.31)).
Next, $I - H_{k+1}$ is nonsingular, since $(I - H_{k+1})z = 0$ implies $\Pi_k Q_{k+1} z = 0$, thus $(I - H_k)z = 0$, and, finally $z = 0$ due to the nonsingularity of $(I - H_k)$.
To complete the proof of Lemma B.1 we have to verify (B.29) and (B.30) for $i = k + 1$, supposing $\ker \Pi_{k-1}M_{k-1} \subseteq \ker \Pi_k M_k$, $\ker \Pi_k M_k = \ker M_0^2 M_1 \cdots M_k$ are valid. From $\Pi_k M_k = \Pi_k M_k \Pi_k$ (cf. (B.37)) and $\ker M_0^2 M_1 \cdots M_k = \ker \Pi_k M_k = \ker \mathcal{U}^{[k+2]}$ we obtain the relations

$$\begin{aligned} \Pi_{k+1}M_{k+1} &= \Pi_k \mathcal{U}^{[k+2]} M_{k+1} = \Pi_k (M_0^2 M_1 \cdots M_k)^- M_0^2 M_1 \cdots M_k M_{k+1}, \\ M_0^2 M_1 \cdots M_{k+1} &= M_0^2 M_1 \cdots M_k \mathcal{U}^{[k+2]} M_{k+1} \\ &= M_0^2 M_1 \cdots M_k (\Pi_k M_k)^- \Pi_k M_k \mathcal{U}^{[k+2]} M_{k+1} \\ &= M_0^2 M_1 \cdots M_k (\Pi_k M_k)^- \Pi_k M_k \Pi_k \mathcal{U}^{[k+2]} M_{k+1} \\ &= M_0^2 M_1 \cdots M_k (\Pi_k M_k)^- \Pi_k M_k \Pi_{k+1} M_{k+1}, \end{aligned}$$

hence $\ker \Pi_{k+1}M_{k+1} = \ker M_0^2 M_1 \cdots M_{k+1}$ holds true. Additionally, from

$$\begin{aligned} \Pi_{k+1}M_{k+1} &= \Pi_{k+1}(I - M_0F_{k+1})^{-1}(I - M_0F_k)M_k \\ &= \Pi_{k+1}[I + (I - M_0F_{k+1})^{-1}M_0(F_{k+1} - F_k)\Pi_k]M_k \\ &= \Pi_{k+1}[I + (I - M_0F_{k+1})^{-1}M_0(F_{k+1} - F_k)\Pi_k]\Pi_k M_k \end{aligned}$$

we conclude the inclusion

$$\ker \Pi_k M_k \subseteq \ker \Pi_{k+1}M_{k+1}.$$

□

Appendix C
Analysis

C.1 A representation result

Proposition C.1. *Let the function $d : \Omega \times \mathcal{I} \to \mathbb{R}^n$, $\Omega \subseteq \mathbb{R}^m$ open, $\mathcal{I} \subseteq \mathbb{R}$ an interval, be continuously differentiable, and let the partial Jacobian $d_x(x,t)$ have constant rank r on $\Omega \times \mathcal{I}$. Let $x_* : \mathcal{I}_* \to \mathbb{R}^m$, $\mathcal{I}_* \subseteq \mathcal{I}$, be a continuous function with values in Ω, i.e., $x_*(t) \in \Omega$, $t \in \mathcal{I}_*$. Put $u_*(t) := d(x_*(t),t)$, $t \in \mathcal{I}_*$.*
Then, if u_* is continuously differentiable the inclusion

$$u_*'(t) - d_t(x_*(t),t) \in \operatorname{im} d_x(x_*(t),t), \quad t \in \mathcal{I}_* \tag{C.1}$$

is valid, and there exists a continuous function $w_* : \mathcal{I}_* \to \mathbb{R}^m$ such that

$$u_*'(t) - d_t(x_*(t),t) = d_x(x_*(t),t)w_*(t), \quad t \in \mathcal{I}_*. \tag{C.2}$$

If $d_x(x_*(t),t)$ is injective, then $w_*(t)$ is uniquely determined by (C.2).

Proof. Derive, for $t \in \mathcal{I}_*$,

$$\begin{aligned} u_*'(t) &= \lim_{\tau \to 0} \frac{1}{\tau}(d(x_*(t+\tau),t+\tau) - d(x_*(t),t)) \\ &= \lim_{\tau \to 0} \Big\{ \frac{1}{\tau}(d(x_*(t+\tau),t+\tau) - d(x_*(t),t+\tau)) \\ &\qquad + \frac{1}{\tau}(d(x_*(t),t+\tau) - d(x_*(t),t)) \Big\}. \end{aligned}$$

Since the expression $\frac{1}{\tau}(d(x_*(t),t+\tau) - d(x_*(t),t))$ has, for $\tau \to 0$, the limit $d_t(x_*(t),t)$, the other difference must possess a limit, too, i.e.

$$e_*(t) := \lim_{\tau \to 0} \frac{1}{\tau}(d(x_*(t+\tau),t+\tau) - d(x_*(t),t+\tau))$$

is well-defined, and $u_*'(t) = e_*(t) + d_t(x_*(t),t)$. Rewrite, for fixed $t \in \mathcal{I}_*$,

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5,

$$\begin{aligned} e_*(t) &= \lim_{\tau\to 0}\int_0^1 d_x(x_*(t)+s(x_*(t+\tau)-x_*(t)),t+\tau)ds\,\frac{1}{\tau}(x_*(t+\tau)-x_*(t)) \\ &=: \lim_{\tau\to 0} E(\tau)g(\tau), \end{aligned}$$

with

$$\begin{aligned} E(\tau) &:= \int_0^1 d_x(x_*(t)+s(x_*(t+\tau)-x_*(t)),t+\tau))ds, \\ g(\tau) &:= \frac{1}{\tau}(x_*(t+\tau)-x_*(t)), \quad \tau\in(-\rho,\rho),\ \rho>0 \text{ small.} \end{aligned}$$

Recall that $g(\tau)$ has not necessarily a limit for $\tau\to 0$, but we can make use of the existing limits $\lim_{\tau\to 0} E(\tau)g(\tau) = e_*(t)$ and $\lim_{\tau\to 0} E(\tau) = d_x(x_*(t),t) = E(0)$. The matrix $E(\tau)\in L(\mathbb{R}^m,\mathbb{R}^n)$ depends continuously on τ, and $E(0)$ has rank r, so that, at least for all sufficiently small τ, it holds that $\operatorname{rank} E(\tau)\geq r$. On the other hand, for all sufficiently small τ and $z\in\mathbb{R}^m$, the decomposition

$$[\operatorname{im} d_x(x_*(t),t)]^\perp \oplus \operatorname{im} d_x(x_*(t)+z,t+\tau) = \mathbb{R}^n \tag{C.3}$$

is valid. If $\mathcal{V}_*\in L(\mathbb{R}^n)$ denotes the projector onto $[\operatorname{im} d_x(x_*(t),t)]^\perp$ according to the decomposition (C.3), then we have, for all sufficiently small τ, that

$$\mathcal{V}_* E(\tau) = \int_0^1 \mathcal{V}_* d_x(x_*(t)+s(x_*(t+\tau)-x_*(t)),t+\tau)ds = 0.$$

$\mathcal{V}_*$ has rank $n-r$, hence $\mathcal{V}_* E(\tau)=0$ implies $\operatorname{rank} E(\tau)\leq r$ for all τ being sufficiently small.
Now, $E(\tau)$ is, for small τ, a constant-rank matrix, so that $E(\tau)^+$ and $\mathcal{U}(\tau) := E(\tau)E(\tau)^+$ are also continuous in τ. This leads to

$$\begin{aligned} e_*(t) &= \lim_{\tau\to 0} E(\tau)g(\tau) = \lim_{\tau\to 0}\mathcal{U}(\tau)E(\tau)g(\tau) \\ &= \lim_{\tau\to 0}\mathcal{U}(\tau)\cdot\lim_{\tau\to 0} E(\tau)g(\tau) = \mathcal{U}(0)\cdot e_*(t), \end{aligned}$$

which means $e_*(t)\in\operatorname{im} d_x(x_*(t),t)$, or, equivalently,

$$u_*'(t) - d_t(x_*(t),t)\in\operatorname{im} d_x(x_*(t),t), \quad t\in\mathcal{I}_*,$$

that is, we are done with the inclusion (C.1). Next, taking any continuous reflexive generalized inverse $d_x(x,t)^-$ to $d_x(x,t)$, the function $w_*:\mathcal{I}_*\to\mathbb{R}^m$,

$$w_*(t) := d_x(x_*(t),t)^-(u_*'(t)-d_t(x_*(t),t)), \quad t\in\mathcal{I}^*,$$

is continuous and satisfies

$$\begin{aligned} d_x(x_*(t),t)w_*(t) &= d_x(x_*(t),t)d_x(x_*(t),t)^-(u_*'(t)-d_t(x_*(t),t)) \\ &= u_*'(t)-d_t(x_*(t),t), \end{aligned}$$

since (C.1) is valid, and $d_x(x_*(t),t)d_x(x_*(t),t)^-$ is a projector onto $\operatorname{im} d_x(x_*(t),t)$. Finally, (C.2) together with (C.1) define $w_*(t)$ uniquely, if $d_x(x_*(t),t)$ is injective, since then $d_x(x_*(t),t)^- d_x(x_*(t),t) = I$, independently of the special choice of the generalized inverse. □

Notice that one can also choose $w_*(t) = x_*'(t)$ to satisfy (C.2) supposing x_* is known to be continuously differentiable.

