Analysis and Numerical Solution of Structured and Switched Differential-Algebraic Systems

vorgelegt von
Dipl. Techn.-Math. Lena Wunderlich
aus Berlin

Von der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Stefan Felsner
Gutachter:	Prof. Dr. Volker Mehrmann
	Prof. Dr. Peter Kunkel
zusätzliche Gutachterin:	Prof. Dr. Caren Tischendorf

Tag der wissenschaftlichen Aussprache: 10. September 2008

Berlin 2008
D 83

Danksagung

An dieser Stelle möchte ich all denen danken, die die Entstehung dieser Dissertation möglich gemacht haben. Zunächst möchte ich vorallem meinem Betreuer Prof. Dr. Volker Mehrmann danken, der diese Arbeit initiiert hat und stets mit wertvollen Ratschlägen und vielen Ideen immer wieder den nötigen Anschwung gegeben hat und mit konstruktiver Kritik für das Gelingen der Arbeit gesorgt hat. Des Weiteren bedanke ich mich recht herzlich bei Prof. Dr. Peter Kunkel und Prof. Dr. Caren Tischendorf für die Begutachtung dieser Arbeit. Auch geht mein Dank an die AG Modellierung, Numerik, Differentialgleichungen der Technischen Universität Berlin, insbesondere an meine Büronachbarn Elena Virnik und Christian Schröder, für die angenehme Arbeitsatmosphäre. Weiter möchte ich mich bei meiner Familie bedanken, die mich immer tatkräftig unterstützt hat. Mein besonderer Dank geht an Holger, für seine Unterstützung während all der Jahre, ohne die ein Studium und eine Doktorarbeit niemals möglich geworden wären.

Zusammenfassung

Die numerische Simulation komplexer dynamischer Systeme spielt heutzutage eine wichtige Rolle in der Entwicklung technischer Komponenten. Die Modellgleichungen dieser dynamischen Systeme werden häufig mit Hilfe von automatisierten Modellierungsprogrammen erstellt und bestehen üblicherweise aus differentiell-algebraischen Gleichungen (DAEs), d.h. aus Differentialgleichungen, die das dynamische Verhalten des Systems beschreiben und daran gekoppelte algebraische Zwangsbedingungen, die diese Dynamik auf eine bestimmte Mannigfaltigkeit zwingen. Neben den bekannten Schwierigkeiten in der numerischen Lösung von DAEs, wie das Auftreten von Ordnungsreduktionen in den numerischen Verfahren, Instabilitäten oder das Abdriften der numerischen Lösung von der Lösungsmannigfaltigkeit, können komplexe Systeme zusätzlich Differentialgleichungen höherer Ordnung enthalten, oder die Modellgleichungen ändern sich mit der Zeit, so daß die Systeme zwischen verschiedenen Systemskonfigurationen schalten, in Abhängigkeit von Schaltbedingungen. Desweiteren treten häufig differentiell-algebraische Systeme mit strukturierten Koeffizienten auf.

Diese Arbeit beschäftigt sich mit der Analyse sowie mit der numerische Lösung von strukturierten und geschalteten differentiell-algebraischen Gleichungen. Im wesentlichen werden drei Schwerpunkte behandelt.

Zunächst werden differentiell-algebraische Systeme zweiter Ordnung betrachtet. Die klassische Ordnungsreduktion, die verwendet wird um gewöhnliche Differentialgleichungen höherer Ordnung in Systeme von Differentialgleichungen ersten Ordnung zu überführen, kann bei der Anwendung auf differentiell-algebraische Gleichungen zu verschiedenen Problemen führen, wie zum Beispiel zu einer Erhöhung des Index der DAE oder sogar zum Verlust der Lösbarkeit. Aufgrund dessen wird in dieser Arbeit ein Indexreduktionsverfahren, sowohl für lineare als auch nichtlineare DAE Systeme zweiter Ordnung entwickelt, das basierend auf Ableitungsfeldern des Systems zweiter Ordnung die Konstruktion eines äquivalenten differentiell-algebraischen Systems bestehend aus entkoppelten Differentialgleichungen erster und zweiter Ordnung sowie davon unabhängigen algebraischen Gleichungen ermöglicht. Dieses reduzierte System besitzt die gleiche Lösung wie das ursprüngliche System. Weiter erlaubt das Verfahren die Transformation in ein reduziertes System erster Ordnung von niedrigem Index sowie eine explizite Lösungsdarstellung im Fall von Zeit-invarianten linearen Systemen zweiter Ordnung.

Der zweite Teil der Arbeit befaßt sich mit strukturierten differentiell-algebraischen Systemen. Da die Strukturen in den Koeffizientenmatrizen die physikalischen Eigenschaften des Systems widerspiegeln, sollten diese Strukturen während der numerischen Lösung erhalten bleiben, um auch die physikalischen Eigenschaften des Systems zu erhalten. In der vorliegenden Arbeit werden lineare DAEs mit symmetrischen und selbstadjungierten Koeffizientenmatrizen untersucht und strukturerhaltende Normalformen für symmetrische und

selbstadjungierte lineare DAE Systeme entwickelt. Es stellt sich heraus, dass eine strukturerhaltende strangeness-freie Formulierung sowohl für symmetrische als auch für selbstadjungierte Systeme nur für Systeme mit Strangeness Index kleiner oder gleich 1 existiert. Für symmetrische Systeme benötigt man außerdem weitere starke Voraussetzungen an die Koeffiezientenmatrizen, um die Struktur erhalten zu können. Desweiteren wird ein strukturerhaltendes Indexreduktionsverfahren für selbstadjungierte lineare DAEs entwickelt, basierend auf minimaler Erweiterung des Originalsystems, welches eine strukturerhaltende numerische Behandlung erlaubt.
Der dritte Teil der Arbeit beschäftigt sich mit geschalteten oder so genannten hybriden differentiell-algebraischen Systemen, welche auf der Basis von Schaltbedingungen zwischen verschiedenen Zustandsbeschreibungen schalten. Zunächst wird die Formulierung dieser Systeme untersucht, sowie die Existenz und Eindeutigkeit von Lösungen nach dem Umschalten. Danach wird die numerische Lösung von hybriden differentiell-algebraischen Systemen behandelt. Hierbei spielt insbesondere die konsistent Re-Initialisierung nach dem Umschalten und die Behandlung von numerischen Schnattern (sogenanntes "Chattering") während der numerischen Simulation eine wichtige Rolle. Für die konsistenten Re-Initialisierung wird ein Verfahren verwendet, welches es erlaubt bestimmte Komponenten des Lösungsvektors an der Stelle des Umschaltens festzuhalten, um so die Lösung des Gesamtsystems auf physikalisch sinnvolle Weise fortzuführen. Unter Verwendung sogenannter "Sliding Mode Simulation" ist es möglich das dynamische Verhalten des Systems während des Schnatterns zu approximieren, um durch die Lösung eines Ersatzmodells ständiges Umschalten zwischen verschiedenen Systembeschreibungen und den damit verbundenen hohen Rechenaufwand zu vermeiden. Eine Modussteuerung für die numerische Simulation hybrider differentiell-algebraischer Systeme, die die numerische Integration der DAEs mit der Organisation der Moduswechsel verbindet und Sliding Mode Simulation erlaubt wurde implementiert. Die Funktionalität der Modussteuerung wird durch eine Anzahl numerischer Beispiele illustriert, insbesondere im Hinblick auf elektrische Schaltkreise mit schaltenden Elementen und mechanische Systeme mit Haft- und Gleitreibung. Desweiteren werden die grundlegenden Ideen zur Steuerung von linearen geschalteten Deskriptorsystemen betrachtet.

Abstract

The numerical simulation of complex dynamical systems nowadays plays an important role in technical applications. Typically, the dynamical systems arising from automatic model generating tools are described by differential-algebraic equations (DAEs), i.e., by differential equations describing the dynamical behavior of the system coupled with algebraic constraints forcing these dynamics onto a specific manifold. Besides the already known difficulties in solving DAEs numerically, as e.g. order reduction of numerical methods, instabilities, or the drift-off from the solution manifold, complex systems additionally can contain higher order differential-algebraic equations, or the system can switch between different system configurations or operation modes based on certain transition conditions. Further, the coefficient matrices of the DAEs can exhibit certain structures.

In this thesis we discuss the analysis as well as the numerical solution of structured and switched differential-algebraic equations. Basically, the thesis focuses on three topics.

First, second order differential-algebraic equations are considered. It is known that the classical order reduction that is used to transform higher order ordinary differential equations into first order systems leads to a number of difficulties when applied to DAEs, as e.g. an increase in the index of the system or even the loss of solvability. In this thesis, an index reduction method for linear as well as nonlinear second order DAEs based on differentiation of the second order system is derived that allows to construct an equivalent second order system of low index in a numerical feasible way. This approach also enables the transformation into so-called trimmed first order form of low index and an explicit representation of solutions in the case of linear time-invariant second order systems.

The second topic involves structured differential-algebraic systems. As the structure of the coefficient matrices represent certain physical properties of the system the symmetry structure should be preserved during the numerical solution. In particular, linear differential-algebraic systems with symmetric and self-adjoint coefficient matrices are considered and structure preserving condensed forms for symmetric and self-adjoint linear DAEs are derived. It turns out that a structure preserving strangeness-free formulation for symmetric and self-adjoint systems only exists if the strangeness index of the system is lower or equal one. For symmetric systems we need in addition strong assumptions on the coefficient matrices in order to preserve the symmetry. Further, a structure preserving index reduction method based on so-called minimal extension is investigated that allows a structure preserving numerical treatment.

The third topic involves switched or so-called hybrid differential-algebraic systems that switch between different modes of operation based on certain transition conditions. First, we examine the formulation of hybrid systems and the existence and uniqueness of solutions after switching. Afterwards, the numerical solution of hybrid systems is considered.

In particular, a consistent reinitialization after mode switching is considered that allows a continuation of the previous solution in a physical reasonable way by fixing certain components of an initial value vector, and the treatment of chattering behavior during the numerical simulation using so-called sliding mode simulation is studied. A hybrid mode controller is implemented for the numerical solution of hybrid differential-algebraic systems that organizes mode switching and allows sliding mode simulation. The functionality of the mode controller is illustrated by several examples, in particular, considering electrical circuits with switching elements and mechanical systems with dry friction phenomena. Further, the basic concepts for the control of linear hybrid descriptor systems are considered.

Contents

Notation

$\dot{x}, \ddot{x}, x^{(i)}$	time derivative of $x(t)$, i.e., $\dot{x}(t) = \frac{d}{dt}x(t)$, $\ddot{x}(t) = \frac{d^2}{dt^2}x(t)$, $x^{(i)}(t) = \frac{d^i}{dt^i}x(t)$, see Chapter 2
$f_{;.} = \frac{\partial f}{\partial .}$	partial derivative of a function f, see Definition 2.1
∇^j	backward difference, see Definition 6.12
$\|.\|$	vector norm, see Definition 2.4
$(.,.)$	sesquilinear form, i.e., $(f,g) = \int_{\mathbb{I}} f^T(t)g(t)dt$ for $f, g \in C^0(\mathbb{I}, \mathbb{R}^n)$, see (4.19) and Definition 4.14
A^T	transpose of a matrix $A \in \mathbb{R}^{m,n}$
A^{-1}	inverse of a matrix $A \in \mathbb{R}^{n,n}$, see Definition 2.14
A^+	Moore-Penrose pseudo-inverse of a matrix $A \in \mathbb{R}^{m,n}$, see Definition 2.17
A^D	Drazin inverse of a matrix $A \in \mathbb{R}^{n,n}$, see Definition 2.20
a_μ, a_i, a_μ^l	number of algebraic variables (in mode l)
α_j	coefficients of a BDF method, see (6.16)
$\alpha_{ij}, \beta_j, \gamma_j$	coefficients of a Runge-Kutta method, see (6.8)
b, b^l	right-hand side of a linear first order DAE (in mode l), see (2.5)
$C^k(\mathbb{I}, V)$	set of k-times continuously differentiable functions $f : \mathbb{I} \to V$
$C(\mathbb{I}, V) = C^0(\mathbb{I}, V)$	set of continuous functions $f : \mathbb{I} \to V$
$\mathcal{C}^m_{imp}(\mathcal{T})$	set of impulsive smooth distributions, see Definition 2.44
$\mathbb{C}$	set of complex numbers
d_μ, d_μ^l	number of differential variables (in mode l), see Theorem 2.36
$d_\mu^{(2)}, d_i^{(2)}$	number of second order differential variables, see Lemma 3.8
$d_\mu^{(1)}, d_i^{(1)}$	number of first order differential variables, see Lemma 3.8
δ_a	Dirac delta distribution at a point $a \in \mathbb{R}$
$\mathbb{D}_x, \mathbb{D}_{\dot{x}}, \mathbb{D}_{x^{(i)}}$	domain of $x, \dot{x}, x^{(i)}$
$\mathcal{D}^n$	set of test functions, see Section 2.2.3
D	differential-algebraic operator $D : \mathbb{X} \to \mathbb{Y}$, see (4.17)
$D^\star$	conjugate differential-algebraic operator $D^\star : \mathbb{Y}^\star \to \mathbb{X}^\star$, see Definition 4.15
D_l	union of subintervals $\mathbb{I}_i$, see Definition 5.3
$\mathcal{E}(T_\tau)$	set of event times for a hybrid time trajectory T_τ, see Section 5.1
f	right-hand side of a linear second order DAE, see (3.6)
$\mathcal{F}_k, \mathcal{F}_k^l$	derivative array of level k (in mode l), see (2.15)
$\hat{F}, \hat{F}_1, \hat{F}_2, \hat{F}_3$	functions describing a reduced nonlinear DAE, see (2.19)
F_N	normal force, see Example 5.2

F, F^l	function describing a nonlinear DAE (in mode l), see (2.3), (5.3)
$\Gamma^l_{j,i}$	switching surface, see (5.9)
Γ^l	boundary of the constraint manifold Λ^l, see (5.10)
$g^l_{j,i}, g^l_j, g^l$	switching functions, see Definition 5.3
h	stepsize of a discretization method, see Section 6.2
$\mathcal{H}$	a hybrid differential-algebraic system, see Definition 5.3
$\mathbb{I} = [t_0, t_f] \subset \mathbb{R}$	interval
$\mathbb{I}_i = [\tau_i, \tau_i')$	subinterval, see Definition 5.3
J^l	set of autonomous transitions in mode l, see Definition 5.3
L^l_j,	transition condition, see Definition 5.3
L_l	Lagrange interpolation polynomial, see (6.12)
$\mathbb{L}_k, \mathbb{L}^l_k$	solution set of the derivative array $\mathcal{F}_k$, $\mathcal{F}^l_k$, see (2.16)
$\mathbb{L}, \mathbb{L}^l$	constraint manifold of a DAE (in mode l), see Section 2.2.2
Λ^l	constraint manifold of a hybrid system (in mode l), see (5.11)
$\mathcal{L}_l$	Jacobian of derivative array, see (3.20), (3.39)
L, L^l	differential part of nonlinear reduced system (in mode l), see (2.21)
m, m_l	number of equations (in mode l)
$\mathbb{M}$	set of modes, see Definition 5.3
$\mathcal{M}_l$	Jacobian of derivative array, see (3.20), (3.39)
n, n_l	number of unknowns (in mode l)
$\mathcal{N}_l$	Jacobian of derivative array, see (3.20), (3.39)
$\mathbb{N}$	set of natural numbers (excluding 0)
$\mathbb{N}_0$	set of natural numbers (including 0)
$N_{\mathbb{I}}$	number of integration intervals, see Definition 5.3
N_F	number of modes, see Definition 5.3
n^l_T	number of transitions of mode l, see Definition 5.3
n^l_j	number of switching functions for transition j in mode l, see Definition 5.3
Π_k	space of polynomials of maximal degree k, see (6.1)
Φ	discretization method, see (6.7)
$\mathbb{R}$	set of real numbers
$\mathbb{R}^{m,n}$	set of real matrices of size $m \times n$
$\mathbb{R}[D_i]$	set of i-th order differential polynomials in $\mathbb{R}$
$\mathcal{R}^l$	reachable set in mode l, see Definition 5.35
R, R^l	algebraic part of a nonlinear reduced system (in mode l), see (2.21)
$\mathcal{R}_{\mathcal{H}}$	set of reachable states of a hybrid system $\mathcal{H}$, see Definition 5.40
$\mathcal{S}$	cost functional of an optimal control problem, see (1.3)
S^l	mode allocation function, see Definition 5.3

$s^{(MCK)}, s_i^{(MCK)}$	strangeness due to (M, C, K), see Lemma 3.8
$s^{(CK)}, s_i^{(CK)}$	strangeness due to (C, K), see Lemma 3.8
$s^{(MC)}, s_i^{(MC)}$	strangeness due to (M, C), see Lemma 3.8
$s^{(MK)}, s_i^{(MK)}$	strangeness due to (M, K), see Lemma 3.8
s_i	part of $s_i^{(MK)}$, see Lemma 3.15
t_i	gridpoint of a discretization
t	independent variable, time
T_l^k	transition function, see Definition 5.3
T_τ	hybrid time trajectory, see Section 5.1
T_m	hybrid mode trajectory, see Section 5.1
τ_i', τ_i	event times, see Definition 5.3
u_μ, u_μ^l, u_i	number of undetermined variables (in mode l), see Theorem 2.36
u_{eq}	equivalent control, see (5.30)
v_μ, v_μ^l, v_i	number of vanishing equations (in mode l), see Theorem 2.36
μ, μ^l	strangeness index (in mode l), see Definition 2.35
μ_{max}	maximal strangeness index of a hybrid system, see Definition 5.8
μ_f, μ_s, μ_k	coefficients of friction, see Example 5.2
ν	index of nilpotency, see Definition 2.19
ν_d	differentiation index, see Chapter 1
x_{imp}	impulsive part of a generalized function, see (2.26)
x	distribution $\mathrm{x} \in \mathcal{C}^n$, see Definition 2.43
$\mathcal{X}_i$	discretized solution, see (6.7)
x_i	approximation of the solution at time t_i, i.e., $x_i \approx x(t_i)$
$\mathbb{X}$	function space, see (4.18)
$\mathbb{X}^\star$	dual function space, see (4.20)
$\mathbb{Y}$	function space, see (4.18)
$\mathbb{Y}^\star$	dual function space, see (4.20)
(M, C, K)	matrix triple or triple of matrix functions describing a linear second order DAE, see (3.6)
$(E, A), (E^l, A^l)$	matrix pair or pair of matrix-valued functions describing a linear DAE (in mode l), see (2.5) or (2.6)
$(\hat{E}, \hat{A})$	pair of matrix functions describing a reduced linear DAE, see (2.24)
$f[t_j, \ldots, t_{j+k}]$	divided difference, see Definition 6.4

Abbreviations

BDF	backward differential formulas, see Section 6.2.2
DAE	differential-algebraic equation, see Section 2.2
d-index	differentiation index, see Chapter 1
GELDA	general linear differential algebraic system solver, see Section 7.1.1
GENDA	general nonlinear differential algebraic system solver, see Section 7.1.1
GESDA	general switched differential algebraic system solver, see Section 7.1
MNA	modified nodal analysis, see Chapter 1
s-index	strangeness index, see Chapter 1
ODE	ordinary differential equation, see Chapter 1
SVD	singular value decomposition, see Chapter 2.1

Chapter 1

Introduction

The numerical simulation of complex dynamical systems nowadays plays an important role for technical inventions and requires reliable mathematical models of the physical systems as well as efficient numerical solution methods. In almost all areas of electrical, mechanical, chemical, or traffic engineering the modeling of the dynamics of complex technical systems is today highly modularized, thus allowing the easy and efficient automatic generation of mathematical models. Modern modeling tools automatically generate models for substructures and link them together via constraints. The numerical simulation of these models, however, exhibits a number of difficulties that have to be dealt with. The dynamical behavior of physical processes is usually modeled via differential equations. If the states of the physical system are in some ways constrained, then the mathematical model also contains algebraic equations to describe these constraints. Such systems, consisting of differential and algebraic equations, are called *differential-algebraic equations (DAEs)*. Differential-algebraic equations arise naturally in the modeling process and are therefore widely used in the simulation and control of constrained dynamical systems in numerous applications, such as mechanical systems, electrical circuit simulation, chemical engineering, fluid dynamics and many other areas. In the following, we present some of the most important examples.

Mechanical Systems. In the industrial simulation and mathematical modeling of mechanical systems the multibody approach is frequently used [34, 130]. A *multibody system* is the result of describing a mechanical system by a finite number of bodies with masses and torques and the interconnections between these bodies. The *equations of motion* of a constrained multibody system are given by

$$\begin{aligned} M(p,t)\ddot{p} &= f_a(p,\dot{p},t) - G^T(p,t)\lambda, \\ 0 &= g(p,t). \end{aligned} \tag{1.1}$$

Here, $p \in \mathbb{R}^{n_p}$ denotes the vector of generalized position coordinates of the mechanical system with n_p degrees of freedom, $M(p,t) \in \mathbb{R}^{n_p,n_p}$ is the mass matrix, which is usually positive semi-definite and symmetric, and the vector $f_a(p,\dot{p},t) \in \mathbb{R}^{n_p}$ describes the applied forces acting on the system. Further, the vector $g(p,t) \in \mathbb{R}^{n_\lambda}$ describes constraints restricting the motion of the system which are coupled via the constraint matrix $G(p,t) := \frac{d}{dp}g(p,t) \in \mathbb{R}^{n_\lambda,n_p}$ and the Lagrange multipliers $\lambda \in \mathbb{R}^{n_\lambda}$ to the dynamical system. Thus, the multibody approach leads to a nonlinear differential-algebraic equation (1.1).

Linearization of the equations of motion (1.1) along the equilibrium solution or the discretization of mechanical structures by finite element methods leads to systems of the form

$$M(t)\ddot{p} + C(t)\dot{p} + K(t)p = f(t), \tag{1.2}$$

where $M(t) \in \mathbb{R}^{n_p,n_p}$ is again the mass matrix, $C(t) \in \mathbb{R}^{n_p,n_p}$ is the damping matrix that can also contain Coriolis forces for gyroscopic systems [65, 89], $K(t) \in \mathbb{R}^{n_p,n_p}$ is the stiffness matrix, and $f(t) \in \mathbb{R}^{n_p}$ denotes time-dependent external forces. When the leading matrix $M(t)$ is singular, then the system (1.2) forms a linear second order differential-algebraic equation.

Optimal control problems. Classical control applications such as stabilization of a system or path following often are formulated as optimal control problems. The linear-quadratic optimal control problem as in [77, 83] is the problem of minimizing a cost functional

$$\mathcal{S}(x,u) = \int_{t_0}^{t_f} \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} Q(t) & S(t) \\ S^T(t) & R(t) \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} dt \tag{1.3a}$$

subject to the initial value problem

$$E(t)\dot{x} = A(t)x + B(t)u, \quad x(t_0) = x_0, \tag{1.3b}$$

with control input u, where $E, A \in C(\mathbb{I}, \mathbb{R}^{m,n})$, $B \in C(\mathbb{I}, \mathbb{R}^{m,k})$, $Q \in C(\mathbb{I}, \mathbb{R}^{n,n})$, $R \in C(\mathbb{I}, \mathbb{R}^{k,k})$, $S \in C(\mathbb{I}, \mathbb{R}^{n,k})$ and $Q(t) = Q^T(t)$, $R(t) = R^T(t)$ for all $t \in \mathbb{I} = [t_0, t_f]$. By application of the Pontryagin maximum principle [83] this linear-quadratic optimal control problem leads to the boundary value problem for a differential-algebraic equation of the form

$$\begin{bmatrix} 0 & E(t) & 0 \\ -E^T(t) & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\lambda} \\ \dot{x} \\ \dot{u} \end{bmatrix} = \begin{bmatrix} 0 & A(t) & B(t) \\ (A(t) + \dot{E}(t))^T & Q(t) & S(t) \\ B^T(t) & S^T(t) & R(t) \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix}, \tag{1.3c}$$

with boundary conditions

$$x(t_0) = x_0, \quad E^T(t_f)\lambda(t_f) = 0.$$

Electrical circuits. In the simulation of electrical circuits the modified nodal analysis (MNA) [35] leads to a quasi-linear differential-algebraic equation of the form

$$\begin{aligned} A_C \frac{dq_C(A_C^T e, t)}{dt} + A_R r(A_R^T e, t) + A_L j_L + A_V j_V + A_I i_S(t) &= 0, \\ \frac{d\Phi_L(j_L, t)}{dt} - A_L^T e &= 0, \\ A_V^T e - v_S(t) &= 0. \end{aligned} \tag{1.4}$$

Here, the vector e denotes the node potentials, j_L and j_V are the currents through inductances and voltage sources, respectively, the input functions i_S and v_S describe the current and voltage sources, the function r describes the resistances, and q_C and Φ_L are the functions describing the charges of the capacitances and the fluxes of the inductances, respectively. Further, the incidence matrix $A = [A_C\ A_L\ A_R\ A_V\ A_I]$ contains the information on the topology of the circuit, with A_C, A_L, A_R, A_V and A_I describing the branch-current relation for capacitive, inductive, resistive branches and branches for voltage sources and current sources, respectively. Thus, the vectors $A_C^T e$, $A_L^T e$, $A_R^T e$ and $A_V^T e$ describe the branch voltages for the capacitive, inductive, resistive and voltage source branches, respectively.

Compared to ordinary differential equations (ODEs) there are many difficulties in solving DAEs analytically as well as numerically. An important property describing the difficulty to solve a DAE is the so-called *index* of the DAE. There are several index concepts such as the *differentiation index (d-index)* [17, 59], the *strangeness index (s-index)* [82], the *perturbation index (p-index)* [24, 59] or the *tractability index (t-index)* [55, 97]. The differentiation index roughly states how often all or part of the DAE have to be differentiated with respect to time t in order to obtain an *ordinary differential equation*, i.e., a system of the form $\dot{x} = \varphi(t, x)$. The concept of the differentiation index is widely-used in the analysis of differential-algebraic equations, but it has the major drawback that it is not defined for over- and underdetermined systems as it is based on a solvability concept that requires unique solvability. Therefore, the concept of the strangeness index was developed in [72, 78, 79, 84] as a generalization of the differentiation index to over- and underdetermined systems. We will use the concept of the strangeness index throughout this thesis. As the differentiation index determines how far the differential-algebraic equation is away from an ordinary differential equation, an ordinary differential equation has d-index zero, while an algebraic equation has d-index one. In contrast, the strangeness index measures the distance to a decoupled system of ordinary differential equations and purely algebraic equations. Hence, ordinary differential equations and purely algebraic equations both have s-index zero. In this way, the index also determines the smoothness that is needed for the inhomogeneity of a linear differential-algebraic equation to guarantee the existence of a classical solution. Also of great importance in the numerical treatment of differential-algebraic equations is the perturbation index that measures the sensitivity of solutions with respect to perturbations of the problem in the initial values and right-hand sides. For a detailed analysis and a comparison of various index concepts with the differentiation index, see [24, 82, 98]. In the following, differential-algebraic equations with a d-index higher than 1 or an s-index higher than 0 are called *higher index problems*. It is well-known that under the standard assumptions that the mass matrix M is symmetric and positive definite and the constraints are independent, meaning that the constraint matrix G has full row rank, the equations of motion (1.1) of multibody systems are of d-index 3 or s-index 2 respectively, see e.g. [34]. Also in electrical circuit analysis it is well-known which influence specific elements and their combination may have on the index, see [56, 57]. Furthermore, in [36, 141] topological methods have been derived that analyze the network topology and show which equations are responsible for a high index. It has been shown

in [36] that for wide classes of circuits the t-index of the MNA equations (1.4) does not exceed 2 and can be determined by topological criteria assuming positive definiteness of the Jacobians of the element-characterizing functions.
The accuracy and stability of the numerical solution of a DAE depends on the index, in such a way that the higher the index of the DAE, the more sensitive is the numerical solution to perturbations and errors in the data. In particular, for higher index problems numerical methods may not converge and instabilities can occur. Further, due to hidden algebraic constraints in higher index problems the discretization errors can cause the numerical solution to drift-off from the constraint manifold that is given by the algebraic relations in the system. To overcome these difficulties in the solution of higher index problems, regularization techniques can be applied to transform the system to an equivalent system of lower index. In many applications like multibody systems or circuit simulation problems the differential-algebraic equations have extra structure that can be used to determine the reduced systems. For mechanical multibody systems there are several regularization techniques, see e.g. [14, 50, 34, 137], all of them involving differentiations of the constraint equations. Further, an index reduction method that allows to conserve certain structural properties of the given problem based on introducing some new variables is the *index reduction by minimal extension* studied in [80] and in [6] for the MNA equations (1.4). Index reduction techniques for general linear and nonlinear over- and underdetermind DAEs are given in [82].
Besides the challenges already given in the numerical solution of general nonlinear DAEs some more difficulties arise that have not yet quite been settled. First of all, the differential-algebraic system can contain certain structures describing physical properties of the system. In the linearized equations of motion (1.2) of mechanical systems the matrices M and K are typically symmetric, positive definite and sparse, and C is symmetric, or skew-symmetric for gyroscopic systems [65, 89]. In linear-quadratic optimal control problems (1.3) the matrices Q and R are usually positive semi-definite and positive definite, respectively, and the corresponding pair of matrix functions in (1.3c) is self-adjoint. These structures present physical properties of the system and numerical integration methods should preserve the structure to meet these characteristic properties. Furthermore, most of the differential-algebraic systems arising in engineering applications are second order systems due to the fact that forces are proportional to accelerations. In the following, differential-algebraic systems where derivatives of the unknown x of order k with $k \geq 2$ occur are called *higher order systems*. The classical approach for the solution of higher order differential-algebraic systems is the transformation into a first order system by introducing new variables for higher order derivatives. While this approach works well for ordinary differential equations it can lead to a number of problems for DAEs. On the one hand, it may increase the index of the DAE as has been shown in [102, 135] and on the other hand numerical methods can fail for which various examples are given in [4, 17, 129, 151, 152]. Furthermore, in practical applications the differential-algebraic system may be badly scaled and perturbations are present in the data, such that the transformation to first order leads to very different solutions in the perturbed system. Moreover, the transformation to first order leads to systems of double dimension and may destroy structures present in the system. Therefore, numer-

ical methods are proposed in [129, 151, 152] which enable the direct numerical solution of higher order differential-algebraic systems. Otherwise, introducing only some derivatives in the transformation to first order systems may avoid an increase of the index, but in general it is not known which derivatives can be included without difficulties. A condensed form for linear higher order differential-algebraic systems is introduced in [102, 135] which allows an identification of those higher order derivatives of variables that can be replaced to obtain a first order system without changing the smoothness requirements or increasing the index. This condensed form allows the analysis of existence and uniqueness of solutions for higher order DAEs and a definition of the strangeness index for higher order linear DAEs, but it is not really numerically computable as it involves derivatives of computed orthogonal transformations.

Another difficulty is that in many technical applications the behavior of the system or the mathematical model can change with time, e.g., due to switching elements in electrical circuits like electric switches or diodes, or friction phenomena and impacts in mechanical systems, see e.g. [34, 88]. Thus, we are faced with so-called *switched* or *hybrid differential-algebraic systems* that were studied e.g. in [12, 60, 61, 94]. In these systems discrete event dynamics and continuous time dynamics interact and influence each other such that they must be analyzed and solved simultaneously. Switching system models also involve discontinuities in the system and in the solution or changes in the index of the system. The analysis of singular points and the behavior of solutions of differential-algebraic systems at critical points, such as singularities or those points where characteristic quantities of the DAE change, as well as the behavior at impasse points, i.e., points beyond which the solution of the DAEs cannot be continued, is still a widely open problem. Only few results in this direction have been obtained, mainly for specially structured systems, see e.g. [99, 120].

In this thesis we treat some of the encountered difficulties concerning the analysis and numerical solution of structured and switched differential-algebraic systems. With regard to higher order differential-algebraic systems, we will present an index reduction method that allows the determination of an equivalent reduced higher order system locally at every time step in a numerical feasible way. The approach also allows to find trimmed first order formulations of s-index 0 for linear second order DAEs and explicit solution representation in the case of time-invariant linear systems. Further, we will consider structure preserving condensed forms for linear differential-algebraic systems with symmetric coefficients as well as self-adjoint system, and derive a structure preserving index reduction method that allows to preserve the structure and therefore also the physical properties of the system during the numerical integration. In the numerical simulation of switched differential-algebraic systems, besides the already existing problems in the numerical integration of DAEs there are new difficulties. An index reduction has to be done in the same way as for DAEs and appropriate numerical methods for DAEs have to be used, but index reduction and integration is often done in small intervals resulting in high computational effort. Further, besides consistent initial values that are needed to start the integration also consistent values at each point where the system behavior changes are needed to restart the integration at those points. The states at the switch points have to be determined exactly, as they are

the basis for the consistent reinitialization and for the restart of the numerical integration. Special phenomena arising in the numerical solution of switched systems like chattering, i.e., cyclic switching between modes of operation, have to be treated in an appropriate way to ensure the termination of the numerical integration. We will consider these difficulties in the numerical integration of switched systems as well as the detection and location of switch points and the restarting of integration methods including consistent reinitialization. We will also consider existence and uniqueness of solutions of switched differential-algebraic systems after switching and the control of linear hybrid DAE systems.

This thesis is organized as follows. Chapter 2 contains some basic definitions and well-known results used throughout the thesis and a short overview of the existing analysis of linear and nonlinear DAEs. For linear DAEs we present condensed forms that allow to transform the system into an equivalent so-called *strangeness-free* system of s-index 0, and for general nonlinear DAEs we present the so-called *derivative array approach* that uses the derivatives of the system to obtain an equivalent strangeness-free system. Further, we introduce the concept of generalized functions and distributions that allows discontinuities in the solution of a differential-algebraic system. In Chapter 3 we consider second order DAEs and we present an derivative array approach for linear second order systems that allows to derive an equivalent strangeness-free systems locally for every time step in a numerical feasible way. For this, the relationships between the local and global characteristic invariants of linear second order DAEs are derived. The approach is also used to determine so-called *trimmed first order formulations* for second order systems and to give explicit representations of solutions in the case of linear time-invariant systems. Concluding, the approach is also generalized to general nonlinear second order systems. In Chapter 4 we derive structure preserving condensed forms for linear systems with symmetric coefficients as well as self-adjoint systems, and we present a structure preserving index reduction method using the ideas of minimal extension. In Chapter 5 we consider switched DAEs. First, we introduce the formulation of so-called *hybrid differential-algebraic systems* and explain the basic properties and behavior of switched differential-algebraic systems. Then, we consider existence and uniqueness of solutions of switched systems. Further, we consider consistent reinitialization of switched systems after switching that allows the continuation of a given solution in a physically reasonable way and we introduce so-called *sliding motion* for switched differential-algebraic systems that allows an efficient treatment of chattering behavior during the numerical integration. In the last part, we show how control theoretical results for DAEs can be extended to switched systems. Chapter 6 describes the numerical methods used during the numerical solution of switched differential-algebraic systems. We describe numerical integration methods for DAEs, in particular BDF and Runge-Kutta methods, that can also be used efficiently to interpolate the numerical solution between the grid points given by the stepsize selection. This is needed to determine the solution of a switched system at a switch point from which the integration is resumed. The switch points are determined as the roots of so-called *switching functions* using a modified secant method as root finding routine. Finally, we show how a mode controller can be realized that uses existing integration methods for DAEs as inner integration routine and organizes the

mode switching and restarts of the integrator. In Chapter 7 we describe the implemented mode controller GESDA that is designed to solve general switched differential-algebraic systems using suitable DAE integration routines. To illustrate the algorithms we present several numerical examples. Finally, in Chapter 8 we summarize and discuss the obtained results and point out several open problems that should be investigated in the future.

Chapter 2

Preliminaries

In this chapter we introduce some basic definitions and general results in analysis and linear algebra used throughout the thesis. We shortly present the existing analysis of linear and general nonlinear differential-algebraic equations that will serve as basis for our investigations. Further, we introduce the concept of generalized functions or distributions that can be used to handle discontinuous solutions of linear differential-algebraic equations.

2.1 Definitions and Basics

In this section we introduce some notations and definitions and review some of the fundamentals that are used throughout the thesis. Most of the proofs can be found in [54, 82].

Definition 2.1 (**Differentiable function).** A function $f : \mathbb{D} \to \mathbb{R}^m$ on an open subset $\mathbb{D} \subset \mathbb{R}^n$ is called *differentiable at a point* $x_0 \in \mathbb{D}$ if there exists a linear function $u : h \mapsto A(x_0)h$ with $A(x_0) \in \mathbb{R}^{m,n}$ such that

$$\lim_{h \to 0} \frac{\|f(x_0 + h) - f(x_0) - u(h)\|}{\|h\|} = 0.$$

Then $A(x_0)$ is called the *derivative of* f *at* x_0. If f is differentiable at every point $x_0 \in \mathbb{D}$ the function f is called *differentiable on* $\mathbb{D}$ and the function $A : \mathbb{D} \to \mathbb{R}^{m,n}$ is called the *derivative of* f. The derivative of the function f (with respect to x) is denoted by $\frac{df}{dx} = f_{;x}$.

Definition 2.2 (**Continuously differentiable).** A differentiable function $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is called *continuously differentiable* if the derivative of f is continuous in $\mathbb{D}$. The set of continuously differentiable functions from $\mathbb{D}$ into $\mathbb{R}^m$ is denoted by $C^1(\mathbb{D}, \mathbb{R}^m)$. A function $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is called *l-times continuously differentiable* if the derivative of f is an $(l-1)$-times continuously differentiable function from $\mathbb{D}$ into $\mathbb{R}^m$. The set of l-times continuously differentiable functions is denoted by $C^l(\mathbb{D}, \mathbb{R}^m)$. Furthermore, the set $C^\infty(\mathbb{D}, \mathbb{R}^m)$ is called the *set of infinitely continuously differentiable functions.*

Let $x : \mathbb{I} \subset \mathbb{R} \to \mathbb{R}^n$ and $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ be differentiable functions. In the following, the derivatives of $x(t)$ with respect to t are denoted by $\dot{x}(t) = \frac{d}{dt}x(t)$, $\ddot{x}(t) = \frac{d^2}{dt^2}x(t)$, $x^{(i)}(t) = \frac{d^i}{dt^i}x(t)$, and partial derivatives of $f(x)$ with respect to selected variables p from $x = [x_i]_{i=1,\dots,n}$ are denoted by $f_{;p}$, e.g.,

$$f_{;x_i}(x) = \frac{\partial}{\partial x_i} f(x), \quad f_{;x_i,\dots,x_{i+j}}(x) = \left[\frac{\partial}{\partial x_i} f(x) \dots \frac{\partial}{\partial x_{i+j}} f(x) \right].$$

Further, we define the directional derivatives of a function f.

Definition 2.3 (**Directional derivative).** Let $\mathbb{D} \subset \mathbb{R}^n$ be open and consider a function $f : \mathbb{D} \to \mathbb{R}$. For a point $x \in \mathbb{D}$ and a vector $v \in \mathbb{R}^n$, the *directional derivative of f in x along v* is defined by

$$D_v f(x) := \frac{d}{ds} f(x + sv)\bigg|_{s=0} = \lim_{s \to 0} \frac{f(x + sv) - f(x)}{s}.$$

If the function f is continuously differentiable then it holds that $D_v f(x) = f_{;x}(x)v$, see e.g. [42, p. 50]. The directional derivative of a function along a vector v at a point x represents the rate of change of the function, moving through x, in the direction of v.

Definition 2.4 (**Vector norm).** A function $\|.\| : \mathbb{R}^n \to \mathbb{R}$ is called a *vector norm* if for all $x, y \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ the following conditions are satisfied:

1. $\|x\| \geq 0$ and $\|x\| = 0 \Leftrightarrow x = 0$,
2. $\|x + y\| \leq \|x\| + \|y\|$,
3. $\|\alpha x\| = |\alpha| \|x\|$.

In the following, we will always consider the so called *Hölder norms* which are defined for $p \in \mathbb{N} \cup \{\infty\}$ by

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

for vectors $x = [x_i]_{i=1,\dots,n} \in \mathbb{R}^n$, where in particular $\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. Furthermore, with respect to matrices we will use the associated norms defined for $p \in \mathbb{N} \cup \{\infty\}$ by

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p},$$

for a matrix $A \in \mathbb{R}^{m,n}$. In the following, it is not necessary to distinguish between different p, so we will use $\|.\|$ instead of $\|.\|_p$.

Definition 2.5 (**Homeomorphism, homeomorphic).** A function $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is called a *homeomorphism* if f is bijective and both f and its inverse function f^{-1} are continuous. Two subsets $\mathbb{D} \subset \mathbb{R}^n$ and $\mathbb{G} \subset \mathbb{R}^m$ are called *homeomorphic* if there exists a homeomorphism $f : \mathbb{D} \to \mathbb{G}$ between them.

Definition 2.6 (**Diffeomorphism).** A function $f : \mathbb{D} \subset \mathbb{R}^n \to \mathbb{R}^m$ is called a *diffeomorphism* if f is bijective and both f and its inverse function f^{-1} are continuously differentiable.

Definition 2.7 (**Manifold).** A subset $M \subset \mathbb{R}^n$ is called a *manifold of dimension r* if every point $x \in M$ has an open neighborhood in M which is homeomorphic to an open subset of $\mathbb{R}^r$.

Theorem 2.8. *Let $H \in C^1(\mathbb{D}, \mathbb{R}^a)$, $\mathbb{D} \subseteq \mathbb{R}^n$ open, $k \in \mathbb{N} \cup \infty$, with $M = H^{-1}(\{0\}) \neq \emptyset$ and suppose that* $\operatorname{rank} H_{;x}(x) = a \leq n$ *for all $x \in M$. Then, M is a manifold of dimension $r = n - a$.*

Proof. See [82, Theorem 4.65]. □

Theorem 2.9 (Implicit Function Theorem). *Suppose that $f : \mathbb{D} \subset \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$ is continuous on an open neighborhood $\mathbb{D}_0 \subset \mathbb{D}$ of a point (x_0, y_0) for which $f(x_0, y_0) = 0$. Assume that $f_{;x}$ exists in a neighborhood of (x_0, y_0) and is continuous at (x_0, y_0), and that $f_{;x}(x_0, y_0)$ is nonsingular. Then there exist open neighborhoods $U_x \subset \mathbb{R}^n$ and $U_y \subset \mathbb{R}^m$ of x_0 and y_0, respectively, such that, for any $y \in \overline{U}_y$, the equation $f(x, y) = 0$ has a unique solution $x = \phi(y) \in \overline{U}_x$ and the mapping $\phi : U_y \to U_x$ is continuous. (Here, $\overline{U}_x$ and $\overline{U}_y$ denote the closure of U_x and U_y, respectively.) Moreover, if $f_{;y}$ exists at (x_0, y_0), then ϕ is continously differentiable at y_0 and its derivative is given by*

$$\phi_{;y}(y_0) = -\left[f_{;x}(x_0, y_0)\right]^{-1} f_{;y}(x_0, y_0).$$

Proof. See [113, p. 128]. □

In the following, we will often use the left- and right-hand limits of functions $x : \mathbb{I} \to \mathbb{R}^n$ that are defined by

$$x(\tau^-) = \lim_{t \to \tau^-} x(t) = \lim_{\substack{t \to \tau \\ t < \tau}} x(t), \quad x(\tau^+) = \lim_{t \to \tau^+} x(t) = \lim_{\substack{t \to \tau \\ t > \tau}} x(t).$$

Next, we introduce some important properties concerning subspaces and matrices. Important subspaces associated with a matrix $A \in \mathbb{R}^{m,n}$ are the range and the nullspace or kernel of the matrix.

Definition 2.10 (Orthogonal complement). The *orthogonal complement* of a subspace $S \subseteq \mathbb{R}^n$ is defined by $S^\perp = \{y \in \mathbb{R}^n : y^T x = 0 \text{ for all } x \in S\}$.

Definition 2.11 (Kernel, cokernel, range and corange). Let $A \in \mathbb{R}^{m,n}$. The *kernel, cokernel, range*, and *corange* of A are defined by

$$\begin{aligned}
&\operatorname{kernel}(A) = \{x \in \mathbb{R}^n \ : \ Ax = 0\},\\
&\operatorname{cokernel}(A) = (\operatorname{kernel}(A))^\perp,\\
&\operatorname{range}(A) = \{y \in \mathbb{R}^m \ : \ \text{there exists an } x \in \mathbb{R}^n \text{ such that } y = Ax\},\\
&\operatorname{corange}(A) = (\operatorname{range}(A))^\perp.
\end{aligned}$$

Lemma 2.12. *Let $A \in \mathbb{R}^{m,n}$, then*

$$\operatorname{cokernel}(A) = \operatorname{range}(A^T) \text{ and } \operatorname{corange}(A) = \operatorname{kernel}(A^T).$$

Proof. See [54, p. 69]. □

Definition 2.13 (**Rank and corank).** Let $A \in \mathbb{R}^{m,n}$. Then, the *rank* of the matrix A is defined by

$$\operatorname{rank}(A) = \dim(\operatorname{range}(A)),$$

where $\dim(S)$ denotes the dimension of a subspace $S \subseteq \mathbb{R}^m$. The *corank* is defined as the dimension of the corange of A, i.e.,

$$\operatorname{corank}(A) = m - \operatorname{rank}(A).$$

Definition 2.14 (**Nonsingular matrix and inverse matrix).** If for a matrix $A \in \mathbb{R}^{n,n}$, a uniquely determined matrix $X \in \mathbb{R}^{n,n}$ exists with $AX = XA = I_n$, where the matrix $I_n \in \mathbb{R}^{n,n}$ denotes the identity matrix, then the matrix A is called *nonsingular or invertible* and the matrix X is called the *inverse of* A and is denoted by $A^{-1} = X$.

Definition 2.15 (**Orthogonal matrix).** A matrix $A \in \mathbb{R}^{n,n}$ is said to be *orthogonal* if $A^T A = I_n$.

Definition 2.16 (**Orthogonal projection).** Let $S \subseteq \mathbb{R}^n$ be a subspace. A matrix $P \in \mathbb{R}^{n,n}$ is called the *orthogonal projection* onto S if $\operatorname{range}(P) = S$, $P^2 = P$, and $P^T = P$.

From Definition 2.16 it follows that $I - P$ is the orthogonal projection onto $S^\perp$. Furthermore, the orthogonal projection onto a subspace S is unique, see [54, p.75].
Generalizations of the inverse of a matrix are given by the Moore-Penrose pseudoinverse or the Drazin inverse, see [27, 54].

Definition 2.17 (**Moore-Penrose pseudo-inverse).** Let $A \in \mathbb{R}^{m,n}$. The *Moore-Penrose pseudo-inverse* of A is defined as the unique matrix $A^+ \in \mathbb{R}^{n,m}$ that satisfies the following Moore-Penrose conditions

$$AA^+A = A, \tag{2.1a}$$
$$A^+AA^+ = A^+, \tag{2.1b}$$
$$(AA^+)^T = AA^+, \tag{2.1c}$$
$$(A^+A)^T = A^+A. \tag{2.1d}$$

Lemma 2.18. *Let $A \in \mathbb{R}^{m,n}$. Then*

$$\begin{aligned} AA^+ &\quad \textit{is a projection onto} \quad \operatorname{range}(A), \\ I - AA^+ &\quad \textit{is a projection onto} \quad \operatorname{corange}(A), \\ I - A^+A &\quad \textit{is a projection onto} \quad \operatorname{kernel}(A), \\ A^+A &\quad \textit{is a projection onto} \quad \operatorname{cokernel}(A). \end{aligned}$$

Proof. See [54, p. 257-258]. □

Definition 2.19 (**Index of nilpotency).** Let $A \in \mathbb{R}^{n,n}$. The quantity

$$\nu = \min\{l \in \mathbb{N}_0 \mid \operatorname{kernel}(A^{l+1}) = \operatorname{kernel}(A^l)\}$$

is called the *index (of nilpotency)* of A and is denoted by $\nu = \operatorname{ind}(A)$.

Definition 2.20 (Drazin inverse). Let $A \in \mathbb{R}^{n,n}$ have the index $\nu = \operatorname{ind} A$. A matrix $X \in \mathbb{R}^{n,n}$ satisfying

$$AX = XA, \tag{2.2a}$$
$$XAX = X, \tag{2.2b}$$
$$XA^{\nu+1} = A^{\nu} \tag{2.2c}$$

is called the *Drazin inverse* of A and is denoted by $A^D = X$.

The Drazin inverse is unique for every matrix $A \in \mathbb{R}^{n,n}$, and for nonsingular matrices it corresponds to the inverse $A^D = A^{-1}$, see e.g. [82, Theorem 2.19, Lemma 2.10]. Further, we have the following properties of the Drazin inverse.

Lemma 2.21. *For a matrix $A \in \mathbb{R}^{n,n}$ and a nonsingular matrix $T \in \mathbb{R}^{n,n}$ we have*

$$(T^{-1}AT)^D = T^{-1}A^DT.$$

Proof. See [82, Lemma 2.20]. □

Lemma 2.22. *Consider two matrices $E, A \in \mathbb{R}^{n,n}$ that commute, i.e., $EA = AE$. Then we have*

$$EA^D = A^DE,$$
$$E^DA = AE^D,$$
$$E^DA^D = A^DE^D.$$

Proof. See [82, Lemma 2.21]. □

Theorem 2.23 (Singular value decomposition (SVD)). *Let $A \in \mathbb{R}^{m,n}$. Then there exist orthogonal matrices $U \in \mathbb{R}^{m,m}$ and $V \in \mathbb{R}^{n,n}$ such that*

$$U^TAV = \Sigma = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{R}^{r,r}$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$.

Proof. See [54, p.70]. □

The singular value decomposition allows the computation of the range, corange, kernel, and cokernel of a matrix A.

Lemma 2.24. *Let $A \in \mathbb{R}^{m,n}$ with $\operatorname{rank}(A) = r$ and suppose that $A = U\Sigma V^T \in \mathbb{R}^{m,n}$ is an SVD of A. If we have the partitioning $U = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \in \mathbb{R}^{m,m}$ with $U_1 \in \mathbb{R}^{m,r}$, $U_2 \in \mathbb{R}^{m,m-r}$ and $V = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \in \mathbb{R}^{n,n}$ with $V_1 \in \mathbb{R}^{n,r}$, $V_2 \in \mathbb{R}^{n,n-r}$, then*

$$\begin{aligned} U_1U_1^T &\quad \textit{is a projection onto } \operatorname{range}(A), \\ U_2U_2^T &\quad \textit{is a projection onto } \operatorname{corange}(A), \\ V_2V_2^T &\quad \textit{is a projection onto } \operatorname{kernel}(A), \\ V_1V_1^T &\quad \textit{is a projection onto } \operatorname{cokernel}(A). \end{aligned}$$

Proof. See [54, p. 75]. □

Note that these orthogonal projections are uniquely determined by U and V although the matrices U and V are in general not unique. There also exists a continuous version of the singular value decompositions for smooth matrix-valued functions.

Theorem 2.25. *Let $E \in C^k(\mathbb{I}, \mathbb{R}^{m,n})$, $k \in \mathbb{N}_0 \cup \{\infty\}$, with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. Then there exist pointwise orthogonal matrix-valued functions $U \in C^k(\mathbb{I}, \mathbb{R}^{m,m})$ and $V \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$, such that*

$$U^T(t)E(t)V(t) = \begin{bmatrix} \Sigma_r(t) & 0 \\ 0 & 0 \end{bmatrix}$$

with pointwise nonsingular $\Sigma_r \in C^k(\mathbb{I}, \mathbb{R}^{r,r})$.

Proof. See [82, Theorem 3.9]. □

2.2 Differential-Algebraic Equations

In this section we review the general theory of linear and nonlinear differential-algebraic equations as presented in [82]. In the most general form a nonlinear differential-algebraic equation (DAE) is given by

$$F(t, x, \dot{x}) = 0, \tag{2.3}$$

where $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{R}^m$ is a sufficiently smooth function on a compact interval $\mathbb{I} = [t_0, t_f] \subset \mathbb{R}$ and $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subseteq \mathbb{R}^n$ are open sets. In addition, an initial condition

$$x(t_0) = x_0 \in \mathbb{R}^n, \tag{2.4}$$

might be given. If not otherwise specified, we will use the classical concept of solvability yielding continuously differentiable solutions as follows.

Definition 2.26 (Solution of a DAE). Consider a nonlinear system (2.3) with a sufficiently smooth function F. A function $x : \mathbb{I} \to \mathbb{R}^n$ is called *a solution* of (2.3) if $x \in C^1(\mathbb{I}, \mathbb{R}^n)$ and x satisfies (2.3) pointwise. It is called *a solution of the initial value problem* (2.3)-(2.4) if x is a solution of (2.3) and satisfies the initial condition (2.4). An initial condition (2.4) is called *consistent* if the corresponding initial value problem has at least one solution.

This notion of solution can be weakend, since the derivative $\dot{x}$ does not occur in the kernel of E such that the solution has to be continuously differentiable only on the cokernel of E. Later on, considering switched systems we will also allow so-called strong solutions and generalized solutions in a distributional setting, see Section 5.3.

2.2.1 Linear Differential-Algebraic Equations

At first, we consider the initial value problem for linear differential-algebraic equations of the form

$$E(t)\dot{x} = A(t)x + b(t), \tag{2.5}$$

with $E, A \in C(\mathbb{I}, \mathbb{R}^{m,n})$ and $b \in C(\mathbb{I}, \mathbb{R}^m)$ together with an initial condition (2.4).
The special case of linear DAEs with constant coefficients $E, A \in \mathbb{R}^{m,n}$ of the form

$$E\dot{x} = Ax + b(t) \tag{2.6}$$

is well-understood using purely algebraic techniques, e.g., through the Kronecker canonical form [71] or the Weierstraß canonical form [150] derived via equivalence transformations of matrix pairs.

Definition 2.27 (Strong equivalence). Two pairs of matrices (E_i, A_i), $E_i, A_i \in \mathbb{R}^{m,n}$, $i = 1, 2$, are called *strongly equivalent* if there exist nonsingular matrices $P \in \mathbb{R}^{m,m}$ and $Q \in \mathbb{R}^{n,n}$, such that

$$E_2 = PE_1Q, \quad A_2 = PA_1Q. \tag{2.7}$$

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

Definition 2.28 (Regularity). Let $E, A \in \mathbb{R}^{m,n}$. The matrix pair (E, A) is called *regular* if $m = n$ and the so-called characteristic polynomial p defined by

$$p(\lambda) = \det(\lambda E - A), \quad \lambda \in \mathbb{C}$$

is not the zero polynomial. Otherwise, the matrix pair is called *singular*.

A canonical form under strong equivalence transformations (2.7) for regular matrix pairs is the *Weierstraß canonical form* given in the following theorem.

Theorem 2.29. *Let $E, A \in \mathbb{R}^{n,n}$ and (E, A) be regular. Then, there exist nonsingular matrices $W, T \in \mathbb{R}^{n,n}$ such that*

$$(E, A) \sim (WET, WAT) = \left(\begin{bmatrix} I_{n_1} & 0 \\ 0 & N \end{bmatrix}, \begin{bmatrix} J & 0 \\ 0 & I_{n-n_1} \end{bmatrix} \right), \tag{2.8}$$

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

Proof. See [82, Theorem 2.7]. □

The eigenvalues of J are called the finite eigenvalues of the pair (E, A) and subspaces $\mathcal{W}, \mathcal{T} \subset \mathbb{R}^n$ are called the *left* and *right deflating subspaces of* (E, A) if $\dim(\mathcal{W}) = \dim(\mathcal{T})$ and $\mathcal{W} = E\mathcal{T} + A\mathcal{T}$. The matrices

$$P_l = W^{-1} \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} W, \qquad P_r = T \begin{bmatrix} I_{n_1} & 0 \\ 0 & 0 \end{bmatrix} T^{-1}, \tag{2.9}$$

are the projections onto the left and right deflating subspace of (E, A) corresponding to the finite eigenvalues, see e.g. [139].

Definition 2.30 (Index of a matrix pair). Consider a pair (E, A) of square matrices that is regular and has a canonical form as in (2.8). The quantity ν defined by $N^\nu = 0$, $N^{\nu-1} \neq 0$, i.e., by the index of nilpotency of N in (2.8), if the nilpotent block in (2.8) is present and by $\nu = 0$ if it is absent, is called the *index of the matrix pair* (E, A), denoted by $\nu = ind(E, A)$.

Using Definition 2.30 the index ν of a matrix $A \in \mathbb{R}^{n,n}$ as in Definition 2.19 is also given by $\nu = \mathrm{ind}(A) = \mathrm{ind}(A, I)$. In this way, the Weierstraß canonical form (2.8) allows to determine the index of a linear DAE (2.6) with constant coefficients and to analyze the existence and uniqueness of solutions. The linear DAE (2.6) is uniquely solvable for a consistent initial value x_0 if the matrix pair (E, A) is regular. In this case, i.e., if the pair is regular, we can find a λ such that $(\lambda E - A)$ is nonsingular. Multiplication of the original system with $(\lambda E - A)^{-1}$ from the left, corresponding to a scaling of the system, does not change the solution and we get the equivalent system

$$\hat{E}\dot{x} = \hat{A}x + \hat{b}(t), \tag{2.10}$$

with $\hat{E} = (\lambda E - A)^{-1}E$, $\hat{A} = (\lambda E - A)^{-1}A$, and $\hat{b} = (\lambda E - A)^{-1}b$. Then, we can give an explicit formula for the solution of the linear system (2.6).

Theorem 2.31. *Let the matrix pair (E, A) be regular and let $b \in C^\nu(\mathbb{I}, \mathbb{R}^n)$ with $\nu = \mathrm{ind}(E)$ and $t_0 \in \mathbb{I}$. Then every solution $x \in C^1(\mathbb{I}, \mathbb{R}^n)$ of (2.6) has the form*

$$x(t) = e^{\hat{E}^D\hat{A}(t-t_0)}\hat{E}^D\hat{E}v + \int_{t_0}^{t} e^{\hat{E}^D\hat{A}(t-s)}\hat{E}^D\hat{b}(s)ds - (I - \hat{E}^D\hat{E})\sum_{i=0}^{\nu-1}(\hat{E}\hat{A}^D)^i\hat{A}^D\hat{b}^{(i)}(t)$$

for some $v \in \mathbb{R}^n$, where $\hat{E} = (\lambda E - A)^{-1}E$, $\hat{A} = (\lambda E - A)^{-1}A$, and $\hat{b} = (\lambda E - A)^{-1}b$ for some λ.

Proof. See [82], Theorem 2.29 and Lemma 2.31. □

Remark 2.32. *For commuting matrices E and A, i.e., $EA = AE$, the solution representation given in Theorem 2.31 can be formulated directly in terms of E and A, see e.g. [82, Theorem 2.29]. To obtain the solution representation for general non-commuting matrices E and A we have used a trick due to Campbell [26]. If we can find a λ such that $(\lambda E - A)$ is nonsingular, then the matrices $\hat{E} = (\lambda E - A)^{-1}E$ and $\hat{A} = (\lambda E - A)^{-1}A$ commute, see e.g. [82, Lemma 2.31].*

In the analysis of linear DAEs some more care has to be taken in the case of time dependent coefficients, see e.g. [82, Examples 3.1, 3.2]. An appropriate equivalence relation in this case can be defined as follows.

Definition 2.33 (Global equivalence). Two pairs (E_i, A_i), $i = 1, 2$, of matrix-valued functions $E_i, A_i \in C(\mathbb{I}, \mathbb{R}^{m,n})$ are called *globally equivalent* if there exist pointwise nonsingular matrix-valued functions $P \in C(\mathbb{I}, \mathbb{R}^{m,m})$, $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$E_2 = PE_1Q, \quad A_2 = PA_1Q - PE_1\dot{Q}. \tag{2.11}$$

There exists a condensed form for pairs of matrix-valued functions via global equivalence transformations (2.11), which allows to extract the characteristic quantities of the corresponding DAE (2.5). In the following, we say that a matrix is a basis of a vector space if this is valid for its columns. We additionally use the convention that the only basis of the vector space $\{0\} \subseteq \mathbb{R}^n$ is given by the empty matrix $\emptyset_{n,0} \in \mathbb{R}^{n,0}$ with the properties $\operatorname{rank} \emptyset_{n,0} = 0$ and $\det \emptyset_{0,0} = 1$. For a given matrix T, we use the notation T' to denote a matrix that completes T to a nonsingular matrix, i.e., $[T\ T']$ forms a nonsingular matrix. This also applies to matrix-valued functions.

Theorem 2.34. *Let $E, A \in C(\mathbb{I}, \mathbb{R}^{m,n})$ be sufficiently smooth and let*

$$\begin{aligned} &T \text{ be a basis of } \operatorname{kernel} E, \\ &Z \text{ be a basis of } \operatorname{corange} E, \\ &T' \text{ be a basis of } \operatorname{cokernel} E, \\ &V \text{ be a basis of } \operatorname{corange}(Z^T A T), \end{aligned}$$

with local characteristic values

$$\begin{aligned} &r = \operatorname{rank} E, && \text{(rank)} \\ &a = \operatorname{rank}(Z^T A T), && \text{(algebraic part)} \\ &s = \operatorname{rank}(V^T Z^T A T'), && \text{(strangeness)} \\ &d = r - s, && \text{(differential part)} \\ &u = n - r - a, && \text{(undetermined variables)} \\ &v = m - r - a - s, && \text{(vanishing equations)} \end{aligned}$$

and suppose that

$$r(t) \equiv r, \quad a(t) \equiv a, \quad s(t) \equiv s \quad \text{for all } t \in \mathbb{I} \tag{2.12}$$

holds for the local characteristic values of (E, A). Then, the pair (E, A) is globally equivalent to a pair of matrix-valued functions of the form

$$\left(\begin{bmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \begin{matrix} s \\ d \\ a \\ s \\ v \end{matrix}. \tag{2.13}$$

Here, all entries A_{ij} are again matrix-valued functions on $\mathbb{I}$ and the last block column in both matrix-valued functions has size u.

Proof. See [82, Theorem 3.11]. □

By a stepwise reduction procedure involving differentiations of some of the algebraic equations we can eliminate the strangeness in the global condensed form (2.13) and finally obtain a so-called *strangeness-free* system where $s = 0$. The number of steps involved in this reduction procedure defines the strangeness index of the DAE.

Definition 2.35 (Strangeness index (s-index)). The minimum number of times μ that all or part of the differential-algebraic equations (2.5) have to be differentiated in order to obatin a system of purely ordinary differential equations and algebraic equations is called the *strangeness index (s-index)* of the DAE.

The strangeness index is well-defined if the assumptions of Theorem 2.34, and in particular the regularity assumptions (2.12) hold.

Theorem 2.36. *Let the strangeness index μ of (E, A) be well-defined and let $b \in C^{\mu}(\mathbb{I}, \mathbb{R}^m)$. Then the linear differential-algebraic equation* (2.5) *is equivalent (in the sense that there is a one-to-one correspondence between the solution sets) to a strangeness-free differential-algebraic system of the form*

$$\begin{aligned} \dot{x}_1 &= A_{13}(t)x_3 + b_1(t), && (d_\mu \text{ differential equations}) \\ 0 &= x_2 + b_2(t), && (a_\mu \text{ algebraic equations}) \\ 0 &= b_3(t), && (v_\mu \text{ consistency conditions}) \end{aligned} \tag{2.14}$$

where $A_{13} \in C(\mathbb{I}, \mathbb{R}^{d_\mu, u_\mu})$ and the inhomogeneities b_1, b_2, b_3 are determined by the derivatives $b^{(0)}, \ldots, b^{(\mu)}$.

Proof. See [82, Theorem 3.17]. □

The strangeness-free form (2.14) allows to decide on existence and uniqueness of solutions of the linear DAE (2.5), see e.g. [82].

2.2.2 Nonlinear Differential-Algebraic Equations

In this section, we consider nonlinear differential-algebraic equations of the form (2.3). As a general approach for the analysis of general nonlinear DAEs, Campbell introduced the derivative array, which summarizes the original equations of the DAE and all its derivatives up to a certain order l in one large system, see [22, 25, 78]. The derivative array $\mathcal{F}_l$ of level $l \in \mathbb{N}_0$ is given by the inflated system

$$\mathcal{F}_l(t, x, \dot{x}, \ldots, x^{(l+1)}) = 0, \tag{2.15}$$

where $\mathcal{F}_l$ has the form

$$\mathcal{F}_l(t, x, \dot{x}, \ldots, x^{(l+1)}) = \begin{bmatrix} F(t, x, \dot{x}) \\ \frac{d}{dt}F(t, x, \dot{x}) \\ \vdots \\ \frac{d^l}{dt^l}F(t, x, \dot{x}) \end{bmatrix}.$$

We introduce the *solution set* of the nonlinear equation associated with the derivative array $\mathcal{F}_l$ for some integer l, given by

$$\mathbb{L}_l = \{(t, x, \dot{x}, \ldots, x^{(l+1)}) \in \mathbb{R}^{(l+2)n+1} \mid \mathcal{F}_l(t, x, \dot{x}, \ldots, x^{(l+1)}) = 0\}, \tag{2.16}$$

and we define the Jacobians

$$\begin{aligned} \mathcal{M}_l(t, x, \dot{x}, \ldots, x^{(l+1)}) &= \mathcal{F}_{l;\dot{x},\ldots,x^{(l+1)}}(t, x, \dot{x}, \ldots, x^{(l+1)}), \\ \mathcal{N}_l(t, x, \dot{x}, \ldots, x^{(l+1)}) &= -[\mathcal{F}_{l;x}(t, x, \dot{x}, \ldots, x^{(l+1)}), 0, \ldots, 0]. \end{aligned} \tag{2.17}$$

The following hypothesis was stated in [79]. Note, that we will use the convention that $\operatorname{corank} \mathcal{F}_{-1;x} = 0$.

Hypothesis 2.37. *Consider a nonlinear differential-algebraic equations* (2.3)*. There exist integers μ, r, a_μ, d_μ and v_μ such that the solution set $\mathbb{L}_\mu$ is nonempty and such that for every point $(t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ there exists a (sufficiently small) neighborhood in which the following properties hold:*

1. *The set $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+2)n+1}$ forms a manifold of dimension $(\mu+2)n+1-r$.*
2. *We have $\operatorname{rank} \mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}} = r$ on $\mathbb{L}_\mu$.*
3. *We have $\operatorname{corank} \mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}} - \operatorname{corank} \mathcal{F}_{\mu-1;x,\dot{x},\ldots,x^{(\mu)}} = v_\mu$ on $\mathbb{L}_\mu$.*
4. *We have $\operatorname{rank} \mathcal{M}_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = r - a_\mu$ on $\mathbb{L}_\mu$ such that there exist smooth matrix functions Z_2 and T_2 defined on $\mathbb{L}_\mu$ of size $((\mu+1)m, a_\mu)$ and $(n, n-a_\mu)$, respectively, and pointwise maximal rank, satisfying*

$$Z_2^T \mathcal{M}_\mu = 0, \quad \operatorname{rank} Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T = a_\mu, \quad Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T T_2 = 0$$

   on $\mathbb{L}_\mu$.

5. *We have $\operatorname{rank} F_{;\dot{x}} T_2 = d_\mu = m - a_\mu - v_\mu$ on $\mathbb{L}_\mu$ such that there exists a smooth matrix function Z_1 defined on $\mathbb{L}_\mu$ of size (m, d_μ) and pointwise maximal rank, satisfying $\operatorname{rank} Z_1^T F_{;\dot{x}} T_2 = d_\mu$.*

In Hypothesis 2.37 we have omitted the function arguments for convenience. Further, note that the matrix functions Z_1, Z_2, and T_2 are smooth along a solution with respect to t, even if the matrix functions formally are defined on $\mathbb{L}_\mu$. When a nonlinear DAE (2.3) satisfies Hypothesis 2.37, then we call the smallest possible μ the *strangeness index* (s-index) of (2.3). A nonlinear system (2.3) with vanishing strangeness index $\mu = 0$ is called *strangeness-free*. The corresponding numbers d_μ and a_μ are the numbers of differential and algebraic equations of the DAE and the quantity v_μ measures the number of redundant equations such that a complete classification of the equations is given.

Definition 2.38 (Regularity). A nonlinear DAE (2.3) that satisfies Hypothesis 2.37 with $n = m = d_\mu + a_\mu$, i.e., $v_\mu = 0$, is called *regular*.

Under Hypothesis 2.37 projection matrices Z_1, Z_2 and T_2 can be computed for every point $z_\mu \in \mathbb{L}_\mu$, which allow locally to construct a reduced strangeness-free system with the same solution set as the original DAE and separated differential and algebraic parts. Let $z_{\mu,0} = (t_0, x_0, \dot{x}_0, \ldots, x_0^{(\mu+1)}) \in \mathbb{L}_\mu$ be fixed. By assumption $\mathbb{L}_\mu$ is a manifold of dimension $(\mu+2)n+1-r$ that can locally be parameterized by $(\mu+2)n+1-r$ parameters. These parameters can be chosen from $(t, x, \dot{x}, \ldots, x^{(\mu+1)})$ in such a way that discarding the associated columns from $\mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}}(z_{\mu,0})$ does not lead to a rank drop. As $\mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+1)}}$ already has maximal rank r, t can always be chosen as a parameter. Since

$$\operatorname{corank} \mathcal{M}_\mu(z_{\mu,0}) = a_\mu, \quad \operatorname{rank}(Z_2(z_{\mu,0})^T \mathcal{N}(z_{\mu,0}) \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T) = a_\mu,$$

we can choose $n - a_\mu$ parameters out of x. Without restriction we can write x as (x_1, x_2, x_3) with $x_1 \in \mathbb{R}^{d_\mu}$, $x_2 \in \mathbb{R}^{n-a_\mu-d_\mu}$, $x_3 \in \mathbb{R}^{a_\mu}$, and choose (x_1, x_2) as these $n - a_\mu$ parameters. In particular, the matrix $Z_2^T \mathcal{F}_{\mu;x_3}$ is then nonsingular. The remaining parameters $p \in \mathbb{R}^{(\mu+1)n+a_\mu-r}$ associated with the columns of $\mathcal{F}_{\mu;t,x,\ldots,x^{(\mu+1)}}(z_{\mu,0})$ that we can remove without having a rank drop, must then be chosen out of $(\dot{x}, \ldots, x^{(\mu+1)})$. Let $(t_0, x_{1,0}, x_{2,0}, p_0)$ be the part of $z_{\mu,0}$ that corresponds to the selected parameters (t, x_1, x_2, p). The implicit function theorem (Theorem 2.9) then implies that there exists a neighborhood $\mathbb{V} \subseteq \mathbb{R}^{(\mu+2)n+1-r}$ of $(t_0, x_{1,0}, x_{2,0}, p_0)$ and a neighborhood $\tilde{\mathbb{U}} \subseteq \mathbb{R}^{(\mu+2)n+1}$ of $z_{\mu,0}$ such that

$$\mathbb{U} = \mathbb{L}_\mu \cap \tilde{\mathbb{U}} = \{\theta(t, x_1, x_2, p) \mid (t, x_1, x_2, p) \in \mathbb{V}\},$$

where $\theta : \mathbb{V} \to \mathbb{U}$ is a diffeomorphism. Thus, the equation

$$\mathcal{F}_\mu(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = 0$$

can be locally solved according to

$$(t, x, \dot{x}, \ldots, x^{(\mu+1)}) = \theta(t, x_1, x_2, p)$$

for some $(t, x_1, x_2, p) \in \mathbb{U}$. In particular, there exist locally defined functions G, corresponding to x_3, and H, corresponding to $(\dot{x}, \ldots, x^{(\mu+1)})$ such that

$$\mathcal{F}_\mu(t, x_1, x_2, G(t, x_1, x_2, p), H(t, x_1, x_2, p)) = 0 \tag{2.18}$$

on $\mathbb{V}$. Setting $y = (\dot{x}, \ldots, x^{(\mu+1)})$, it follows with Z_2 as defined by Hypothesis 2.37 that

$$\frac{d}{dp}(Z_2^T \mathcal{F}_\mu) = (Z_{2;x_3}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;x_3}) G_{;p} + (Z_{2;y}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;y}) H_{;p} = Z_2^T \mathcal{F}_{\mu;x_3} G_{;p} = 0,$$

on $\mathbb{V}$, since $\mathcal{F}_\mu = 0$ and $Z_2^T \mathcal{F}_{\mu;y} = Z_2^T \mathcal{M}_\mu = 0$. By construction the variables in x_3 were selected such that $Z_2^T \mathcal{F}_{\mu;x_3}$ is nonsingular. Hence,

$$G_{;p}(t, x_1, x_2, p) = 0$$

for all $(t, x_1, x_2, p) \in \mathbb{V}$, implying the existence of a function R such that

$$x_3 = G(t, x_1, x_2, p) = G(t, x_1, x_2, p_0) = R(t, x_1, x_2),$$

and

$$\mathcal{F}_\mu(t, x_1, x_2, R(t, x_1, x_2), H(t, x_1, x_2, p)) = 0$$

on $\mathbb{V}$. In a similar way we then get that

$$\begin{aligned}\frac{d}{dx_1}(Z_2^T \mathcal{F}_\mu) &= (Z_{2;x_1}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;x_1}) + (Z_{2;x_3}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;x_3}) R_{;x_1} + (Z_{2;y}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;y}) H_{;x_1} \\ &= Z_2^T \mathcal{F}_{\mu;x_1} + Z_2^T \mathcal{F}_{\mu;x_3} R_{;x_1} = 0\end{aligned}$$

and

$$\begin{aligned}\frac{d}{dx_2}(Z_2^T \mathcal{F}_\mu) &= (Z_{2;x_2}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;x_2}) + (Z_{2;x_3}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;x_3}) R_{;x_2} + (Z_{2;y}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;y}) H_{;x_2} \\ &= Z_2^T \mathcal{F}_{\mu;x_2} + Z_2^T \mathcal{F}_{\mu;x_3} R_{;x_2} = 0\end{aligned}$$

on $\mathbb{V}$, again using that $\mathcal{F}_\mu = 0$ and $Z_2^T \mathcal{F}_{\mu;y} = 0$. Thus,

$$Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T \begin{bmatrix} I_{n-a_\mu} \\ R_{;x_1,x_2} \end{bmatrix} = 0.$$

Following Hypothesis 2.37 we can therefore choose T_2 as

$$T_2(t, x_1, x_2) = \begin{bmatrix} I_{n-a_\mu} \\ R_{;x_1,x_2}(t, x_1, x_2) \end{bmatrix}.$$

Thus, Hypothesis 2.37 yields a matrix function Z_1 which only depends on $(t, x, \dot{x})$. Due to the full rank assumption, we can choose the neighborhood $\mathbb{V}$ so small that we can take a constant Z_1. The corresponding *reduced differential-algebraic equation* therefore reads

$$\hat{F}(t, x, \dot{x}) = \begin{bmatrix} \hat{F}_1(t, x, \dot{x}) \\ \hat{F}_2(t, x) \end{bmatrix} = 0, \tag{2.19}$$

with

$$\hat{F}_1(t, x, \dot{x}) = Z_1^T F(t, x, \dot{x}), \quad \hat{F}_2(t, x) = Z_2^T \mathcal{F}_\mu(t, x, H(t, x)).$$

Theorem 2.39. *The reduced system* (2.19) *satisfies Hypothesis 2.37 with characteristic values* $\mu = 0$, $r = a_\mu + d_\mu$, a_μ, d_μ, *and* v_μ.