C.2 ODEs

Proposition C.2. *Let the function $g \in C(\mathcal{I},\mathbb{R}^m)$, $\mathcal{I} = [0,\infty)$, satisfy the one-sided Lipschitz condition*

$$\langle g(x,t) - g(\bar{x},t), x-\bar{x}\rangle \le \gamma |x-\bar{x}|^2,\ x,\bar{x} \in \mathbb{R}^m,\ t \in \mathcal{I}, \tag{C.4}$$

with a constant $\gamma \le 0$.
Then the ODE

$$x'(t) = g(x(t),t) \tag{C.5}$$

has the following properties:

(1) *The IVP for (C.5) with the initial condition*

$$x(t_0) = x_0,\ t_0 \in \mathcal{I},\ x_0 \in \mathbb{R}^m,$$

is uniquely solvable, and the solution is defined on $\mathcal{I}$.

(2) *Each pair of solutions $x(.)$, $\bar{x}(.)$ satisfies the inequality*

$$|x(t) - \bar{x}(t)| \le e^{\gamma t} |x(0) - \bar{x}(0)|,\ t \in \mathcal{I}.$$

(3) *The ODE has at most one stationary solution.*

Proof. (1), (2): Let $x(.)$, $\bar{x}(.)$ be arbitrary solutions defined on $[0,\tau)$, with $\tau > 0$. Derive for $t \in [0,\tau)$:

$$\begin{aligned}\frac{d}{dt}|x(t)-\bar{x}(t)|^2 &= 2\langle g(x(t),t) - g(\bar{x}(t),t), x(t)-\bar{x}(t)\rangle \\ &\le 2\gamma |x(t)-\bar{x}(t)|^2.\end{aligned}$$

By means of Gronwall's lemma we find

$$|x(t) - \bar{x}(t)| \le e^{\gamma t} |x(0) - \bar{x}(0)|,\ t \in [0,\tau). \tag{C.6}$$

The inequality

$$|x(t)| - |x(0) - \bar{x}(0)| \le |\bar{x}(t)| \le |x(t)| + |x(0) - \bar{x}(0)|,\ t \in [0,\tau)$$

is a particular consequence of (C.6). It shows that $x(t)$ grows unboundedly for $t \to \tau$ if $\bar{x}(t)$ does, and vice versa.
This means that all solutions of the ODE can simultaneously be continued through τ or not.
Assume that $\tau_* > 0$ exists such that all IVPs for (C.5) and $x(0) = x_0$ have solutions $x(.,x_0)$ defined on $[0,\tau_*)$, but $x(t,x_0)$ grows unboundedly, if $t \to \tau_*$. Fix $x_* \in \mathbb{R}^m$ and put $x_{**} := x(\frac{1}{2}\tau_*, x_*)$.
The solution $x(.,x_{**})$ is also defined on $[0,\tau_*)$, in particular at $t = \frac{1}{2}\tau_*$. However, this contradicts the property $x(\frac{1}{2}\tau_*, x_{**}) = x(\tau_*, x_*)$. In consequence, such a value τ_* does not exist, and all solutions can be continued on the infinite interval.
(3): For two stationary solutions c and $\bar{c}$, (2) implies

$$|c - \bar{c}| \le e^{\gamma t}|c - \bar{c}| \to 0 \quad (t \to \infty),$$

and hence $c = \bar{c}$. □

Lemma C.3. *Given a real valued $m \times m$ matrix C, then:*

(1) *If all eigenvalues of C have strictly negative real parts, then there exist a constant $\beta < 0$ and an inner product $\langle \cdot, \cdot \rangle$ for $\mathbb{R}^m$, such that*

$$\langle Cz, z \rangle \le \beta |z|^2, \text{ for all } z \in \mathbb{R}^m, \tag{C.7}$$

and vice versa.

(2) *If all eigenvalues of C have nonpositive real parts, and the eigenvalues on the imaginary axis are nondefective, then there is an inner product $\langle \cdot, \cdot \rangle$ for $\mathbb{R}^m$, such that*

$$\langle Cz, z \rangle \le 0, \text{ for all } z \in \mathbb{R}^m, \tag{C.8}$$

and vice versa.

Proof. Let $\sigma_1, \ldots, \sigma_m \in \mathbb{C}$ denote the eigenvalues of C, and let $T \in L(\mathbb{C}^m)$ be the transformation into Jordan canonical form J such that, with entries $\delta_1, \ldots, \delta_{m-1}$ being 0 or 1,

$$J = T^{-1}CT = \begin{bmatrix} \sigma_1 & \delta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \delta_{m-1} \\ & & & \sigma_m \end{bmatrix}$$

is given. For $\varepsilon > 0$, we form further

$$J_\varepsilon = D_\varepsilon^{-1} J D_\varepsilon = \begin{bmatrix} \sigma_1 & \varepsilon\delta_1 & & \\ & \ddots & \ddots & \\ & & \ddots & \varepsilon\delta_{m-1} \\ & & & \sigma_m \end{bmatrix}, \quad D_\varepsilon = \begin{bmatrix} \varepsilon & & & \\ & \varepsilon^2 & & \\ & & \ddots & \\ & & & \varepsilon^m \end{bmatrix}.$$

J_ε and C are similar, $J_\varepsilon = D_\varepsilon^{-1}T^{-1}CTD_\varepsilon = (TD_\varepsilon)^{-1}C(TD_\varepsilon)$.
By $\langle z,y\rangle_\varepsilon := \langle (TD_\varepsilon)^{-1}z,(TD_\varepsilon)^{-1}y\rangle_2$ and $|z|_\varepsilon := |(TD_\varepsilon)^{-1}z|_2,\ z,y\in\mathbb{C}^m$, we introduce an inner product and the corresponding norm for C^m. Moreover, the expression

$$a_\varepsilon(u,v) := \operatorname{Re}\langle (TD_\varepsilon)^{-1}u,(TD_\varepsilon)^{-1}v\rangle_2,\ u,v\in\mathbb{R}^m,$$

defines an inner product for $\mathbb{R}^m$.
Recall that the relation

$$\operatorname{Re}\langle Mz,z\rangle_2 = \langle \frac{1}{2}(M+M^*)z,z\rangle_2 \le \lambda_{max}(\frac{1}{2}(M+M^*))|z|_2^2,\ z\in\mathbb{C}^m,$$

is valid for each arbitrary matrix $M\in L(\mathbb{C}^m)$, and, in particular,

$$\operatorname{Re}\langle J_\varepsilon z,z\rangle_2 \le \lambda_{max}(\frac{1}{2}(J_\varepsilon+J_\varepsilon^*))|z|_2^2,\ z\in\mathbb{R}^m.$$

We have

$$\frac{1}{2}(J_\varepsilon+J_\varepsilon^*) = \begin{bmatrix} \operatorname{Re}\sigma_1 & \frac{\varepsilon}{2}\delta_1 & & \\ \frac{\varepsilon}{2}\delta_1 & \ddots & \ddots & \\ & \ddots & \ddots & \frac{\varepsilon}{2}\delta_{m-1} \\ & & \frac{\varepsilon}{2}\delta_{m-1} & \operatorname{Re}\sigma_m \end{bmatrix} \xrightarrow[\varepsilon\to 0]{} \begin{bmatrix} \operatorname{Re}\sigma_1 & & \\ & \ddots & \\ & & \operatorname{Re}\sigma_m \end{bmatrix}$$

and $\lambda_{max}(\frac{1}{2}(J_\varepsilon+J_\varepsilon^*)) \xrightarrow[\varepsilon\to 0]{} \max\limits_{i=1,\dots,m} \operatorname{Re}\sigma_i$.
If all eigenvalues of C have strictly negative real parts, that is, $\max\limits_{i=1,\dots,m} \operatorname{Re}\sigma_i =: 2\beta < 0$, then choose a value $\varepsilon>0$ such that $\lambda_{max}(\frac{1}{2}(J_\varepsilon+J_\varepsilon^*)) \le \beta < 0$.
If the eigenvalues $\sigma_1,\dots,\sigma_m$ have zero real part, but these eigenvalues are nondefective, and $\operatorname{Re}\sigma_j < 0,\ j=s,\dots,m$, then it holds that

$$\frac{1}{2}(J_\varepsilon+J_\varepsilon^*) = \left[\begin{array}{ccc|c} 0 & & & \\ & \ddots & & \\ & & 0 & \\ \hline & & & \frac{1}{2}(\check{J}_\varepsilon+\check{J}_\varepsilon^*) \end{array}\right] \xrightarrow[\varepsilon\to 0]{} \left[\begin{array}{ccc|c} 0 & & & \\ & \ddots & & \\ & & 0 & \\ \hline & & & \operatorname{Re}\sigma_s \\ & & & & \ddots \\ & & & & & \operatorname{Re}\sigma_m \end{array}\right],$$

$$\check{J}_\varepsilon = \begin{bmatrix} \sigma_s & \delta_s & & \\ & \ddots & \ddots & \\ & & \ddots & \delta_{m-1} \\ & & & \sigma_m \end{bmatrix} \in L(\mathbb{C}^{m-s+1}).$$