Proof. See [82, p. 208]. □

The condition $\hat{F}_2(t, x) = 0$ is locally equivalent via the implicit function theorem to a relation $x_3 = R(t, x_1, x_2)$ such that we get the system

$$\hat{F}_1(t, x_1, x_2, x_3, \dot{x}_1, \dot{x}_2, \dot{x}_3) = 0, \tag{2.20a}$$

$$x_3 - R(t, x_1, x_2) = 0. \tag{2.20b}$$

We can eliminate x_3 and $\dot{x}_3$ in (2.20a) using (2.20b) and its time derivative to obtain

$$\hat{F}_1(t, x_1, x_2, R(t, x_1, x_2), \dot{x}_1, \dot{x}_2, R_{;t}(t, x_1, x_2) + R_{;x_1}(t, x_1, x_2)\dot{x}_1 + R_{;x_2}(t, x_1, x_2)\dot{x}_2) = 0.$$

By Hypothesis 2.37 we may assume without loss of generality that this system can be locally solved for $\dot{x}_1$. In this way, we obtain a decoupled differential-algebraic system of the form

$$\begin{aligned} \dot{x}_1 &= L(t, x_1, x_2, \dot{x}_2), \\ x_3 &= R(t, x_1, x_2), \end{aligned} \tag{2.21}$$

which has a vanishing strangeness index $\mu = 0$ with d_μ differential and a_μ algebraic equations. In this system $x_2 \in C^1(\mathbb{I}, \mathbb{R}^{u_\mu})$ with $u_\mu = n - d_\mu - a_\mu$ can be chosen arbitrarily. Then, the resulting system has locally a unique solution for x_1 and x_3, provided that consistent initial values are given. This means that x_2 can be interpreted as a control.

Theorem 2.40. *Let F as in (2.3) be sufficiently smooth and satisfy Hypothesis 2.37 with characteristic values $\mu, r, a_\mu, d_\mu, v_\mu$, and $u_\mu = n - d_\mu - a_\mu$. Then every sufficiently smooth solution of (2.3) also solves the reduced DAE (2.19) and (2.21) consisting of d_μ differential and a_μ algebraic equations.*

Proof. See [82, Theorem 4.31]. □

In the case of linear DAEs (2.5) with sufficiently smooth matrix-valued functions E, A and inhomogeneity b the inflated system (2.15) takes the form

$$\mathcal{M}_l(t)\dot{z}_l = \mathcal{N}_l(t)z_l + g_l(t), \quad l = 0, \ldots, \mu, \tag{2.22}$$

where for $i, j = 0, \ldots, l$ we have

$$\begin{aligned} [\mathcal{M}_l]_{i,j} &= \binom{i}{j} E^{(i-j)} - \binom{i}{j+1} A^{(i-j-1)}, \quad i, j = 0, \ldots, l, \\ [\mathcal{N}_l]_{i,j} &= \begin{cases} A^{(i)} & \text{for } i = 0, \ldots, l,\ j = 0, \\ 0 & \text{otherwise}, \end{cases} \\ [z_l]_j &= x^{(j)}, \quad j = 0, \ldots, l, \\ [g_l]_i &= b^{(i)}, \quad i = 0, \ldots, l, \end{aligned} \tag{2.23}$$

using the convention that $\binom{i}{j} = 0$ for $i < 0$, $j < 0$ or $j > i$. In this case the Hypothesis 2.37 can be formulated as a Theorem.

Theorem 2.41. *Let the strangeness-index μ of (E, A) in (2.5) be well-defined. Then there exist integers μ, a_μ, d_μ, $\hat{v}$ and u_μ such that the inflated pair $(\mathcal{M}_\mu, \mathcal{N}_\mu)$ associated with (E, A) has the following properties:*

1. *For all* $t \in \mathbb{I}$ *it holds that* $\operatorname{rank} \mathcal{M}_\mu(t) = (\mu+1)n - a_\mu - \hat{v}$*, such that there exists a smooth matrix function* Z *with orthonormal columns and size* $((\mu+1)m, a_\mu + \hat{v})$ *and pointwise maximal rank satisfying* $Z^T \mathcal{M}_\mu = 0$*.*

2. *For all* $t \in \mathbb{I}$ *we have* $\operatorname{rank} Z^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T = a_\mu$ *and without loss of generality* Z *can be partitioned as* $[Z_2, Z_3]$*, with* Z_2 *of size* $((\mu+1)n, a_\mu)$ *and* Z_3 *of size* $((\mu+1)n, \hat{v})$*, such that* $\hat{A}_2 = Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T$ *has full row rank* a_μ *and* $Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T = 0$*. Furthermore, there exists a smooth matrix function* T_2 *with orthonormal columns and size* $(n, d_\mu + u_\mu)$*, satisfying* $\hat{A}_2 T_2 = 0$*.*

3. *For all* $t \in \mathbb{I}$ *it holds that* $\operatorname{rank} E T_2 = d_\mu$*, such that there exists a smooth matrix function* Z_1 *with orthonormal columns and size* (m, d_μ) *yielding that* $Z_1^T E$ *has constant rank* d_μ*.*

Proof. See [84, Theorem 11]. □

Remark 2.42. *The integers* μ*,* a_μ*,* d_μ *and* u_μ *in Theorem 2.41 correspond to the characteristic values of the strangeness-free system* (2.14)*.*

Using Theorem 2.41, it is possible to derive an equivalent strangeness-free system of the form

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{b}(t), \tag{2.24}$$

with $\hat{E} = \begin{bmatrix} \hat{E}_1 \\ 0 \\ 0 \end{bmatrix}$, $\hat{A} = \begin{bmatrix} \hat{A}_1 \\ \hat{A}_2 \\ 0 \end{bmatrix}$ and $\hat{b} = \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix}$ defined by

$$\begin{aligned} &\hat{E}_1 = Z_1^T E, \quad \hat{A}_1 = Z_1^T A, \quad \hat{A}_2 = Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T, \\ &\hat{b}_1 = Z_1^T f, \quad \hat{b}_i = Z_i^T g_\mu \quad \text{for } i = 2, 3. \end{aligned}$$

In general, the solution of a DAE is restricted to a subset $\mathbb{L} \subseteq \mathbb{I} \times \mathbb{R}^n$ by certain algebraic constraints implicitly contained in the DAE. In the following, this subset is called the *constraint manifold*. The constraint manifold of a DAE is defined by all constraints, i.e., explicitly given algebraic constraints in the DAE and additionally hidden constraints that are not stated explicitly as equations but results from differentiations of certain parts of the DAE in higher index problems. For nonlinear DAEs given in the reduced strangeness-free forms (2.19) or (2.21) the constraint manifold is explicitly given by

$$\mathbb{L} = \{(t, x) \in \mathbb{I} \times \mathbb{R}^n \mid \hat{F}_2(t, x) = 0\},$$

or

$$\mathbb{L} = \{(t, x_1, x_2, x_3) \in \mathbb{I} \times \mathbb{R}^n \mid R(t, x_1, x_2) = x_3\},$$

respectively, and the corresponding differential parts $\hat{F}_1(t, x, \dot{x}) = 0$ in (2.19) or $\dot{x}_1 = L(t, x_1, x_2, \dot{x}_2)$ in (2.21) then describe the dynamical behavior inside the constraint manifold. Thus, the constraint manifold $\mathbb{L}$ is a subset of the solution set $\mathbb{L}_\mu$ of the derivative array defined in (2.16). For a linear DAE in reduced form (2.24) the constraint manifold is given by

$$\mathbb{L} = \{(t, x) \in \mathbb{I} \times \mathbb{R}^n \mid \hat{A}_2(t)x + \hat{b}_2(t) = 0\}.$$

In particular, initial values (2.4) have to lie in the constrained manifold $\mathbb{L}$ in order to be consistent with the DAE.

2.2.3 Generalized Functions and Distributional Solutions

In the previous section is was assumed that the solution of a DAE is a continuously differentiable function $x : \mathbb{I} \to \mathbb{R}^n$. It is also possible to allow jumps or discontinuities in the solution at a number of distinct points in a distributional setting. In particular, this will be important when considering switched differential-algebraic systems. In this section, we recall a few important facts about generalized functions, see e.g. [82, 133]. In the distributional setting we follow an approach based on the space of impulsive smooth distributions introduced in [62] and for DAEs in [51, 52, 118, 119] for a single point of discontinuity. In [82] ideas for the extension to the case, where nonsmooth behavior occurs at a countable number of points is proposed. This approach allows generalized functions (or distributions) as solution of a differential-algebraic system, relaxed smoothness requirements for the inhomogeneities, as well as non-differentiability or discontinuous inhomogeneities, and non-consistent initial values. The proofs of the theorems given in this section are all given in [82] for a single point of discontinuity at $t = 0$. If nonsmooth behavior of the solution occurs at points $\tau_j \in \mathcal{T}$, where the set $\mathcal{T}$ has no accumulation points, then all results can be obtained in an analogous way. In the following, we denote by $\mathcal{D}^n = C_0^\infty(\mathbb{R}, \mathbb{R}^n)$ the set of infinitely differentiable functions with values in $\mathbb{R}^n$ and compact support in $\mathbb{R}$. The elements of $\mathcal{D}^n$ are also called *test functions*.

Definition 2.43 (Generalized function/Distribution). A linear function $\mathrm{f} : \mathcal{D}^n \to \mathbb{R}^n$ with

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

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

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

$$\mathrm{f}^{(q)}(\phi) = (-1)^q\mathrm{f}(\phi^{(q)}) \text{ for all } \phi \in \mathcal{D}^n.$$

The so obtained functional $\mathrm{f}^{(q)}$ is linear and continuous, hence, every distribution has derivatives of arbitrary order in $\mathcal{C}^n$, see e.g. [133]. For a given distribution $\mathrm{f} \in \mathcal{C}^n$ any distribution $\mathrm{x} \in \mathcal{C}^n$ which satisfies

$$\dot{\mathrm{x}}(\phi) = \mathrm{f}(\phi) \text{ for every } \phi \in \mathcal{D}^n$$

is called a *primitive* of f, i.e., x is a solution of $\dot{\mathrm{x}} = \mathrm{f}$. Further, the *Dirac delta distribution* $\delta_a \in \mathcal{C}^n$ is defined by

$$\delta_a(\phi) = \phi(a) \quad \text{for all} \quad \phi \in \mathcal{D}^n, \quad a \in \mathbb{R},$$

and we can use multiplication by matrix-valued functions in the form

$$A\mathrm{x}(\phi) = \mathrm{x}(A^T\phi) \text{ for all } \phi \in \mathcal{D}^n,$$

where $A \in C^\infty(\mathbb{R}, \mathbb{R}^{m,n})$ and $\mathrm{x} \in \mathcal{C}^n$. Note the difference in notation for continuous functions $x \in C(\mathbb{R}, \mathbb{R}^n)$ and distributions $\mathrm{x} \in \mathcal{C}^n$. For ease of notation we treat every function $x : \mathbb{I} \to \mathbb{R}^n$, $\mathbb{I} \subseteq \mathbb{R}$ as being defined on $\mathbb{R}$ by trivially extending it by zero, i.e., setting $x(t) = 0$ for $t \notin \mathbb{I}$. Nonsmooth behavior of the solution is restricted to happen at a countable number of points $\tau_j \in \mathcal{T} \subset \mathbb{R}$. Away from τ_j the solution should be as smooth as the solution in the classical sense.

Definition 2.44 (Impulsive smooth distribution). Suppose that the set $\mathcal{T} = \{\tau_j \in \mathbb{R} \,|\, \tau_j < \tau_{j+1}, j \in \mathbb{Z}\}$ has no accumulation point. A generalized function $\mathrm{x} \in \mathcal{C}^n$ is called *impulsive smooth* if it can be written in the form

$$\mathrm{x} = \hat{\mathrm{x}} + \mathrm{x}_{imp}, \quad \hat{\mathrm{x}} = \sum_{j\in\mathbb{Z}} \hat{x}_j, \tag{2.25}$$

where $\hat{x}_j \in C^\infty([\tau_j, \tau_{j+1}], \mathbb{R}^n)$ for all $j \in \mathbb{Z}$ and the *impulsive part* x_{imp} has the form

$$\mathrm{x}_{imp} = \sum_{j\in\mathbb{Z}} \mathrm{x}_{imp,j}, \quad \mathrm{x}_{imp,j} = \sum_{i=0}^{q_j} c_{ij}\delta_{\tau_j}^{(i)}, \qquad c_{ij} \in \mathbb{C},\ q_j \in \mathbb{N}_0. \tag{2.26}$$

The set of impulsive smooth distributions is denoted by $\mathcal{C}^n_{imp}(\mathcal{T})$.

Lemma 2.45. *Impulsive smooth distributions in $\mathcal{C}^n_{imp}(\mathcal{T})$, where $\mathcal{T} \subset \mathbb{R}$ has no accumulation point, have the following properties:*

1. *A distribution $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ uniquely determines the decomposition* (2.25).
2. *With a distribution $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$, we can assign a function value $x(t)$ for every $t \in \mathbb{R}\backslash\mathcal{T}$ by*

$$x(t) = \hat{x}_j(t) \quad \text{for } t \in (\tau_j, \tau_{j+1}),$$

and limits

$$\mathrm{x}(\tau_j^-) = \lim_{t\to\tau_j^-} \hat{x}_{j-1}(t), \quad \mathrm{x}(\tau_j^+) = \lim_{t\to\tau_j^+} \hat{x}_j(t),$$

for every $\tau_j \in \mathcal{T}$.

3. *All derivatives and primitives of* $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ *are again in* $\mathcal{C}^n_{imp}(\mathcal{T})$.

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

$$A\mathrm{x} = A\hat{\mathrm{x}} + \sum_{j\in\mathbb{Z}} \sum_{i=0}^{q_j} \sum_{k=0}^{q_j - i} (-1)^k \binom{k+i}{k} A^{(k)}(\tau_j) c_{i+k,j} \delta^{(i)}_{\tau_j} \tag{2.27}$$

for x *as in* (2.25).

Proof. See [82, Lemma 2.38]. □

Further, we introduce a measure for the smoothness of impulsive smooth distributions.

Definition 2.46 (Impulse order). Let the impulsive part of $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ have the form (2.26). The *impulse order of* x *at* $\tau_j \in \mathcal{T}$ is defined as $\operatorname{iord} \mathrm{x}|_{\tau_j} = -q - 2$ if x can be associated with a continuous function in $[\tau_{j-1}, \tau_{j+1}]$ and q with $0 \leq q \leq \infty$ is the largest integer such that $\mathrm{x}|_{[\tau_{j-1},\tau_{j+1}]} \in C^q([\tau_{j-1}, \tau_{j+1}], \mathbb{R}^n)$. It is defined as $\operatorname{iord} \mathrm{x}|_{\tau_j} = -1$ if x can be associated with a function that is continuous in $[\tau_{j-1}, \tau_{j+1}]$ except at $t = \tau_j$, and it is defined as

$$\operatorname{iord} \mathrm{x}|_{\tau_j} = \max\{i \in \mathbb{N}_0 |\ 0 \leq i \leq q_j,\ c_{ij} \neq 0\}$$

otherwise. Further, the *impulse order* of x is defined as

$$\operatorname{iord} \mathrm{x} = \max_{\tau_j \in \mathcal{T}} \operatorname{iord} \mathrm{x}|_{\tau_j}.$$

Lemma 2.47. *Let* $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ *and* $A \in C^\infty(\mathbb{R}, \mathbb{R}^{m,n})$. *Then*

$$\operatorname{iord} A\mathrm{x} \leq \operatorname{iord} \mathrm{x}$$

with equality for $m = n$ *and* $A(\tau_j)$ *invertible for each* $\tau_j \in \mathcal{T}$.

Proof. For the impulse order of x at every $\tau_j \in \mathcal{T}$ the inequality $\operatorname{iord} A\mathrm{x}|_{\tau_j} \leq \operatorname{iord} \mathrm{x}|_{\tau_j}$ follows directly from (2.27), see also [82, Lemma 2.40]. Thus, the inequality holds also for the maximum over all $\tau_j \in \mathcal{T}$. □

To describe impulsive smooth solutions $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ for a linear distributional DAE

$$E(t)\dot{\mathrm{x}} = A(t)\mathrm{x} + \mathrm{b}, \tag{2.28}$$

with $\mathrm{b} \in \mathcal{C}^m_{imp}(\mathcal{T})$, we must require that $E, A \in C^\infty(\mathbb{R}, \mathbb{R}^{m,n})$ in order to have well-defined products $E\dot{\mathrm{x}}$ and $A\mathrm{x}$.

Theorem 2.48. *Let $E, A \in C^\infty(\mathbb{R}, \mathbb{R}^{m,n})$ and let the strangeness index μ of (E, A) be well-defined. Furthermore, let* $\mathrm{b} \in \mathcal{C}^m_{imp}(\mathcal{T})$ *with* $\operatorname{iord} \mathrm{b} = q \in \mathbb{Z} \cup \{-\infty\}$. *Then the differential-algebraic system* (2.28) *is equivalent (in the sense that there is a one-to-one correspondence between the solution sets) to a strangeness-free system of the form*

$$\begin{aligned} \dot{\mathrm{x}}_1 &= A_{13}(t)\mathrm{x}_3 + \mathrm{b}_1(t), && (d_\mu \text{ differential equations}) \\ 0 &= \mathrm{x}_2 + \mathrm{b}_2(t), && (a_\mu \text{ algebraic equations}) \\ 0 &= \mathrm{b}_3(t), && (v_\mu \text{ vanishing equations}) \end{aligned} \tag{2.29}$$

where $A_{13} \in C^\infty(\mathbb{R}, \mathbb{R}^{d_\mu, u_\mu})$ *and* $\operatorname{iord} \begin{bmatrix} \mathrm{b}_1 \\ \mathrm{b}_2 \\ \mathrm{b}_3 \end{bmatrix} \leq q + \mu$.

Proof. The constructions that lead to the condensed form (2.13) can be done by using infinitely often differentiable matrix functions due to Theorem 2.25. Then the form (2.29) follows directly from Theorem 2.36, where the inhomogeneities are determined from the derivatives of b via transformations with infinitely often differentiable matrix functions. See also [82, Theorem 3.72]. □

Corollary 2.49. *Let $E, A \in C^\infty(\mathbb{R}, \mathbb{R}^{m,n})$ satisfy the assumptions of Theorem 2.48. Then we have:*

1. *The problem* (2.28) *has a solution in $\mathcal{C}^n_{imp}(\mathcal{T})$ if and only if the v_μ distributional conditions*

$$\mathrm{b}_3 = 0 \tag{2.30}$$

   are fulfilled.

2. *Let $t_0 \in \mathbb{R} \backslash \mathcal{T}$ and $x_0 \in \mathbb{R}^n$. There exists a solution* $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ *satisfying one of the initial conditions*

$$\mathrm{x}(t_0) = x_0, \quad \mathrm{x}(\tau_j^-) = x_0, \quad \mathrm{x}(\tau_j^+) = x_0, \quad \text{for some } \tau_j \in \mathcal{T}, \tag{2.31}$$

   if and only if in addition to (2.30) *the corresponding condition out of*

$$\mathrm{x}_2(t_0) = -\mathrm{b}_2(t_0), \quad \mathrm{x}_2(\tau_j^-) = -\mathrm{b}_2(\tau_j^-), \quad \mathrm{x}_2(\tau_j^+) = -\mathrm{b}_2(\tau_j^+)$$

   is implied by the initial condition.

3. *The corresponding initial value problem has a unique solution in $\mathcal{C}^n_{imp}(\mathcal{T})$ if and only if in addition it holds that*

$$u_\mu = 0.$$

Moreover, all solutions x *satisfy* $\operatorname{iord} \mathrm{x} \leq \max\{q + \mu, \operatorname{iord} \mathrm{x}_3\}$.

Proof. The proof follows directly from the strangeness-free form (2.29) in Theorem 2.48. □

In the following, we will call a solution $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$ of (2.28) a *generalized solution* of the DAE. To treat initial conditions for the DAE (2.28) in the distributional setting we decompose the system as in (2.25) into

$$\mathrm{x} = \hat{\mathrm{x}} + \mathrm{x}_{imp}, \quad \hat{\mathrm{x}} = \sum_{j\in\mathbb{Z}} \hat{x}_j, \quad \text{and} \quad \mathrm{b} = \hat{\mathrm{b}} + \mathrm{b}_{imp}, \quad \hat{\mathrm{b}} = \sum_{j\in\mathbb{Z}} \hat{b}_j, \tag{2.32}$$

and therefore consider the DAEs

$$E(t)\dot{\mathrm{x}} = A(t)\mathrm{x} + \hat{b}_j, \quad t \in [\tau_j, \tau_{j+1}], \quad \text{for all } j \in \mathbb{Z}, \tag{2.33}$$

with $\hat{b}_j \in C^\infty([\tau_j, \tau_{j+1}], \mathbb{R}^m)$. The problem of solving (2.33) with $x(\tau_j) = x_{j,0}$ for an arbitrary $x_{j,0} \in \mathbb{R}^n$ that is not necessarily consistent with the DAE at τ_j often arises when a known function is to be extended into a solution of the current DAE. If $x(\tau_j^-) = x_{j,0}$ exists for some past solution $x(t)$ defined for $t < \tau_j$, then this would be a natural initial condition at time τ_j, but it does not have to be consistent with the DAE (2.33). Inconsistent initial conditions can be treated as impulses in the inhomogeneity. We can change the inhomogeneity of the DAE such that it satisfies a given history and solve the modified system. Let the distributional DAE (2.28) satisfies the assumptions of Theorem 2.48 with $u_\mu = v_\mu = 0$ and suppose that a function $x_0^j \in C^\infty([\tau_{j-1}, \tau_j], \mathbb{R}^n)$ is given that describes a certain history for the system. The initial condition $x_{j,0} = x_0^j(\tau_j)$, however, may not be consistent for (2.33). Setting

$$\hat{b}_j = E(t)\dot{x}_0^j - A(t)x_0^j \tag{2.34}$$

forces x_0^j to be a solution for (2.33), thus making the initial condition consistent. For an initial value $x_{j,0}$ at τ_j the problem under consideration therefore should be

$$E(t)\dot{\mathrm{x}} = A(t)\mathrm{x} + \mathrm{b}, \quad \hat{x}_{j-1} = x_0^j, \tag{2.35}$$

where $\mathrm{b} = \hat{\mathrm{b}} + \mathrm{b}_{imp}$, $\hat{\mathrm{b}} = \sum_{i\in\mathbb{Z}} \hat{b}_i$ and $\hat{b}_j$ satisfies (2.34). Then, according to Corollary 2.49, the problem (2.35) with initial condition $x_{j,0}$ at τ_j has a unique solution $\mathrm{x} \in \mathcal{C}^n_{imp}(\mathcal{T})$. Since

$$\dot{\mathrm{x}} = \dot{\hat{\mathrm{x}}} + \dot{\mathrm{x}}_{imp} + \sum_{i\in\mathbb{Z}} (\hat{x}_i(\tau_i) - \hat{x}_{i-1}(\tau_i))\delta_{\tau_i}$$

the system (2.35) can be written as

$$E(t)(\dot{\hat{\mathrm{x}}} + \dot{\mathrm{x}}_{imp} + \sum_{i\in\mathbb{Z}} (\hat{x}_i(\tau_i) - \hat{x}_{i-1}(\tau_i))\delta_{\tau_i}) = A(t)(\hat{\mathrm{x}} + \mathrm{x}_{imp}) + \hat{\mathrm{b}} + \mathrm{b}_{imp},$$

with $\hat{x}_{j-1} = x_0^j$ and due to (2.34) we get

$$E(t)(\dot{\hat{\mathrm{x}}} + \dot{\mathrm{x}}_{imp} + \sum_{i\in\mathbb{Z}}(\hat{x}_i(\tau_i) - \hat{x}_{i-1}(\tau_i))\delta_{\tau_i}) = A(t)(\hat{\mathrm{x}} + \mathrm{x}_{imp}) + E(t)\dot{x}_0^j \\ - A(t)x_0^j + \sum_{i\neq j} b_i + \mathrm{b}_{imp}.$$

Setting $\tilde{\mathrm{x}} = \mathrm{x} - \hat{x}_{j-1}$ and $\tilde{\mathrm{b}} = \mathrm{b} - \hat{b}_j$ this can be expressed in the form

$$E(t)\dot{\tilde{\mathrm{x}}} = A(t)\tilde{\mathrm{x}} + \tilde{\mathrm{b}} + E(t)x_{j,0}\delta_{\tau_j}, \quad \tilde{x}_{j-1} = 0, \tag{2.36}$$

where $x_{j,0} = \hat{x}_{j-1}(\tau_j)$. The initial condition does not occur as it is stated in the classical formulation, as we cannot prescribe values of distributions, but as part of the inhomogeneity. The general formulation of an initial value problem

$$E(t)\dot{\mathrm{x}} = A(t)\mathrm{x} + \mathrm{b} + E(t)x_{j,0}\delta_{\tau_j}, \quad x_{j-1} = 0, \tag{2.37}$$

suggests that for sufficiently smooth b the smoothness of x will depend on the initial condition. Thus, the impulsive behavior and the future smooth development of the system does not depend on the whole history but only on the initial condition.

2.2.4 Remarks

A drawback of the presented DAE theory is the restrictive constant rank assumption (2.12) that needs to be applied in each step of the index reduction procedure leading to the reduced system (2.14). Thus, the strangeness index (as in Definition 2.35) is only defined on a dense subset of the given closed interval $\mathbb{I}$ which can be shown using the following properties of the rank of continuous matrix-valued functions.

Theorem 2.50. *Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $M \in C(\mathbb{I}, \mathbb{R}^{m,n})$. Then there exist open intervals $\mathbb{I}_j \subseteq \mathbb{I}$, $j \in \mathbb{N}$, with*

$$\overline{\bigcup_{j\in\mathbb{N}} \mathbb{I}_j} = \mathbb{I}, \quad \mathbb{I}_i \cap \mathbb{I}_j = \emptyset \text{ for } i \neq j, \tag{2.38}$$

and integers $r_j \in \mathbb{N}_0$, $j \in \mathbb{N}$, such that

$$\operatorname{rank} M(t) = r_j \text{ for all } t \in \mathbb{I}_j.$$

Proof. See [82, Theorem 3.25] or [27, Ch. 10]. □

Applying this property to a continuous matrix-valued function we immediately obtain the following result.

Corollary 2.51. *Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $E, A \in C(\mathbb{I}, \mathbb{R}^{m,n})$ be sufficiently smooth. Then there exist open intervals $\mathbb{I}_j$, $j \in \mathbb{N}$, as in Theorem 2.50, such that the strangeness index of (E, A) restricted to $\mathbb{I}_j$ is well-defined for every $j \in \mathbb{N}$.*

Proof. See [82, Corollary 3.26]. □

Thus, the consequence of the constant rank assumption in Theorem 2.34 is that the strangeness index is defined on a dense subset of a given closed interval $\mathbb{I} \subseteq \mathbb{R}$. Note, that the proof of Theorem 2.50 and Corollary 2.51 is given in [82] only for the case of complex matrix-valued functions. Nevertheless, the proof can be given in the case of real-valued continuous matrix functions under the additional assumption that no accumulation of critical points, i.e., points where the constant rank assumption is not fulfilled, occur. In the following, we will exclude the case that accumulation of critical points occur from our examinations. As a consequence, we can transform to the global canonical form (2.13) on each component $\mathbb{I}_j$ separately, but the theory does not allow to treat jumps in the index and in the characteristic values between the intervals $\mathbb{I}_j$ of (2.38). Even within the framework of impulsive smooth distributions it is not straightforward to define impulsive smooth distributions with impulses allowed at every point of a set

$$\mathbb{T} = \mathbb{I} \setminus \bigcup_{j \in \mathbb{N}} \mathbb{I}_j$$

as the set $\mathbb{T}$ does not need to be countable. Further, jumps in the characteristic values may affect the solvability within the set of impulsive smooth distributions as can be seen in the following example.

Example 2.52. [82] Consider the initial value problem

$$tx = 0, \quad x(0) = 0.$$

For this DAE the strangeness index is not defined, but there exits a unique smooth solution of the initial value problem in $C^1(\mathbb{R}, \mathbb{R})$, namely $x = 0$. Within the solution space $\mathcal{C}_{imp}$ a possible decomposition according to Corollary 2.51 is given by

$$\mathbb{R} = \overline{(-\infty, 0) \cup (0, \infty)}.$$

Obviously, all distributions of the form $x = c\delta$ with $c \in \mathbb{R}$ solve the initial value problem. Hence, we may loose unique solvability when we turn to distributional solutions. Moreover, there is no initial condition of the form (2.31) that fixes a unique solution.

A further drawback of the distributional solution theory presented in Section 2.2.3 is that we have to require infinitely often differentiable matrix-valued functions $E(t)$ and $A(t)$. Another distributional solution theory for linear DAEs considering so-called *piecewise-smooth distributions* is presented in [144] that also allows discontinuities in the coefficient matrices $E(t)$ and $A(t)$. In this case, a suitable multiplication for distributions need to be defined. Note, that the space of impulsive smooth distributions is a subspace of piecewise smooth distributions where jumps and Dirac impulse can only occur at times τ_i.

Chapter 3

Higher Order Differential-Algebraic Systems

General nonlinear k-th order differential-algebraic systems of the form

$$F(t, x, \dot{x}, \ldots, x^{(k)}) = 0, \tag{3.1}$$

with $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \times \cdots \times \mathbb{D}_{x^{(k)}} \to \mathbb{R}^m$ sufficiently smooth on a compact interval $\mathbb{I} \subseteq \mathbb{R}$ and open sets $\mathbb{D}_x, \mathbb{D}_{\dot{x}}, \ldots, \mathbb{D}_{x^{(k)}} \subseteq \mathbb{R}^n$, as well as linear k-th order differential-algebraic equations of the form

$$A_k(t)x^{(k)} + A_{k-1}(t)x^{(k-1)} + \cdots + A_0(t)x = f(t), \tag{3.2}$$

where $A_i \in C(\mathbb{I}, \mathbb{R}^{m\times n})$ for $i = 0, 1, \ldots, k$, $k \in \mathbb{N}_0$ and $f \in C(\mathbb{I}, \mathbb{R}^m)$ naturally arise in many technical applications. In particular, second order differential-algebraic systems with $k = 2$ play a key role in the modeling and simulation of constrained dynamical systems, e.g., in the simulation of mechanical multibody systems or in electrical circuit simulation, as we have seen in Chapter 1.

In the classical theory of differential equations, higher order systems are turned into first order systems by introducing new variables for the derivatives. For DAEs this classical approach has to be performed with great care since it may lead to a number of mathematical difficulties as has been discussed in several publications, see [4, 32, 102, 129, 135]. In [102, 135] several examples are presented that show that the classical approach of introducing the derivatives of the unknown vector-valued function $x(t)$ as new variables may lead to higher smoothness requirements for the inhomogeneity $f(t)$ that are needed to ensure the existence of a solution, which corresponds to an increase in the index of the DAE. By introducing only some new variables, however, this difficulty can be circumvented.

Example 3.1. [102] Consider the linear second order constant coefficient DAE

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} x = f(t), \quad t \in \mathbb{I}, \tag{3.3}$$

where $x = [x_1, x_2]^T$, and $f = [f_1, f_2]^T$. System (3.3) has the unique solution

$$\begin{aligned} x_1 &= f_2, \\ x_2 &= f_1 - \dot{f}_2 - \ddot{f}_2, \end{aligned}$$

and hence the minimum requirement for the existence of a continuous solution is that f_1 is continuous and f_2 is twice continuously differentiable. Using the classical transformation to first order by introducing

$$v = [v_1, v_2]^T = [\dot{x}_1, \dot{x}_2]^T, \quad y = [v_1, v_2, x_1, x_2]^T,$$

we get the additional solution components

$$\begin{aligned} v_1 &= \dot{f}_2, \\ v_2 &= \dot{f}_1 - \ddot{f}_2 - f_2^{(3)}, \end{aligned}$$

and thus, f_2 has to be three times continuously differentiable to obtain a continuous solution. If, however, we only introduce $v_1 = \dot{x}_2$, then no extra smoothness requirements are needed.

Another difficulty that arises in practical numerical applications is that the system may be badly scaled and that there are disturbances and perturbations in the data, such that the transformation to first order may lead to very different solutions in the perturbed system.

Example 3.2. [102] Consider the second order system

$$\epsilon_1 \ddot{x} + \epsilon_2 \dot{x} + \epsilon_3 x = \epsilon_4 f(t), \quad t \in \mathbb{I}, \tag{3.4}$$

with coefficients ϵ_i, $i = 1, \ldots, 4$ of absolute value close to the machine precision and f of norm approximately 1. If we transform (3.4) to first order in the classical way by introducing

$$y = [y_1, y_2]^T := [\dot{x}, x]^T,$$

then we obtain the system

$$\begin{bmatrix} \epsilon_1 & 0 \\ 0 & 1 \end{bmatrix} \dot{y} + \begin{bmatrix} \epsilon_2 & \epsilon_3 \\ -1 & 0 \end{bmatrix} y = \begin{bmatrix} \epsilon_4 f(t) \\ 0 \end{bmatrix}. \tag{3.5}$$

For different values of the ϵ_i, in finite precision arithmetic, we may decide that the system (3.5) has a unique solution, no solutions at all, or is actually underdetermined.

Recently, it has been shown in [129, 152] that the direct discretization of the second order system may yield better numerical results and is able to prevent certain numerical difficulties as the failure of numerical methods, see also [4, 17, 151]. Therefore, a proper treatment of higher order differential-algebraic systems requires either the direct numerical solution of the high order system by appropriate numerical methods as proposed in [129, 152], or carefully chosen first order formulations.

Example 3.3. Consider the example of a multibody system

$$\begin{aligned} M(p,t)\ddot{p} &= f_a(p, \dot{p}, t) - G^T(p,t)\lambda, \\ 0 &= g(p,t). \end{aligned}$$

In such systems it is common practice to derive a first order formulation by just introducing new variables $v = \dot{p}$ but not the derivative of λ, i.e.,

$$\begin{aligned} \dot{p} &= v, \\ M(p,t)\dot{v} &= f_a(p, v, t) - G^T(p,t)\lambda, \\ 0 &= g(p,t). \end{aligned}$$

In this way an unnecessary derivative of the Lagrange multiplier λ is avoided.

The theoretical analysis of linear high order differential-algebraic equations of the form (3.2) regarding existence and uniqueness of solutions has been studied in [102, 135], where condensed forms and corresponding invariants under equivalence transformations are derived and a definition of the strangeness-index is given. Further, a stepwise index reduction procedure allows to transform the original system to a strangeness-free system that enables an identification of those higher order derivatives of variables that can be replaced to obtain a first order formulation without changing the smoothness requirements. However, the computation of these condensed forms is not numerical feasible as it involves the derivatives of computed transformation matrices.

In this chapter, we first give a brief survey of the relevant results for linear second order DAEs obtained in [102, 135] and introduce the characteristic invariants. Then, we present a numerically computable method to determine the strangeness index as well as the characteristic invariants using derivative arrays, following the ideas that were presented in Section 2.2.2. An equivalent strangeness-free differential-algebraic system can be obtained from the original system and its higher derivatives that has the same solution behavior as the original DAE. In Section 3.2 the ideas are extended to nonlinear DAEs. Further, in Section 3.3, we discuss first order formulations for linear second order DAEs and present a trimmed first order formulation, see also [21] for further trimmed first order formulations. The trimmed first order formulation also allows an explicit representation of solutions for linear second order DAEs with constant coefficient matrices. In the following, we restrict to second order systems for ease of representation and since they are most frequently used. In principle, all ideas can also be extended to arbitrary k-th order systems.

3.1 Linear Second Order Differential-Algebraic Systems

In this section we consider linear second order differential-algebraic equations with variable coefficients of the form

$$M(t)\ddot{x} + C(t)\dot{x} + K(t)x = f(t), \tag{3.6}$$

where $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m\times n})$ and $f \in C(\mathbb{I}, \mathbb{R}^m)$ are sufficiently smooth functions, together with the initial conditions

$$x(t_0) = x_0 \in \mathbb{R}^n, \quad \dot{x}(t_0) = \dot{x}_0 \in \mathbb{R}^n, \quad \text{for } t_0 \in \mathbb{I}. \tag{3.7}$$

At first, we introduce the condensed forms that have been derived in [102, 135] for linear second order systems (3.6) that allow to describe the characteristic quantities of the DAE. In Section 3.1.2 these condensed forms and the relationships between the characteristic quantities are used to extract a strangeness-free reduced system using derivative arrays.

3.1.1 Condensed Forms

To analyze linear second order DAEs of the form (3.6) we first derive condensed forms for the corresponding triple $(M(t), C(t), K(t))$ of matrix-valued functions under appropriate

equivalence transformations. An equivalence relation for triples of matrix-valued functions can be defined in the same way as for pairs of matrix-valued functions in Definition 2.33.

Definition 3.4 (Global equivalence of matrix triples). Two triples of matrix-valued functions (M_1, C_1, K_1) and (M_2, C_2, K_2), with $M_i, C_i, K_i \in C(\mathbb{I}, \mathbb{R}^{m\times n})$, $i = 1, 2$ are called *globally equivalent* if there exist pointwise nonsingular matrix functions $P \in C(\mathbb{I}, \mathbb{R}^{m\times m})$ and $Q \in C^2(\mathbb{I}, \mathbb{R}^{n\times n})$ such that

$$M_2 = PM_1Q, \quad C_2 = 2PM_1\dot{Q} + PC_1Q, \quad K_2 = PM_1\ddot{Q} + PC_1\dot{Q} + PK_1Q. \tag{3.8}$$

For equivalent matrix triples we write $(M_1, C_1, K_1) \sim (M_2, C_2, K_2)$.