Now we fix an $\varepsilon>0$ such that $\lambda_{max}(\frac{1}{2}(\check{J}_\varepsilon+\check{J}_\varepsilon^*)) \le 0,\ \lambda_{max}(\frac{1}{2}(J_\varepsilon+J_\varepsilon^*)) = 0$.
In both cases it results that

$$\begin{aligned} a_\varepsilon(Cx,x) &= \mathrm{Re}\,\langle (TD_\varepsilon)^{-1}Cx, (TD_\varepsilon)^{-1}x\rangle_2 \\ &= \mathrm{Re}\,\langle \underbrace{(TD_\varepsilon)^{-1}C(TD_\varepsilon)}_{J_\varepsilon}(TD_\varepsilon)^{-1}x, (TD_\varepsilon)^{-1}x\rangle_2 \\ &\le \lambda_{max}(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*))|(TD_\varepsilon)^{-1}x|_2^2 \\ &= \lambda_{max}(\frac{1}{2}(J_\varepsilon + J_\varepsilon^*))|x|_\varepsilon^2, \end{aligned}$$

and hence the inequalities (C.7) and (C.8) are proved to follow from the given properties of C.
The converse assertions become evident if one considers the homogeneous ODE $x'(t) = Cx(t)$. All its solutions satisfy

$$\frac{d}{dt}|x(t)|^2 = 2\langle Cx(t), x(t)\rangle \le 2\beta|x(t)|^2,\ t \ge 0,$$

thus $|x(t)| \le e^{\beta t}|x(0)|$, $t \ge 0$ in the first case, and

$$\frac{d}{dt}|x(t)|^2 = 2\langle Cx(t), x(t)\rangle \le 0,\ t \ge 0,$$

thus $|x(t)| \le |x(0)|$, $t \ge 0$ in the second case. □

C.3 Basics for evolution equations

This section summarizes basic spaces and their properties for the treatment of evolution equations (see, e.g., [217]).

1. Dual space. Let V be a real Banach space. Then, V^* denotes the set of all linear continuous functionals on V, i.e., the set of all linear continuous maps $f : V \to \mathbb{R}$. Furthermore,
$$\langle f, v\rangle := f(v) \qquad \text{for all} \quad v \in V$$
and
$$\|f\|_{V^*} := \sup_{\|v\|_V \le 1} |\langle f, v\rangle|.$$
In this way, V^* becomes a real Banach space. It is called the dual space to V.
2. Reflexive Banach space. Let V be a real Banach space. Then, V is called *reflexive* if $V = V^{**}$.
3. Evolution triple. The spaces $V \subseteq H \subseteq V^*$ are called an evolution triple if
   (i) V is a real, separable, and reflexive Banach space,
   (ii) H is a real, separable Hilbert space,
   (iii) the embedding $V \subseteq H$ is continuous, i.e.,

$$\|v\|_H \leq \text{const}\|v\|_V \qquad \text{for all} \quad v \in V,$$

and V is dense in H.

Below, Proposition C.5 explains how the inclusion $H \subseteq V^*$ is to be understood.

4. The Lebesgue space $L_p(t_0,T;V)$ of vector valued functions. Let V be a Banach space, $1 < p < \infty$, and $t_0 < T < \infty$. The space $L_p(t_0,T;V)$ consists of all measurable functions $v:\ (t_0,T) \to V$ for which

$$\|v\|_p := \left(\int_{t_0}^{T} \|v(t)\|_V^p \, \mathrm{d}t \right)^{\frac{1}{p}} < \infty.$$

The dual space of $L_p(t_0,T;V)$ is given by $L_q(t_0,T;V^*)$ where $p^{-1} + q^{-1} = 1$.

5. Generalized derivatives. Let X and Y be Banach spaces. Furthermore, let $u \in L_1(t_0,T;X)$ and $w \in L_1(t_0,T;Y)$. Then, the function w is called the generalized derivative of the function u on (t_0,T) if

$$\int_{t_0}^{T} \varphi'(t)u(t) \, \mathrm{d}t = -\int_{t_0}^{T} \varphi(t)w(t) \, \mathrm{d}t \qquad \text{for all} \quad \varphi \in C_0^\infty(t_0,T).$$

The last equation includes the requirement that the integrals on both sides belong to $X \cap Y$.

6. The Sobolev space $W_2^1(t_0,T;V,H)$. Let $V \subseteq H \subseteq V^*$ be an evolution triple and $t_0 < T < \infty$. Then, the Sobolev space

$$W_2^1(t_0,T;V,H) := \{u \in L_2(t_0,T;V):\ u' \in L_2(t_0,T;V^*)\}$$

forms a Banach space with the norm

$$\|u\|_{W_2^1} = \|u\|_{L_2(t_0,T;V)} + \|u'\|_{L_2(t_0,T;V^*)}.$$

The following proposition is a consequence of the Riesz theorem.

Proposition C.4. *Let H be a Hilbert space. Then for each $u \in H$, there is a unique linear continuous functional Ju on V with*

$$\langle Ju, v\rangle = (u|v) \qquad \textit{for all} \quad u,v \in V,$$

where $(\cdot|\cdot)$ denotes the scalar product of H. The operator $J: V \to V^*$ is linear, bijective, and norm isomorphic, i.e.,

$$\|Ju\|_{V^*} = \|u\|_V \qquad \textit{for all} \quad u \in V.$$

Therefore, one can identify Ju with u for all $u \in V$. This way we get $H = H^*$ and

$$\langle u, v\rangle = (u|v) \qquad \text{for all} \quad u,v \in V.$$

The next proposition explains how the relation $H \subseteq V^*$ is to be understood.

Proposition C.5. *Let $V \subseteq H \subseteq V^*$ be an evolution triple. Then, the following is satisfied*

(i) *To each $u \in H$, there corresponds a linear continuous functional $\bar{u} \in V^*$ with*

$$\langle \bar{u}, v \rangle_V = (u|v)_H \quad \text{for all} \quad v \in V.$$

(ii) *The mapping $u \mapsto \bar{u}$ from H into V^* is linear, injective, and continuous.*

Proof. (i) Let $u \in H$. Then:

$$|(u|v)_H| \leq \|u\|_H \|v\|_H \leq \text{const} \|u\|_H \|v\|_V$$

is fulfilled for all $v \in V$. Therefore, there exists a $\bar{u} \in V^*$ with

$$\langle \bar{u}, v \rangle_V = (u|v)_H \qquad \text{and} \qquad \|\bar{u}\|_{V^*} \leq \text{const} \|u\|_H.$$

(ii) The mapping $u \mapsto \bar{u}$ is obviously linear and continuous. In order to show injectivity, we assume that $\bar{u} = 0$. This implies

$$(u|v)_H = 0 \qquad \text{for all} \quad v \in V.$$

Since V is dense in H, we get $u = 0$. □

This allows us to identify $\bar{u}$ with u such that

$$\begin{aligned} \langle u, v \rangle_V &= (u|v)_H \qquad \text{for all} \quad u \in H,\ v \in V, \\ \|u\|_{V^*} &\leq \text{const} \|u\|_H \qquad \text{for all} \quad u \in H. \end{aligned}$$

In this sense, the relation $H \subseteq V^*$ is to be understood. Obviously, this embedding is continuous.