At first, we consider the action of the equivalence relation (3.8) locally at a fixed point $\hat{t} \in \mathbb{I}$, taking into account that for given matrices $\hat{P}$, $\hat{Q}$, $\hat{R}_1$ and $\hat{R}_2$ of appropriate size, using Hermite interpolation, we can always find matrix functions P and Q, such that at a given value $t = \hat{t}$ we have $P(\hat{t}) = \hat{P}$, $Q(\hat{t}) = \hat{Q}$, $\dot{Q}(\hat{t}) = \hat{R}_1$ and $\ddot{Q}(\hat{t}) = \hat{R}_2$, i.e., we can choose $Q(\hat{t}), \dot{Q}(\hat{t}), \ddot{Q}(\hat{t})$ independently. Therefore, we can define local equivalence of matrix triples in the following way.

Definition 3.5 (Local equivalence of matrix triples). Two matrix triples (M_1, C_1, K_1) and (M_2, C_2, K_2) with $M_i, C_i, K_i \in \mathbb{R}^{m\times n}$, $i = 1, 2$, are called *locally equivalent* if there exist nonsingular matrices $P \in \mathbb{R}^{m\times m}$ and $Q \in \mathbb{R}^{n\times n}$ and matrices $R_1, R_2 \in \mathbb{R}^{n\times n}$ such that

$$M_2 = PM_1Q, \quad C_2 = 2PM_1R_1 + PC_1Q, \quad K_2 = PM_1R_2 + PC_1R_1 + PK_1Q. \tag{3.9}$$

Again, we write $(M_1, C_1, K_1) \sim (M_2, C_2, K_2)$ if the context is clear.

It has be shown in [135, Propositions 3.2 and 3.4] that the relations (3.8) and (3.9) are equivalence relations on the set of tuples of matrix-valued functions, and on the set of tuples of matrices, respectively. Further, condensed forms for matrix triples under strong equivalence, i.e., for $R_1 = 0$ and $R_2 = 0$ in (3.9), are considered in [102, 135], that correspond to the case of linear time-invariant second order differential-algebraic systems. For a linear second order differential-algebraic system of the form (3.6) a condensed form under local equivalence transformation (3.9) of the corresponding matrix triple $(M(\hat{t}), C(\hat{t}), K(\hat{t}))$ at a fixed point $\hat{t} \in \mathbb{I}$ has also been derived in [102, 135]. This local condensed form if given in the following Lemma.

Lemma 3.6. *Consider matrices $M, C, K \in \mathbb{R}^{m\times n}$. Then the matrix triple (M, C, K) is locally equivalent via equivalence transformation (3.9) to a matrix triple $(\hat{M}, \hat{C}, \hat{K})$ of the*

following local condensed form

$$
\left(\begin{bmatrix}
I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{s^{(MK)}} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{d^{(2)}} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix},
\begin{bmatrix}
0 & 0 & C & C & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & C & C \\
0 & 0 & 0 & C & I_{s^{(CK)}} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & I_{d^{(1)}} & 0 & 0 \\
I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix},
\begin{bmatrix}
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K \\
0 & 0 & 0 & 0 & 0 & 0 & I_a & 0 \\
0 & 0 & 0 & 0 & I_{s^{(CK)}} & 0 & 0 & 0 \\
0 & 0 & I_{s^{(MK)}} & 0 & 0 & 0 & 0 & 0 \\
I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}\right)
\begin{matrix}
s^{(MCK)} \\ s^{(MC)} \\ s^{(MK)} \\ d^{(2)} \\ s^{(CK)} \\ d^{(1)} \\ s^{(MCK)} \\ s^{(MC)} \\ a \\ s^{(CK)} \\ s^{(MK)} \\ s^{(MCK)} \\ v
\end{matrix}, \tag{3.10}
$$

where the quantities $s^{(MCK)}$, $s^{(MC)}$, $s^{(MK)}$, $s^{(CK)}$, $d^{(2)}$, $d^{(1)}$, a and v are nonnegative inte-

gers and the last column in each matrix is of width u.

Proof. See [102, 135]. □

Remark 3.7. *For convenience of expression, in Lemma 3.6 and in the following, we drop the subscripts of the elements of block matrices unless they are needed for clarification.*

The quantities $s^{(MCK)}$, $s^{(MC)}$, $s^{(MK)}$, $s^{(CK)}$, $d^{(2)}$, $d^{(1)}$, a, v and u in (3.10) are called the *(local) characteristic values of the linear second order DAE* (3.6). Each of these characteristic values appearing in Lemma 3.6 can be expressed in terms of ranks of matrices and dimensions of column spaces. In addition, it can be shown that the quantities are invariant under the local equivalence relation (3.9).

Lemma 3.8. *Let* $M, C, K \in \mathbb{R}^{m\times n}$ *and let*

$$\begin{aligned}
&V_1 \text{ be a basis of } \operatorname{kernel}(M^T),\\
&V_2 \text{ be a basis of } \operatorname{kernel}(M),\\
&V_3 \text{ be a basis of } \operatorname{kernel}(M^T) \cap \operatorname{kernel}(C^T),\\
&V_4 \text{ be a basis of } \operatorname{kernel}(M) \cap \operatorname{kernel}(V_1^T C).
\end{aligned}$$

Then, the quantities

$r = \operatorname{rank}(M)$	*(rank of M)*
$a = \operatorname{rank}(V_3^T K V_4)$	*(algebraic part)*
$s^{(MCK)} = \dim(\operatorname{range}(M^T) \cap \operatorname{range}(C^T V_1) \cap \operatorname{range}(K^T V_3))$	*(strangeness of M, C, K)*
$s^{(CK)} = \operatorname{rank}(V_3^T K V_2) - a$	*(strangeness of C, K)*
$d^{(1)} = \operatorname{rank}(V_1^T C V_2) - s^{(CK)}$	*(1st-order diff. part)*
$s^{(MC)} = \operatorname{rank}(V_1^T C) - s^{(MCK)} - s^{(CK)} - d^{(1)}$	*(strangeness of M, C)*
$s^{(MK)} = \operatorname{rank}(V_3^T K) - a - s^{(MCK)} - s^{(CK)}$	*(strangeness of M, K)*
$d^{(2)} = r - s^{(MCK)} - s^{(MC)} - s^{(MK)}$	*(2nd-order diff. part)*
$v = m - r - 2s^{(CK)} - d^{(1)} - 2s^{(MCK)} - s^{(MC)} - a - s^{(MK)}$	*(vanishing equations)*
$u = n - r - s^{(CK)} - d^{(1)} - a$	*(undetermined part)*

are invariant under the local equivalence relation (3.9).

Proof. See [102, 135]. □

Thus, in contrast to linear first order DAEs which are characterized by their differential, algebraic and strangeness parts, see Theorem 2.34, second order differential-algebraic systems require a distinction into first and second order differential parts, algebraic parts, and the strangeness $s^{(MCK)}$, $s^{(MC)}$, $s^{(MK)}$, $s^{(CK)}$ due to the different possible couplings between the matrices M, C and K. For triples $(M(t), C(t), K(t))$ of matrix-valued functions we can compute the local condensed form (3.10) at any fixed value $\hat{t} \in \mathbb{I}$ and determine the characteristic quantities given in Lemma 3.8 for the triple $(M(\hat{t}), C(\hat{t}), K(\hat{t}))$, so that we obtain the functions

$$r, a, d^{(2)}, d^{(1)}, s^{(MCK)}, s^{(CK)}, s^{(MC)}, s^{(MK)}, u, v : \mathbb{I} \to \mathbb{N}_0.$$

In the following, we assume that these functions are constant over the considered interval $\mathbb{I}$, i.e.,

$$\begin{aligned} &r(t) \equiv r, \quad a(t) \equiv a, \quad d^{(1)}(t) \equiv d^{(1)}, \quad s^{(MCK)}(t) \equiv s^{(MCK)}, \\ &s^{(CK)}(t) \equiv s^{(CK)}, \quad s^{(MC)}(t) \equiv s^{(MC)}, \quad s^{(MK)}(t) \equiv s^{(MK)}, \quad \text{for all } t \in \mathbb{I}. \end{aligned} \tag{3.11}$$

From Lemma 3.8 it then follows that $d^{(2)}(t)$, $u(t)$ and $v(t)$ are also constant in $\mathbb{I}$. We will call (3.11) the *regularity conditions* for the triple $(M(t), C(t), K(t))$. Under this assumption we can obtain a global condensed form for triples of matrix-valued functions via the global equivalence transformations (3.8).

Lemma 3.9. *Let the matrix-valued functions $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m, \times n})$ be sufficiently smooth, and suppose that the regularity conditions* (3.11) *hold for the local characteristic values of (M, C, K). Then, (M, C, K) is globally equivalent to a triple of matrix-valued functions $(\tilde{M}, \tilde{C}, \tilde{K})$ of the condensed form*

$$\left(\begin{bmatrix} I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{s^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{s^{(MK)}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{d^{(2)}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \right.$$

$$\begin{bmatrix} 0 & 0 & C & C & 0 & 0 & C & C \\ 0 & 0 & C & C & 0 & 0 & C & C \\ 0 & 0 & C & C & 0 & 0 & C & C \\ 0 & 0 & C & C & 0 & 0 & C & C \\ 0 & 0 & 0 & 0 & I_{s^{(CK)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{d^{(1)}} & 0 & 0 \\ I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{s^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{3.12}$$

$$\left.\begin{bmatrix} 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & 0 & 0 & 0 & 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & 0 & I_{s^{(CK)}} & 0 & 0 & 0 \\ 0 & 0 & I_{s^{(MK)}} & 0 & 0 & 0 & 0 & 0 \\ I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s^{(MCK)} \\ s^{(MC)} \\ s^{(MK)} \\ d^{(2)} \\ s^{(CK)} \\ d^{(1)} \\ s^{(MCK)} \\ s^{(MC)} \\ a \\ s^{(CK)} \\ s^{(MK)} \\ s^{(MCK)} \\ v \end{matrix} .$$

Here, all blocks are again functions on $\mathbb{I}$ and the last block columns have size u.

Proof. See [102, 135]. □

The global condensed form (3.12) can now be used to derive an equivalent strangeness-free second order DAE by a stepwise index reduction procedure. The associated differential-algebraic system with coefficient matrices $(\tilde{M}, \tilde{C}, \tilde{K})$ in global condensed form (3.12) is given by

$$\begin{aligned}
\ddot{x}_1 + C_{13}\dot{x}_3 + C_{14}\dot{x}_4 + C_{17}\dot{x}_7 + C_{18}\dot{x}_8 + K_{12}x_2 + K_{14}x_4 + K_{16}x_6 + K_{18}x_8 &= f_1, && (a)\\
\ddot{x}_2 + C_{23}\dot{x}_3 + C_{24}\dot{x}_4 + C_{27}\dot{x}_7 + C_{28}\dot{x}_8 + K_{22}x_2 + K_{24}x_4 + K_{26}x_6 + K_{28}x_8 &= f_2, && (b)\\
\ddot{x}_3 + C_{33}\dot{x}_3 + C_{34}\dot{x}_4 + C_{37}\dot{x}_7 + C_{38}\dot{x}_8 + K_{32}x_2 + K_{34}x_4 + K_{36}x_6 + K_{38}x_8 &= f_3, && (c)\\
\ddot{x}_4 + C_{43}\dot{x}_3 + C_{44}\dot{x}_4 + C_{47}\dot{x}_7 + C_{48}\dot{x}_8 + K_{42}x_2 + K_{44}x_4 + K_{46}x_6 + K_{48}x_8 &= f_4, && (d)\\
\dot{x}_5 + K_{52}x_2 + K_{54}x_4 + K_{56}x_6 + K_{58}x_8 &= f_5, && (e)\\
\dot{x}_6 + K_{62}x_2 + K_{64}x_4 + K_{66}x_6 + K_{68}x_8 &= f_6, && (f)\\
\dot{x}_1 + K_{72}x_2 + K_{74}x_4 + K_{76}x_6 + K_{78}x_8 &= f_7, && (g)\\
\dot{x}_2 + K_{82}x_2 + K_{84}x_4 + K_{86}x_6 + K_{88}x_8 &= f_8, && (h)\\
x_7 &= f_9, && (i)\\
x_5 &= f_{10}, && (j)\\
x_3 &= f_{11}, && (k)\\
x_1 &= f_{12}, && (l)\\
0 &= f_{13}. && (m)
\end{aligned}$$

If we differentiate the equations $(g) - (l)$ once, we can eliminate the corresponding derivatives $\dot{x}_1, \dot{x}_3, \dot{x}_5, \dot{x}_7, \ddot{x}_1, \ddot{x}_2$ in the equations $(a) - (e)$ and (g). This, yields the differential-algebraic system

$$-K_{72}\dot{x}_2 + (C_{14} - K_{74})\dot{x}_4 - K_{76}\dot{x}_6 + (C_{18} - K_{78})\dot{x}_8 + (K_{12} - \dot{K}_{72})x_2$$

$$
\begin{aligned}
+(K_{14}-\dot{K}_{74})x_4+(K_{16}-\dot{K}_{76})x_6+(K_{18}-\dot{K}_{78})x_8 &= \tilde{f}_1,\\
-K_{82}\dot{x}_2+(C_{24}-K_{84})\dot{x}_4-K_{86}\dot{x}_6+(C_{28}-K_{88})\dot{x}_8+(K_{22}-\dot{K}_{82})x_2 & \\
+(K_{24}-\dot{K}_{84})x_4+(K_{26}-\dot{K}_{86})x_6+(K_{28}-\dot{K}_{88})x_8 &= \tilde{f}_2,\\
\ddot{x}_3+C_{34}\dot{x}_4+C_{38}\dot{x}_8+K_{32}x_2+K_{34}x_4+K_{36}x_6+K_{38}x_8 &= \tilde{f}_3,\\
\ddot{x}_4+C_{44}\dot{x}_4+C_{48}\dot{x}_8+K_{42}x_2+K_{44}x_4+K_{46}x_6+K_{48}x_8 &= \tilde{f}_4,\\
K_{52}x_2+K_{54}x_4+K_{56}x_6+K_{58}x_8 &= \tilde{f}_5,\\
\dot{x}_6+K_{62}x_2+K_{64}x_4+K_{66}x_6+K_{68}x_8 &= f_6,\\
K_{72}x_2+K_{74}x_4+K_{76}x_6+K_{78}x_8 &= \tilde{f}_7,\\
\dot{x}_2+K_{82}x_2+K_{84}x_4+K_{86}x_6+K_{88}x_8 &= f_8,\\
x_7 &= f_9,\\
x_5 &= f_{10},\\
x_3 &= f_{11},\\
x_1 &= f_{12},\\
0 &= f_{13},
\end{aligned}
$$

where

$$
\begin{aligned}
\tilde{f}_1 &= f_1-\dot{f}_7-C_{17}\dot{f}_9-C_{13}\dot{f}_{11},\\
\tilde{f}_2 &= f_2-C_{27}\dot{f}_9-\dot{f}_8-C_{23}\dot{f}_{11},\\
\tilde{f}_3 &= f_3-C_{37}\dot{f}_9-C_{33}\dot{f}_{11},\\
\tilde{f}_4 &= f_4-C_{47}\dot{f}_9-C_{43}\dot{f}_{11},\\
\tilde{f}_5 &= f_5-\dot{f}_{10},\\
\tilde{f}_7 &= f_7-\dot{f}_{12}.
\end{aligned}
$$

Thus, after one differentiation and elimination step we get an equivalent second order system of DAEs

$$M^{\langle 1\rangle}(t)\ddot{\tilde{x}}+C^{\langle 1\rangle}(t)\dot{\tilde{x}}+K^{\langle 1\rangle}(t)\tilde{x}=f^{\langle 1\rangle}(t),$$

with $(M^{\langle 1\rangle},C^{\langle 1\rangle},K^{\langle 1\rangle};f^{\langle 1\rangle})$ being of the form

$$
\left(\begin{bmatrix}
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&I_{s(MK)}&0&0&0&0&0\\
0&0&0&I_{d^{(2)}}&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0
\end{bmatrix},\;
\begin{bmatrix}
0&C&0&C&0&C&0&C\\
0&C&0&C&0&C&0&C\\
0&0&0&C&0&0&0&C\\
0&0&0&C&0&0&0&C\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&I_{d^{(1)}}&0&0\\
0&0&0&0&0&0&0&0\\
0&I_{s(MC)}&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0
\end{bmatrix}\right),
$$

$$\left[\begin{array}{cccccccc} 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & K & 0 & K & 0 & K & 0 & K \\ 0 & 0 & 0 & 0 & 0 & 0 & I_a & 0 \\ 0 & 0 & 0 & 0 & I_{s^{(CK)}} & 0 & 0 & 0 \\ 0 & 0 & I_{s^{(MK)}} & 0 & 0 & 0 & 0 & 0 \\ I_{s^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] ; \left[\begin{array}{c} \tilde{f}_1 \\ \tilde{f}_2 \\ \tilde{f}_3 \\ \tilde{f}_4 \\ \tilde{f}_5 \\ f_6 \\ \tilde{f}_7 \\ f_8 \\ f_9 \\ f_{10} \\ f_{11} \\ f_{12} \\ f_{13} \end{array}\right) . \tag{3.13}$$

The triple $(M^{<1>}, C^{<1>}, K^{<1>})$ in (3.13) can then again be transformed to the global condensed form (3.12) and the reduction step is repeated, i.e., again, the derivatives of certain equations are added to other equations to eliminate the coupling between equations. Note, that the identity block of size $s^{(MK)}$ in the matrix $\tilde{M}$ is not eliminated in the first reduction step, as this would require the second derivative of equation (k). This block is eliminated in the second reduction step by differentiating equation (k) twice and eliminating $\ddot{x}_3$ in equation (c). Continuing this process, we obtain a sequence of triples of matrix-valued functions $(M^{<i>}(t), C^{<i>}(t), K^{<i>}(t))$, $i \in \mathbb{N}_0$, with corresponding characteristic values $(r_i, d_i^{(1)}, a_i, s_i^{(MC)}, s_i^{(CK)}, s_i^{(MK)}, s_i^{(MCK)}, u_i, v_i)$, where $(M^{<0>}(t), C^{<0>}(t), K^{<0>}(t)) = (M(t), C(t), K(t))$. During this reduction procedure the identity blocks of size $s_i^{(MK)}$ are decomposed into two blocks of size s_i and s_{i-1}, with $s_i^{(MK)} = s_i + s_{i-1}$, and $s_0 = s_0^{(MK)}$, in such a way that s_{i-1} denotes the part of $s_i^{(MK)}$ that can be eliminated in the $(i+1)$-th reduction step, while s_i denotes the part that cannot be eliminated until reduction step $i+2$. Here, we set $s_{-1} = 0$ and for the following we use the convention that characteristic quantities with negative subscript are zero. The relations

$$\begin{aligned} \operatorname{rank}(M^{<i+1>}) &= r_{i+1} = r_i - s_i^{(MCK)} - s_i^{(MC)} - s_{i-1}, \\ \operatorname{rank}(K^{<i+1>}) &\geq a_{i+1} \geq a_i + s_i^{(CK)} + s_i^{(MCK)} \end{aligned}$$

guarantee that after a finite number of steps μ, the strangeness $s_\mu^{(MCK)}$, $s_\mu^{(MK)}$, $s_\mu^{(CK)}$ and $s_\mu^{(MC)}$ corresponding to $(M^{<\mu>}(t), C^{<\mu>}(t), K^{<\mu>}(t))$ vanish and the process becomes stationary. We call μ the *strangeness index* or *s-index* of the second order system of DAEs (3.6) and we call the final equivalent second order system of DAEs *strangeness-free*.

Remark 3.10. *Here, we differ slightly from the index reduction procedure described in [102, 135], where the identity block of size $s^{(MK)}$ is also completely eliminated in every reduction step. Thus, for one reduction step one or two differentiations of equations are required, depending on the occurrence of strangeness blocks. In this way, the index definition does not correspond to the differentiability requirements for the right hand side. In our*

approach the right-hand side is only differentiated once in each elimination step before the system is again transformed to global condensed form such that the strangeness index corresponds to the differentiability requirements for the right hand side, which is the case for all general index concepts.

Theorem 3.11. *Consider the linear second order system* (3.6)*, suppose that the regularity conditions* (3.11) *hold, and let* μ *be the strangeness index of* (3.6)*. If* $f \in C^\mu(\mathbb{I}, \mathbb{R}^m)$*, then system* (3.6) *is equivalent (in the sense that there is a one-to-one correspondence between the solution sets) to a strangeness-free system of second order differential-algebraic equations of the form*

$$\begin{aligned} \ddot{\tilde{x}}_1 + \tilde{C}_{11}(t)\dot{\tilde{x}}_1 + \tilde{C}_{14}(t)\dot{\tilde{x}}_4 + \tilde{K}_{11}(t)\tilde{x}_1 + \tilde{K}_{12}(t)\tilde{x}_2 + \tilde{K}_{14}(t)\tilde{x}_4 &= \tilde{f}_1(t), \quad (\ d_\mu^{(2)}\ ) \\ \dot{\tilde{x}}_2 + \tilde{K}_{21}(t)\tilde{x}_1 + \tilde{K}_{22}(t)\tilde{x}_2 + \tilde{K}_{24}(t)\tilde{x}_4 &= \tilde{f}_2(t), \quad (\ d_\mu^{(1)}\ ) \\ \tilde{x}_3 &= \tilde{f}_3(t), \quad (\ a_\mu\ ) \\ 0 &= \tilde{f}_4(t), \quad (\ v_\mu\ ) \end{aligned} \tag{3.14}$$

where the inhomogeneity $\tilde{f} := [\tilde{f}_1^T, \ldots, \tilde{f}_4^T]^T$ *is determined by* $f^{(0)}, \ldots, f^{(\mu)}$*. In particular,* $d_\mu^{(2)}$*,* $d_\mu^{(1)}$ *and* a_μ *are the number of second order differential, first order differential, and algebraic components of the unknown* $\tilde{x} := [\tilde{x}_1^T, \ldots, \tilde{x}_4^T]^T$*, while* u_μ *is the dimension of the undetermined vector* $\tilde{x}_4$*, and* v_μ *is the number of conditions in the last equation.*

Proof. The proof is similar to the proof of [135, Theorem 2.12] with slight modifications concerning the definition of the strangeness-index and the counting of the differentiations (see Remark 3.10). □

Using the strangeness-free form (3.14) we can analyze existence and uniqueness of solutions and consistency of initial conditions for linear second order differential-algebraic systems (3.6), see [102, 135]. Further, Theorem 3.11 allows an identification of those second order derivatives of variables that can be replaced to obtain a first order system without changing the smoothness requirements or increasing the index.

Corollary 3.12. *Under the assumptions of Theorem 3.11, let* μ *be the strangeness index of the matrix triple associated with the system (3.6) and let* $f \in C^\mu(\mathbb{I}, \mathbb{R}^m)$*. Then, the solution set of system (3.6) is in one-to-one correspondence (without further smoothness requirements) to the partial solution set given by the components* $\tilde{x}_1, \ldots, \tilde{x}_4$ *of the system of first order differential-algebraic equations*

$$\begin{aligned} \dot{\tilde{x}}_5 + \tilde{C}_{11}(t)\dot{\tilde{x}}_1 + \tilde{C}_{14}(t)\dot{\tilde{x}}_4 + \tilde{K}_{11}(t)\tilde{x}_1 + \tilde{K}_{12}(t)\tilde{x}_2 + \tilde{K}_{14}(t)\tilde{x}_4 &= \tilde{f}_1(t), \\ \dot{\tilde{x}}_2 + \tilde{K}_{21}(t)\tilde{x}_1 + \tilde{K}_{22}(t)\tilde{x}_2 + \tilde{K}_{24}(t)\tilde{x}_4 &= \tilde{f}_2(t), \\ \tilde{x}_3 &= \tilde{f}_3(t), \\ 0 &= \tilde{f}_4(t), \\ \dot{\tilde{x}}_1 - \tilde{x}_5 &= 0. \end{aligned} \tag{3.15}$$

Proof. See [102]. □

Remark 3.13. *The linear first order DAE* (3.15) *is strangeness-free and has the characteristic values* $d_\mu = 2d_\mu^{(2)} + d_\mu^{(1)}, a_\mu, v_\mu$ *and* u_μ.

We give an example to illustrate how we can derive an equivalent strangeness-free system using the described index reduction procedure.

Example 3.14. Consider the linear second order DAE

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ \ddot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & t \\ t & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} f_1(t) \\ f_2(t) \\ f_3(t) \\ f_4(t) \end{bmatrix}, \tag{3.16}$$

for $t \in [t_0, t_f]$ with $t_0 > 0$, $x = [x_1, x_2, x_3, x_4]^T$ and $f = [f_1, f_2, f_3, f_4]^T$. System (3.16) is already given in global condensed form (3.12) with characteristic values

$$d_0^{(2)} = 2, \quad s_0^{(CK)} = 1, \quad s_0^{(MCK)} = s_0^{(MC)} = s_0^{(MK)} = a_0 = d_0^{(1)} = v_0 = u_0 = 0.$$

One index redcution step consisting of differentiation of the last equation and elimination of $\dot{x}_3$ in the third equation yields the system

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ \ddot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & t \\ t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ f_3 - \dot{f}_4 \\ f_4 \end{bmatrix},$$

with characteristic values

$$d_1^{(2)} = 1, \quad s_1^{(MK)} = 1, \quad a_1 = 1, \quad s_1^{(MCK)} = s_1^{(MC)} = s_1^{(CK)} = d_1^{(1)} = v_1 = u_1 = 0.$$

In the second reduction step nothing is changed, i.e.,

$$d_2^{(2)} = 1, \quad s_2^{(MK)} = 1, \quad a_2 = 1, \quad s_2^{(MCK)} = s_2^{(MC)} = s_2^{(CK)} = d_2^{(1)} = v_2 = u_2 = 0,$$

and in the last reduction step differentiating the third equation twice and eliminating $\ddot{x}_1$ in the first equation yields the system

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \\ \ddot{x}_3 \\ \ddot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & t \\ t & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} f_1 - \ddot{f}_3 - \dddot{f}_4 \\ f_2 \\ f_3 - \dot{f}_4 \\ f_4 \end{bmatrix},$$

which is strangeness-free with characteristic values

$$d_3^{(2)} = 1, \quad d_2^{(1)} = 1, \quad a_3 = 2, \quad s_3^{(MCK)} = s_3^{(MC)} = s_3^{(CK)} = s_3^{(MK)} = v_2 = u_2 = 0,$$

and thus $\mu = 3$.

The sequence of characteristic values $(r_i,\, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(MC)}, s_i^{(MK)}, s_i^{(CK)}, u_i, v_i)$ that is obtained during the stepwise index reduction procedure can be further characterized in terms of ranks of block matrices of the matrix triple in order to get some insight into the index reduction procedure described above.

Lemma 3.15. *Let the functions $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m,\times n})$ be sufficiently smooth and let the strangeness index μ be well-defined. Further, let the process leading to Theorem 3.11 yield a sequence $(M^{<i>}, C^{<i>}, K^{<i>})$, $i \in \mathbb{N}_0$, with $(M^{<0>}, C^{<0>}, K^{<0>}) = (M, C, K)$ and characteristic values $(r_i, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(MC)}, s_i^{(MK)}, s_i^{(CK)}, u_i, v_i)$ according to Lemma 3.8. The triple $(M^{<i>}, C^{<i>}, K^{<i>})$ of matrix-valued functions is globally equivalent to the triple*

$$
\left(\begin{bmatrix}
I_{s_i^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s_i^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{s_{i-1}} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{s_i} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_{d_i^{(2)}} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}\right.,
$$

$$
\begin{bmatrix}
0 & 0 & C_{13}^{<i>} & C_{14}^{<i>} & C_{15}^{<i>} & 0 & 0 & C_{18}^{<i>} & C_{19}^{<i>} \\
0 & 0 & C_{23}^{<i>} & C_{24}^{<i>} & C_{25}^{<i>} & 0 & 0 & C_{28}^{<i>} & C_{29}^{<i>} \\
0 & 0 & C_{33}^{<i>} & C_{34}^{<i>} & C_{35}^{<i>} & 0 & 0 & C_{38}^{<i>} & C_{39}^{<i>} \\
0 & 0 & C_{43}^{<i>} & C_{44}^{<i>} & C_{45}^{<i>} & 0 & 0 & C_{48}^{<i>} & C_{49}^{<i>} \\
0 & 0 & C_{53}^{<i>} & C_{54}^{<i>} & C_{55}^{<i>} & 0 & 0 & C_{58}^{<i>} & C_{59}^{<i>} \\
0 & 0 & 0 & 0 & 0 & I_{s_i^{(CK)}} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{d_i^{(1)}} & 0 & 0 \\
I_{s_i^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s_i^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}, \tag{3.17}
$$

$$\left[\begin{array}{ccccccccc}
0 & K_{12}^{<i>} & 0 & 0 & K_{15}^{<i>} & 0 & K_{17}^{<i>} & 0 & K_{19}^{<i>} \\
0 & K_{22}^{<i>} & 0 & 0 & K_{25}^{<i>} & 0 & K_{27}^{<i>} & 0 & K_{29}^{<i>} \\
0 & K_{32}^{<i>} & 0 & 0 & K_{35}^{<i>} & 0 & K_{37}^{<i>} & 0 & K_{39}^{<i>} \\
0 & K_{42}^{<i>} & 0 & 0 & K_{45}^{<i>} & 0 & K_{47}^{<i>} & 0 & K_{49}^{<i>} \\
0 & K_{52}^{<i>} & 0 & 0 & K_{55}^{<i>} & 0 & K_{57}^{<i>} & 0 & K_{59}^{<i>} \\
0 & K_{62}^{<i>} & 0 & 0 & K_{65}^{<i>} & 0 & K_{67}^{<i>} & 0 & K_{69}^{<i>} \\
0 & K_{72}^{<i>} & 0 & 0 & K_{75}^{<i>} & 0 & K_{77}^{<i>} & 0 & K_{79}^{<i>} \\
0 & K_{82}^{<i>} & 0 & 0 & K_{85}^{<i>} & 0 & K_{87}^{<i>} & 0 & K_{89}^{<i>} \\
0 & K_{92}^{<i>} & 0 & 0 & K_{95}^{<i>} & 0 & K_{97}^{<i>} & 0 & K_{99}^{<i>} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{a_i} & 0 \\
0 & 0 & 0 & 0 & 0 & I_{s_i^{(CK)}} & 0 & 0 & 0 \\
0 & 0 & I_{s_{i-1}} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{s_i} & 0 & 0 & 0 & 0 & 0 \\
I_{s_i^{(MCK)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right)
\begin{array}{l}
s_i^{(MCK)} \\
s_i^{(MC)} \\
s_{i-1} \\
s_i \\
d_i^{(2)} \\
s_i^{(CK)} \\
d_i^{(1)} \\
s_i^{(MCK)} \\
s_i^{(MC)} \\
a_i \\
s_i^{(CK)} \\
s_{i-1} \\
s_i \\
s_i^{(MCK)} \\
v_i
\end{array},$$

where $s_i^{(MK)}$ *is separated into* $s_i^{(MK)} = s_i + s_{i-1}$ *and the last block columns have size* u_i. *We define*

$$\begin{aligned}
\tilde{C}_{1j}^{<i>} &:= C_{1j}^{<i>} - K_{8j}^{<i>}, \quad j = 5, 9, \\
\tilde{C}_{2j}^{<i>} &:= C_{2j}^{<i>} - K_{9j}^{<i>}, \quad j = 5, 9, \\
\tilde{K}_{1j}^{<i>} &:= K_{1j}^{<i>} - \dot{K}_{8j}^{<i>} + K_{82}^{<i>} K_{9j}^{<i>} + K_{87}^{<i>} K_{7j}^{<i>}, \quad j = 2, 5, 7, 9, \\
\tilde{K}_{2j}^{<i>} &:= K_{2j}^{<i>} - \dot{K}_{9j}^{<i>} + K_{92}^{<i>} K_{9j}^{<i>} + K_{97}^{<i>} K_{7j}^{<i>}, \quad j = 2, 5, 7, 9, \\
\tilde{C}_1 &:= \begin{bmatrix} \tilde{C}_{15}^{<i>T} & \tilde{C}_{25}^{<i>T} & C_{35}^{<i>T} \end{bmatrix}^T, \\
\tilde{C}_2 &:= \begin{bmatrix} \tilde{C}_{19}^{<i>T} & \tilde{C}_{29}^{<i>T} & C_{39}^{<i>T} \end{bmatrix}^T,
\end{aligned}$$

as well as

$$\begin{aligned}
k_0 &= d_0^{(1)} + s_0^{(CK)}, & k_{i+1} &= \operatorname{rank} \tilde{C}_2, \\
e_0 &= d_0^{(1)} + s_0^{(MC)} + s_0^{(CK)} + s_0^{(MCK)}, & e_{i+1} &= \operatorname{rank} \begin{pmatrix} \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \end{pmatrix}.
\end{aligned}$$

Then, let U *and* V *be nonsingular matrix-valued functions of size* $(s_i^{(MCK)} + s_i^{(MC)} + s_{i-1}, s_i^{(MCK)} + s_i^{(MC)} + s_{i-1})$ *and* $(d_i^{(2)} + u_i, d_i^{(2)} + u_i)$, *respectively, such that*

$$U^T \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} V = \begin{bmatrix} I_{e_{i+1}} & 0 \\ 0 & 0 \end{bmatrix}.$$

Further, let U *and* V *be partitioned into* $U = \begin{bmatrix} U_1 & U_2 & U_3 \end{bmatrix}$ *and* $V = \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix}$ *such that*

$$\begin{bmatrix} U_1^T \\ U_2^T \\ U_3^T \end{bmatrix} \begin{bmatrix} \tilde{C}_1 & \tilde{C}_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 & V_3 \end{bmatrix} = \begin{bmatrix} I_{e_{i+1}-k_{i+1}} & 0 & 0 \\ 0 & I_{k_{i+1}} & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and with a splitting of V_3 *into* $V_3 = \begin{bmatrix} V_{31} & V_{32} \end{bmatrix}$ *with* V_{31} *of size* $(d_i^{(2)} + u_i, d_i^{(2)} - e_{i+1} + k_{i+1})$ *and* V_{32} *of size* $(d_i^{(2)} + u_i, u_i - k_{i+1})$ *we can define*

$$\begin{bmatrix} K_1 & K_2 & K_3 & K_4 & K_5 & K_6 \end{bmatrix} := \begin{bmatrix} U_3^T & 0 \\ 0 & I \end{bmatrix} \left[\begin{array}{c|cc|c} \tilde{K}_{15}^{<i>} & \tilde{K}_{17}^{<i>} & \tilde{K}_{12}^{<i>} & \tilde{K}_{19}^{<i>} \\ \tilde{K}_{25}^{<i>} & \tilde{K}_{27}^{<i>} & \tilde{K}_{22}^{<i>} & \tilde{K}_{29}^{<i>} \\ K_{35}^{<i>} & K_{37}^{<i>} & K_{32}^{<i>} & K_{39}^{<i>} \\ \hline K_{65}^{<i>} & K_{67}^{<i>} & K_{62}^{<i>} & K_{69}^{<i>} \\ K_{85}^{<i>} & K_{87}^{<i>} & K_{82}^{<i>} & K_{89}^{<i>} \end{array}\right] \begin{bmatrix} [V_1\ V_{31}] & 0 & 0 & 0 \\ 0 & I_{d_i^{(1)}} & 0 & 0 \\ 0 & 0 & I_{s_i^{(MC)}} & 0 \\ 0 & 0 & 0 & [V_2\ V_{32}] \end{bmatrix}, \tag{3.18}$$

where the identity matrix on the left-hand side is of size $s_i^{(CK)} + s_i^{(MCK)}$*. Further, we define*

$$\begin{aligned}
b_0 &= a_0, & b_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_6 \end{bmatrix}\right), \\
p_0 &= a_0 + s_0^{(CK)}, & p_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_5 & K_6 \end{bmatrix}\right), \\
t_0 &= a_0 + s_0^{(CK)} - s_0^{(MK)}, & t_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_4 & K_5 & K_6 \end{bmatrix}\right), \\
d_0 &= a_0 + s_0^{(CK)}, & d_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_3 & K_4 & K_5 & K_6 \end{bmatrix}\right), \\
h_0 &= a_0 + s_0^{(CK)} + s_0^{(MK)}, & h_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_2 & K_3 & K_4 & K_5 & K_6 \end{bmatrix}\right), \\
c_0 &= a_0 + s_0^{(MCK)} + s_0^{(CK)} + s_0^{(MK)}, & c_{i+1} &= \operatorname{rank}\left(\begin{bmatrix} K_1 & K_2 & K_3 & K_4 & K_5 & K_6 \end{bmatrix}\right), \\
w_0 &= v_0, & w_{i+1} &= v_{i+1} - v_i, \\
q_0 &= e_0, & q_{i+1} &= e_{i+1} + c_i - s_i^{(CK)} - s_i^{(MCK)}.
\end{aligned}$$