The next theorem extends the solvability results for linear systems from Chapter 2 to distributions on the right-hand side. We consider DAEs of the form

$$A(t)(D(t)x(t))' + B(t)x(t) = q(t), \tag{C.9}$$

$$D(t_0)x_0 = z_0 \in \text{im} D(t_0). \tag{C.10}$$

Theorem C.6. *If $q \in L^2(t_0, T; \mathbb{R}^n)$, then the index-1 IVP (C.9)–(C.10) has a unique solution x in*

$$L_D^2(t_0, T; \mathbb{R}^n) := \{x \in L^2(t_0, T; \mathbb{R}^n) : \ Dx \in C([t_0, T], \mathbb{R}^m)\}.$$

Equation (C.9) holds for almost all $t \in [t_0, T]$. Furthermore, Dx is differentiable for almost all $t \in [t_0, T]$ and there is a constant $C > 0$ such that

$$\|x\|_{L^2(t_0,T;\mathbb{R}^n)} + \|Dx\|_{C([t_0,T],\mathbb{R}^m)} + \|(Dx)'\|_{L^2(t_0,T;\mathbb{R}^m)} \leq C \left(\|z_0\| + \|q\|_{L^2(t_0,T;\mathbb{R}^n)} \right).$$

For continuous solutions, the right-hand side belonging to the nondynamical part has to be continuous. The next theorem describes this more precisely.

Theorem C.7. *If $q \in L^2(t_0,T;\mathbb{R}^n)$ and $Q_0G_1^{-1}q \in C([t_0,T];\mathbb{R}^n)$, then the solution x of the index-1 IVP (C.9)–(C.10) belongs to $C([t_0,T];\mathbb{R}^n)$ and we find a constant $C>0$ such that*

$$\|x\|_{C([t_0,T],\mathbb{R}^n)} + \|(Dx)'\|_{L^2(t_0,T;\mathbb{R}^m)} \le C\left(\|z_0\| + \|q\|_{L^2(t_0,T;\mathbb{R}^n)} + \|Q_0G_1^{-1}q\|_{C([t_0,T],\mathbb{R}^n)}\right).$$

Proof (of Theorems C.6 and C.7). The proof is straightforward. We simply have to combine standard techniques from DAE and Volterra operator theory. Due to the index-1 assumption, the matrix

$$G_1(t) = A(t)D(t) + B(t)Q_0(t)$$

is nonsingular for all $t \in [t_0,T]$. Recall that $Q_0(t)$ is a projector onto $\ker A(t)D(t)$. Multiplying (C.9) by $D(t)G^{-1}(t)$ and $Q_0(t)G^{-1}(t)$, respectively, we obtain the system

$$(Dx)'(t) - R'(t)(Dx)(t) + (DG_1^{-1}BD^-)(t)(Dx)(t) = (DG_1^{-1}r)(t), \tag{C.11}$$
$$(Q_0x)(t) + (Q_0G_1^{-1}BD^-)(t)(Dx)(t) = (Q_0G_1^{-1}r)(t), \tag{C.12}$$

which is equivalent to (C.9). Here, we have used the properties

$$(DG_1^{-1}A)(t) = R(t), \qquad (G_1^{-1}BQ_0)(t) = Q_0(t)$$

for all $t \in [t_0,T]$. Recall that $R(t) = D(t)D^-(t)$ is a continuously differentiable projector onto $\operatorname{im} D(t)$ along $\ker A(t)$ and $D^-(t)$ is a generalized inverse that satisfies $D^-(t)D(t) = P_0(t)$.
For $z := Dx$, equation (C.11) together with (C.10) represents an ordinary initial value problem of the form

$$z'(t) = \hat{A}(t)z(t) + b(t), \qquad z(t_0) = z_0 \tag{C.13}$$

with $\hat{A} \in C([t_0,T], L(\mathbb{R}^m,\mathbb{R}^m))$ and $b \in L^2(t_0,T;\mathbb{R}^m)$. Since $\hat{A}$ is linear and continuous, the map

$$x \mapsto \hat{A}(t)x$$

is Lipschitz continuous as a map from $L^2(t_0,T;\mathbb{R}^m)$ into $L^2(t_0,T;\mathbb{R}^m)$ with a Lipschitz constant that is independent of t. Consequently (see, e.g., [79], pp. 166–167), the IVP (C.13) has a unique solution $z \in C([t_0,T],\mathbb{R}^m)$ with $z' \in L^2(t_0,T;\mathbb{R}^m)$. The solution z satisfies (C.13) for almost all $t \in [t_0,T]$ and it is differentiable for almost all $t \in [t_0,T]$. Furthermore, there is a constant $C_1 > 0$ such that

$$\|z\|_{C([t_0,T],\mathbb{R}^m)} + \|z'\|_{L^2(t_0,T;\mathbb{R}^m)} \le C_1\left(\|z_0\| + \|b\|_{L^2(t_0,T;\mathbb{R}^m)}\right). \tag{C.14}$$

In [79], this was proven not only for maps into the finite-dimensional space $\mathbb{R}^m$ but also for maps into any Banach space. In the finite-dimensional case, the unique solvability of (C.13) and the validity of the estimation (C.14) follow also from the theorem of Carathéodory (see, e.g., [218], [121]), an a priori estimate and the gen-

eralized Gronwall lemma (see, e.g., [216]). For convenience, we omit an extended explanation of the second way.
Multiplying (C.11) by $I - R(t)$, we obtain that

$$((I-R)z)'(t) = -R'(t)((I-R)z)(t)$$

for the solution z and almost all $t \in [t_0, T]$. Since z_0 belongs to $\operatorname{im} D(t_0)$, we get

$$((I-R)z)(t_0) = 0.$$

Using again the unique solvability, we obtain that

$$((I-R)z)(t) = 0 \qquad \text{for almost all } t \in [t_0, T]. \tag{C.15}$$

From (C.12), we see that all solutions of (C.9)–(C.10) are given by

$$x(t) = D^-(t)z(t) - (Q_0 G_1^{-1} B D^-)(t)z(t) + (Q_0 G_1^{-1} r)(t), \tag{C.16}$$

where z is the unique solution of (C.13). Obviously, $Dx = z$ belongs to $C([t_0, T], \mathbb{R}^m)$. Since D, R and P_0 are continuous on $[t_0, T]$, the generalized inverse D^- is continuous. This implies $x \in L^2(t_0, T; \mathbb{R}^n)$ since $r \in L^2(t_0, T; \mathbb{R}^n)$. Recall that G_1 is continuous due to the index-1 assumption. If, additionally, $Q_0 G^{-1} r$ is continuous on $[t_0, T]$, then the whole solution x belongs to $C([t_0, T], \mathbb{R}^n)$. The estimations of Theorems C.6 and C.7 are a simple conclusion of the solution representation (C.16) and the estimation (C.14). □

References

1. Abramov, A.A., Balla, K., Ulyanova, V.I., Yukhno, L.F.: On a nonlinear selfadjoint eigenvalue problem for certain differential–algebraic equations of index 1. J. Comput. Math. **42**(7), 957–973 (2002)
2. Ascher, U., Mattheij, R., Russell, R.: Numerical Solution of Boundary Value Problems for Ordinary Differential Equations. Computational Mathematics. Prentice Hall, Englewood Cliffs (1988)
3. Ascher, U., Petzold, L.: Stability of computational methods for constrained dynamics systems. SIAM J. Sci. Stat. Comput. **1**, 95–120 (1991)
4. Ascher, U.M., Petzold, L.R.: Projected implicit Runge–Kutta methods for differential–algebraic equations. SIAM J. Numer. Anal. **28**, 1097–1120 (1991)
5. Ascher, U.M., Petzold, L.R.: Computer Methods for Ordinary Differential Equations and Differential–Algebraic Equations. SIAM, Philadelphia (1998)
6. Backes, A.: Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen. Logos Verlag Berlin (2006). Dissertation, Humboldt-University Berlin, October 2005/January 2006
7. Balla, K.: Differential–algebraic equations and their adjoints. Ph.D. thesis, Hungarian Academy of Sciences, Budapest (2004). Doctor of the Hungarian Academy of Sciences
8. Balla, K., Kurina, G., März, R.: Index criteria for differential–algebraic equations arising from linear-quadratic optimal control problems. Journal of Dynamical and Control Systems **12**(3), 289–311 (2006)
9. Balla, K., März, R.: Linear differential–algebraic equations of index 1 and their adjoint equations. Result. Math. **37**(1), 13–35 (2000)
10. Balla, K., März, R.: An unified approach to linear differential–algebraic equations and their adjoint equations. Preprint 2018, Humboldt-Universität zu Berlin, Inst. für Mathematik (2000)
11. Balla, K., März, R.: A unified approach to linear differential–algebraic equations and their adjoints. Z. Anal. Anwend. **21**(3), 783–802 (2002)
12. Balla, K., März, R.: A unified approach to linear differential–algebraic equations and their adjoints. Z. Anal. Anwend. **21**(3), 783–802 (2002)
13. Barbu, V., Favini, A.: Existence for an implicit nonlinear differential equations. Nonlinear Anal. **32**(1), 33–40 (1998)
14. Baumanns, S.: Coupled electromagnetic field/circuit simulation: modeling and numerical analysis. Ph.D. thesis, University of Cologne (2012)
15. Ben-Israel, A., Greville, T.N.: Generalized Inverses. Springer-Verlag, New York (2003)
16. Bender, D.J., Laub, A.J.: The linear-quadratic optimal regulator for descriptor systems. IEEE Trans. Automat. Control **AC-32**(8), 672–688 (1987)
17. Berger, T., Ilchmann, A.: On stability of time-varying linear differential–algebraic equations. J. of Dynamics and Differential Equations (2010). Submitted to

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5,

18. Berger, T., Ilchmann, A., Trenn, S.: The quasi-Weierstraß form for regular matrix pencils. Linear Algebra Appl. **436**(10), 4052–4069 (2012)
19. Bickart, T., Picel, Z.: High order stiffly stable composite multistep methods for numerical integration of stiff differential equations. BIT **13**, 272–286 (1973)
20. Biegler, L., Campbell, S., Mehrmann, V.: Control and optimization with differential–algebraic constraints. Advances in design and control. SIAM (2012)
21. Boyarincev, J.E.: Regular and singular systems of linear ordinary differential equations (in Russian). Nauka Siberian Branch, Novosibirsk (1980)
22. Braess, D.: Finite Elemente. Springer, Berlin, Heidelberg (1992)
23. Brenan, K., Petzold, L.: The numerical solution of higher index differential/algebraic equations by implicit methods. SIAM J. Numer. Anal. **26**, 976–996 (1989)
24. Brenan, K.E.: Stability and convergence of difference approximations for higher index differential–algebraic systems with applications in trajectory control. Ph.D. thesis, University of California, Los Angeles (1983)
25. Brenan, K.E., Campbell, S.L., Petzold, L.R.: The Numerical Solution of Initial Value Problems in Ordinary Differential–Algebraic Equations. North Holland Publishing Co., New York (1989)
26. Brunner, H.: (2000). Private comm.
27. Burrage, K., Butcher, J.: Non-linear stability for a general class of differential equation methods. BIT **20**, 326–340 (1980)
28. Butcher, J.: On the convergence of numerical solutions of ordinary differential equations. Math. Comp. **20**, 1–10 (1966)
29. Butcher, J.: Numerical Methods for Ordinary Differential Equations, Second Edition. John Wiley & Sons, Chichester (2003)
30. Butcher, J., Burrage, K.: Stability criteria for implicit methods. SIAM J. Numer. Anal **16**, 46–57 (1979)
31. Butcher, J., Podhaisky, H.: On error estimation in general linear methods for stiff odes. Appl. Numer. Math. **56**(3–4), 345–357 (2006)
32. Butcher, J.C.: Coefficients for the study of Runge–Kutta integration processes. J. Austral. Math. Soc. **3**, 185–201 (1963)
33. Butcher, J.C.: The numerical analysis of ordinary differential equations, Runge–Kutta and general linear methods. Wiley, Chichester and New York (1987)
34. Butcher, J.C., Heard, A.D.: Stability of numerical methods for ordinary differential equations. Numer. Algorithms **31**(1–4), 59–73 (2002)
35. Callies, R.: Some aspects of optimal control of nonlinear differential–algebraic equations. In: S.L. Campbell, R. März, L.R. Petzold, P. Rentrop (eds.) Differential–Algebraic Equations, pp. 19–21. Mathematisches Forschungsinstitut Oberwolfach, Report No. 18/2006 (2006)
36. Campbell, S., März, R.: Direct transcription solution of high index optimal control problems and regular Euler–Lagrange equations. J. Comput. Appl. Math. **202**(2), 186–202 (2007)
37. Campbell, S.L.: Singular Systems of Differential Equations I. Research Notes in Math.; 40. Pitman, Marshfield (1980)
38. Campbell, S.L. (ed.): Recent applications of generalized inverses. Pitman (1982)
39. Campbell, S.L.: One canonical form for higher index linear time varying singular systems. Circuits, Systems & Signal Processing **2**, 311–326 (1983)
40. Campbell, S.L.: The numerical solution of higher index linear time varying singular systems of differential equations. SIAM J. Sci. Stat. Comput. **6**, 334–348 (1985)
41. Campbell, S.L.: A general form for solvable linear time varying singular systems of differential equations. SIAM J. Math. Anal. **18**(4), 1101–1115 (1987)
42. Campbell, S.L.: A General Method for Nonlinear Descriptor Systems: An Example from Robotic Path Control. Tech. Rep. CRSC 090588-01, North Carolina State University Raleigh (1988)
43. Campbell, S.L.: A computational method for general higher index singular systems of differential equations. In: IMACS Transactions on Scientific Computing 1988, vol. 1.2, pp. 555–560 (1989)

44. Campbell, S.L., Gear, C.W.: The index of general nonlinear DAEs. Report, North Carolina State Univ., NC, U.S.A. (1993)
45. Campbell, S.L., Gear, C.W.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)
46. Campbell, S.L., Griepentrog, E.: Solvability of general differential–algebraic equations. SIAM J. Sci. Comp. **16**, 257–270 (1995)
47. Campbell, S.L., Marszalek, W.: The index of an infinite dimensional implicit system. Math. Comput. Model. Dyn. Syst. **5**(1), 18–42 (1999)
48. Campbell, S.L., März, R., Petzold, L.R., Rentrop, P. (eds.): Differential–Algebraic Equations. No. 18 in Oberwolfach Rep. European Mathematical Society Publishing House (2006)
49. Campbell, S.L., Meyer, C.D.: Generalized inverses of linear transformations. Dover Publications (1991)
50. Campbell, S.L., Moore, E.: Progress on a general numerical method for nonlinear higher index deas. Circuits System Signal Process **13**, 123–138 (1994)
51. Chua, L.O., Desoer, C.A., Kuh, E.S.: Linear and nonlinear circuits. McGraw-Hill Book Company, New York (1987)
52. Clark, K.D., Petzold, L.R.: Numerical Methods for Differential-Algebraic Equations in Conservative Form. Tech. Rep. UCRL-JC-103423, Lawrence Livermore National Laboratory (1990)
53. Cobb, D.: On the solution of linear differential equations with singular coefficients. J. Differential Equations **46**, 311–323 (1982)
54. Crouzeix, M.: Sur la B-stabilité des méthodes de Runge–Kutta. Numer. Math. **32**(1), 75–82 (1979)
55. Curtiss, C., Hirschfelder, J.: Integration of stiff equations. Proc. Nat. Acad. Sci. **38**, 235–243 (1952)
56. Dai, L.: Singular Control Systems, Lecture Notes on Control and Information, vol. 118. Springer-Verlag, New York (1989)
57. De Luca, A., Isidori, A.: Feedback Linearization of Invertible Systems. 2nd Colloq. Aut. & Robots, Duisburg (1987)
58. Desoer, C.A., Kuh, E.S.: Basic Circuit Theory. McGraw-Hill Book Company (1969)
59. Dokchan, R.: Numerical intergration of differential–algebraic equations with harmless critical points. Ph.D. thesis, Humboldt-University of Berlin (2011)
60. Dolezal, V.: Zur Dynamik der Linearsysteme. ACTA TECHNICA **1**, 19–33 (1960)
61. Dolezal, V.: The existence of a continuous basis of a certain linear subspace of E_r which depends on a parameter. Časopis pro pěstováni matematiky **89**, 466–468 (1964)
62. Döring, H.: Traktabilitätsindex und Eigenschaften von matrixwertigen Riccati-Typ Algebrodifferentialgleichungen. Master's thesis, Humboldt University of Berlin, Inst. of Math. (2004)
63. Eich-Soellner, E., Führer, C.: Numerical Methods in Multibody Dynamics. B.G.Teubner, Stuttgart (1998)
64. Elschner, H., Möschwitzer, A., Reibiger, A.: Rechnergestützte Analyse in der Elektronik. VEB Verlag Technik, Berlin (1977)
65. England, R., Gómez, S., Lamour, R.: Expressing optimal control problems as differential–algebraic equations. Comput. Chem. Eng. **29**(8), 1720–1730 (2005)
66. England, R., Gómez, S., Lamour, R.: The properties of differential–algebraic equations representing optimal control problems. Appl. Numer. Math. **59**(10), 2357–2373 (2009)
67. Estévez Schwarz, D.: Topological analysis for consistent initialization in circuit simulation. Tech. Rep. 99-3, Fachbereich Mathematik, Humboldt-Univ. zu Berlin (1999)
68. Estévez Schwarz, D.: Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation. Ph.D. thesis, Humboldt-Univ. zu Berlin (2000)
69. Estévez Schwarz, D., Lamour, R.: The computation of consistent initial values for nonlinear index-2 differential–algebraic equations. Numer. Algorithms **26**(1), 49–75 (2001)
70. Estévez Schwarz, D., Tischendorf, C.: Structural analysis of electric circuits and consequences for MNA. Int. J. Circ. Theor. Appl. **28**(2), 131–162 (2000)