Then, we have

$$\begin{aligned}
r_{i+1} &= r_i - s_i^{(MCK)} - s_i^{(MC)} - s_{i-1}, \\
c_{i+1} &= b_{i+1} + s_{i+1}^{(MCK)} + s_{i+1}^{(CK)} + s_{i+1}^{(MK)} - s_i, \\
e_{i+1} &= k_{i+1} + s_{i+1}^{(MC)} + s_{i+1}^{(MCK)}, \\
a_{i+1} &= a_i + s_i^{(CK)} + s_i^{(MCK)} + s_{i-1} + b_{i+1} = c_0 + \cdots + c_{i+1} - s_{i+1}^{(CK)} - s_{i+1}^{(MK)} - s_{i+1}^{(MCK)}, \\
s_{i+1}^{(MCK)} &= c_{i+1} - h_{i+1}, \\
s_{i+1}^{(MC)} &= e_{i+1} - k_{i+1} - c_{i+1} + h_{i+1}, \\
s_{i+1} &= h_{i+1} - d_{i+1}, \\
s_{i+1}^{(MK)} &= s_{i+1} + s_i, \\
s_{i+1}^{(CK)} &= d_{i+1} - b_{i+1}, \\
d_{i+1}^{(2)} &= r_{i+1} - s_{i+1}^{(MCK)} - s_{i+1}^{(MC)} - s_{i+1}^{(MK)} = d_i^{(2)} - e_{i+1} + k_{i+1} - s_{i+1}, \\
d_{i+1}^{(1)} &= d_i^{(1)} + s_i^{(MC)} + k_{i+1} - s_{i+1}^{(CK)} \\
&\quad = q_0 + \cdots + q_{i+1} - c_0 - \cdots - c_i - s_{i+1}^{(MCK)} - s_{i+1}^{(MC)} - s_{i+1}^{(CK)}, \\
w_{i+1} &= 2s_i^{(MCK)} + s_i^{(CK)} + s_i^{(MC)} + s_{i-1} - e_{i+1} - c_{i+1}, \\
u_{i+1} &= u_0 - b_1 - \cdots - b_{i+1}, \\
v_{i+1} &= v_0 + w_1 + \cdots + w_{i+1} = 2s_i^{(MCK)} + s_i^{(CK)} + s_i^{(MC)} + s_{i-1} - e_{i+1} - c_{i+1} + v_i.
\end{aligned}$$

Proof. Form Lemma 3.9 it directly follows that each triple $(M^{<i>}, C^{<i>}, K^{<i>})$, $i \in \mathbb{N}_0$ is globally equivalent to the form (3.17). The identity blocks of size $s_i^{(MK)}$ are decomposed into two identity blocks of size s_i and s_{i-1}, such that in one differentiation and elimination step the block of size s_{i-1} can be eliminated. To prove the relations for the characteristic quantities we first perform one differentiation and elimination step for system (3.17) and thus get the matrix triple

$$
\left(
\begin{bmatrix}
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&I_{s_i}&0&0&0&0&0\\
0&0&0&0&I_{d_i^{(2)}}&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0
\end{bmatrix},
\begin{bmatrix}
0&0&0&0&\tilde{C}_{15}^{<i>}&0&0&0&\tilde{C}_{19}^{<i>}\\
0&0&0&0&\tilde{C}_{25}^{<i>}&0&0&0&\tilde{C}_{29}^{<i>}\\
0&0&0&0&C_{35}^{<i>}&0&0&0&C_{39}^{<i>}\\
0&0&0&0&C_{45}^{<i>}&0&0&0&C_{49}^{<i>}\\
0&0&0&0&C_{55}^{<i>}&0&0&0&C_{59}^{<i>}\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&I_{d_i^{(1)}}&0&0\\
0&0&0&0&0&0&0&0&0\\
0&I_{s_i^{(MC)}}&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0
\end{bmatrix},
\right.
$$

$$
\left.
\begin{bmatrix}
0&\tilde{K}_{12}^{<i>}&0&0&\tilde{K}_{15}^{<i>}&0&\tilde{K}_{17}^{<i>}&0&\tilde{K}_{19}^{<i>}\\
0&\tilde{K}_{22}^{<i>}&0&0&\tilde{K}_{25}^{<i>}&0&\tilde{K}_{27}^{<i>}&0&\tilde{K}_{29}^{<i>}\\
0&K_{32}^{<i>}&0&0&K_{35}^{<i>}&0&K_{37}^{<i>}&0&K_{39}^{<i>}\\
0&K_{42}^{<i>}&0&0&K_{45}^{<i>}&0&K_{47}^{<i>}&0&K_{49}^{<i>}\\
0&K_{52}^{<i>}&0&0&K_{55}^{<i>}&0&K_{57}^{<i>}&0&K_{59}^{<i>}\\
0&K_{62}^{<i>}&0&0&K_{65}^{<i>}&0&K_{67}^{<i>}&0&K_{69}^{<i>}\\
0&K_{72}^{<i>}&0&0&K_{75}^{<i>}&0&K_{77}^{<i>}&0&K_{79}^{<i>}\\
0&K_{82}^{<i>}&0&0&K_{85}^{<i>}&0&K_{87}^{<i>}&0&K_{89}^{<i>}\\
0&K_{92}^{<i>}&0&0&K_{95}^{<i>}&0&K_{97}^{<i>}&0&K_{99}^{<i>}\\
0&0&0&0&0&0&0&I_{a_i}&0\\
0&0&0&0&0&I_{s_i^{(CK)}}&0&0&0\\
0&0&I_{s_{i-1}}&0&0&0&0&0&0\\
0&0&0&I_{s_i}&0&0&0&0&0\\
I_{s_i^{(MCK)}}&0&0&0&0&0&0&0&0\\
0&0&0&0&0&0&0&0&0
\end{bmatrix}
\right).
$$

In the following, we omit the subscript i. Permutation of the resulting triple of matrix-valued functions yields

$$\left(\left[\begin{array}{cccc|ccccc} I_{d_i^{(2)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{s_i} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cccc|ccccc} C_{55} & 0 & 0 & C_{59} & 0 & 0 & 0 & 0 & 0 \\ C_{45} & 0 & 0 & C_{49} & 0 & 0 & 0 & 0 & 0 \\ \tilde{C}_{15} & 0 & 0 & \tilde{C}_{19} & 0 & 0 & 0 & 0 & 0 \\ \tilde{C}_{25} & 0 & 0 & \tilde{C}_{29} & 0 & 0 & 0 & 0 & 0 \\ C_{35} & 0 & 0 & C_{39} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{d_i^{(1)}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_{s_i^{(MC)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cccc|ccccc} K_{55} & K_{57} & K_{52} & K_{59} & 0 & 0 & 0 & 0 & 0 \\ K_{45} & K_{47} & K_{42} & K_{49} & 0 & 0 & 0 & 0 & 0 \\ \tilde{K}_{15} & \tilde{K}_{17} & \tilde{K}_{12} & \tilde{K}_{19} & 0 & 0 & 0 & 0 & 0 \\ \tilde{K}_{25} & \tilde{K}_{27} & \tilde{K}_{22} & \tilde{K}_{29} & 0 & 0 & 0 & 0 & 0 \\ K_{35} & K_{37} & K_{32} & K_{39} & 0 & 0 & 0 & 0 & 0 \\ K_{75} & K_{77} & K_{72} & K_{79} & 0 & 0 & 0 & 0 & 0 \\ K_{95} & K_{97} & K_{92} & K_{99} & 0 & 0 & 0 & 0 & 0 \\ K_{65} & K_{67} & K_{62} & K_{69} & 0 & 0 & 0 & 0 & 0 \\ K_{85} & K_{87} & K_{82} & K_{89} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{s_i} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{s_{i-1}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I_{a_i} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_i^{(CK)}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_i^{(MCK)}} \end{array}\right]\right).$$

The so obtained triple has to be transformed to global condensed form (3.12). We can restrict ourselves to the upper left blocks and use global equivalence transformations to separate the corresponding nullspaces. In the following, we only specify the blocks we are using in the transformations for convenience. Thus, we consider the triple of matrix-valued functions

$$\left(\left[\begin{array}{cccc} I_{d^{(2)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cccc} C_{55} & 0 & 0 & C_{59} \\ C_{45} & 0 & 0 & C_{49} \\ \tilde{C}_{15} & 0 & 0 & \tilde{C}_{19} \\ \tilde{C}_{25} & 0 & 0 & \tilde{C}_{29} \\ C_{35} & 0 & 0 & C_{39} \\ 0 & I_{d^{(1)}} & 0 & 0 \\ 0 & 0 & I_{s^{(MC)}} & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cccc} K_{55} & K_{57} & K_{52} & K_{59} \\ K_{45} & K_{47} & K_{42} & K_{49} \\ \tilde{K}_{15} & \tilde{K}_{17} & \tilde{K}_{12} & \tilde{K}_{19} \\ \tilde{K}_{25} & \tilde{K}_{27} & \tilde{K}_{22} & \tilde{K}_{29} \\ K_{35} & K_{37} & K_{32} & K_{39} \\ K_{75} & K_{77} & K_{72} & K_{79} \\ K_{95} & K_{97} & K_{92} & K_{99} \\ K_{65} & K_{67} & K_{62} & K_{69} \\ K_{85} & K_{87} & K_{82} & K_{89} \end{array}\right]\right)$$

$$= \left(\begin{bmatrix} I_{d^{(2)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} C & 0 & 0 & C \\ C & 0 & 0 & C \\ \tilde{C}_1 & 0 & 0 & \tilde{C}_2 \\ 0 & I_{d^{(1)}} & 0 & 0 \\ 0 & 0 & I_{s^{(MC)}} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} K & K & K & K \\ K & K & K & K \\ K & K & K & K \\ K & K & K & K \\ K & K & K & K \\ K & K & K & K \end{bmatrix} \right),$$

with $\tilde{C}_1 = \begin{bmatrix} \tilde{C}_{15}^T & \tilde{C}_{25}^T & C_{35}^T \end{bmatrix}^T$ and $\tilde{C}_2 = \begin{bmatrix} \tilde{C}_{19}^T & \tilde{C}_{29}^T & C_{39}^T \end{bmatrix}^T$. Using equivalence transformations this triple is equivalent to the following triple

$$\left(\left[\begin{array}{cc|cccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ \hline 0 & 0 & 0 & 0 & I_{e_1} & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{s^{(MC)}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & K & K \end{array}\right] \right),$$

with $\operatorname{rank}[\tilde{C}_1\ \tilde{C}_2] = e_{i+1} = e_1 + e_2$, $e_1 = k_{i+1}$, and $r_1 = d_i^{(2)} - e_2$. Rearanging the block rows and summarizing the last two block rows using the definition in (3.18) yields

$$\sim \left(\left[\begin{array}{cc|cccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ \hline 0 & 0 & 0 & 0 & I_{e_1} & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{s^{(MC)}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ K_1 & K_2 & K_3 & K_4 & K_5 & K_6 \end{array}\right] \right),$$

$$\sim \left(\left[\begin{array}{cc|cccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ 0 & C & 0 & 0 & 0 & C \\ \hline 0 & 0 & 0 & 0 & I_{e_1} & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{s^{(MC)}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccc} K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ \hline K & K & K & K & K & K \\ K & K & K & K & K & K \\ K & K & K & K & \tilde{K}_5 & 0 \end{array}\right] \right), \quad \text{with } \operatorname{rank} K_6 = r_2,$$

$$\sim \left(\left[\begin{array}{cc|cccccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccccc} 0 & C & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & C & C \\ \hline 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{s_2} & 0 & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{s^{(MC)}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \right.$$

$$\left. \left[\begin{array}{cc|cccccc} K & K & K & K & 0 & K & 0 & K \\ K & K & K & K & 0 & K & 0 & K \\ K & K & K & K & 0 & K & 0 & K \\ \hline K & K & K & K & 0 & K & 0 & K \\ K & K & K & K & 0 & K & 0 & K \\ K & K & K & K & 0 & K & 0 & K \\ \hline K & K & K & K & 0 & K & 0 & K \\ K & K & K & K & 0 & K & 0 & K \\ 0 & 0 & 0 & 0 & 0 & 0 & I_{r_2} & 0 \\ 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\ K & K & K & \tilde{K}_4 & 0 & 0 & 0 & 0 \end{array}\right] \right), \quad \text{with } \operatorname{rank} \tilde{K}_5 = r_3, \text{ and } s_2 = e_1 - r_3,$$

$$\sim \left(\left[\begin{array}{cc|ccccccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|ccccccc} 0 & C & 0 & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & 0 & C & C \\ \hline 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I_{s_2} & 0 & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{d^{(1)}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{s_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \right.$$

$$\left. \left[\begin{array}{cc|ccccccc} K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ \hline K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ \hline K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ K & K & K & 0 & K & 0 & K & 0 & K \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_2} & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\ K & K & \tilde{K}_3 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \right), \quad \text{with } \operatorname{rank} \tilde{K}_4 = r_4, \text{ and } s_3 = s_i^{(MC)} - r_4,$$

$$\sim \left(\left[\begin{array}{cc|cccccccc} I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & I_{r_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cccccccc} 0 & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\ 0 & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_2} & 0 & 0 \\ I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{s_4} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_{s_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \right.$$

$$
\left.\left.\left[\begin{array}{cc|cccccccc}
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
K & \tilde{K}_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]\right)\right), \quad \text{with rank } \tilde{K}_3 = r_5, \text{ and } s_4 = d_i^{(1)} - r_5,
$$

$$
\sim \left(\left[\begin{array}{ccc|cccccccc}
I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{r_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{r_8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right], \left[\begin{array}{ccc|cccccccc}
0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_2} & 0 & 0 \\
I_{e_2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_{s_4} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{s_3} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right],\right.
$$

$$
\left.\left.\left[\begin{array}{ccc|cccccccc}
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{r_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\tilde{K}_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array}\right]\right)\right), \quad \text{with rank } \tilde{K}_2 = r_6,\ r_8 = r_1 - r_6,
$$

$$
\sim \left(\left[\begin{array}{cccc|cccccccc}
I_{r_7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{r_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & I_{r_8} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array} \right], \right.
$$

$$
\left[\begin{array}{cccc|cccccccc}
0 & 0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
0 & 0 & C & C & 0 & 0 & 0 & 0 & 0 & 0 & C & C \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_2} & 0 & 0 \\
I_{r_7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & I_{s_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & I_{s_4} & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{s_3} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array} \right],
$$

$$
\left. \left[\begin{array}{cccc|cccccccc}
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
\hline
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & K & 0 & K & 0 & K & 0 & K & 0 & K & 0 & K \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_2} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & I_{r_3} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & I_{r_4} & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & I_{r_5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & I_{r_6} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
I_{r_7} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{array} \right] \right),
$$

with $\operatorname{rank} \tilde{K}_1 = r_7$, and $s_6 = e_2 - r_7$. With the given definitions we then have that

$$
\begin{aligned}
r_2 &= b_{i+1}, & r_6 &= h_{i+1} - d_{i+1},\\
r_3 &= p_{i+1} - b_{i+1}, & r_7 &= c_{i+1} - h_{i+1},\\
r_4 &= t_{i+1} - p_{i+1}, & e_1 &= k_{i+1},\\
r_5 &= d_{i+1} - t_{i+1}, & e_2 &= e_{i+1} - k_{i+1},
\end{aligned}
$$

and thus we get

$$
\begin{aligned}
a_{i+1} &= a_i + s_i^{(CK)} + s_i^{(MCK)} + s_{i-1} + r_2,\\
s_{i+1}^{(MCK)} &= r_7,\\
s_{i+1}^{(MC)} &= s_6 = e_2 - r_7,\\
s_{i+1}^{(MK)} &= r_6 + s_i,\\
s_{i+1}^{(CK)} &= r_5 + r_4 + r_3,\\
d_{i+1}^{(2)} &= r_8 = d_i^{(2)} - e_2 - r_6,\\
d_{i+1}^{(1)} &= s_2 + s_3 + s_4 = d_i^{(1)} + s_i^{(MC)} + e_1 - r_3 - r_4 - r_5,\\
w_{i+1} &= v_{i+1} - v_i\\
&= 2s_i^{(MCK)} + s_i^{(CK)} + s_i^{(MC)} + s_{i-1} - e_1 - e_2 - r_2 - r_3 - r_4 - r_5 - r_6 - r_7.
\end{aligned}
$$

□

3.1.2 Derivative Array Approach

The algebraic approach described in the previous section allows for the theoretical analysis of linear second order DAEs (3.6), but it cannot be used for the development of numerical methods as neither the inductive process of the reduction to the strangeness-free formulation (3.14) nor the condensed form (3.12) are obtained in a way that is feasible for numerical methods, since we would need derivatives of computed transformation matrices. Therefore, we look for other ways to compute the characteristic invariants of a given DAE as well as a condensed form similar to (3.14) in a numerically stable procedure. The basic idea due to Campbell [22] is to differentiate the differential-algebraic equation (3.6) a number of times and put the original DAE and its derivatives into a large system. Then, purely local invariants can be constructed via local equivalence transformations, which allow to determine the global invariants including the strangeness index, wherever they are defined. Further, it is also possible to derive a strangeness-free formulation using only local informations. The great advantage of derivative arrays is that we can deal with them numerically, since only the original data functions together with their derivatives are included.

In the following, we consider matrix-valued functions $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m,n})$ that are sufficiently smooth and we assume that the strangeness index μ is well-defined, i.e., the ranks are constant in the considered interval and none of the invariant values changes its value during the process. Differentiating the differential-algebraic equation (3.6) and putting the

original DAE and its derivatives up to a sufficiently high order into a large system, in a similar way as in Section 2.2.2, we obtain the *derivative array*, or *inflated differential-algebraic equation* associated with the linear second order DAE (3.6) of the form

$$\mathcal{M}_l(t)\ddot{z}_l + \mathcal{L}_l(t)\dot{z}_l + \mathcal{N}_l(t)z_l = g_l(t), \quad l \in \mathbb{N}_0, \tag{3.19}$$

where $\mathcal{M}_l$, $\mathcal{L}_l$, $\mathcal{N}_l$, z_l and g_l are defined as follows

$$\begin{aligned}
[\mathcal{M}_l]_{i,j} &:= \binom{i}{j} M^{(i-j)} + \binom{i}{j+1} C^{(i-j-1)} + \binom{i}{j+2} K^{(i-j-2)}, \quad i,j = 0,\dots,l,\\
[\mathcal{L}_l]_{i,j} &:= \begin{cases} C^{(i)} + iK^{(i-1)} & \text{for } i = 0,\dots,l,\ j = 0,\\ 0 & \text{otherwise,} \end{cases}\\
[\mathcal{N}_l]_{i,j} &:= \begin{cases} K^{(i)} & \text{for } i = 0,\dots,l,\ j = 0,\\ 0 & \text{otherwise,} \end{cases}\\
[z_l]_i &:= x^{(i)}, \quad i = 0,\dots,l,\\
[g_l]_i &:= f^{(i)}, \quad i = 0,\dots,l.
\end{aligned} \tag{3.20}$$

For $l = 3$, for example, the extended system (3.19) is of the form

$$\begin{bmatrix} M & 0 & 0 & 0\\ \dot{M}+C & M & 0 & 0\\ \ddot{M}+2\dot{C}+K & 2\dot{M}+C & M & 0\\ M^{(3)}+3\ddot{C}+3\dot{K} & 3\ddot{M}+3\dot{C}+K & 3\dot{M}+C & M \end{bmatrix} \begin{bmatrix} x\\ \dot{x}\\ \ddot{x}\\ x^{(3)} \end{bmatrix}^{(2)}$$
$$+ \begin{bmatrix} C & 0 & 0 & 0\\ \dot{C}+K & 0 & 0 & 0\\ \ddot{C}+2\dot{K} & 0 & 0 & 0\\ C^{(3)}+3\ddot{K} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x\\ \dot{x}\\ \ddot{x}\\ x^{(3)} \end{bmatrix}^{(1)} + \begin{bmatrix} K & 0 & 0 & 0\\ \dot{K} & 0 & 0 & 0\\ \ddot{K} & 0 & 0 & 0\\ K^{(3)} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x\\ \dot{x}\\ \ddot{x}\\ x^{(3)} \end{bmatrix} = \begin{bmatrix} f\\ \dot{f}\\ \ddot{f}\\ f^{(3)} \end{bmatrix}.$$

For every $l \in \mathbb{N}_0$ and every $t \in \mathbb{I}$, we can now determine the local characteristic values of the triple $(\mathcal{M}_l(t), \mathcal{L}_l(t), \mathcal{N}_l(t))$ by transforming to the local condensed form (3.10). We can show that these local quantities at a fixed point $\hat{t} \in \mathbb{I}$ are invariant under global equivalence transformations of the original triple $(M(t), C(t), K(t))$ of matrix-valued functions. To do this, we need the following Lemmas.

Lemma 3.16. *Let $D = ABC$ be the product of three sufficiently smooth matrix valued functions of appropriate dimensions. Then*

$$D^{(i)} = \sum_{j=0}^{i} \sum_{k=0}^{i-j} \binom{i}{j} \binom{i-j}{k} A^{(j)} B^{(k)} C^{(i-j-k)}.$$

Proof. See [82, Lemma 3.28]. □

Lemma 3.17. *For all integers i, j, k, l with $i \geq 0$, $i \geq j \geq 0$, $i - j \geq k \geq 0$, we have the identities*

$$\begin{aligned}
\binom{i}{k}\binom{i-k}{l}\binom{i-k-l}{j} &= \binom{i}{j}\binom{i-j}{k}\binom{i-j-k}{l},\\
\binom{i}{k}\binom{i-k}{l}\binom{i-k-l+2}{j+2} &= \binom{i}{j}\binom{i-j}{k}\binom{i-j-k}{l} + 2\binom{i}{j+1}\binom{i-j-1}{k}\binom{i-j-k-1}{l}\\
&\quad + \binom{i}{j+2}\binom{i-j-2}{k}\binom{i-j-k-2}{l},\\
\binom{i}{k}\binom{i-k}{l}\binom{i-k-l+1}{j+2} &= \binom{i}{j+1}\binom{i-j-1}{k}\binom{i-j-1-k}{l}\\
&\quad + \binom{i}{j+2}\binom{i-j-2}{k}\binom{i-j-k-2}{l},\\
\binom{i}{k}\binom{i-k}{l}\binom{i-k-l}{j+2} &= \binom{i}{j+2}\binom{i-j-2}{k}\binom{i-j-k-2}{l}.
\end{aligned}$$

Proof. The proof follows by straightforward calculations using the definition of the binomial coefficient. □

Now, we can show that the local quantities of the triple $(\mathcal{M}_l(\hat{t}), \mathcal{L}_l(\hat{t}), \mathcal{N}_l(\hat{t}))$ are invariant under global equivalence transformations of the original triple $(M(t), C(t), K(t))$ of matrix-valued functions.

Theorem 3.18. *Consider two triples (M, C, K) and $(\tilde{M}, \tilde{C}, \tilde{K})$ of sufficiently smooth matrix-valued functions that are globally equivalent via the transformation*

$$\tilde{M} = PMQ, \quad \tilde{C} = PCQ + 2PM\dot{Q}, \quad \tilde{K} = PKQ + PC\dot{Q} + PM\ddot{Q}$$

according to Definition 3.4, with sufficiently smooth matrix-valued functions P and Q. Let $(\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l)$ and $(\tilde{\mathcal{M}}_l, \tilde{\mathcal{L}}_l, \tilde{\mathcal{N}}_l)$, $l \in \mathbb{N}_0$, be the corresponding inflated triples constructed as in (3.20) *and introduce the block matrix functions*

$$\begin{aligned}
[\Pi_l]_{i,j} &= \tbinom{i}{j} P^{(i-j)}, \qquad [\Psi_l]_{i,j} = \begin{cases} \frac{i+2}{2} Q^{(i+1)} & \text{for } i = 0, \ldots, l,\ j = 0,\\ 0 & \text{otherwise,} \end{cases}\\
[\Theta_l]_{i,j} &= \tbinom{i+2}{j+2} Q^{(i-j)}, \quad [\Sigma_l]_{i,j} = \begin{cases} Q^{(i+2)} & \text{for } i = 0, \ldots, l,\ j = 0,\\ 0 & \text{otherwise.} \end{cases}
\end{aligned} \tag{3.21}$$

Then,

$$[\tilde{\mathcal{M}}_l(t), \tilde{\mathcal{L}}_l(t), \tilde{\mathcal{N}}_l(t)] = \Pi_l(t)[\mathcal{M}_l(t), \mathcal{L}_l(t), \mathcal{N}_l(t)] \begin{bmatrix} \Theta_l(t) & 2\Psi_l(t) & \Sigma_l(t) \\ 0 & \Theta_l(t) & \Psi_l(t) \\ 0 & 0 & \Theta_l(t) \end{bmatrix} \tag{3.22}$$

for every $t \in \mathbb{I}$, and the corresponding matrix triples are locally equivalent.

Proof. First, we note that all matrix-valued functions $\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l, \tilde{\mathcal{M}}_l, \tilde{\mathcal{L}}_l, \tilde{\mathcal{N}}_l, \Pi_l, \Psi_l, \Theta_l$ and Σ_l are block lower triangular with the same block structure. Furthermore, $\mathcal{N}_l$, $\tilde{\mathcal{N}}_l$,$\mathcal{L}_l$, $\tilde{\mathcal{L}}_l$, Ψ_l and Σ_l have nonzero blocks only in the first block column. Using Lemma 3.16 we obtain

$$\begin{aligned}
\tilde{M}^{(i)} &= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} \tbinom{i-k_1}{k_2} P^{(k_1)} M^{(k_2)} Q^{(i-k_1-k_2)}, \\
\tilde{C}^{(i)} &= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} \tbinom{i-k_1}{k_2} \left[P^{(k_1)} C^{(k_2)} Q^{(i-k_1-k_2)} + 2P^{(k_1)} M^{(k_2)} Q^{(i+1-k_1-k_2)} \right], \\
\tilde{K}^{(i)} &= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} \tbinom{i-k_1}{k_2} \left[P^{(k_1)} K^{(k_2)} Q^{(i-k_1-k_2)} + P^{(k_1)} C^{(k_2)} Q^{(i+1-k_1-k_2)} \right. \\
&\qquad \left. + P^{(k_1)} M^{(k_2)} Q^{(i+2-k_1-k_2)} \right].
\end{aligned}$$

Inserting the definitions, shifting and inverting the summations and applying Lemma 3.17 leads to

$$\begin{aligned}
&[\Pi_l \mathcal{M}_l \Theta_l]_{i,j} = \sum_{l_1=j}^{i} \sum_{l_2=j}^{l_1} [\Pi_l]_{i,l_1} [\mathcal{M}_l]_{l_1,l_2} [\Theta_l]_{l_2,j} \\
&= \sum_{l_1=j}^{i} \sum_{l_2=j}^{l_1} \tbinom{i}{l_1} P^{(i-l_1)} \left[\tbinom{l_1}{l_2} M^{(l_1-l_2)} + \tbinom{l_1}{l_2+1} C^{(l_1-l_2-1)} + \tbinom{l_1}{l_2+2} K^{(l_1-l_2-2)} \right] \tbinom{l_2+2}{j+2} Q^{(l_2-j)} \\
&= \sum_{k_1=0}^{i-j} \sum_{l_2=j}^{k_1+j} \tbinom{i}{k_1+j} P^{(i-k_1-j)} \left[\binom{k_1+j}{l_2} M^{(k_1+j-l_2)} + \binom{k_1+j}{l_2+1} C^{(k_1+j-l_2-1)} \right. \\
&\qquad \left. + \tbinom{k_1+j}{l_2+2} K^{(k_1+j-l_2-2)} \right] \tbinom{l_2+2}{j+2} Q^{(l_2-j)} \\
&= \tbinom{i}{j} \sum_{k_1=0}^{i-j} \sum_{k_2=0}^{i-j-k_1} \tbinom{i-j}{k_1} \tbinom{i-j-k_1}{k_2} P^{(k_1)} M^{(k_2)} Q^{(i-j-k_1-k_2)} \\
&+ \tbinom{i}{j+1} \sum_{k_1=0}^{i-j-1} \sum_{k_2=0}^{i-j-1-k_1} \tbinom{i-j-1}{k_1} \tbinom{i-j-1-k_1}{k_2} \left[P^{(k_1)} C^{(k_2)} Q^{(i-j-1-k_1-k_2)} + 2P^{(k_1)} M^{(k_2)} Q^{(i-j-k_1-k_2)} \right] \\
&+ \tbinom{i}{j+2} \sum_{k_1=0}^{i-j-2} \sum_{k_2=0}^{i-j-2-k_1} \tbinom{i-j-2}{k_1} \tbinom{i-j-2-k_1}{k_2} \left[P^{(k_1)} K^{(k_2)} Q^{(i-j-2-k_1-k_2)} \right. \\
&\qquad \left. + P^{(k_1)} C^{(k_2)} Q^{(i-j-1-k_1-k_2)} + P^{(k_1)} M^{(k_2)} Q^{(i-j-k_1-k_2)} \right] \\
&= \tbinom{i}{j} \tilde{M}^{(i-j)} + \tbinom{i}{j+1} \tilde{C}^{(i-j-1)} + \tbinom{i}{j+2} \tilde{K}^{(i-j-2)} = [\tilde{\mathcal{M}}_l]_{i,j}.
\end{aligned}$$

In the same way we get

$$[\Pi_l \mathcal{L}_l \Theta_l]_{i,0} + [2\Pi_l \mathcal{M}_l \Psi_l]_{i,0} = \sum_{l_1=0}^{i} [\Pi_l]_{i,l_1} [\mathcal{L}_l]_{l_1,0} [\Theta_l]_{0,0} + 2 \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} [\Pi_l]_{i,l_1} [\mathcal{M}_l]_{l_1,l_2} [\Psi_l]_{l_2,0}$$

$$
\begin{aligned}
&= \sum_{l_1=0}^{i} \tbinom{i}{l_1} P^{(i-l_1)} \left[C^{(l_1)} + l_1 K^{(l_1-1)} \right] Q \\
&\quad + 2 \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} \tbinom{i}{l_1} P^{(i-l_1)} \left[\tbinom{l_1}{l_2} M^{(l_1-l_2)} + \tbinom{l_1}{l_2+1} C^{(l_1-l_2-1)} + \tbinom{l_1}{l_2+2} K^{(l_1-l_2-2)} \right] \frac{l_2+2}{2} Q^{(l_2+1)} \\
&= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} \tbinom{i-k_1}{k_2} \left[P^{(k_1)} C^{(k_2)} Q^{(i-k_1-k_2)} + 2 P^{(k_1)} M^{(k_2)} Q^{(i+1-k_1-k_2)} \right] \\
&\quad + i \sum_{k_1=0}^{i-1} \sum_{k_2=0}^{i-1-k_1} \tbinom{i-1}{k_1} \tbinom{i-1-k_1}{k_2} \Big[P^{(k_1)} K^{(k_2)} Q^{(i-1-k_1-k_2)} \\
&\qquad\qquad + P^{(k_1)} C^{(k_2)} Q^{(i-k_1-k_2)} + P^{(k_1)} M^{(k_2)} Q^{(i+1-k_1-k_2)} \Big] \\
&= \tilde{C}^{(i)} + i \tilde{K}^{(i-1)} = [\tilde{\mathcal{L}}_l]_{i,0},
\end{aligned}
$$

and

$$
\begin{aligned}
&[\Pi_l \mathcal{N}_l \Theta_l]_{i,0} + [\Pi_l \mathcal{L}_l \Psi_l]_{i,0} + [\Pi_l \mathcal{M}_l \Sigma_l]_{i,0} = \\
&= \sum_{l_1=0}^{i} [\Pi_l]_{i,l_1} [\mathcal{N}_l]_{l_1,0} [\Theta_l]_{0,0} + \sum_{l_1=0}^{i} [\Pi_l]_{i,l_1} [\mathcal{L}_l]_{l_1,0} [\Psi_l]_{0,0} + \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} [\Pi_l]_{i,l_1} [\mathcal{M}_l]_{l_1,l_2} [\Sigma_l]_{l_2,0} \\
&= \sum_{l_1=0}^{i} \tbinom{i}{l_1} P^{(i-l_1)} K^{(l_1)} Q + \sum_{l_1=0}^{i} \tbinom{i}{l_1} P^{(i-l_1)} \left[C^{(l_1)} + l_1 K^{(l_1-1)} \right] Q^{(1)} \\
&\quad + \sum_{l_1=0}^{i} \sum_{l_2=0}^{l_1} \tbinom{i}{l_1} P^{(i-l_1)} \left[\tbinom{l_1}{l_2} M^{(l_1-l_2)} + \tbinom{l_1}{l_2+1} C^{(l_1-l_2-1)} + \tbinom{l_1}{l_2+2} K^{(l_1-l_2-2)} \right] Q^{(l_2+2)} \\
&= \sum_{k_1=0}^{i} \sum_{k_2=0}^{i-k_1} \tbinom{i}{k_1} \tbinom{i-k_1}{k_2} \left[P^{(k_1)} (K^{(k_2)} Q^{(i-k_1-k_2)} + C^{(k_2)} Q^{(i+1-k_1-k_2)} + M^{(k_2)} Q^{(i+2-k_1-k_2)}) \right] \\
&= \tilde{K}^{(i)} = [\tilde{\mathcal{N}}_l]_{i,0}.
\end{aligned}
$$

□

As a consequence of Theorem 3.18 we have shown that the local characteristic values of the inflated triple $(\mathcal{M}_l(\hat{t}), \mathcal{L}_l(\hat{t}), \mathcal{N}_l(\hat{t}))$ at a fixed point $\hat{t}$, that in the following will be denoted by $(\tilde{r}_l,\, \tilde{d}_l^{(1)},\, \tilde{a}_l,\, \tilde{s}_l^{(MCK)},\, \tilde{s}_l^{(MK)},\, \tilde{s}_l^{(CK)},\, \tilde{s}_l^{(MC)},\, \tilde{u}_l,\, \tilde{v}_l)$, are well-defined for equivalent triples of matrix-valued functions and for each $l \in \mathbb{N}_0$. These quantities are numerically computable via a number of numerical rank decisions using e.g. a singular value decomposition or a rank revealing QR decomposition, see [54]. Next, we show how these local quantities of the inflated triple $(\mathcal{M}_l(\hat{t}), \mathcal{L}_l(\hat{t}), \mathcal{N}_l(\hat{t}))$ are related to the global characteristic values $(r_i, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(MK)}, s_i^{(CK)}, s_i^{(MC)}, u_i, v_i)$ of the original triple (M, C, K) at the point $\hat{t}$. For convenience of representation we restrict ourselves in the following to the case that the strangeness index is restricted by $\mu \leq 2$.