71. Favini, A., Plazzi, P.: Some results concerning the abstract nonlinear equation $D_t Mu(t) + Lu(t) = f(t, Ku(t))$. Circuits Systems Signal Process **5**, 261–274 (1986)
72. Favini, A., Plazzi, P.: On some abstract degenerate problems of parabolic type 2: The nonlinear case. Nonlinear Anal. **13**, 23–31 (1989)
73. Favini, A., Rutkas, A.: Existence and uniqueness of solutions of some abstract degenerate nonlinear equations. Diff. Int. Eqs **12**(3), 373–394 (1999)
74. Favini, A., Yagi, A.: Degenerate Differential Equations in Banach Spaces. Pure and Applied Mathematics, Marcel Dekker, New York (1999)
75. Feldmann, U., Wever, U., Zheng, Q., Schultz, R., Wriedt, H.: Algorithms for modern circuit simulation. AEÜ Archiv für Elektronik und Übertragungstechnik, Int. J. Electron. Commun. **46**(4), 274–285 (1992)
76. Flügel, J.: Lösung von Algebro-Differentialgleichungen mit properem Hauptterm durch die BDF. Master's thesis, Humboldt University of Berlin, Inst. of Math. (2002)
77. Fosséprez, M.: Non-linear Circuits: Qualitative Analysis of Non-linear, Non-reciprocal Circuits. John Wiley & Sons, Chichester (1992)
78. Führer, C., Schwertassek, R.: Generation and solution of multibody system equations. Int. Journal of Nonl. Mechanics **25**(2/3), 127–141 (1990)
79. Gajewski, H., Gröger, K., Zacharias, K.: Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen. Akademie-Verlag, Berlin (1974)
80. Gajshun: Vvedenie v teoriyu linejnykh nestatsionarnykh sistem. Institut matematiki NAN Belarusi, Minsk (1999). (Russian)
81. Gantmacher, F.R.: Teoriya matrits (Russian). Gostekhizdat (1953)
82. Gantmacher, F.R.: Matrizenrechnung I+II. VEB Deutcher Verlag der Wissenchaften, Berlin (1970)
83. Garcia-Celayeta, B., Higueras, I.: Runge–Kutta methods for DAEs. a new approach. J. Computational and Applied Mathematics **111**(1–2), 49–61 (1999)
84. Gear, C., Petzold, L.: Differential/algebraic systems and matrix pencils. In: B. Kagstrom, A. Ruhe (eds.) Matrix Pencils, Lecture Notes in Mathematics, vol. 973, pp. 75–89. Springer-Verlag, Berlin, New York (1983)
85. Gear, C.W.: Hybrid methods for initial value problems in ordinary differential equations. SIAM J. Numer. Anal. **2**, 69–86 (1965)
86. Gear, C.W.: Simultaneous numerical solution of differential–algebraic equations. IEEE Trans. Circuit Theory **CT-18**(1), 89–95 (1971)
87. Gear, C.W.: Maintaining solution invariants in the numerical solution of ODEs. SIAM J. Sci. Statist. Comput. **7**, 734–743 (1986)
88. Gear, C.W., Gupta, G.K., Leimkuhler, B.J.: Automatic integration of the Euler-Lagrange equations with constraints. J. Comp. Appl. Math. **12,13**, 77–90 (1985)
89. Gear, C.W., Hsu, H.H., Petzold, L.: Differential–algebraic equations revisited. In: Proc. Oberwolfach Conf. on Stiff Equations, Bericht des Instituts für Geom. und Prakt. Math.; 9. Rhein.-Westfälische Techn. Hochschule, Aachen (1981)
90. Gear, C.W., Petzold, L.R.: ODE methods for the solution of differential/algebraic systems. SIAM J. Numer. Anal. **21**, 716–728 (1984)
91. Gerdts, M.: Local minimum principle for optimal control problems subject to index-two differential–algebraic equations. J. Opt. Th. Appl. **130**(3), 443–462 (2006)
92. Gerdts, M.: Representation of Lagrange multipliers for optimal control problems subject to index-two differential–algebraic equations. J. Opt. Th. Appl. **130**(2), 231–251 (2006)
93. Golub, G., Kahan, W.: Calculating the singular values and pseudo-inverse of a matrix. Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis **2**(2), 205–224 (1965). DOI 10.1137/0702016. URL http://link.aip.org/link/?SNA/2/205/1
94. Golub, G.H., van Loan, C.F.: Matrix Computations. The Johns Hopkins University Press, Baltimore, London (1991)
95. Griepentrog, E.: Index reduction methods for differential–algebraic equations. Tech. Rep. 91-12, Fachbereich Mathematik, Humboldt-Univ. zu Berlin (1991)

96. Griepentrog, E., März, R.: Differential–Algebraic Equations and Their Numerical Treatment. Teubner-Texte zur Mathematik No. 88. BSB B.G. Teubner Verlagsgesellschaft, Leipzig (1986)
97. Griepentrog, E., März, R.: Basic properties of some differential–algebraic equations. Zeitschrift für Analysis und ihre Anwendungen **8**(1), 25–40 (1989)
98. Griewank, A., Walter, A.: Evaluating Derivatives: Principles and Techniques of Algorithmic Differentiation, second edn. SIAM, Philadelphia (2008)
99. Grigorieff, R.: Stability of multistep-methods on variable grids. Numer. Math. **42**, 359–377 (1983)
100. Guglielmi, N., Zennaro, M.: On the zero-stability of variable stepsize multistep methods: the spectral radius approach. Tech. rep., Dip. di Matematica Pura e Applicata, Universita dell'Aquila (1999)
101. Günther, M., Feldmann, U.: CAD based electric modeling in industry. Part I: Mathematical structure and index of network equations. Surv. Math. Ind. **8**, 97–129 (1999)
102. Hairer, E.: RADAU5: Implicit Runge–Kutta method of order 5 (Radau IIA) for semi-implicit DAEs. URL http://www.unige.ch/~hairer/software.html
103. Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential–Algebraic Equations by Runge–Kutta Methods. Lecture Notes in Mathematics Vol. 1409. Springer-Verlag, Heidelberg (1989)
104. Hairer, E., Nørsett, S.P., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer Series in Computational Mathematics 8. Springer-Verlag, Berlin, Heidelberg (1987)
105. Hairer, E., Wanner, G.: Solving Ordinary Differential Equations II: Stiff and Differential–Algebraic Problems. Springer Series in Computational Mathematics 14. Springer-Verlag, Berlin, Heidelberg (1991)
106. Hanke, M.: Beiträge zur Regularisierung von Randwertaufgaben für Algebro-Differentialgleichungen mit höherem Index. Dissertation(B), Habilitation, Humboldt-Universität zu Berlin, Inst. für Mathematik (1989)
107. Hanke, M.: On the asymptotic representation of a regularization approach to nonlinear semiexplicit higher index differential–algebraic equations. IMA J. Appl. Math. (1990)
108. Hanke, M., Lamour, R.: Consistent initialization for nonlinear index-2 differential–algebraic equation: large sparse systems in MATLAB. Numer. Algorithms **32**, 67–85 (2003)
109. Hanke, M., Macana, E.I., März, R.: On asymptotics in case of linear index-2 DAE's. SIAM J. Numer. Anal. **35**(4), 1326–1346 (1998)
110. Hanke, M., März, R., Neubauer, A.: On the regularization of a certain class of nontransferable differential–algebraic equations. J. of Differential Equations **73**(1), 119–132 (1988)
111. Hansen, B.: Differentiell-algebraic equations.- Consistent initial values for index-k-tractable linear equations and nonlinear index-2 equations. Ph.D. thesis, Humboldt-University of Berlin, Institute of Mathematics (1990)
112. Higueras, I., Celayeta, B.G.: Logarithmic norms for matrix pencils. SIAM J. Matrix. Anal. **20**(3), 646–666 (1999)
113. Higueras, I., März, R.: Formulating differential–algebraic equations properly. Preprint 2020, Humboldt-Universität zu Berlin, Inst. für Mathematik (2000)
114. Higueras, I., März, R.: Differential–algebraic equations with properly stated leading terms. Comput. Math. Appl. **48**, 215–235 (2004)
115. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-1 DAEs. Appl. Numer. Math. **45**(2–3), 175–200 (2003)
116. Higueras, I., März, R., Tischendorf, C.: Stability preserving integration of index-2 DAEs. Appl. Numer. Math. **45**(2–3), 201–229 (2003)
117. Horn, R., Johnson, C.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
118. Hoschek, M., Rentrop, P., Wagner, Y.: Network approach and differential–algebraic systems in technical applications. Surv. Math. Ind. **9**, 49–76 (1999)
119. de Jalón, J.G., Bayo, E.: Kinematic and Dynamic Simulation of Multibody Systems: The Real-Time challenge. Springer-Verlag, New York (1994)

120. Butcher, J.C., Chartier, P.C., Jackiewicz, Z.: Nordsieck representation of dimsims. Numer. Algorithms **16**, 209–230 (1997)
121. Kamke, E.: Das Lebesgue-Stieltjes Integral. Teubner, Leipzig (1960)
122. Kaps, P., Wanner, G.: A study of Rosenbrock-type methods of high order. Numer. Math. **38**, 279–298 (1981)
123. Koch, O., Kofler, P., Weinmüller, E.: Initial value problems for systems of ordinary first and second order differential equations with a singularity of the first kind. Analysis **21**, 373–389 (2001)
124. Koch, O., März, R., Praetorius, D., Weinmüller, E.: Collocation methods for index 1 DAEs with a singularity of the first kind. Math. Comp. **79**(269), 281–304 (2010)
125. König, D.: Indexcharakterisierung bei nichtlinearen Algebro-Differentialgleichungen. Master's thesis, Humboldt-Universität zu Berlin, Inst. für Mathematik (2006)
126. Kronecker, L.: Gesammelte Werke, vol. III, chap. Reduktion der Scharen bilinearer Formen, pp. 141–155. Akad. d. Wiss., Berlin (1890)
127. Kunkel, P., Mehrmann, V.: Canonical forms for linear differential–algebraic equations with variable coefficients. J. Comput. Appl. Math. **56**, 225–251 (1994)
128. Kunkel, P., Mehrmann, V.: The linear quadratic optimal control problem for linear descriptor systems with variable coefficients. Math. Control, Signals, Sys. **10**, 247–264 (1997)
129. Kunkel, P., Mehrmann, V.: Analysis of over- and underdetermined nonlinear differential–algebraic systems with application to nonlinear control problems. Math. Control, Signals, Sys. **14**, 233–256 (2001)
130. Kunkel, P., Mehrmann, V.: Differential–Algebraic Equations - Analysis and Numerical Solution. EMS Publishing House, Zürich, Switzerland (2006)
131. Kunkel, P., Mehrmann, V.: Necessary and sufficient conditions in the optimal control for general nonlinear differential–algebraic equations. Tech. Rep. 355, Matheon (2006)
132. Kunkel, P., Mehrmann, V., Seufer, I.: GENDA: A software package for the numerical solution of general nonlinear differential–algebraic equations. Tech. Rep. 730-02, Technische Universität Berlin (2002)
133. Kurina, G.: Singular perturbations of control problems with equation of state not solved for the derivative (a survey). Journal of Computer and System Science International **31**(6), 17–45 (1993)
134. Kurina, G.A.: On operator Riccati equation unresolved with respect to derivative (in Russian). Differential. Uravnen. **22**, 1826–1829 (1986)
135. Kurina, G.A., März, R.: Feedback solutions of optimal control problems with DAE constraints. SIAM J. Control Optim. **46**(4), 1277–1298 (2007)
136. Kutta, W.: Beitrag zur näherungsweisen Integration totaler Differentialgleichungen. Z. Math. Phys. **46**, 435–453 (1901)
137. Lamour, R.: Index determination and calculation of consistent initial values for DAEs. Comput. Math. Appl. **50**, 1125–1140 (2005)
138. Lamour, R.: A projector based representation of the strangeness index concept. Tech. Rep. 07-03, Humboldt-Universität zu Berlin, Inst. für Mathematik (2007)
139. Lamour, R.: Tractability index - strangeness index. Tech. rep., Humboldt-University of Berlin, Dep. of Mathematics (2008). In preparation
140. Lamour, R., März, R., Tischendorf, C.: PDAEs and further mixed systems as abstract differential algebraic systems. Tech. Rep. 01-11, Inst. of Math., Humboldt Univ. of Berlin (2001). URL http://www2.math.hu-berlin.de/publ/publ01.html
141. Lamour, R., März, R., Winkler, R.: How Floquet theory applies to index-1 differential algebraic equations. J. Math. Appl. **217**(2), 372–394 (1998)
142. Lamour, R., Mazzia, F.: Computations of consistent initial values for properly stated index 3 DAEs. BIT Numerical Mathematics **49**, 161–175 (2009)
143. Lamour, R., Monett Diaz, D.: A new algorithm for the index determination in DAEs by Taylor series using algorithmic differentiation. Tech. Rep. 10-04, Humboldt-University, Dep. of Mathematics (2010)
144. Lamour, R., Monett Diaz, D.: A new algorithm for the index determination in DAEs using Algorithmic Differentiation. Numer. Algorithms **58**(2), 261–292 (2011)

145. Le Vey, G.: Some remarks on solvability and various indices for implicit differential equations. Numer. Algorithms **19**, 127–145 (1998)
146. Lentini, M., März, R.: Conditioning and dichotomy in differential–algebraic equations. SIAM J. Numer. Anal. **27**(6), 1519–1526 (1990)
147. Lewis, F.L.: A survey of linear singular systems. Circuits, systems and signal processing **5**, 3–36 (1986)
148. Li, S., Petzold, L.: Design of new DASPK for sensitivity analysis. Tech. rep., UCSB (1999)
149. Lucht, W., Strehmel, K., Eichler-Liebenow, C.: Indexes and special discretization methods for linear partial differential–algebraic equations. BIT **39**(3), 484–512 (1999)
150. Luenberger, D.G.: Dynamic equations in descriptor form. IEEE Trans. Autom. Control **22**, 312–321 (1977)
151. Lyusternik, L.A.: Ob uslovnykh ehkhstremumakh funktsionalov. Matematiceskij Sbornik **41**, 390–401 (1934). In Russian
152. März, R.: Multistep methods for initial value problems in implicit differential–algebraic equations. Beiträge zur Num. Mathem. **12** (1984)
153. März, R.: On difference and shooting methods for boundary value problems in differential–algebraic equations. ZAMM **64**(11), 463–473 (1984)
154. März, R.: On initial value problems in differential–algebraic equations and their numerical treatment. Computing **35**, 13–37 (1985)
155. März, R.: On correctness and numerical treatment of boundary value problems in DAEs. Zhurnal Vychisl. Matem. i Matem. Fiziki **26**(1), 50–64 (1986)
156. März, R.: A matrix chain for analyzing differential–algebraic equations. Preprint (Neue Folge) 162, Humboldt-Universtät zu Berlin, Sektion Mathematik (1987)
157. März, R.: Index-2 differential–algebraic equations. Results in Mathematics **15**, 149–171 (1989)
158. März, R.: Once more on backward differentiation formulas applied to higher index differential–algebraic equations. ZAMM **69**, T37–T39 (1989)
159. März, R.: Some new results concerning index-3 differential–algebraic equations. J. Mathem. Analysis and Applications **140**(1), 177–199 (1989)
160. März, R.: Numerical methods for differential–algebraic equations. Acta Numerica pp. 141–198 (1992)
161. März, R.: On quasilinear index 2 differential–algebraic equations. In: E. Griepentrog, M. Hanke, R. März (eds.) Berlin Seminar on Differential–Algebraic Equations, Fachbereich Mathematik, Humboldt-Univ. Berlin (1992)
162. März, R.: Practical Lyapunov stability criteria for differential algebraic equations. Num. Anal. Math. Model. **29**, 245–266 (1994)
163. März, R.: On linear differential–algebraic equations and linearizations. APNUM **18**, 267–292 (1995)
164. März, R.: Canonical projectors for linear differential–algebraic equations. Comput. Math. Appl. **31**(4–5), 121–135 (1996)
165. März, R.: Criteria for the trivial solution of differential–algebraic equations with small nonlinearities to be asymptotically stable. J. Math. Anal. Appl. **225**, 587–607 (1998)
166. März, R.: Differential–algebraic systems anew. Appl. Num. Math. **42**, 315–335 (2002)
167. März, R.: The index of linear differential–algebraic equations with properly stated leading terms. Results Math. **42**(3–4), 308–338 (2002)
168. März, R.: Differential–algebraic systems with properly stated leading term and MNA equations. In: K. Anstreich, R. Bulirsch, A. Gilg, P. Rentrop (eds.) Modelling, Simulation and Optimization of Integrated Circuits, pp. 135–151. Birkhäuser (2003)
169. März, R.: Fine decouplings of regular differential–algebraic equations. Results Math. **46**, 57–72 (2004)
170. März, R.: Solvability of linear differential–algebraic equations with property stated leading terms. Results Math. **45**(1–2), 88–105 (2004)
171. März, R.: Characterizing differential–algebraic equations without the use of derivative arrays. Comput. Math. Appl. **50**(7), 1141–1156 (2005)