Theorem 3.19. *Let the functions $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m,n})$ be sufficiently smooth with well-defined strangeness index $\mu \leq 2$ and global characteristic values $(r_i, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(CK)}, s_i^{(MC)}, s_i^{(MK)}, u_i, v_i)$, $i \in \mathbb{N}_0$. Furthermore, let $(\mathcal{M}_l(\hat{t}), \mathcal{L}_l(\hat{t}), \mathcal{N}_l(\hat{t}))$ be the corresponding inflated matrix triple at a fixed $\hat{t} \in \mathbb{I}$ with local characteristic values $(\tilde{r}_l, \tilde{d}_l^{(1)}, \tilde{a}_l, \tilde{s}_l^{(MCK)}, \tilde{s}_l^{(CK)}, \tilde{s}_l^{(MC)}, \tilde{s}_l^{(MK)}, \tilde{u}_l, \tilde{v}_l)$. Then, for $l = 0, 1, 2$, we have*

$$\begin{aligned}
\operatorname{rank}[\mathcal{M}_l] &= \tilde{r}_l = (l+1)m - \sum_{i=0}^{l} q_i - \sum_{i=0}^{l} c_i - \sum_{i=0}^{l} v_i, \\
\operatorname{rank}[\mathcal{M}_l, \mathcal{L}_l] &= (l+1)m - \sum_{i=0}^{l} c_i - \sum_{i=0}^{l} v_i, \\
\operatorname{rank}[\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l] &= (l+1)m - \sum_{i=0}^{l} v_i,
\end{aligned} \tag{3.23}$$

using the definitions as in Lemma 3.15, and

$$\begin{aligned}
\tilde{d}_l^{(1)} &= k_l - p_l + b_l, \\
\tilde{a}_l &= b_l = c_l - s_l^{(MCK)} - s_l^{(CK)} - s_l^{(MK)} + s_{l-1}, \\
\tilde{s}_l^{(MCK)} &= \sum_{i=0}^{l} c_i - b_{l-1} - p_l - \sum_{i=1}^{l} (d_i - t_i), \\
\tilde{s}_l^{(CK)} &= b_{l-1} - b_l + p_l, \\
\tilde{s}_l^{(MC)} &= \sum_{i=0}^{l} q_i - \sum_{i=0}^{l} c_i + \sum_{i=1}^{l} (d_i - t_i) + p_l - k_l, \\
\tilde{s}_l^{(MK)} &= \sum_{i=1}^{l} (d_i - t_i), \\
\tilde{v}_l &= \sum_{i=0}^{l} v_i, \\
\tilde{u}_l &= (l+1)u_0 + (l+1)a_0 + lk_0 - \sum_{i=0}^{l} k_i - \sum_{i=0}^{l} b_i.
\end{aligned} \tag{3.24}$$

Proof. Due to Theorem 3.18 we may assume without loss of generality that the triple (M, C, K) is already given in the global condensed form (3.12). Then, for fixed $\hat{t} \in \mathbb{I}$, we have to determine the local characteristic quantities of $(\mathcal{M}_l(\hat{t}), \mathcal{L}_l(\hat{t}), \mathcal{N}_l(\hat{t}))$. In the following, we will omit the argument $\hat{t}$ and we use local equivalence transformations of the form (3.9) on the inflated triple. For $l = 0$ it is immediately clear that

$$\tilde{r}_0 = \operatorname{rank} \mathcal{M}_0 = m - q_0 - c_0 - v_0,$$

$$\begin{aligned}
&\operatorname{rank}[\mathcal{M}_0, \mathcal{L}_0] = m - a_0 - s_0^{(CK)} - s_0^{(MCK)} - s_0^{(MK)} - v_0 = m - c_0 - v_0,\\
&\operatorname{rank}[\mathcal{M}_0, \mathcal{L}_0, \mathcal{N}_0] = m - v_0,
\end{aligned}$$

and the local characteristic values of $(\mathcal{M}_0, \mathcal{L}_0, \mathcal{N}_0)$ correspond to the global characteristic values of (M, C, K). For $l = 1$ we have to consider the inflated triple

$$(\mathcal{M}_1, \mathcal{L}_1, \mathcal{N}_1) = \left(\begin{bmatrix} M & 0 \\ \dot{M} + C & M \end{bmatrix}, \begin{bmatrix} C & 0 \\ \dot{C} + K & 0 \end{bmatrix}, \begin{bmatrix} K & 0 \\ \dot{K} & 0 \end{bmatrix} \right),$$

with (M, C, K) in global condensed form (3.12). The identity blocks in the matrix M allow to eliminate all other entries in the corresponding block rows of $\mathcal{M}_1$ by local equivalence transformations. Further eliminations using the identity blocks of the global condensed form and block decompositions yield the following matrix triple, where we only state the first block columns of $\mathcal{L}_1$ and $\mathcal{N}_1$ since all other entries are zero

$$\left(\left[\begin{array}{cccccccccc|cccccc}
0 & . & . & . & . & . & . & . & . & . & & & & & & \\
. & 0 & . & . & . & . & . & . & . & . & & & & & & \\
. & 0 & . & . & . & . & . & . & . & . & & & & & & \\
. & . & I & . & . & . & . & . & . & . & & & & & & \\
. & . & . & I & . & . & . & . & . & . & & & & & & \\
. & . & . & . & I & . & . & . & . & . & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & & & & & & & & & & & \\
\hline
. & . & 0 & 0 & 0 & . & . & 0 & 0 & 0 & I & & & & & \\
. & . & 0 & 0 & 0 & . & . & 0 & 0 & 0 & & I & & & & \\
. & . & 0 & 0 & 0 & . & . & 0 & 0 & 0 & & & I & & & \\
. & . & 0 & 0 & 0 & . & . & 0 & 0 & 0 & & & & I & & \\
. & . & . & . & . & I & . & . & . & . & & & & & & \\
. & . & . & . & . & . & I & . & . & . & & & & & & \\
I & . & & & & & & & & & & & & & & \\
. & I & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & &
\end{array}\right] \right), \tag{3.25}$$

$$
\left.
\left[\begin{array}{cccccccccc}
. & 0 & 0 & 0 & 0 & . & 0 & 0 & I_{k_1} & 0 \\
. & 0 & 0 & I_{\tilde{e}_1} & 0 & . & 0 & 0 & 0 & 0 \\
. & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
. & . & 0 & 0 & C & . & . & 0 & 0 & C \\
. & . & 0 & 0 & C & . & . & 0 & 0 & C \\
. & . & 0 & 0 & C & . & . & 0 & 0 & C \\
. & . & . & . & . & 0 & . & . & . & . \\
. & . & . & . & . & . & I & . & . & . \\
0 & . & & & & & & & & \\
. & I & & & & & & & & \\
. & . & & & & & & & & \\
. & . & & & & & & & & \\
. & . & & & & & & & & \\
. & . & & & & & & & & \\
. & . & & & & & & & & \\
\hline
. & K & 0 & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} \\
. & K & 0 & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} \\
. & K & 0 & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} \\
. & K & 0 & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & . & . & . & . & . & . & I & . & . \\
. & . & . & . & . & I & . & . & . & . \\
. & . & I & . & & . & . & . & . & . \\
I & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & .
\end{array}\right],
\left[\begin{array}{cccccccccc}
. & \tilde{K} & . & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & K & . & K & K & . & K & . & K & K \\
. & . & . & . & . & . & . & I & . & . \\
. & . & . & . & . & I & . & . & . & . \\
. & . & I & . & . & . & . & . & . & . \\
I & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
\hline
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & .
\end{array}\right]\right),
\quad
\begin{array}{c}
k_1 \\ \tilde{e}_1 \\ \tilde{s}_5 \\ \tilde{s}_6 \\ \tilde{e}_1 \\ \tilde{d}_2 \\ \tilde{s}_4 \\ \tilde{d}_1 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ a_0 \\ \tilde{s}_4 \\ \tilde{s}_3 \\ \tilde{s}_1 \\ v_0 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \\ \tilde{d}_2 \\ \tilde{s}_4 \\ \tilde{d}_1 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ a_0 \\ \tilde{s}_4 \\ \tilde{s}_3 \\ \tilde{s}_1 \\ v_0
\end{array}
$$

with dimensions $\tilde{s}_1 = s_0^{(MCK)}$, $\tilde{s}_2 = s_0^{(MC)}$, $\tilde{s}_3 = s_0^{(MK)}$, $\tilde{s}_4 = s_0^{(CK)}$, $\tilde{s}_5 = s_0^{(MCK)} + s_0^{(MC)} - e_1$, $\tilde{s}_6 = s_0^{(MK)} - e_1 + k_1$, $\tilde{e}_1 = e_1 - k_1$, $\tilde{d}_1 = d_0^{(1)}$, and $\tilde{d}_2 = d_0^{(2)}$. Now, we can read off the corresponding ranks as

$$
\begin{aligned}
\tilde{r}_1 = \operatorname{rank} \mathcal{M}_1 &= 2m - c_0 - 2v_0 - e_0 - 2s_0^{(MCK)} - s_0^{(MC)} - s_0^{(CK)} - s_0^{(MK)} - a_0 \\
&= 2m - c_0 - c_1 - v_0 - v_1 - e_0 - e_1 - a_0 - s_0^{(MK)} \\
&= 2m - c_0 - c_1 - v_0 - v_1 - e_0 - (e_1 + c_0 - s_0^{(CK)} - s_0^{(MCK)}), \\
\operatorname{rank}\,[\mathcal{M}_1, \mathcal{L}_1] &= 2m - c_0 - s_0^{(CK)} - 2s_0^{(MCK)} - 2v_0 - s_0^{(MC)} + e_1 \\
&= 2m - c_0 - c_1 - v_0 - v_1, \\
\operatorname{rank}\,[\mathcal{M}_1, \mathcal{L}_1, \mathcal{N}_1] &= 2m - 2v_0 - 2s_0^{(MCK)} - s_0^{(CK)} - s_0^{(MC)} + e_1 + c_1 \\
&= 2m - v_0 - v_1.
\end{aligned}
$$

Let $\Pi \in \mathbb{R}^{2m \times 2m}$ and $\Theta, \Psi, \Sigma \in \mathbb{R}^{2n \times 2n}$ be block matrices, such that

$$
(\tilde{\mathcal{M}}_1, \tilde{\mathcal{L}}_1, \tilde{\mathcal{N}}_1) = (\Pi \mathcal{M}_1 \Theta, \Pi \mathcal{L}_1 \Theta + 2\Pi \mathcal{M}_1 \Psi, \Pi \mathcal{N}_1 \Theta + \Pi \mathcal{L}_1 \Psi + \Pi \mathcal{M}_1 \Sigma)
$$

is in the locally transformed form (3.25). Further, let Π and Θ be partitioned as

$$
\Pi := \left[\Pi_1^T, \ldots, \Pi_{28}^T\right]^T, \quad \Theta := [\Theta_1, \ldots, \Theta_{18}],
$$

corresponding to the row and column structure of the block matrices in (3.25). Then we have

$$\begin{aligned}
[\Pi_1^T, \Pi_2^T, \Pi_3^T, \Pi_7^T, \ldots, \Pi_{15}^T, \Pi_{24}^T, \ldots, \Pi_{28}^T]^T \mathcal{M}_1 &= 0,\\
\mathcal{M}_1[\Theta_8, \Theta_9, \Theta_{10}, \Theta_{15}, \ldots, \Theta_{18}] &= 0,\\
[\Pi_3^T, \Pi_7^T, \Pi_9^T, \Pi_{11}^T, \ldots, \Pi_{15}^T, \Pi_{28}^T]^T \mathcal{L}_1 &= 0,\\
[\Pi_1^T, \Pi_2^T, \Pi_3^T, \Pi_7^T, \ldots, \Pi_{15}^T, \Pi_{24}^T, \ldots, \Pi_{28}^T]^T \mathcal{L}_1 [\Theta_{10}, \Theta_{15}, \ldots, \Theta_{18}] &= 0.
\end{aligned}$$

This means that the columns of the matrix $V_1 := [\Pi_1^T, \Pi_2^T, \Pi_3^T, \Pi_7^T, \ldots, \Pi_{15}^T, \Pi_{24}^T, \ldots, \Pi_{28}^T]$ form a basis of kernel$(\mathcal{M}_1^T)$, the columns of $V_2 := [\Theta_8, \Theta_9, \Theta_{10}, \Theta_{15}, \ldots, \Theta_{18}]$ form a basis of kernel$(\mathcal{M}_1)$, the columns of $V_3 := [\Pi_3^T, \Pi_7^T, \Pi_9^T, \Pi_{11}^T, \ldots, \Pi_{15}^T, \Pi_{28}^T]$ form a basis of kernel$(\mathcal{M}_1^T) \cap$ kernel$(\mathcal{L}_1^T)$, and the columns of $V_4 := [\Theta_{10}, \Theta_{15}, \ldots, \Theta_{18}]$ form a basis of kernel$(\mathcal{M}_1) \cap$kernel$(V_1^T \mathcal{L}_1)$. Therefore, using Lemma 3.8 and with the definitions of Lemma 3.15 we have

$$\begin{aligned}
&\tilde{a}_1 = \operatorname{rank}(V_3^T \mathcal{N}_1 V_4) = b_1,\\
&\tilde{s}_1^{(MCK)} = \dim(\operatorname{range}(\mathcal{M}_1^T) \cap \operatorname{range}(\mathcal{L}_1^T V_1) \cap \operatorname{range}(\mathcal{N}_1^T V_3)) = c_0 + c_1 - b_0 - d_1 + t_1 - p_1,\\
&\tilde{s}_1^{(CK)} = \operatorname{rank}(V_3^T \mathcal{N}_1 V_2) - \tilde{a}_1 = a_0 + p_1 - b_1,\\
&\tilde{d}_1^{(1)} = \operatorname{rank}(V_1^T \mathcal{L}_1 V_2) - \tilde{s}_1^{(CK)} = k_1 + a_0 - p_1 - a_0 + b_1 = k_1 - p_1 + b_1,\\
&\tilde{s}_1^{(MC)} = \operatorname{rank}(V_1^T \mathcal{L}_1) - \tilde{s}_1^{(MCK)} - \tilde{s}_1^{(CK)} - \tilde{d}_1^{(1)} = q_0 + q_1 - c_0 - c_1 + d_1 - t_1 + p_1 - k_1,\\
&\tilde{s}_1^{(MK)} = \operatorname{rank}(V_3^T \mathcal{N}_1) - \tilde{a}_1 - \tilde{s}_1^{(MCK)} - \tilde{s}_1^{(CK)} = d_1 - t_1,\\
&\tilde{v}_1 = 2m - \tilde{r}_1 - 2\tilde{s}_1^{(CK)} - \tilde{d}_1^{(1)} - 2\tilde{s}_1^{(MCK)} - \tilde{s}_1^{(MC)} - \tilde{a}_1 - \tilde{s}_1^{(MK)} = v_0 + v_1,\\
&\tilde{u}_1 = 2n - \tilde{r}_1 - \tilde{s}_1^{(CK)} - \tilde{d}_1^{(1)} - \tilde{a}_1 = 2u_0 + a_0 + k_0 - k_1 - b_1.
\end{aligned}$$

Finally, for $l = 2$ we have to consider the inflated triple

$$(\mathcal{M}_2, \mathcal{L}_2, \mathcal{N}_2) = \left(\begin{bmatrix} M & 0 & 0 \\ \dot{M} + C & M & 0 \\ \ddot{M} + 2\dot{C} + K & 2\dot{M} + C & M \end{bmatrix}, \begin{bmatrix} C & 0 & 0 \\ \dot{C} + K & 0 & 0 \\ \ddot{C} + 2\dot{K} & 0 & 0 \end{bmatrix}, \begin{bmatrix} K & 0 & 0 \\ \dot{K} & 0 & 0 \\ \ddot{K} & 0 & 0 \end{bmatrix} \right).$$

Again, the identities in the diagonal blocks of $\mathcal{M}_2$ allow to eliminate all another entries in the corresponding block rows of $\mathcal{M}_2$ without altering $\mathcal{L}_2$ or $\mathcal{N}_2$. Further eliminations using identity blocks in the global condensed form and block decompositions using local equivalence transformations yield the matrix triple $(\tilde{\mathcal{M}}_2, \tilde{\mathcal{L}}_2, \tilde{\mathcal{N}}_2)$ of the form

$$
\left(\left[\begin{array}{cccccccccccc|ccc|ccc}
0 & . & . & . & . & . & . & . & . & . & . & . & & & & & & \\
. & 0 & . & . & . & . & . & . & . & . & . & . & & & & & & \\
. & . & 0 & . & . & . & . & . & . & . & . & . & & & & & & \\
. & . & 0 & . & . & . & . & . & . & . & . & . & & & & & & \\
. & . & 0 & . & . & . & . & . & . & . & . & . & & & & & & \\
. & . & 0 & . & . & . & . & . & . & . & . & . & & & & & & \\
. & . & . & I & . & . & . & . & . & . & . & . & & & & & & \\
. & . & . & . & I & . & . & . & . & . & . & . & & & & & & \\
. & . & . & . & . & I & . & . & . & . & . & . & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
. & . & . & . & . & . & & & & & & & & & & & & \\
\hline
. & . & 0 & 0 & 0 & 0 & . & . & 0 & I_{k_1} & 0 & 0 & 0 & & & & & & \\
. & . & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 & & 0 & & & & & \\
. & . & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 & & & I & & & & \\
. & . & 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 & & & & I & & & \\
. & . & . & . & . & . & 0 & . & . & . & . & . & & & & & & & \\
. & . & . & . & . & . & . & I & . & . & . & . & & & & & & & \\
0 & . & & & & & & & & & & & & & & & & & \\
. & I & & & & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & & & & \\
. & . & & & & & & & & & & & & & & & & & \\
\hline
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & . & . & 0 & 0 & I & & & & \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & . & . & 0 & 0 & & I & & & \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & . & . & 0 & 0 & & & I & & \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 & . & . & 0 & 0 & . & . & 0 & 0 & & & & I & \\
. & 0 & . & 0 & 0 & 0 & . & 0 & . & 0 & 0 & 0 & . & . & . & . & I & . & . & . & & & & & \\
. & 0 & . & 0 & 0 & 0 & . & 0 & . & 0 & 0 & 0 & . & . & . & . & . & I & . & . & & & & & \\
. & 0 & . & 0 & 0 & 0 & . & 0 & . & 0 & 0 & 0 & I & . & . & . & . & . & . & . & & & & & \\
. & 0 & . & 0 & 0 & 0 & . & 0 & . & 0 & 0 & 0 & . & I & & & & & & & & & & & \\
. & . & . & . & . & . & . & . & I & . & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & . & I & . & . & . & . & . & . & . & & & & & & & & & & & \\
. & . & I & . & . & . & . & . & . & . & . & . & . & . & & & & & & & & & & & \\
I & . & . & . & . & . & . & . & . & . & . & . & . & . & & & & & & & & & & & \\
. & . & . & . & . & . & . & . & . & . & . & . & . & . & & & & & & & & & & &
\end{array}\right]\right),
$$

$$\left[\begin{array}{cccccccccccc}
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & I_{k_1} & 0 & 0 \\
. & 0 & 0 & I_{\tilde{e}_1} & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & I_{k_2} & 0 \\
. & 0 & 0 & 0 & I_{\tilde{e}_2} & 0 & . & 0 & 0 & 0 & 0 & 0 \\
. & 0 & 0 & 0 & 0 & 0 & . & 0 & 0 & 0 & 0 & 0 \\
. & . & 0 & C & C & C & . & . & 0 & C & C & C \\
. & . & 0 & C & C & C & . & . & 0 & C & C & C \\
. & . & 0 & C & C & C & . & . & 0 & C & C & C \\
. & . & . & . & . & . & 0 & . & . & . & . & . \\
. & . & . & . & . & . & . & I & . & . & . & . \\
0 & . & . & . & . & . & . & . & . & . & . & . \\
. & I & . & . & . & . & . & . & . & . & . & . \\
. & . & & & & & & & & & & \\
. & . & & & & & & & & & & \\
. & . & & & & & & & & & & \\
. & . & & & & & & & & & & \\
. & . & & & & & & & & & & \\
\hline
. & \tilde{K} & 0 & \dot{\tilde{C}} & \dot{\tilde{C}} & \dot{\tilde{C}} & . & \tilde{K} & 0 & \dot{\tilde{C}} & \dot{\tilde{C}} & \dot{\tilde{C}} \\
. & \tilde{K} & 0 & \dot{\tilde{C}} & \dot{\tilde{C}} & \dot{\tilde{C}} & . & \tilde{K} & 0 & \dot{\tilde{C}} & \dot{\tilde{C}} & \dot{\tilde{C}} \\
. & K & 0 & \dot{C} & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} & \dot{C} \\
. & K & 0 & \dot{C} & \dot{C} & \dot{C} & . & K & 0 & \dot{C} & \dot{C} & \dot{C} \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & . & . & . & . & . & . & . & I & . & . & . \\
. & . & . & . & . & . & I & . & . & . & . & . \\
. & . & I & . & . & . & . & . & . & . & . & . \\
I & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} & . & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} \\
. & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} & . & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} \\
. & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} & . & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} \\
. & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} & . & \dot{K} & 0 & \ddot{C} & \ddot{C} & \ddot{C} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right], \tag{3.26}$$

$$
\left[\begin{array}{cccccccccccc}
. & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} \\
. & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} & . & \tilde{K} & . & \tilde{K} & \tilde{K} & \tilde{K} \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & K & . & K & K & K & . & K & . & K & K & K \\
. & . & . & . & . & . & . & . & I & . & . & . \\
. & . & . & . & . & . & I & . & . & . & . & . \\
 & . & I & . & . & . & . & . & . & . & . & . \\
I & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & \dot{\tilde{K}} & . & \dot{\tilde{K}} & \dot{\tilde{K}} & \dot{\tilde{K}} & . & \dot{\tilde{K}} & . & \dot{\tilde{K}} & \dot{\tilde{K}} & \dot{\tilde{K}} \\
. & \dot{\tilde{K}} & . & \dot{\tilde{K}} & \dot{\tilde{K}} & \dot{\tilde{K}} & . & \dot{\tilde{K}} & . & \dot{\tilde{K}} & \dot{\tilde{K}} & \dot{\tilde{K}} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} & . & \dot{K} & . & \dot{K} & \dot{K} & \dot{K} \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
\hline
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} & . & \ddot{K} & . & \ddot{K} & \ddot{K} & \ddot{K} \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & . \\
. & . & . & . & . & . & . & . & . & . & . & .
\end{array}\right]\Bigg) , \quad
\begin{array}{c}
k_1 \\ \tilde{e}_1 \\ \tilde{s}_7 \\ k_2 \\ \tilde{e}_2 \\ \tilde{s}_8 \\ \tilde{e}_1 \\ \tilde{e}_2 \\ \tilde{d}_2 \\ \tilde{s}_4 \\ \tilde{d}_1 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ a_0 \\ \tilde{s}_4 \\ \tilde{s}_3 \\ \tilde{s}_1 \\ v_0 \\
\tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \\ \tilde{d}_2 \\ \tilde{s}_4 \\ \tilde{d}_1 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ a_0 \\ \tilde{s}_4 \\ \tilde{s}_3 \\ \tilde{s}_1 \\ v_0 \\
\tilde{s}_1 \\ \tilde{s}_2 \\ \tilde{s}_3 \\ \tilde{d}_2 \\ \tilde{s}_4 \\ \tilde{d}_1 \\ \tilde{s}_1 \\ \tilde{s}_2 \\ a_0 \\ \tilde{s}_4 \\ \tilde{s}_3 \\ \tilde{s}_1 \\ v_0
\end{array}
$$

with dimensions $\tilde{e}_2 = e_2 - k_2$, $\tilde{s}_7 = s_0^{(MCK)} + s_0^{(MC)} - e_1 - e_2$, $\tilde{s}_8 = s_0^{(MK)} - e_1 + k_1 - e_2 + k_2$. We can read off the corresponding ranks to be

$$
\begin{aligned}
\tilde{r}_2 = \operatorname{rank} \tilde{\mathcal{M}}_2 &= 3m - q_0 - 2c_0 - 3v_0 - 3s_0^{(MCK)} - 2s_0^{(MC)} - s_0^{(CK)} - s_0^{(MK)} + k_1 \\
&= 3m - c_0 - c_1 - c_2 - v_0 - v_1 - v_2 - q_0 - q_1 - q_2,
\end{aligned}
$$

$$\begin{aligned}
\operatorname{rank}[\tilde{\mathcal{M}}_2, \tilde{\mathcal{L}}_2] &= 3m - c_0 - 3v_0 - 4s_0^{(MCK)} - 2s_0^{(MC)} - 2s_0^{(CK)} - s_0^{(MK)} + e_1 + e_2 + k_1 + b_1 \\
&= 3m - c_0 - c_1 - c_2 - v_0 - v_1 - v_2, \\
\operatorname{rank}[\tilde{\mathcal{M}}_2, \tilde{\mathcal{L}}_2, \tilde{\mathcal{N}}_2] & \\
&= 3m - 3v_0 - 4s_0^{(MCK)} - 2s_0^{(MC)} - 2s_0^{(CK)} - s_0^{(MK)} + c_1 + b_1 + k_1 + c_2 + e_1 + e_2 \\
&= 3m - v_0 - v_1 - v_2.
\end{aligned}$$

Again, let $\tilde{\Pi} \in \mathbb{R}^{3m\times 3m}$ and $\tilde{\Theta}, \tilde{\Psi}, \tilde{\Sigma} \in \mathbb{R}^{3n\times 3n}$ be the corresponding block matrices that locally transform the inflated triple $(\mathcal{M}_2, \mathcal{L}_2, \mathcal{N}_2)$ to the form (3.26), i.e.,

$$(\tilde{\mathcal{M}}_2, \tilde{\mathcal{L}}_2, \tilde{\mathcal{N}}_2) = (\tilde{\Pi}\mathcal{M}_2\tilde{\Theta}, \tilde{\Pi}\mathcal{L}_2\tilde{\Theta} + 2\tilde{\Pi}\mathcal{M}_2\tilde{\Psi}, \tilde{\Pi}\mathcal{N}_2\tilde{\Theta} + \tilde{\Pi}\mathcal{L}_2\tilde{\Psi} + \tilde{\Pi}\mathcal{M}_2\tilde{\Sigma}),$$

and $\tilde{\Pi}$ and $\tilde{\Theta}$ be partitioned as

$$\tilde{\Pi} := \left[\Pi_1^T, \ldots, \Pi_{44}^T\right]^T, \quad \tilde{\Theta} := [\Theta_1, \ldots, \Theta_{28}],$$

corresponding to the row and column structure of the block matrices in (3.26). Then we have

$$\begin{aligned}
[\Pi_1^T, ..., \Pi_6^T, \Pi_{10}^T, ..., \Pi_{18}^T, \Pi_{20}^T, \Pi_{23}^T, \Pi_{25}^T, \Pi_{27}^T, ..., \Pi_{31}^T, \Pi_{44}^T]^T \mathcal{M}_2 = 0, \\
\mathcal{M}_2[\Theta_{11}, \Theta_{12}, \Theta_{19}, \Theta_{20}, \Theta_{25}, ..., \Theta_{28}] = 0,
\end{aligned}$$

such that the columns of $V_1 := [\Pi_1^T, ..., \Pi_6^T, \Pi_{10}^T, ..., \Pi_{18}^T, \Pi_{20}^T, \Pi_{23}^T, \Pi_{25}^T, \Pi_{27}^T, ..., \Pi_{31}^T, \Pi_{44}^T]$ form a basis of kernel$(\mathcal{M}_2^T)$ and the columns of $V_2 := [\Theta_{11}, \Theta_{12}, \Theta_{19}, \Theta_{20}, \Theta_{25}, ..., \Theta_{28}]$ form a basis of kernel$(\mathcal{M}_2)$. Further decompositions of rows and columns of $\tilde{\mathcal{L}}_2$, where $\Pi_{.,2}$, $\Theta_{.,2}$ denote the parts of the rows and columns of $\tilde{\Pi}$ and $\tilde{\Theta}$, respectively, that corresponds to null-rows or null-columns after the decomposition, yield

$$[\Pi_3^T, \Pi_6^T, \Pi_{10}^T, \Pi_{12}^T, \Pi_{14}^T, ..., \Pi_{18}^T, \Pi_{20,2}^T, \Pi_{23,2}^T, \Pi_{25,2}^T, \Pi_{31}^T, \Pi_{44}^T]^T \mathcal{L}_2 = 0,$$

as well as

$$V_1^T \mathcal{L}_2[\Theta_{12,2}, \Theta_{19}, \Theta_{20}, \Theta_{25}, ..., \Theta_{28}] = 0,$$

such that the columns of

$$V_3 := [\Pi_3^T, \Pi_6^T, \Pi_{10}^T, \Pi_{12}^T, \Pi_{14}^T, ..., \Pi_{18}^T, \Pi_{20,2}^T, \Pi_{23,2}^T, \Pi_{25,2}^T, \Pi_{31}^T, \Pi_{44}^T]$$

form a basis of kernel$(\mathcal{M}_2^T) \cap$ kernel$(\mathcal{L}_2^T)$, and the columns of the matrix V_4 given by $V_4 := [\Theta_{12,2}, \Theta_{19}, \Theta_{20}, \Theta_{25}, ..., \Theta_{28}]$ form a basis of kernel$(\mathcal{M}_2) \cap$ kernel$(V_1^T\mathcal{L}_2)$. Therefore, we have

$$\begin{aligned}
\operatorname{rank}(V_1^T\mathcal{L}_2) &= q_0 + q_1 + q_2, \\
\operatorname{rank}(V_3^T\mathcal{N}_2) &= c_0 + c_1 + c_2,
\end{aligned}$$

and

$$\begin{aligned}
\tilde{a}_2 &= \operatorname{rank}(V_3^T \mathcal{N}_2 V_4) = b_2,\\
\tilde{s}_2^{(MCK)} &= \dim(\operatorname{range}(\mathcal{M}_2^T) \cap \operatorname{range}(\mathcal{L}_2^T V_1) \cap \operatorname{range}(\mathcal{N}_2^T V_3))\\
&= c_0 + c_1 + c_2 - b_1 - d_1 + t_1 - d_2 + t_2 - p_2,\\
\tilde{s}_2^{(CK)} &= \operatorname{rank}(V_3^T \mathcal{N}_2 V_2) - \tilde{a}_2 = b_1 + p_2 - b_2,\\
\tilde{d}_2^{(1)} &= \operatorname{rank}(V_1^T \mathcal{L}_2 V_2) - \tilde{s}_2^{(CK)} = k_2 + b_1 - b_1 - p_2 + b_2,\\
\tilde{s}_2^{(MC)} &= \operatorname{rank}(V_1^T \mathcal{L}_2) - \tilde{s}_2^{(MCK)} - \tilde{s}_2^{(CK)} - \tilde{d}_2^{(1)}\\
&= q_0 + q_1 + q_2 - c_0 - c_1 - c_2 + d_1 - t_1 + d_2 - t_2 - k_2 + p_2,\\
\tilde{s}_2^{(MK)} &= \operatorname{rank}(V_3^T \mathcal{N}_2) - \tilde{a}_2 - \tilde{s}_2^{(MCK)} - \tilde{s}_2^{(CK)} = d_1 - t_1 + d_2 - t_2,\\
\tilde{v}_2 &= 3m - \tilde{r}_2 - 2\tilde{s}_2^{(CK)} - \tilde{d}_2^{(1)} - 2\tilde{s}_2^{(MCK)} - \tilde{s}_2^{(MC)} - \tilde{a}_2 - \tilde{s}_2^{(MK)} = v_0 + v_1 + v_2,\\
\tilde{u}_2 &= 3n - \tilde{r}_2 - \tilde{s}_2^{(CK)} - \tilde{d}_2^{(1)} - \tilde{a}_2 = 3u_0 + 2a_0 + k_0 - k_1 - k_2 - b_1 - b_2.
\end{aligned}$$

□

From the formulas (3.24) for the local characteristic values of the inflated triple $(\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l)$ we can determine the global characteristic values of the original matrix triple (M, C, K).

Corollary 3.20. *Let the strangeness index μ of the matrix triple (M, C, K) be well-defined with $\mu \leq 2$ and let $(\tilde{r}_l, \tilde{d}_l^{(1)}, \tilde{a}_l, \tilde{s}_l^{(MCK)}, \tilde{s}_l^{(CK)}, \tilde{s}_l^{(MC)}, \tilde{s}_l^{(MK)}, \tilde{u}_l, \tilde{v}_l)$, $l = 0, \ldots, \mu$ be the sequence of the local characteristic values of $(\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l)$ for some $t \in \mathbb{I}$. Then for the sequence $(r_i, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(CK)}, s_i^{(MC)}, s_i^{(MK)}, u_i, v_i)$ of the global characteristic values of (M, C, K) it holds that*

$$\begin{aligned}
&c_0 = \tilde{a}_0 + \tilde{s}_0^{(MCK)} + \tilde{s}_0^{(CK)} + \tilde{s}_0^{(MK)},\\
&c_{i+1} = (\tilde{a}_{i+1} - \tilde{a}_i) + (\tilde{s}_{i+1}^{(MCK)} - \tilde{s}_i^{(MCK)}) + (\tilde{s}_{i+1}^{(CK)} - \tilde{s}_i^{(CK)}) + (\tilde{s}_{i+1}^{(MK)} - \tilde{s}_i^{(MK)}),\\
&q_0 = \tilde{d}_0^{(1)} + \tilde{s}_0^{(MCK)} + \tilde{s}_0^{(CK)} + \tilde{s}_0^{(MC)},\\
&q_{i+1} = (\tilde{d}_{i+1}^{(1)} - \tilde{d}_i^{(1)}) + (\tilde{s}_{i+1}^{(MCK)} - \tilde{s}_i^{(MCK)}) + (\tilde{s}_{i+1}^{(CK)} - \tilde{s}_i^{(CK)}) + (\tilde{s}_{i+1}^{(MC)} - \tilde{s}_i^{(MC)}),\\
&v_0 = m - c_0 - q_0 - \tilde{r}_0,\\
&v_{i+1} = m - c_{i+1} - q_{i+1} - (\tilde{r}_{i+1} - \tilde{r}_i),\\
&s_i^{(MCK)} + s_i^{(CK)} + s_i = c_i - \tilde{a}_i,\\
&s_i^{(MCK)} + s_i^{(MC)} + s_{i-1} = q_i - \tilde{d}_i^{(1)} - \tilde{s}_i^{(CK)}.
\end{aligned} \tag{3.27}$$

Proof. The relations follow directly from Theorem 3.19 and from the definitions in Theorem

3.15 since

$$\begin{aligned}
&\tilde{a}_{i+1} - \tilde{a}_i = b_{i+1} - b_i,\\
&\tilde{s}_{i+1}^{(MCK)} - \tilde{s}_i^{(MCK)} + \tilde{s}_{i+1}^{(CK)} - \tilde{s}_i^{(CK)} = c_{i+1} - d_{i+1} + t_{i+1} - b_{i+1} + b_i,\\
&\tilde{s}_{i+1}^{(MK)} - \tilde{s}_i^{(MK)} = d_{i+1} - t_{i+1},\\
&\tilde{d}_{i+1}^{(1)} - \tilde{d}_i^{(1)} = k_{i+1} - k_i + p_i - p_{i+1} + b_{i+1} - b_i,\\
&\tilde{s}_{i+1}^{(MC)} - \tilde{s}_i^{(MC)} = q_{i+1} - c_{i+1} + d_{i+1} - t_{i+1} + p_{i+1} - p_i + k_i - k_{i+1},\\
&m - c_{i+1} - q_{i+1} - (\tilde{r}_{i+1} - \tilde{r}_i) = m - c_{i+1} - q_{i+1} - m + q_{i+1} + c_{i+1} + v_{i+1} = v_{i+1},\\
&c_i - \tilde{a}_i = c_i - c_i + s_i^{(MCK)} + s_i^{(CK)} + s_i^{(MK)} - s_{i-1} = s_i^{(MCK)} + s_i^{(CK)} + s_i,\\
&q_i - \tilde{d}_i^{(1)} - \tilde{s}_i^{(CK)} = q_i - k_i - b_{i-1} = s_i^{(MCK)} + s_i^{(MC)} + s_{i-1}.
\end{aligned}$$

□

The recursive formulas (3.27) enable the determination of the strangeness index μ in a numerically computable way by determining the local characteristic values of the inflated triple $(\mathcal{M}_l, \mathcal{L}_l, \mathcal{N}_l)$ for each time $t \in \mathbb{I}$. The system is strangeness-free if all strangeness parts vanish, and we have $s_i^{(MCK)} = s_i^{(CK)} = s_i^{(MK)} = s_i^{(MC)} = 0$ if and only if the sums $s_i^{(MCK)} + s_i^{(MC)} + s_{i-1}$ and $s_i^{(MCK)} + s_i^{(CK)} + s_i$ vanish since all summands are nonnegative integer values. For the characteristic values of the strangeness-free system we then get

$$\begin{aligned}
a_\mu &= \sum_{i=0}^{\mu} c_i = \operatorname{rank}[\mathcal{M}_\mu, \mathcal{L}_\mu, \mathcal{N}_\mu] - \operatorname{rank}[\mathcal{M}_\mu, \mathcal{L}_\mu],\\
d_\mu^{(1)} &= \sum_{i=0}^{\mu} q_i - \sum_{i=0}^{\mu-1} c_i\\
&= \operatorname{rank}[\mathcal{M}_\mu, \mathcal{L}_\mu] - \tilde{r}_\mu + \operatorname{rank}[\mathcal{M}_{\mu-1}, \mathcal{L}_{\mu-1}] - \operatorname{rank}[\mathcal{M}_{\mu-1}, \mathcal{L}_{\mu-1}, \mathcal{N}_{\mu-1}],\\
v_\mu &= \tilde{v}_\mu - \tilde{v}_{\mu-1},\\
d_\mu^{(2)} &= m - a_\mu - d_\mu^{(1)} - v_\mu.
\end{aligned} \tag{3.28}$$

Next, we want to extract a strangeness-free triple $(\hat{M}, \hat{C}, \hat{K})$ from the inflated system with characteristic values $\hat{r} = d_\mu^{(2)}$, $\hat{d}^{(1)} = d_\mu^{(1)}$, $\hat{a} = a_\mu$, $\hat{u} = u_\mu$, $\hat{v} = v_\mu$ and $\hat{s}^{(MCK)} = \hat{s}^{(CK)} = \hat{s}^{(MK)} = \hat{s}^{(MC)} = 0$ using only local information from $(\mathcal{M}_\mu(t), \mathcal{L}_\mu(t), \mathcal{N}_\mu(t))$.

Theorem 3.21. *Consider a linear second order differential-algebraic system* (3.6) *with well-defined strangeness-index $\mu \leq 2$. Then the inflated triple $(\mathcal{M}_\mu, \mathcal{L}_\mu, \mathcal{N}_\mu)$ associated with (M, C, K) has the following properties:*

1. *For all $t \in \mathbb{I}$ it holds that*

$$\operatorname{rank} \mathcal{M}_\mu(t) = (\mu+1)m - a_\mu - \tilde{v}_\mu - d_\mu^{(1)} - \sum_{i=0}^{\mu-1} c_i,$$

such that there exists a smooth matrix function Z with orthonormal columns and size $((\mu+1)m, a_\mu + \tilde{v}_\mu + d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i)$ satisfying

$$Z^T \mathcal{M}_\mu = 0.$$

2. *For all $t \in \mathbb{I}$ it holds that*

$$\begin{aligned} \operatorname{rank}\left[\mathcal{M}_\mu(t), \mathcal{L}_\mu(t)\right] &= (\mu+1)m - a_\mu - \tilde{v}_\mu, \\ \operatorname{rank}\left[\mathcal{M}_\mu(t), \mathcal{L}_\mu(t), \mathcal{N}_\mu(t)\right] &= (\mu+1)m - \tilde{v}_\mu, \end{aligned}$$

such that without loss of generality Z can be partitioned into $Z = [Z_2, Z_3, Z_4]$, with Z_2 of size $((\mu+1)m, d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i)$, Z_3 of size $((\mu+1)m, a_\mu)$ and Z_4 of size $((\mu+1)m, \tilde{v}_\mu)$ such that

$$Z_3^T \mathcal{L}_\mu = 0, \quad Z_4^T \mathcal{L}_\mu = 0, \quad Z_4^T \mathcal{N}_\mu = 0.$$

3. *For all $t \in \mathbb{I}$ we have*

$$\begin{aligned} \operatorname{rank}\left(Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T\right) &= a_\mu, \\ \operatorname{rank}\left(Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T\right) &= d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i, \end{aligned}$$

such that there exists a smooth matrix function T_3 with orthonormal columns and size $(n, n-a_\mu)$, with $n - a_\mu = d_\mu^{(2)} + d_\mu^{(1)} + u_\mu$ satisfying

$$Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T T_3 = 0.$$

4. *For all $t \in \mathbb{I}$ we have*

$$\operatorname{rank}\left(Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T T_3\right) = d_\mu^{(1)},$$

such that there exists a smooth matrix function Z_1 of size $(d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i, d_\mu^{(1)})$ with orthonormal columns such that

$$\operatorname{rank}\left(Z_1^T Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T\right) = d_\mu^{(1)}.$$

Furthermore, there exists a smooth matrix function T_2 of size $(n - a_\mu, n - a_\mu - d_\mu^{(1)})$ with orthonormal columns, such that

$$Z_1^T Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T T_3 T_2 = 0.$$

5. *For all $t \in \mathbb{I}$ it holds that $\operatorname{rank}(M T_3 T_2) = d_\mu^{(2)}$. This implies the existence of a smooth matrix function Z_0 with orthonormal columns and size $(m, d_\mu^{(2)})$ such that $Z_0^T M$ has constant rank $d_\mu^{(2)}$.*

Proof. By assumption, the strangeness index is well-defined and the ranks of $\mathcal{M}_\mu$, $\mathcal{L}_\mu$ and $\mathcal{N}_\mu$ are constant in $\mathbb{I}$ with

$$\begin{aligned}
&\operatorname{rank}[\mathcal{M}_\mu, \mathcal{L}_\mu, \mathcal{N}]_\mu = (\mu+1)m - \tilde{v}_\mu,\\
&\operatorname{rank}[\mathcal{M}_\mu, \mathcal{L}_\mu] = (\mu+1)m - a_\mu - \tilde{v}_\mu,\\
&\operatorname{rank}\mathcal{M}_\mu = (\mu+1)m - a_\mu - \tilde{v}_\mu - d_\mu^{(1)} - \sum_{i=0}^{\mu-1} c_i,
\end{aligned}$$

due to Theorem 3.19 and Corollary 3.20 (see also the relations (3.28)). Thus, there exists a continuous matrix-valued function Z of size $((\mu+1)m, a_\mu + \tilde{v}_\mu + d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i)$, whose columns form a basis of corange $\mathcal{M}_\mu$, i.e., $Z^T\mathcal{M}_\mu = 0$. Without loss of generality the matrix Z can be partitioned into $Z = [Z_2, Z_3, Z_4]$, with Z_2 of size $((\mu+1)m, d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i)$, Z_3 of size $((\mu+1)m, a_\mu)$ and Z_4 of size $((\mu+1)m, \tilde{v}_\mu)$, such that

$$Z_3^T\mathcal{L}_\mu = 0, \quad Z_4^T\mathcal{L}_\mu = 0, \quad Z_4^T\mathcal{N}_\mu = 0,$$

i.e., the columns of the matrices Z_4 and Z_3 form bases of corange $([\mathcal{M}_\mu, \mathcal{L}_\mu, \mathcal{N}_\mu])$, and corange $([\mathcal{M}_\mu, \mathcal{L}_\mu])$, respectively. First, we note that multiplication of (3.19) for $l = \mu$ by Z_3^T gives

$$Z_3^T\mathcal{N}_\mu z_\mu = Z_3^T g_\mu.$$

The only nontrivial entries in $\mathcal{N}_\mu$ are in the first block column belonging to the original unknown x. Hence, we get purely algebraic equations for x. Lemma 3.8 and (3.24) give that

$$\operatorname{rank}(Z_3^T\mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T) = \tilde{a}_\mu + \tilde{s}_\mu^{(MCK)} + \tilde{s}_\mu^{(CK)} + \tilde{s}_\mu^{(MK)} = a_\mu, \tag{3.29}$$

thus, with Z_3 we obtain the complete set of algebraic equations. Next, we must get $d_\mu^{(1)}$ first order differential equations and $d_\mu^{(2)}$ second order differential equations to complete these algebraic equations to a strangeness-free differential-algebraic system. In a similar way multiplication of (3.19) with the matrix Z_2^T yields

$$Z_2^T\mathcal{L}_\mu\dot{z}_\mu + Z_2^T\mathcal{N}_\mu z_\mu = Z_2^T g_\mu.$$

Again, the only non-zero entries of $\mathcal{L}_\mu$ are in the first block column belonging to the first order derivative $\dot{x}$. Lemma 3.8 and (3.24) give

$$\operatorname{rank}(Z_2^T\mathcal{L}_\mu) = \tilde{d}_\mu^{(1)} + \tilde{s}_\mu^{(CK)} + \tilde{s}_\mu^{(MC)} + \tilde{s}_\mu^{(MCK)} = \sum_{i=0}^{\mu} q_i = d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i.$$

Note that in each step of the iterative procedure the number of equations with second order derivatives of the unknown function is reduced. Therefore, the second order differential equations we are looking for must be already present in the original system.

So far we have shown the first three parts of Theorem 3.21. To show part 4 and 5 of the Theorem we distinguish between systems of strangeness index $\mu = 0, 1$ and 2. Let $(\tilde{M}, \tilde{C}, \tilde{K})$ be a normal form of the triple (M, C, K) according to (3.12) with corresponding inflated triples $(\tilde{\mathcal{M}}_\mu, \tilde{\mathcal{L}}_\mu, \tilde{\mathcal{N}}_\mu)$. Due to Theorem 3.18, there exist matrices Π, Θ, Ψ, and Σ such that

$$\tilde{\mathcal{M}}_\mu = \Pi \mathcal{M}_\mu \Theta, \quad \tilde{\mathcal{L}}_\mu = \Pi \mathcal{L}_\mu \Theta + 2\Pi \mathcal{M}_\mu \Psi, \quad \tilde{\mathcal{N}}_\mu = \Pi \mathcal{N}_\mu \Theta + \Pi \mathcal{L}_\mu \Psi + \Pi \mathcal{M}_\mu \Sigma,$$

according to (3.22). For $\mu = 0$ the triple $(\tilde{\mathcal{M}}_0, \tilde{\mathcal{L}}_0, \tilde{\mathcal{N}}_0)$ is of the form

$$\left(\begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} C & 0 & C & C \\ 0 & I_{d_\mu^{(1)}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} K & K & 0 & K \\ K & K & 0 & K \\ 0 & 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{3.30}$$

Let Π and Θ be partitioned as $\Pi := [\Pi_1^T, \Pi_2^T, \Pi_3^T, \Pi_4^T]^T$ and $\Theta := [\Theta_1, \Theta_2, \Theta_3, \Theta_4]$ according to (3.30). Then, setting

$$Z_2 = \Pi_2^T, \quad Z_3 = \Pi_3^T, \quad Z_4 = \Pi_4^T,$$

yields

$$\begin{aligned} Z_4^T \mathcal{M}_0 = 0, \quad Z_4^T \mathcal{L}_0 = 0, \quad Z_4^T \mathcal{N}_0 = 0, \\ Z_3^T \mathcal{M}_0 = 0, \quad Z_3^T \mathcal{L}_0 = 0, \quad Z_2^T \mathcal{M}_0 = 0, \end{aligned}$$

as well as

$$\begin{aligned} \operatorname{rank}(Z_3^T \mathcal{N}_0) &= \operatorname{rank} \begin{bmatrix} 0 & 0 & I_{a_\mu} & 0 \end{bmatrix} = a_\mu, \\ \operatorname{rank}(Z_2^T \mathcal{L}_0) &= \operatorname{rank} \begin{bmatrix} 0 & I_{d_\mu^{(1)}} & 0 & 0 \end{bmatrix} = d_\mu^{(1)}. \end{aligned}$$

Further, setting $T_3 = [\Theta_1, \Theta_2, \Theta_4]$ we get

$$\operatorname{rank}(Z_2^T \mathcal{L}_0 T_3) = \operatorname{rank} \begin{bmatrix} 0 & I_{d_\mu^{(1)}} & 0 \end{bmatrix} = d_\mu^{(1)},$$

and with $Z_1 = I_{d_\mu^{(1)}}$ and $T_2 = \begin{bmatrix} I_{d_\mu^{(2)}} & 0 \\ 0 & 0 \\ 0 & I_{u_\mu} \end{bmatrix}$ we get

$$Z_1^T Z_2^T \mathcal{L}_0 T_3 T_2 = 0,$$

and

$$\operatorname{rank}(M T_3 T_2) = \operatorname{rank} \begin{bmatrix} I_{d_\mu^{(2)}} & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = d_\mu^{(2)}.$$

Finally, setting $Z_0^T = \begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 & 0 \end{bmatrix}$ yields

$$\operatorname{rank}(Z_0^T M) = \operatorname{rank} \begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 & 0 \end{bmatrix} = d_\mu^{(2)}.$$

In the case $\mu = 1$ the triple $(\tilde{\mathcal{M}}_1, \tilde{\mathcal{L}}_1, \tilde{\mathcal{N}}_1)$ is in the form (3.25) given in the proof of Theorem 3.19. Let Π and Θ be partitioned as

$$\Pi := \begin{bmatrix} \Pi_1^T, \ldots, \Pi_{28}^T \end{bmatrix}^T, \quad \Theta := [\Theta_1, \ldots, \Theta_{18}],$$

corresponding to the structure of block matrices in (3.25). Then, setting

$$\begin{aligned} Z &= [\Pi_1^T, \Pi_2^T, \Pi_3^T, \Pi_7^T, \ldots, \Pi_{15}^T, \Pi_{24}^T, \ldots, \Pi_{28}^T], \\ Z_4 &= [\Pi_{3,2}^T, \Pi_{7,2}^T, \Pi_{9,2}^T, \Pi_{15}^T, \Pi_{28}^T], \\ Z_3 &= [\Pi_{3,1}^T, \Pi_{7,1}^T, \Pi_{9,1}^T, \Pi_{11}^T, \ldots, \Pi_{14}^T], \\ Z_2 &= [\Pi_1^T, \Pi_2^T, \Pi_8^T, \Pi_{10}^T, \Pi_{24}^T, \ldots, \Pi_{27}^T], \end{aligned}$$

where again $\Pi_{\cdot,1}$ and $\Pi_{\cdot,2}$ denote the parts of $\Pi_\cdot$ that after one more block decomposition of the matrices in (3.25) corresponds to the range and nullspace of block rows, respectively (see also the proof of Theorem 3.19), we have

$$\begin{aligned} \operatorname{rank}(Z_3^T \mathcal{N}_1) &= c_0 + c_1 = a_\mu, \\ \operatorname{rank}(Z_2^T \mathcal{L}_1) &= e_1 + d_0^{(1)} + s_0^{(MC)} + c_0 = d_\mu^{(1)} + c_0. \end{aligned}$$

Setting $T_3 = [\Theta_{2,2}, \Theta_{4,2}, \Theta_{5,2}, \Theta_{7,2}, \Theta_{9,2}, \Theta_{10,2}]$ we get

$$\operatorname{rank}(Z_2^T \mathcal{L}_1 \begin{bmatrix} I_n & 0 \end{bmatrix}^T T_3) = d_0^{(1)} + s_0^{(MC)} + e_1 = d_\mu^{(1)},$$

and further choosing Z_1 such that $Z_2 Z_1 = [\Pi_1^T, \Pi_2^T, \Pi_8^T, \Pi_{10}^T]$ we have

$$\operatorname{rank}(Z_1^T Z_2^T \mathcal{L}_1) = d_\mu^{(1)}.$$

If we choose T_2 such that $T_3 T_2 = [\Theta_{5,2}, \Theta_{10,2}]$, then we have

$$Z_1^T Z_2^T \mathcal{L}_1 \begin{bmatrix} I_n & 0 \end{bmatrix}^T T_3 T_2 = 0,$$

and

$$\operatorname{rank}(M T_3 T_2) = d_0^{(2)} = d_\mu^{(2)}.$$

Finally, there exists a smooth matrix function Z_0 of size $(m, d_\mu^{(2)})$ with orthonormal columns such that

$$\operatorname{rank}(Z_0^T M) = d_\mu^{(2)}.$$

For s-index $\mu = 2$ the triple $(\tilde{\mathcal{M}}_2, \tilde{\mathcal{L}}_2, \tilde{\mathcal{N}}_2)$ is in the form (3.26) and Π and Θ are partitioned into

$$\Pi := \left[\Pi_1^T, \ldots, \Pi_{44}^T\right]^T, \quad \Theta := [\Theta_1, \ldots, \Theta_{28}]$$

according to the block structure of (3.26). Then, by setting

$$\begin{aligned}
Z &= [\Pi_1^T, \ldots, \Pi_6^T, \Pi_{10}^T, \ldots, \Pi_{18}^T, \Pi_{20}^T, \Pi_{23}^T, \Pi_{25}^T, \Pi_{27}^T, \ldots, \Pi_{31}^T, \Pi_{44}^T],\\
Z_4 &= [\Pi_{3,2}^T, \Pi_{6,2}^T, \Pi_{10,2}^T, \Pi_{12,2}^T, \Pi_{18}^T, \Pi_{20,3}^T, \Pi_{23,3}^T, \Pi_{25,3}^T, \Pi_{31}^T, \Pi_{44}^T],\\
Z_3 &= [\Pi_{3,1}^T, \Pi_{6,1}^T, \Pi_{10,1}^T, \Pi_{12,1}^T, \Pi_{14}^T, \ldots, \Pi_{17}^T, \Pi_{20,2}^T, \Pi_{23,2}^T, \Pi_{25,2}^T],\\
Z_2 &= [\Pi_1^T, \Pi_2^T, \Pi_4^T, \Pi_5^T, \Pi_{11}^T, \Pi_{13}^T, \Pi_{20,1}^T, \Pi_{23,1}^T, \Pi_{25,1}^T, \Pi_{27}^T, \ldots, \Pi_{30}^T],
\end{aligned}$$

we have

$$\begin{aligned}
&\operatorname{rank}(Z_3^T \mathcal{N}_2) = c_0 + c_1 + c_2 = a_\mu,\\
&\operatorname{rank}(Z_2^T \mathcal{L}_2) = e_1 + e_2 + d_0^{(1)} + s_0^{(MC)} + c_0 + b_1 = d_2^{(1)} + c_0 + c_1 - s_1 = d_\mu^{(1)} + c_0 + c_1.
\end{aligned}$$

Further, setting $T_3 = [\Theta_{2,2}, \Theta_{4,2}, \Theta_{5,2}, \Theta_{6,2}, \Theta_{8,2}, \Theta_{10,2}, \Theta_{11,2}, \Theta_{12,2}]$ yields

$$Z_3^T \mathcal{N}_2 \begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T T_3 = 0$$

and we get

$$\operatorname{rank}(Z_2^T \mathcal{L}_2 \begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T T_3) = d_2^{(1)}.$$

In a similar way as before, choosing Z_1 such that $Z_2 Z_1 = [\Pi_1^T, \Pi_2^T, \Pi_4^T, \Pi_5^T, \Pi_{11}^T, \Pi_{13}^T]$ we have

$$\operatorname{rank}(Z_1^T Z_2^T \mathcal{L}_2) = d_\mu^{(1)},$$

and if we choose T_2 such that $T_3 T_2 = [\Theta_{5,2}, \Theta_{10,2}]$ we have

$$Z_1^T Z_2^T \mathcal{L}_2 \begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T T_3 T_2 = 0,$$

as well as

$$\operatorname{rank}(M T_3 T_2) = d_\mu^{(2)}.$$

Again, there exists a smooth matrix function Z_0 of size $(m, d_\mu^{(2)})$ with orthonormal columns such that

$$\operatorname{rank}(Z_0^T M) = d_\mu^{(2)}.$$

□

From the results of Theorem 3.21 we can construct a triple of matrix-valued functions

$$(\hat{M}, \hat{C}, \hat{K}) = \left(\begin{bmatrix} \hat{M}_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{C}_1 \\ \hat{C}_2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{K}_1 \\ \hat{K}_2 \\ \hat{K}_3 \\ 0 \end{bmatrix} \right), \tag{3.31}$$

with entries

$$\begin{aligned} &\hat{M}_1 = Z_0^T M, \quad \hat{C}_1 = Z_0^T C, \quad \hat{K}_1 = Z_0^T K, \\ &\hat{C}_2 = Z_1^T Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T, \quad \hat{K}_2 = Z_1^T Z_2^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T, \\ &\hat{K}_3 = Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T, \end{aligned}$$

which has the same size as the original triple (M, C, K). We can show that this triple is strangeness-free with the same characteristic values as the strangeness-free system (3.14).

Theorem 3.22. *Let the strangeness index μ of (M, C, K) be well-defined with $\mu \leq 2$ and global characteristic values $(r_i, d_i^{(1)}, a_i, s_i^{(MCK)}, s_i^{(MC)}, s_i^{(MK)}, s_i^{(CK)}, u_i, v_i)$, $i = 0, \ldots, \mu$. Then, the triple $(\hat{M}, \hat{C}, \hat{K})$, constructed as in (3.31), has a well-defined strangeness index $\hat{\mu} = 0$ and the global characteristic values of $(\hat{M}(t), \hat{C}(t), \hat{K}(t))$ are given by*

$$(\hat{r}, \hat{d}^{(1)}, \hat{a}, \hat{s}^{(MCK)}, \hat{s}^{(MC)}, \hat{s}^{(MK)}, \hat{s}^{(CK)}, \hat{v}) = (d_\mu^{(2)}, d_\mu^{(1)}, a_\mu, 0, 0, 0, 0, v_\mu)$$

uniformly for all $t \in \mathbb{I}$.

Proof. In the following, we omit the argument t. By construction the columns of T_3 defined in Theorem 3.21 form a basis of kernel $\hat{K}_3$ and the columns of T_2 form a basis of kernel $(\hat{C}_2 T_3)$. We consider the matrix $T = T_3 T_2$. Because $\hat{M}_1$ has full row rank we can split T without loss of generality into $T = \begin{bmatrix} T_1' & T_4' \end{bmatrix}$ in such a way that $\hat{M}_1 T_1'$ is nonsingular. Choosing T_3' such that $\hat{K}_3 T_3'$ is also nonsingular and T_2' such that $\hat{C}_2 T_2'$ is nonsingular and $\hat{K}_3 T_2' = 0$ we get a nonsingular matrix $\hat{T} = \begin{bmatrix} T_1' & T_2' & T_3' & T_4' \end{bmatrix}$. By multiplication with this matrix from the right we get the following local equivalence

$$\begin{aligned} (\hat{M}, \hat{C}, \hat{K}) &= \left(\begin{bmatrix} \hat{M}_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{C}_1 \\ \hat{C}_2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{K}_1 \\ \hat{K}_2 \\ \hat{K}_3 \\ 0 \end{bmatrix} \right) \\ &\sim \left(\begin{bmatrix} \hat{M}_1 T_1' & \hat{M}_1 T_2' & \hat{M}_1 T_3' & \hat{M}_1 T_4' \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{C}_1 T_1' & \hat{C}_1 T_2' & \hat{C}_1 T_3' & \hat{C}_1 T_4' \\ \hat{C}_2 T_1' & \hat{C}_2 T_2' & \hat{C}_2 T_3' & \hat{C}_2 T_4' \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \right. \\ &\qquad \left. \begin{bmatrix} \hat{K}_1 T_1' & \hat{K}_1 T_2' & \hat{K}_1 T_3' & \hat{K}_1 T_4' \\ \hat{K}_2 T_1' & \hat{K}_2 T_2' & \hat{K}_2 T_3' & \hat{K}_2 T_4' \\ \hat{K}_3 T_1' & \hat{K}_3 T_2' & \hat{K}_3 T_3' & \hat{K}_3 T_4' \\ 0 & 0 & 0 & 0 \end{bmatrix} \right) \end{aligned}$$

$$\sim \left(\begin{bmatrix} \hat{M}_1 T_1' & \hat{M}_1 T_2' & \hat{M}_1 T_3' & \hat{M}_1 T_4' \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{C}_1 T_1' & \hat{C}_1 T_2' & \hat{C}_1 T_3' & \hat{C}_1 T_4' \\ 0 & \hat{C}_2 T_2' & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \hat{K}_1 T_1' & \hat{K}_1 T_2' & \hat{K}_1 T_3' & \hat{K}_1 T_4' \\ \hat{K}_2 T_1' & \hat{K}_2 T_2' & \hat{K}_2 T_3' & \hat{K}_2 T_4' \\ 0 & 0 & \hat{K}_3 T_3' & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right)$$
$$\sim \left(\begin{bmatrix} \hat{M}_1 T_1' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \star & \star & \star & \star \\ 0 & \hat{C}_2 T_2' & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \star & \star & \star & \star \\ \star & \star & \star & \star \\ 0 & 0 & \hat{K}_3 T_3' & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right)$$
$$\sim \left(\begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \star & \star & \star & \star \\ 0 & I_{d_\mu^{(1)}} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \star & \star & \star & \star \\ \star & \star & \star & \star \\ 0 & 0 & I_{a_\mu} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right).$$

From the last triple we obtain $\hat{r} = d_\mu^{(2)}$, $\hat{d}^{(1)} = d_\mu^{(1)}$, $\hat{a} = a_\mu$, $\hat{s}^{(MCK)} = \hat{s}^{(MC)} = \hat{s}^{(MK)} = \hat{s}^{(CK)} = 0$ and $\hat{v} = v_\mu$ from Lemma 3.8. □

Thus, we have derived an index reduction method that allows us to extract a strangeness-free triple from the original triple of matrix-valued functions and its derivatives. The matrix-valued functions Z_0, Z_1, Z_2 and Z_3 as given in Theorem 3.21 can be determined numerically via numerical rank decisions, e.g., using a singular value decomposition or a rank revealing QR decomposition, see e.g. [54]. Setting the inhomogeneities $\hat{f}_1 = Z_0^T f$, $\hat{f}_2 = Z_1^T Z_2^T g_\mu$, $\hat{f}_3 = Z_3^T g_\mu$ and $\hat{f}_4 = 0$ accordingly (assuming that the system is solvable) we obtain a differential-algebraic system

$$\hat{M}(t)\ddot{x} + \hat{C}(t)\dot{x} + \hat{K}(t)x = \hat{f}(t), \tag{3.32}$$

from the inflated differential-algebraic equation (3.19). This system is strangeness-free and has the same size and also the same solution set as the original system (3.6) since only transformations from the left are involved. Setting $\hat{f}_4 = 0$ in (3.32) can be seen as a regularization, since we replace an probably unsolvable problem by a solvable one. Concluding, we give an example to illustrate the index reduction procedure.

Example 3.23. Consider the linear second order system

$$\begin{bmatrix} t & 0 & 0 \\ 0 & 1 & 1 \\ 0 & t & t \end{bmatrix} \ddot{x} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1+t & 1 \end{bmatrix} x = f(t), \tag{3.33}$$

for $t \in [t_0, t_1]$ with $t_0 > 0$. This system has a strangeness index $\mu = 2$ and the characteristic values of (3.33) are $d_\mu^{(2)} = 1$, $d_\mu^{(1)} = 0$, $a_\mu = 2$, $v_\mu = 0$ and $u_\mu = 0$. The matrix triple

corresponding to the extended system (3.19) is given by

$$(\mathcal{M}_2(t),\mathcal{L}_2(t),\mathcal{N}_2(t)) =$$
$$\left(\left[\begin{array}{ccc|ccc|ccc} t & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & t & t & 0 & 0 & 0 & 0 & 0 & 0\\ \hline 2 & 0 & 0 & t & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0\\ 0 & 1 & 1 & 0 & t & t & 0 & 0 & 0\\ \hline 1 & 0 & 0 & 3 & 0 & 0 & t & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 1\\ 0 & 1+t & 1 & 0 & 2 & 2 & 0 & t & t \end{array}\right], \left[\begin{array}{ccc|c|c} 1 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ \hline 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ 0 & 1+t & 1 & & \\ \hline 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ 0 & 2 & 0 & & \end{array}\right], \left[\begin{array}{ccc|c|c} 1 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ 0 & 1+t & 1 & & \\ \hline 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ 0 & 1 & 0 & & \\ \hline 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \\ 0 & 0 & 0 & & \end{array}\right]\right).$$

We have

$$\begin{aligned} \operatorname{rank}[\mathcal{M}_2(t),\mathcal{L}_2(t),\mathcal{N}_2(t)] &= 9 = (\mu+1)m - \tilde{v}_\mu,\\ \operatorname{rank}[\mathcal{M}_2(t),\mathcal{L}_2(t)] &= 7 = (\mu+1)m - a_\mu - \tilde{v}_\mu,\\ \operatorname{rank}[\mathcal{M}_2(t)] &= 6 = (\mu+1)m - d_\mu^{(1)} - c_0 - c_1 - a_\mu - \tilde{v}_\mu, \end{aligned}$$

independent of $t \in \mathbb{I}$ and we can choose

$$\begin{aligned} Z_3^T &= \begin{bmatrix} 0 & -1 & 0 & 0 & -2 & 0 & 0 & -t & 1\\ 0 & -t & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},\\ Z_2^T &= \begin{bmatrix} 0 & -1 & 0 & 0 & -t & 1 & 0 & 0 & 0 \end{bmatrix},\\ T_3 &= \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T, \quad T_2 = 1. \end{aligned}$$

Then, we have

$$\begin{aligned} &\operatorname{rank}\left(Z_3^T\mathcal{N}_2\begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T\right) = \operatorname{rank}\left(\begin{bmatrix} 0 & -1 & 0\\ 0 & 1 & 1 \end{bmatrix}\right) = 2 = a_\mu,\\ &\operatorname{rank}\left(Z_2^T\mathcal{L}_2\begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T\right) = \operatorname{rank}\left(\begin{bmatrix} 0 & 1 & 1 \end{bmatrix}\right) = 1 = d_\mu^{(1)} + c_0 + c_1,\\ &\operatorname{rank}\left(Z_2^T\mathcal{L}_2\begin{bmatrix} I_n & 0 & 0 \end{bmatrix}^T T_3\right) = \operatorname{rank}\left(\begin{bmatrix} 0 \end{bmatrix}\right) = 0 = d_\mu^{(1)},\\ &\operatorname{rank}(MT_3T_2) = \operatorname{rank}\left(\begin{bmatrix} t & 0 & 0 \end{bmatrix}^T\right) = 1 = d_\mu^{(2)}. \end{aligned}$$

Finally, choosing $Z_0^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ we get a strangeness-free system of the form

$$\begin{bmatrix} t & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0 \end{bmatrix}\ddot{x} + \begin{bmatrix} 1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0 \end{bmatrix}\dot{x} + \begin{bmatrix} 1 & 0 & 0\\ 0 & -1 & 0\\ 0 & 1 & 1 \end{bmatrix}x = \begin{bmatrix} f_1\\ -f_2 - 2\dot{f}_2 - t\ddot{f}_2 + \ddot{f}_3\\ -tf_2 + f_3 \end{bmatrix},$$

with the same solution as the original system (3.33).

Remark 3.24. *The derivative array approach presented in this section can also be extended to arbitrary high order differential-algebraic systems of the form* (3.2)*. The theoretical analysis and the condensed forms given in Section 3.1.1 have been generalized to linear k-th order systems in [102, 135]. The inflated system corresponding to* (3.19) *can be obtained in the same way by differentiating the original k-th order system and ordering the derivatives of the coefficient matrices in such a way that only the leading coefficient matrix has a lower triangular block structure and all the other coefficient matrices of the inflated system have entries only in the first block columns. Then, the results of Theorem 3.18 can also be proven for k-th order systems and a Hypothesis similar to Theorem 3.21 can be formulated that allows an index reduction for linear k-th order systems by choosing suitable projections in the same way as for linear second order systems.*

3.2 Nonlinear Second Order Differential-Algebraic Equations

With the results that we have obtained in Section 3.1.2 we can now study general nonlinear second order DAEs of the form

$$F(t, x, \dot{x}, \ddot{x}) = 0, \tag{3.34}$$

where $F \in C(\mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \times \mathbb{D}_{\ddot{x}}, \mathbb{R}^m)$ with open sets $\mathbb{D}_x, \mathbb{D}_{\dot{x}}, \mathbb{D}_{\ddot{x}} \subseteq \mathbb{R}^n$. To analyze nonlinear problems of the form (3.34) we linearize the system along a solution in the same way as it is done for nonlinear first order systems, see e.g. [23], and apply the ideas derived for linear second order DAEs. We consider the linearization of the nonlinear DAE (3.34) in a function space along a solution $\bar{x}$. For $x = \bar{x} + \hat{x}$ we get

$$F(t, \bar{x} + \hat{x}, \dot{\bar{x}} + \dot{\hat{x}}, \ddot{\bar{x}} + \ddot{\hat{x}}) = 0,$$

and from the Taylor expansion it follows that

$$F(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) + F_{;x}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}})\hat{x} + F_{;\dot{x}}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}})\dot{\hat{x}} + F_{;\ddot{x}}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}})\ddot{\hat{x}} + \Phi = 0.$$

Here, Φ sums up all expressions that contain higher order terms, i.e., all terms containing nonlinear expressions of $\hat{x}$ and $\hat{x}^{(i)}$, $i \geq 0$. Neglecting the higher order terms Φ, we get a linearization of (3.34) in the form (3.6) with

$$\begin{aligned} M(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) &= F_{;\ddot{x}}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}), \quad C(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) = F_{;\dot{x}}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}), \\ K(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) &= F_{;x}(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}), \quad f(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) = -F(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}), \end{aligned} \tag{3.35}$$

and x in (3.6) now corresponds to $\hat{x}$. Note, that $f(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}) = 0$ if $\bar{x}$ is a solution of the system (3.34). We can show that the linearization of the nonlinear system (3.34) along a solution and differentiation of the system commute.

Theorem 3.25. *Consider a nonlinear second order DAE* (3.34) *that is sufficiently smooth on a compact interval* $\mathbb{I}$ *and a solution* $\bar{x}$ *of* (3.34)*. Further, suppose that* μ *is well-defined*

for (3.34) *in a neighborhood of* $(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}}, \bar{w})$ *for* $t \in \mathbb{I}$*, where* $\bar{w} = (\bar{x}^{(3)}, \ldots, \bar{x}^{(\mu+2)})$*. Then, the derivatives of the linearized DAE*

$$M(t)\ddot{\hat{x}} + C(t)\dot{\hat{x}} + K(t)\hat{x} = f(t), \tag{3.36}$$

where $\hat{x} = x - \bar{x}$ *and* M, C, K *given as in* (3.35)*, are well-defined and identical to the linearized derivatives of the original DAE* (3.34) *along the solution* $\bar{x}$*.*

Proof. The derivatives of the linearized equation (3.36) are given by

$$\begin{aligned}\frac{d^i}{dt^i}(M\ddot{\hat{x}} + C\dot{\hat{x}} + K\hat{x} - f) = \sum_{j=0}^{i} \left[\tbinom{i}{j} M^{(i-j)} + \tbinom{i}{j+1} C^{(i-j-1)} + \tbinom{i}{j+2} K^{(i-j-2)} \right] \hat{x}^{(j+2)} \\ + [C^{(i)} + iK^{(i-1)}]\dot{\hat{x}} + K^{(i)}\hat{x} - f^{(i)},\end{aligned} \tag{3.37}$$

using the formulas (3.20). On the other hand, the first time derivative of the original equation (3.34) is given by

$$\frac{d}{dt}F(t, x, \dot{x}, \ddot{x}) = F_{;t}(t, x, \dot{x}, \ddot{x}) + F_{;x}(t, x, \dot{x}, \ddot{x})\dot{x} + F_{;\dot{x}}(t, x, \dot{x}, \ddot{x})\ddot{x} + F_{;\ddot{x}}(t, x, \dot{x}, \ddot{x})x^{(3)}.$$

Setting $x = \bar{x} + \hat{x}$ and linearization along $\bar{x}$ yields

$$\begin{aligned}\frac{d}{dt}F(t, x, \dot{x}, \ddot{x}) =& F_{;t} + F_{;tx}\hat{x} + F_{;t\dot{x}}\dot{\hat{x}} + F_{;t\ddot{x}}\ddot{\hat{x}} + F_{;x}(\dot{\bar{x}} + \dot{\hat{x}}) + F_{;xx}\hat{x}(\dot{\bar{x}} + \dot{\hat{x}}) \\ &+ F_{;x\dot{x}}\dot{\hat{x}}(\dot{\bar{x}} + \dot{\hat{x}}) + F_{;x\ddot{x}}\ddot{\hat{x}}(\dot{\bar{x}} + \dot{\hat{x}}) + F_{;\dot{x}}(\ddot{\bar{x}} + \ddot{\hat{x}}) + F_{;\dot{x}x}\hat{x}(\ddot{\bar{x}} + \ddot{\hat{x}}) \\ &+ F_{;\dot{x}\dot{x}}\dot{\hat{x}}(\ddot{\bar{x}} + \ddot{\hat{x}}) + F_{;\dot{x}\ddot{x}}\ddot{\hat{x}}(\ddot{\bar{x}} + \ddot{\hat{x}}) + F_{;\ddot{x}}(\bar{x}^{(3)} + \hat{x}^{(3)}) + F_{;\ddot{x}x}\hat{x}(\bar{x}^{(3)} + \hat{x}^{(3)}) \\ &+ F_{;\ddot{x}\dot{x}}\dot{\hat{x}}(\bar{x}^{(3)} + \hat{x}^{(3)}) + F_{;\ddot{x}\ddot{x}}\ddot{\hat{x}}(\bar{x}^{(3)} + \hat{x}^{(3)}) + \Phi \\ =& F_{;\ddot{x}}\hat{x}^{(3)} + [F_{;\dot{x}} + F_{;\ddot{x}t} + F_{;\ddot{x}x}\dot{\bar{x}} + F_{;\ddot{x}\dot{x}}\ddot{\bar{x}} + F_{;\ddot{x}\ddot{x}}\bar{x}^{(3)}]\ddot{\hat{x}} \\ &+ [F_{;x} + F_{;\dot{x}t} + F_{;\dot{x}x}\dot{\bar{x}} + F_{;\dot{x}\dot{x}}\ddot{\bar{x}} + F_{;\dot{x}\ddot{x}}\bar{x}^{(3)}]\dot{\hat{x}} \\ &+ [F_{;tx} + F_{;xx}\dot{\bar{x}} + F_{;x\dot{x}}\ddot{\bar{x}} + F_{;x\ddot{x}}\bar{x}^{(3)}]\hat{x} \\ &+ [F_{;t} + F_{;x}\dot{\bar{x}} + F_{;\dot{x}}\ddot{\bar{x}} + F_{;\ddot{x}}\bar{x}^{(3)}] + \Phi, \\ =& M\hat{x}^{(3)} + (C + \dot{M})\ddot{\hat{x}} + (K + \dot{C})\dot{\hat{x}} + \dot{K}\hat{x} - \dot{f} + \Phi,\end{aligned}$$

where we have omitted the function arguments, i.e., all terms are functions in $(t, \bar{x}, \dot{\bar{x}}, \ddot{\bar{x}})$, and higher order terms are again summarized in Φ. Neglecting the higher order terms this is just the derivative of the linearized equation (3.36) given in (3.37) for $i = 1$. The proof

for arbitrary $i > 1$ follows by induction as

$$\begin{aligned}
\frac{d^i}{dt^i}F(t,x,\dot{x},\ddot{x}) &= \frac{d}{dt}\left(\frac{d^{i-1}}{dt^{i-1}}F(t,x,\dot{x},\ddot{x})\right)\\
&= \frac{d}{dt}\left\{\sum_{j=0}^{i-1}\left[\tbinom{i-1}{j}M^{(i-1-j)} + \tbinom{i-1}{j+1}C^{(i-j-2)} + \tbinom{i-1}{j+2}K^{(i-j-3)}\right]\hat{x}^{(j+2)}\right.\\
&\quad \left.+[C^{(i-1)} + (i-1)K^{(i-2)}]\dot{\hat{x}} + K^{(i-1)}\hat{x} - f^{(i-1)} + \Phi\right\}\\
&= \sum_{j=0}^{i}\left[\tbinom{i}{j}M^{(i-j)} + \tbinom{i}{j+1}C^{(i-j-1)} + \tbinom{i}{j+2}K^{(i-j-2)}\right]\hat{x}^{(j+2)}\\
&\quad + [C^{(i)} + iK^{(i-1)}]\dot{\hat{x}} + K^{(i)}\hat{x} - f^{(i)} + \Phi.
\end{aligned}$$

□

We can now use the derivative array approach derived in Section 3.1.2 to analyze the nonlinear system (3.34) similar as in Section 2.2.2. First of all, we gather the original equation (3.34) and its derivatives up to order $l \in \mathbb{N}_0$ into an inflated system

$$\mathcal{F}_l(t,x,\dot{x},\ldots,x^{(l+2)}) = 0, \tag{3.38}$$

where the derivative array $\mathcal{F}_l$ of level l has the form

$$\mathcal{F}_l(t,x,\dot{x},\ldots,x^{(l+2)}) = \begin{bmatrix} F(t,x,\dot{x},\ddot{x}) \\ \frac{d}{dt}F(t,x,\dot{x},\ddot{x}) \\ \vdots \\ \frac{d^l}{dt^l}F(t,x,\dot{x},\ddot{x}) \end{bmatrix}.$$

Further, we define the Jacobians

$$\begin{aligned}
\mathcal{M}_l(t,x,\dot{x},\ldots,x^{(l+2)}) &= \mathcal{F}_{l;\ddot{x},\ldots,x^{(l+2)}}(t,x,\dot{x},\ldots,x^{(l+2)}),\\
\mathcal{L}_l(t,x,\dot{x},\ldots,x^{(l+2)}) &= [\mathcal{F}_{l;\dot{x}}(t,x,\dot{x},\ldots,x^{(l+2)}),0,\ldots,0],\\
\mathcal{N}_l(t,x,\dot{x},\ldots,x^{(l+2)}) &= [\mathcal{F}_{l;x}(t,x,\dot{x},\ldots,x^{(l+2)}),0,\ldots,0]
\end{aligned} \tag{3.39}$$

analogous to (3.19). Then, we can formulate the following Hypothesis, as a generalization of Theorem 3.21, that contains the requirements on the nonlinear system such that a reformulation as reduced system with separated differential and algebraic parts is possible.

Hypothesis 3.26. *Consider a nonlinear second order differential-algebraic system* (3.34). *There exist integers* $\mu, r, a_\mu, d_\mu^{(2)}, d_\mu^{(1)}, v_\mu$ *and* u_μ *such that the solution set*

$$\mathbb{L}_\mu = \{(t,x,\dot{x},\ldots,x^{(\mu+2)}) \in \mathbb{R}^{(\mu+3)n+1} | \mathcal{F}_\mu(t,x,\dot{x},\ldots,x^{(\mu+2)}) = 0\} \tag{3.40}$$

associated with (3.34) *is nonempty and such that for every point* $(t_0,x_0,\ldots,x_0^{(\mu+2)}) \in \mathbb{L}_\mu$, *there exists a (sufficiently small) neighborhood in which the following properties hold:*

1. *The set* $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+3)n+1}$ *forms a manifold of dimension* $(\mu+3)n+1-r$.

2. *We have* $\operatorname{rank} \mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+2)}} = r$ *and* $\operatorname{rank} \mathcal{F}_{\mu;\dot{x},\ldots,x^{(\mu+2)}} = r - a_\mu$ *on* $\mathbb{L}_\mu$.

3. *We have* $\operatorname{corank} \mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+2)}} - \operatorname{corank} \mathcal{F}_{\mu-1;x,\dot{x},\ldots,x^{(\mu+1)}} = v_\mu$ *on* $\mathbb{L}_\mu$.