172. März, R.: Differential–algebraic equations in optimal control problems. In: Proceedings of the International Conference "Control Problems and Applications (Technology, Industry, Economics)", pp. 22–31. Minsk (2005)
173. März, R., Riaza, R.: Linear differential–algebraic equations with properly stated leading term: Regular points. J. Math. Anal. Appl. **323**, 1279–1299 (2006)
174. März, R., Riaza, R.: Linear differential–algebraic equations with properly stated leading term: A-critical points. Mathematical and Computer Modelling of Dynamical Systems **13**(3), 291–314 (2007)
175. Matthes, M.: Numerical analysis of nonlinear partial differential–algebraic equations: A coupled and an abstract systems approach (2012). Dissertation, Universität zu Köln, in preparation
176. Matthes, M., Tischendorf, C.: Convergence of Galerkin solutions for linear differential–algebraic equations in Hilbert spaces. In: ICNAAM 2010: International Conference on Numerical Analysis and Applied Mathematics 2010: AIP Conference Proceedings, vol. 1281, pp. 247–250 (2010)
177. Matthes, M., Tischendorf, C.: Private communication (2012)
178. Menrath, M.: Stability criteria for nonlinear fully implicit differential-algebraic systems. Ph.D. thesis, University of Cologne (2011). URL http://kups.ub.uni-koeln.de/id/eprint/3303
179. Nordsieck, A.: On numerical integration of ordinary differential equations. Math. Comp. **16**, 22–49 (1962)
180. Petzold, L.: Differential/algebraic equations are not ODE's. SIAM J. Sci. Stat. Comput. **3**, 367–384 (1982)
181. Petzold, L.R.: DASPK: Large scale differential–algebraic equation solver. http://www.cs.ucsb.edu/~cse/software.html
182. Petzold, L.R.: A description of DASSL: A differential/algebraic system solver. In: Proc. 10th IMACS World Congress, August 8–13 Montreal 1982 (1982). URL http://www.cs.ucsb.edu/~cse/software.html
183. Petzold, L.R.: Order results for implicit Runge–Kutta methods applied to differential/algebraic systems. SIAM J. Numer. Anal. **23**, 837–852 (1986)
184. Prato, G.D., Grisvard, P.: On an abstract singular Cauchy problem. Comm. Partial. Diff. Eqs. **3**, 1077–1082 (1978)
185. Pryce, J.: Solving high-index DAEs by Taylor series. Numer. Algorithms **19**, 195–211 (1998)
186. Pryce, J.D.: A simple structural analysis method for DAEs. BIT **41**(2), 364–394 (2001)
187. Rabier, P., Rheinboldt, W.: Classical and generalized solutions of time-dependent linear differential–algebraic equations. Linear Algebra and its Applications **245**, 259–293 (1996)
188. Rabier, P., Rheinboldt, W.: Nonholonomic motion of rigid mechanical systems from a DAE viewpoint. SIAM, Society for Industrial and Applied Mathematics (1999)
189. Rabier, P., Rheinboldt, W.: Techniques of scientific computing (part 4) - theoretical and numerical analysis of differential–algebraic equations. In: Ciarlet, P.G. et al.(eds.) Handbook of Numerical Analysis, vol. VIII, pp. 183–540. North Holland/Elsevier, Amsterdam (2002)
190. Rang, J., Angermann, L.: Perturbation index of linear partial differential–algebraic equations. Appl. Numer. Math. **53**(2/4), 437–456 (2005)
191. Reis, T.: Systems theoretic aspects of PDAEs and applications to electrical circuits. Ph.D. thesis, Technische Universiät Kaiserslautern (2006)
192. Reis, T.: Consistent initialization and perturbation analysis for abstract differential-algebraic equations. Mathematics of Control, Signals, and Systems (MCSS) **19**, 255–281 (2007). URL http://dx.doi.org/10.1007/s00498-007-0013-9
193. Reis, T., Tischendorf, C.: Frequency domain methods and decoupling of linear infinite dimensional differential–algebraic systems. J. Evol. Equ. **5**(3), 357–385 (2005)
194. Riaza, R.: Differential–algebraic systems. Analytical aspects and circuit applications. World Scientific, Singapore (2008)
195. Rump, S.: Intlab - interval laboratory. In: T. Csendes (ed.) Developments in Reliable Computing, pp. 77–104. SCAN-98, Kluwer Academic Publishers (1999)

196. Runge, C.: Über die numerische Auflösung von Differentialgleichungen. Math. Ann. **46**, 167–178 (1895)
197. Schulz, S.: Convergence of general linear methods for numerically qualified DAEs. manuscript (2003)
198. Schulz, S.: Four lectures on differential–algebraic equations. Tech. Rep. 497, University of Auckland, Department of Mathematics (2003)
199. Schulz, S.: General linear methods for numerically qualified DAEs. manuscript (2003)
200. Schumilina, I.: Charakterisierung der Algebro–Differentialgleichungen mit Traktabilitätsindex 3. Ph.D. thesis, Humboldt-Universität zu Berlin (2004)
201. Schwartz, L.: Théorie des Distributions. Herrmann, Paris (1966)
202. Sincovec, R.F., Erisman, A.M., Yip, E.L., Epton, M.A.: Analysis of descriptor systems using numerical algorithms. IEEE Trans. Automatic Control **26**, 139–147 (1981)
203. Soto, M.S., Tischendorf, C.: Numerical analysis of DAEs from coupled circuit and semiconductor simulation. Appl. Numer. Math. **53**(2–4), 471–488 (2005)
204. Stuart, A., Humphres, A.: Dynamical Systems and Numerical Analysis. Cambridge University Press, Cambridge (1998)
205. Tischendorf, C.: Solution of index-2 differential–algebraic equations and its application in circuit simulation. Ph.D. thesis, Humboldt-Univ. of Berlin (1996)
206. Tischendorf, C.: Topological index calculation of DAEs in circuit simulation. Surv. Math. Ind. **8**(3–4), 187–199 (1999)
207. Tischendorf, C.: Coupled systems of differential–algebraic and partial differential equations in circuit and device simulation. Modeling and numerical analysis (2004). Habilitation thesis at Humboldt Univ. of Berlin
208. Trenn, S.: Distributional differential–algebraic equations. Ph.D. thesis, TU Ilmenau (2009)
209. Unger, J., Kröner, A., Marquardt, W.: Structural analysis of differential–algebraic equation systems - theory and applications. Computers chem. Engng **19**(8), 867–882 (1995)
210. Voigtmann, S.: GLIMDA - A General LInear Method solver for Differential Algebraic equations. URL http://www.mathematik.hu-berlin.de/~steffen/software.html
211. Voigtmann, S.: General linear methods for integrated circuit design. Ph.D. thesis, Humboldt-Universität zu Berlin, Inst. für Mathematik (2006)
212. Weierstraß, K.: Gesammelte Werke, vol. II, chap. Zur Theorie der bilinearen und quadratischen Formen, pp. 19–44. Akad. d. Wiss., Berlin (1868)
213. Wong, N.: An efficient passivity test for descriptor systems via canonical projector techniques. In: Proceedings of the 46th Annual Design Automation Conference, DAC '09, pp. 957–962. ACM, New York, NY, USA (2009). URL http://doi.acm.org/10.1145/1629911.1630157
214. Wright, W.: General linear methods with inherent Runge-Kutta stability. Ph.D. thesis, The University of Auckland, New Zealand (2003). URL www.math.auckland.ac.nz/~butcher/theses/willwright.pdf
215. Yosida, K.: Functional Analysis, sixth edn. Springer-Verlag, Berlin Heidelberg New York (1980)
216. Zeidler, E.: Nonlinear Functional Analysis and its Applications I. Fixed-Point Theorems. Springer-Verlag, New York (1986)
217. Zeidler, E.: Nonlinear Functional Analysis and its Applications II/A. Linear Monotone Operators. Springer-Verlag, New York (1990)
218. Zeidler, E.: Nonlinear Functional Analysis and its Applications II/B. Nonlinear Monotone Operators. Springer-Verlag, New York (1990)
219. Zielke, G.: Motivation und Darstellung von verallgemeinerten Matrixinversen. Beiträge zur Numerischen Mathematik **7**, 177–218 (1979)

Index

R. Lamour et al., *Differential-Algebraic Equations: A Projector Based Analysis*, Differential-Algebraic Equations Forum, DOI 10.1007/978-3-642-27555-5,