4. *We have*

$$\operatorname{rank} \mathcal{M}_\mu = r - a_\mu - d_\mu^{(1)} - \sum_{i=0}^{\mu-1} c_i,$$

on $\mathbb{L}_\mu$ *such that there exist smooth matrix functions* Z_2, Z_3 *of pointwise maximal rank defined on* $\mathbb{L}_\mu$, *with* Z_2 *of size* $((\mu+1)m, d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i)$, *and* Z_3 *of size* $((\mu+1)m, a_\mu)$ *satisfying*

$$Z_2^T \mathcal{M}_\mu = 0, \quad Z_3^T \mathcal{M}_\mu = 0, \quad Z_3^T \mathcal{L}_\mu = 0,$$

on $\mathbb{L}_\mu$.

5. *We have*

$$\operatorname{rank} (Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T) = a_\mu,$$
$$\operatorname{rank} (Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T) = d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i,$$

on $\mathbb{L}_\mu$ *such that there exists a smooth matrix function* T_3 *with orthonormal columns and size* $(n, n-a_\mu)$, *with* $n - a_\mu = d_\mu^{(2)} + d_\mu^{(1)} + u_\mu$, *satisfying*

$$Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T T_3 = 0.$$

6. *We have*

$$\operatorname{rank} (Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T T_3) = d_\mu^{(1)},$$

on $\mathbb{L}_\mu$ *such that there exists a smooth matrix function* Z_1 *defined on* $\mathbb{L}_\mu$ *of size* $(d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i, d_\mu^{(1)})$ *and with orthonormal columns such that*

$$\operatorname{rank} (Z_1^T Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T) = d_\mu^{(1)}$$

on $\mathbb{L}_\mu$. *Furthermore, there exists a smooth matrix function* T_2 *of size* $(n-a_\mu, n-a_\mu-d_\mu^{(1)})$ *with orthonormal columns such that*

$$Z_1^T Z_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \cdots & 0 \end{bmatrix}^T T_3 T_2 = 0.$$

7. *We have* $\operatorname{rank}(F_{;\ddot{x}}T_3T_2) = d_\mu^{(2)} = m - d_\mu^{(1)} - a_\mu - v_\mu$ *on* $\mathbb{L}_\mu$ *such that there exists a smooth matrix function* Z_0 *defined on* $\mathbb{L}_\mu$ *of size* $(m, d_\mu^{(2)})$ *and pointwise maximal rank satisfying* $\operatorname{rank} Z_0^T F_{;\ddot{x}}T_3T_2 = d_\mu^{(2)}$.

Again, we call the smallest possible μ for which the DAE (3.34) satisfies Hypothesis 3.26 the *strangeness index* of (3.34) and a nonlinear system (3.34) with vanishing strangeness index $\mu = 0$ is called *strangeness-free*. Also in the nonlinear case the Hypothesis 3.26 is invariant under equivalence transformations of the original system (3.34).

Lemma 3.27. *Let* F *as in* (3.34) *satisfy Hypothesis 3.26 with characteristic values* μ, a_μ, $d_\mu^{(2)}$, $d_\mu^{(1)}$, *and* v_μ, *and let* $\tilde{F}$ *be given by*

$$\tilde{F}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) = F(t, x, \dot{x}, \ddot{x}), \tag{3.41}$$

with

$$\begin{aligned} x &= Q(t, \tilde{x}), \\ \dot{x} &= Q_{;t}(t, \tilde{x}) + Q_{;\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}, \\ \ddot{x} &= Q_{;tt}(t, \tilde{x}) + 2Q_{;t\tilde{x}}(t, \tilde{x})\dot{\tilde{x}} + Q_{;\tilde{x}\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}^2 + Q_{;\tilde{x}}(t, \tilde{x})\ddot{\tilde{x}}, \end{aligned} \tag{3.42}$$

with sufficiently smooth function $Q \in C(\mathbb{I} \times \mathbb{R}^n, \mathbb{R}^n)$, *where* $Q(t, \cdot)$ *is bijective for every* $t \in \mathbb{I}$ *and the Jacobian* $Q_{;\tilde{x}}(t, \tilde{x})$ *is nonsingular for every* $(t, \tilde{x}) \in \mathbb{I} \times \mathbb{R}^n$. *Then,* $\tilde{F}$ *satisfies Hypothesis 3.26 with characteristic values* μ, a_μ, $d_\mu^{(2)}$, $d_\mu^{(1)}$, *and* v_μ.

Proof. Let $\mathbb{L}_\mu$ and $\tilde{\mathbb{L}}_\mu$ be the solution sets as defined in Hypothesis 3.26 corresponding to F and $\tilde{F}$, respectively. Since $Q(t, \cdot)$ is bijective and smooth, for every $\tilde{z} = (t, \tilde{x}, \ldots, \tilde{x}^{(\mu+2)}) \in \tilde{\mathbb{L}}_\mu$ we have that $z = (t, x, \ldots, x^{(\mu+2)}) \in \mathbb{L}_\mu$ and vice versa. Setting

$$\begin{aligned} \tilde{M}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= \tilde{F}_{;\ddot{\tilde{x}}}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}), \\ \tilde{C}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= \tilde{F}_{;\dot{\tilde{x}}}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}), \\ \tilde{K}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= \tilde{F}_{;\tilde{x}}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}), \end{aligned}$$

according to (3.35) and using (3.42), we get

$$\begin{aligned} \tilde{M}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= F_{;\ddot{x}}(t, x, \dot{x}, \ddot{x})Q_{;\tilde{x}}(t, \tilde{x}), \\ \tilde{C}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= F_{;\dot{x}}(t, x, \dot{x}, \ddot{x})Q_{;\tilde{x}}(t, \tilde{x}) + F_{;\ddot{x}}(t, x, \dot{x}, \ddot{x})[2Q_{;t\tilde{x}}(t, \tilde{x}) + 2Q_{;\tilde{x}\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}], \\ \tilde{K}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) &= F_{;x}(t, x, \dot{x}, \ddot{x})Q_{;\tilde{x}}(t, \tilde{x}) + F_{;\dot{x}}(t, x, \dot{x}, \ddot{x})[Q_{;t\tilde{x}}(t, \tilde{x}) + Q_{;\tilde{x}\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}] \\ &\quad + F_{;\ddot{x}}(t, x, \dot{x}, \ddot{x})[Q_{;tt\tilde{x}}(t, \tilde{x}) + 2Q_{;t\tilde{x}\tilde{x}}(t, \tilde{x})\dot{\tilde{x}} + Q_{;\tilde{x}\tilde{x}\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}^2 + Q_{;\tilde{x}\tilde{x}}(t, \tilde{x})\ddot{\tilde{x}}]. \end{aligned}$$

Together with (3.35) this can be written as

$$\begin{aligned} &\begin{bmatrix} \tilde{M}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) & \tilde{C}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) & \tilde{K}(t, \tilde{x}, \dot{\tilde{x}}, \ddot{\tilde{x}}) \end{bmatrix} \\ &= \begin{bmatrix} M(t, x, \dot{x}, \ddot{x}) & C(t, x, \dot{x}, \ddot{x}) & K(t, x, \dot{x}, \ddot{x}) \end{bmatrix} \begin{bmatrix} Q_{;\tilde{x}}(t, \tilde{x}) & 2\frac{d}{dt}Q_{;\tilde{x}}(t, \tilde{x}) & \frac{d^2}{dt^2}Q_{;\tilde{x}}(t, \tilde{x}) \\ 0 & Q_{;\tilde{x}}(t, \tilde{x}) & \frac{d}{dt}Q_{;\tilde{x}}(t, \tilde{x}) \\ 0 & 0 & Q_{;\tilde{x}}(t, \tilde{x}) \end{bmatrix}. \end{aligned}$$

This relation has exactly the form of a global equivalence transformation (3.8) replacing Q by $Q_{;\tilde{x}}(t,\tilde{x})$. Since the corresponding inflated triples $(\mathcal{M}_\mu,\mathcal{L}_\mu,\mathcal{N}_\mu)$ and $(\tilde{\mathcal{M}}_\mu,\tilde{\mathcal{L}}_\mu,\tilde{\mathcal{N}}_\mu)$ are built according to (3.20), we get

$$\begin{bmatrix}\tilde{\mathcal{M}}_\mu(\tilde{z}) & \tilde{\mathcal{L}}_\mu(\tilde{z}) & \tilde{\mathcal{N}}_\mu(\tilde{z})\end{bmatrix} = \begin{bmatrix}\mathcal{M}_\mu(z) & \mathcal{L}_\mu(z) & \mathcal{N}_\mu(z)\end{bmatrix} \begin{bmatrix}\Theta_\mu(\tilde{z}) & 2\Psi_\mu(\tilde{z}) & \Sigma_\mu(\tilde{z}) \\ 0 & \Theta_\mu(\tilde{z}) & \Psi_\mu(\tilde{z}) \\ 0 & 0 & \Theta_\mu(\tilde{z})\end{bmatrix}$$

according to (3.22), where we only have to replace Q by $Q_{;\tilde{x}}(t,\tilde{x})$ in (3.21). Then, the invariance of Hypothesis 3.26 follows immediately from Theorem 3.21. □

Lemma 3.28. *Let F as in* (3.34) *satisfy Hypothesis 3.26 with characteristic values μ, a_μ, $d_\mu^{(2)}$, $d_\mu^{(1)}$, and v_μ, and let $\tilde{F}$ be given by*

$$\tilde{F}(t,x,\dot{x},\ddot{x}) = P(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x})), \tag{3.43}$$

with sufficiently smooth function $P \in C(\mathbb{I}\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n,\mathbb{R}^n)$, where $P(t,x,\dot{x},\ddot{x},\cdot)$ is bijective with $P(t,x,\dot{x},\ddot{x},0)=0$, and $P_{;w}(t,x,\dot{x},\ddot{x},\cdot)$ is nonsingular for every $(t,x,\dot{x},\ddot{x}) \in \mathbb{I}\times\mathbb{R}^n\times\mathbb{R}^n\times\mathbb{R}^n$, where $P_{;w}$ denotes the derivative of P with respect to the fifth argument. Then, $\tilde{F}$ satisfies Hypothesis 3.26 with characteristic values μ, a_μ, $d_\mu^{(2)}$, $d_\mu^{(1)}$, and v_μ.

Proof. Let $\mathbb{L}_\mu$ and $\tilde{\mathbb{L}}_\mu$ be the solution sets as defined in Hypothesis 3.26 corresponding to F and $\tilde{F}$, respectively. For every $(t,x,\dot{x},\ddot{x}) \in \mathbb{L}_0$ with $F(t,x,\dot{x},\ddot{x})=0$ it follows that $\tilde{F}(t,x,\dot{x},\ddot{x}) = P(t,x,\dot{x},\ddot{x},0) = 0$. In the same way, for $(t,x,\dot{x},\ddot{x},x^{(3)}) \in \mathbb{L}_1$ with

$$\mathcal{F}_1(t,x,\dot{x},\ddot{x},x^{(3)}) = \begin{bmatrix}F(t,x,\dot{x},\ddot{x}) \\ \frac{d}{dt}F(t,x,\dot{x},\ddot{x})\end{bmatrix} = 0,$$

it follows that

$$\tilde{\mathcal{F}}_1(t,x,\dot{x},\ddot{x},x^{(3)}) = \begin{bmatrix}\tilde{F}(t,x,\dot{x},\ddot{x}) \\ \frac{d}{dt}\tilde{F}(t,x,\dot{x},\ddot{x})\end{bmatrix} = \begin{bmatrix}P(t,x,\dot{x},\ddot{x},0) \\ \frac{d}{dt}P(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x}))\end{bmatrix} = 0.$$

Thus, by induction it follows that $\tilde{\mathbb{L}}_\mu = \mathbb{L}_\mu$. Setting

$$\begin{aligned}\tilde{M}(t,x,\dot{x},\ddot{x}) &= \tilde{F}_{;\ddot{x}}(t,x,\dot{x},\ddot{x}),\\ \tilde{C}(t,x,\dot{x},\ddot{x}) &= \tilde{F}_{;\dot{x}}(t,x,\dot{x},\ddot{x}),\\ \tilde{K}(t,x,\dot{x},\ddot{x}) &= \tilde{F}_{;x}(t,x,\dot{x},\ddot{x}),\end{aligned}$$

it follows that

$$\begin{aligned}\tilde{M}(t,x,\dot{x},\ddot{x}) &= P_{;\ddot{x}}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x})) + P_{;w}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x}))F_{;\ddot{x}}(t,x,\dot{x},\ddot{x}),\\ \tilde{C}(t,x,\dot{x},\ddot{x}) &= P_{;\dot{x}}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x})) + P_{;w}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x}))F_{;\dot{x}}(t,x,\dot{x},\ddot{x}),\\ \tilde{K}(t,x,\dot{x},\ddot{x}) &= P_{;x}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x})) + P_{;w}(t,x,\dot{x},\ddot{x},F(t,x,\dot{x},\ddot{x}))F_{;x}(t,x,\dot{x},\ddot{x}).\end{aligned}$$

If we restrict to the set $\mathbb{L}_\mu$, we have $P_{;\ddot{x}}(t,x,\dot{x},\ddot{x},0) = 0$, $P_{;\dot{x}}(t,x,\dot{x},\ddot{x},0) = 0$, and $P_{;x}(t,x,\dot{x},\ddot{x},0) = 0$, such that we obtain

$$\begin{aligned}&\begin{bmatrix}\tilde{M}(t,x,\dot{x},\ddot{x}) & \tilde{C}(t,x,\dot{x},\ddot{x}) & \tilde{K}(t,x,\dot{x},\ddot{x})\end{bmatrix}\\ &= P_{;w}(t,x,\dot{x},\ddot{x},0)\begin{bmatrix}M(t,x,\dot{x},\ddot{x}) & C(t,x,\dot{x},\ddot{x}) & K(t,x,\dot{x},\ddot{x})\end{bmatrix}\end{aligned}$$

on $\mathbb{L}_\mu$, which is a global equivalence transformation of the form (3.8). The corresponding inflated triples $(\mathcal{M}_\mu,\mathcal{L}_\mu,\mathcal{N}_\mu)$ and $(\tilde{\mathcal{M}}_\mu,\tilde{\mathcal{L}}_\mu,\tilde{\mathcal{N}}_\mu)$ are built according to (3.20) such that we get

$$\begin{bmatrix}\tilde{\mathcal{M}}_\mu(z) & \tilde{\mathcal{L}}_\mu(z) & \tilde{\mathcal{N}}_\mu(z)\end{bmatrix} = \Pi_\mu(z)\begin{bmatrix}\mathcal{M}_\mu(z) & \mathcal{L}_\mu(z) & \mathcal{N}_\mu(z)\end{bmatrix},$$

with $z = (t,x,\dot{x},\ldots,x^{(\mu+2)}) \in \mathbb{L}_\mu$ according to (3.22), where we only must replace P by $P_{;w}(t,x,\dot{x},\ddot{x},0)$ in (3.21). Then, the invariance of Hypothesis 3.26 follows immediately from Theorem 3.21. □

With the help of Hypothesis 3.26 we can now extract a strangeness-free system similar as in Section 2.2.2. Let $z_{\mu,0} = (t_0,x_0,\dot{x}_0,\ldots,x_0^{(\mu+2)}) \in \mathbb{L}_\mu$ be fixed. Then we have $\mathcal{F}_\mu(z_{\mu,0}) = 0$ by definition. By assumption $\mathbb{L}_\mu \subseteq \mathbb{R}^{(\mu+3)n+1}$ is a manifold of dimension $(\mu+3)n+1-r$ that can locally be parameterized by $(\mu+3)n+1-r$ parameters. These parameters can be chosen from $(t,x,\dot{x},\ldots,x^{(\mu+2)})$ in such a way that discarding the associated columns from $\mathcal{F}_{\mu;t,x,\dot{x},\ldots,x^{(\mu+2)}}(z_{\mu,0})$ does not lead to a rank drop. Because of Part 2 of Hypothesis 3.26 $\mathcal{F}_{\mu;x,\dot{x},\ldots,x^{(\mu+2)}}$ already has maximal rank r, such that t can always be chosen as a parameter. Since

$$\begin{aligned}\operatorname{corank}([\mathcal{L}_\mu(z_{\mu,0})\begin{bmatrix}I_n & 0 & \ldots & 0\end{bmatrix}^T\ \mathcal{M}_\mu(z_{\mu,0})]) &= a_\mu,\\ \operatorname{rank}(Z_3(z_{\mu,0})^T\mathcal{N}_\mu(z_{\mu,0})\begin{bmatrix}I_n & 0 & \ldots & 0\end{bmatrix}^T) &= a_\mu,\end{aligned}$$

we can choose $n-a_\mu$ parameters out of x. Without restriction x can be written as (x_1,x_2,x_3,x_4), with $x_1 \in \mathbb{R}^{d_\mu^{(2)}}$, $x_2 \in \mathbb{R}^{d_\mu^{(1)}}$, $x_3 \in \mathbb{R}^{a_\mu}$, $x_4 \in \mathbb{R}^{n-a_\mu-d_\mu^{(2)}-d_\mu^{(1)}}$, and we can choose (x_1,x_2,x_4) as these $n-a_\mu$ parameters. Note, that discarding the columns of $\mathcal{F}_{\mu;x,\ldots,x^{(\mu+2)}}$ belonging to x_1,x_2,x_4 does not lead to a rank drop. In particular, due to the full rank assumption the matrix $Z_3^T\mathcal{F}_{\mu;x_3}$ is then nonsingular. The remaining parameters $q \in \mathbb{R}^{(\mu+2)n+a_\mu-r}$ associated with the columns of $\mathcal{F}_{\mu;t,x,\ldots,x^{(\mu+2)}}(z_{\mu,0})$ that we can remove without having a rank drop, must then be chosen out of $(\dot{x},\ddot{x},\ldots,x^{(\mu+2)})$.
Let $(t_0,x_{1,0},x_{2,0},x_{4,0},q_0)$ be that part of $z_{\mu,0}$ that corresponds to the selected parameters (t,x_1,x_2,x_4,q). Then, the implicit function theorem (Theorem 2.9) implies that there exists a neighborhood $\mathbb{V} \subseteq \mathbb{R}^{(\mu+3)n+1-r}$ of $(t_0,x_{1,0},x_{2,0},x_{4,0},q_0)$, and a neighborhood $\tilde{\mathbb{U}} \subseteq \mathbb{R}^{(\mu+3)n+1}$ of $z_{\mu,0}$ such that

$$\mathbb{U} = \mathbb{L}_\mu \cap \tilde{\mathbb{U}} = \{\theta(t,x_1,x_2,x_4,q) \mid (t,x_1,x_2,x_4,q) \in \mathbb{V}\},$$

where $\theta : \mathbb{V} \to \mathbb{U}$ is a diffeomorphism. Thus, the equation

$$\mathcal{F}_\mu(t,x,\dot{x},\ldots,x^{(\mu+2)}) = 0$$

can be locally solved according to

$$(t, x, \dot{x}, \ldots, x^{(\mu+2)}) = \theta(t, x_1, x_2, x_4, q),$$

for some $(t, x_1, x_2, x_4, q) \in \mathbb{U}$. In particular, there exist locally defined functions G corresponding to x_3, and J corresponding to $(\dot{x}, \ddot{x}, \ldots, x^{(\mu+2)})$ such that

$$\mathcal{F}_\mu(t, x_1, x_2, G(t, x_1, x_2, x_4, q), x_4, J(t, x_1, x_2, x_4, q)) = 0$$

on $\mathbb{V}$. Setting $v = (\dot{x}, \ddot{x}, \ldots, x^{(\mu+2)})$ and with Z_3 as defined by Hypothesis 3.26, it follows that

$$\frac{d}{dq}(Z_3^T \mathcal{F}_\mu) = (Z_{3;x_3}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_3}) G_{;q} + (Z_{3;v}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;v}) J_{;q} = Z_3^T \mathcal{F}_{\mu;x_3} G_{;q} = 0,$$

on $\mathbb{V}$, since $\mathcal{F}_\mu = 0$ and $Z_3^T \mathcal{F}_{\mu;v} = Z_3^T [\mathcal{L}_\mu [I_n \, 0 \ldots 0]^T \;\; \mathcal{M}_\mu] = 0$. By construction the variables in x_3 were selected such that $Z_3^T \mathcal{F}_{\mu;x_3}$ is nonsingular. Hence,

$$G_{;q}(t, x_1, x_2, x_4, q) = 0$$

for all $(t, x_1, x_2, x_4, q) \in \mathbb{V}$, implying the existence of a function R such that

$$x_3 = G(t, x_1, x_2, x_4, q) = G(t, x_1, x_2, x_4, q_0) = R(t, x_1, x_2, x_4),$$

and

$$\mathcal{F}_\mu(t, x_1, x_2, R(t, x_1, x_2, x_4), x_4, J(t, x_1, x_2, x_4, q)) = 0$$

on $\mathbb{V}$. In a similar way we get that

$$\begin{aligned}
\frac{d}{dx_1}(Z_3^T \mathcal{F}_\mu) =& (Z_{3;x_1}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_1}) + (Z_{3;x_3}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_3}) R_{;x_1} + (Z_{3;v}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;v}) J_{;x_1} \\
=& Z_3^T \mathcal{F}_{\mu;x_1} + Z_3^T \mathcal{F}_{\mu;x_3} R_{;x_1} = 0, \\
\frac{d}{dx_2}(Z_3^T \mathcal{F}_\mu) =& (Z_{3;x_2}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_2}) + (Z_{3;x_3}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_3}) R_{;x_2} + (Z_{3;v}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;v}) J_{;x_2} \\
=& Z_3^T \mathcal{F}_{\mu;x_2} + Z_3^T \mathcal{F}_{\mu;x_3} R_{;x_2} = 0, \\
\frac{d}{dx_4}(Z_3^T \mathcal{F}_\mu) =& (Z_{3;x_4}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_4}) + (Z_{3;x_3}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;x_3}) R_{;x_4} + (Z_{3;v}^T \mathcal{F}_\mu + Z_3^T \mathcal{F}_{\mu;v}) J_{;x_4} \\
=& Z_3^T \mathcal{F}_{\mu;x_4} + Z_3^T \mathcal{F}_{\mu;x_3} R_{;x_4} = 0
\end{aligned}$$

on $\mathbb{V}$, again using that $\mathcal{F}_\mu = 0$, and $Z_3^T \mathcal{F}_{\mu;v} = 0$. Thus,

$$Z_3^T \mathcal{N}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T \begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 \\ 0 & I_{d_\mu^{(1)}} & 0 \\ R_{;x_1} & R_{;x_2} & R_{;x_4} \\ 0 & 0 & I_{u_\mu} \end{bmatrix} = 0,$$

with $u_\mu = n - d_\mu^{(2)} - d_\mu^{(1)} - a_\mu$. Following Hypothesis 3.26 we can therefore choose T_3 as

$$T_3(t, x_1, x_2, x_4) = \begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 \\ 0 & I_{d_\mu^{(1)}} & 0 \\ R_{;x_1} & R_{;x_2} & R_{;x_4} \\ 0 & 0 & I_{u_\mu} \end{bmatrix}.$$

Further, since

$$\operatorname{corank}(\mathcal{M}_\mu(z_{\mu,0})) = a_\mu + d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i,$$

$$\operatorname{rank}\left(Z_2(z_{\mu,0})^T \mathcal{L}_\mu(z_{\mu,0}) \begin{bmatrix} I_n & 0 & \dots & 0 \end{bmatrix}^T\right) = d_\mu^{(1)} + \sum_{i=0}^{\mu-1} c_i,$$

we can choose $n - d_\mu^{(1)} - \sum_{i=0}^{\mu-1} c_i$ parameters out of $\dot{x}$. Without restriction we can write x_3 as $(x_{3,0}, x_{3,1}, \dots, x_{3,\mu})$ with $x_{3,i} \in \mathbb{R}^{c_i}$, $i = 0, \dots, \mu$, and choose $(\dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4)$ as these $n - d_\mu^{(1)} - \sum_{i=0}^{\mu-1} c_i$ parameters. Then again, discarding the columns of $\mathcal{F}_{\mu;x,\dot{x},\dots,x^{(\mu+2)}}$ belonging to $\dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4$ does not lead to a rank drop, and due to the full rank assumptions the matrix $Z_2^T \mathcal{F}_{\mu;\dot{x}_2,\dot{x}_{3,0},\dots,\dot{x}_{3,\mu-1}}$ is nonsingular. Now, the remaining parameters $p \in \mathbb{R}^{(\mu+1)n+a_\mu+d_\mu^{(1)}+\sum_{i=0}^{\mu-1} c_i - r}$ must be chosen out of $(\ddot{x}, \dots, x^{(\mu+2)})$.
Let $(t_0, x_{1,0}, x_{2,0}, x_{4,0}, \dot{x}_{1,0}, \dot{x}_{3\mu,0}, \dot{x}_{4,0}, p_0)$ be that part of $z_{\mu,0}$ that corresponds to the selected parameters $(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)$. The implicit function theorem then implies that there exists a neighborhood $\mathbb{V}_2 \subseteq \mathbb{R}^{(\mu+3)n+1-r}$ of $(t_0, x_{1,0}, x_{2,0}, x_{4,0}, \dot{x}_{1,0}, \dot{x}_{3\mu,0}, \dot{x}_{4,0}, p_0)$, and a neighborhood $\tilde{\mathbb{U}}_2 \subseteq \mathbb{R}^{(\mu+3)n+1}$ of $z_{\mu,0}$ such that

$$\mathbb{U}_2 = \mathbb{L}_\mu \cap \tilde{\mathbb{U}}_2 = \{\theta_2(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p) \mid (t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p) \in \mathbb{V}_2\},$$

where $\theta_2 : \mathbb{V}_2 \to \mathbb{U}_2$ is a diffeomorphism. In particular, there exist locally defined functions H corresponding to $(\dot{x}_2, \dot{x}_{3,0}, \dots, \dot{x}_{3,\mu-1})$, and W corresponding to $(\ddot{x}, \dots, x^{(\mu+2)})$ such that

$$\begin{aligned} \mathcal{F}_\mu(t, x_1, x_2, &R(t, x_1, x_2, x_4), x_4, \dot{x}_1, \\ &H(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p), \dot{x}_{3,\mu}, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)) = 0 \end{aligned} \tag{3.44}$$

on $\mathbb{V}_2$. Setting $y = (\ddot{x}, \dots, x^{(\mu+2)})$ and with Z_2 as defined by Hypothesis 3.26 it follows that

$$\begin{aligned} \frac{d}{dp}(Z_2^T \mathcal{F}_\mu) &= (Z_{2;\dot{x}_2,\dot{x}_{3,0},\dots,\dot{x}_{3,\mu-1}}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;\dot{x}_2,\dot{x}_{3,0},\dots,\dot{x}_{3,\mu-1}}) H_{;p} + (Z_{2;y}^T \mathcal{F}_\mu + Z_2^T \mathcal{F}_{\mu;y}) W_{;p} \\ &= Z_2^T \mathcal{F}_{\mu;\dot{x}_2,\dot{x}_{3,0},\dots,\dot{x}_{3,\mu-1}} H_{;p} = 0, \end{aligned}$$

on $\mathbb{V}_2$, since $\mathcal{F}_\mu = 0$ and $Z_2^T \mathcal{F}_{\mu;y} = 0$. By construction $Z_2^T \mathcal{F}_{\mu;\dot{x}_2,\dot{x}_{3,0},\dots,\dot{x}_{3,\mu-1}}$ is nonsingular. Hence,

$$H_{;p}(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p) = 0$$

for all $(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p) \in \mathbb{V}_2$. From Part 6 of Hypothesis 3.26 we have

$$\operatorname{rank}\left(Z_2^T \mathcal{F}_{\mu;\dot{x}} T_3\right) = d_\mu^{(1)},$$

and there exists a matrix function Z_1 such that

$$\operatorname{rank}\left(Z_1^T Z_2^T \mathcal{F}_{\mu;\dot{x}}\right) = d_\mu^{(1)}.$$

Defining $\tilde{Z}_2 = Z_2 Z_1$, then $\tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}$ is nonsingular due to construction and

$$\frac{d}{dp}(\tilde{Z}_2^T \mathcal{F}_\mu) = \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2,\dot{x}_{3,0},\ldots,\dot{x}_{3,\mu-1}} H_{;p} = 0$$

on $\mathbb{V}_2$. This implies the existence of a function S such that

$$\begin{aligned}
\dot{x}_2 &= H(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p) \\
&= H(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p_0) \\
&= S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4),
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{F}_\mu(t, x_1, x_2, R(t, x_1, x_2, x_4), x_4, \dot{x}_1, S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4), R_{;t}(t, x_1, x_2, x_4) \\
&\quad + R_{;x_1}(t, x_1, x_2, x_4)\dot{x}_1 + R_{;x_2}(t, x_1, x_2, x_4) S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4) \\
&\quad + R_{;x_4}(t, x_1, x_2, x_4)\dot{x}_4, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)) = 0
\end{aligned}$$

on $\mathbb{V}_2$, replacing $\dot{x}_3$ by $\frac{d}{dt} R(t, x_1, x_2, x_4)$. Further, we have

$$\begin{aligned}
\tfrac{d}{d\dot{x}_1}(\tilde{Z}_2^T \mathcal{F}_\mu) &= (\tilde{Z}_{2;\dot{x}_1}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_1}) + (\tilde{Z}_{2;\dot{x}_2}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}) S_{;\dot{x}_1} \\
&\quad + (\tilde{Z}_{2;\dot{x}_3}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_3})(R_{;x_1} + R_{;x_2} S_{;\dot{x}_1}) + (\tilde{Z}_{2;y}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;y}) W_{;\dot{x}_1} \\
&= \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_1} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2} S_{;\dot{x}_1} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_3}(R_{;x_1} + R_{;x_2} S_{;\dot{x}_1}) = 0, \\
\tfrac{d}{d\dot{x}_4}(\tilde{Z}_2^T \mathcal{F}_\mu) &= (\tilde{Z}_{2;\dot{x}_4}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_4}) + (\tilde{Z}_{2;\dot{x}_2}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}) S_{;\dot{x}_4} \\
&\quad + (\tilde{Z}_{2;\dot{x}_3}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_3})(R_{;x_4} + R_{;x_2} S_{;\dot{x}_4}) + (\tilde{Z}_{2;y}^T \mathcal{F}_\mu + \tilde{Z}_2^T \mathcal{F}_{\mu;y}) W_{;\dot{x}_4} \\
&= \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_4} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2} S_{;\dot{x}_4} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_3}(R_{;x_4} + R_{;x_2} S_{;\dot{x}_4}) = 0,
\end{aligned}$$

on $\mathbb{V}_2$, again using that $\mathcal{F}_\mu = 0$ and $\tilde{Z}_2^T \mathcal{F}_{\mu;y} = 0$. Thus,

$$\tilde{Z}_2^T \mathcal{L}_\mu \begin{bmatrix} I_n & 0 & \ldots & 0 \end{bmatrix}^T \begin{bmatrix} I_{d_\mu^{(2)}} & 0 \\ S_{;\dot{x}_1} & S_{;\dot{x}_4} \\ R_{;x_1} + R_{;x_2} S_{\dot{x}_1} & R_{;x_4} + R_{;x_2} S_{\dot{x}_4} \\ 0 & I_{u_\mu} \end{bmatrix} = 0,$$

and following Hypothesis 3.26 we can choose T_2 as

$$T_2(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4) = \begin{bmatrix} I_{d_\mu^{(2)}} & 0 \\ S_{;\dot{x}_1} & S_{;\dot{x}_4} \\ 0 & I_{u_\mu} \end{bmatrix}.$$

Finally, Part 7 of Hypothesis 3.26 yields a matrix-valued function Z_0 which only depends on the original variables $(t, x, \dot{x}, \ddot{x})$. Due to the full rank assumption we can choose the neighborhood $\mathbb{V}_2$ so small that we can take a constant Z_0. Altogether, setting

$$\begin{aligned}
\hat{F}_1(t, x, \dot{x}, \ddot{x}) &= Z_0^T F(t, x_1, x_2, x_3, x_4, \dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4, \ddot{x}_1, \ddot{x}_2, \ddot{x}_3, \ddot{x}_4),\\
\hat{F}_2(t, x, \dot{x}) &= \tilde{Z}_2^T \mathcal{F}_\mu(t, x_1, x_2, x_3, x_4, \dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p_0)),\\
\hat{F}_3(t, x) &= Z_3^T \mathcal{F}_\mu(t, x_1, x_2, x_3, x_4, J(t, x_1, x_2, x_4, q_0)),
\end{aligned}$$

we then get the corresponding *reduced differential-algebraic equation*

$$\hat{F}(t, x, \dot{x}, \ddot{x}) = \begin{bmatrix} \hat{F}_1(t, x, \dot{x}, \ddot{x}) \\ \hat{F}_2(t, x, \dot{x}) \\ \hat{F}_3(t, x) \end{bmatrix} = 0. \tag{3.45}$$

We can show that this reduced system is strangeness-free.

Theorem 3.29. *The reduced differential-algebraic system* (3.45) *satisfies Hypothesis 3.26 with characteristic values* $\mu = 0$, $r = a_\mu + d_\mu^{(1)} + d_\mu^{(2)}$, a_μ, $d_\mu^{(2)}$, $d_\mu^{(1)}$, *and* v_μ.

Proof. By construction, we have $\hat{F}(t_0, x_0, \dot{x}_0, \ddot{x}_0) = 0$ for all $(t_0, x_0, \dot{x}_0, \ddot{x}_0)$ part of $z_{\mu,0} \in \mathbb{L}_\mu$, thus the system (3.45) has at least one solution. Moreover, for all $(t, x, \dot{x}, \ddot{x})$ satisfying $\hat{F}(t, x, \dot{x}, \ddot{x}) = 0$ it follows that

$$\begin{aligned}
\hat{F}_{;\ddot{x}}(t, x, \dot{x}, \ddot{x}) &= \begin{bmatrix} Z_0^T F_{;\ddot{x}}(t, x, \dot{x}, \ddot{x}) \\ 0 \\ 0 \end{bmatrix},\\
\hat{F}_{;\dot{x}}(t, x, \dot{x}, \ddot{x}) &= \begin{bmatrix} Z_0^T F_{;\dot{x}}(t, x, \dot{x}, \ddot{x}) \\ \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}}(t, x, \dot{x}, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p_0)) \\ 0 \end{bmatrix},\\
\hat{F}_{;x}(t, x, \dot{x}, \ddot{x}) &= \begin{bmatrix} Z_0^T F_{;x}(t, x, \dot{x}, \ddot{x}) \\ \tilde{Z}_2^T \mathcal{F}_{\mu;x}(t, x, \dot{x}, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p_0)) \\ Z_3^T \mathcal{F}_{\mu;x}(t, x, J(t, x_1, x_2, x_4, q_0)) \end{bmatrix}.
\end{aligned}$$

We have

$$\begin{aligned}
\operatorname{rank} \hat{F}_{;x,\dot{x},\ddot{x}} &= \operatorname{rank} (Z_0^T F_{;\ddot{x}} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}} + Z_3^T \mathcal{F}_{\mu;x}) = d_\mu^{(2)} + d_\mu^{(1)} + a_\mu,\\
\operatorname{rank} \hat{F}_{;\dot{x},\ddot{x}} &= \operatorname{rank} (Z_0^T F_{;\ddot{x}} + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}}) = d_\mu^{(2)} + d_\mu^{(1)},\\
\operatorname{rank} \hat{F}_{;\ddot{x}} &= \operatorname{rank} (Z_0^T F_{;\ddot{x}}) = d_\mu^{(2)},
\end{aligned}$$

which gives Part 2, Part 3 and Part 4 of Hypothesis 3.26. Since

$$\operatorname{rank} \hat{F}_{3;x} = \operatorname{rank} (Z_3^T \mathcal{F}_{\mu;x}) = a_\mu,$$

and since

$$\frac{d}{dx_3}\hat{F}_3(t,x) = Z_3^T \mathcal{F}_{\mu;x_3}(t,x,J(t,x_1,x_2,x_4,q_0))$$

is nonsingular, the implicit function theorem implies that

$$\hat{F}_3(t,x_1,x_2,x_3,x_4) = Z_3^T \mathcal{F}_\mu(t,x_1,x_2,x_3,x_4,J(t,x_1,x_2,x_4,q_0)) = 0$$

holds if and only if $x_3 = R(t,x_1,x_2,x_4)$. Hence,

$$\hat{F}_{;x}(t,x,\dot{x},\ddot{x}) = \begin{bmatrix} Z_0^T F_{;x}(t,x,\dot{x},\ddot{x}) \\ \tilde{Z}_2^T \mathcal{F}_{\mu;x}(t,x,\dot{x},W(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_{3,\mu},\dot{x}_4,p_0)) \\ Z_3^T \mathcal{F}_{\mu;x}(t,x_1,x_2,R(t,x_1,x_2,x_4),x_4,J(t,x_1,x_2,x_4,q_0)) \end{bmatrix},$$

provided that $\hat{F}(t,x,\dot{x},\ddot{x}) = 0$, and the kernel of the third block row is given by the span of the columns of T_3. Further, since

$$\operatorname{rank} \hat{F}_{2;\dot{x}} = \operatorname{rank}(\tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}}) = d_\mu^{(1)},$$

and

$$\hat{F}_{2;\dot{x}_2} = \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}(t,x,\dot{x},W(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_{3,\mu},\dot{x}_4,p_0))$$

is nonsingular, the implicit function theorem implies that

$$\begin{aligned}\hat{F}_2(t,x_1,x_2,&x_3,x_4,\dot{x}_1,\dot{x}_2,\dot{x}_3,\dot{x}_4) = \\ &\tilde{Z}_2^T \mathcal{F}_\mu(t,x_1,x_2,x_3,x_4,\dot{x}_1,\dot{x}_2,\dot{x}_3,\dot{x}_4,W(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_{3,\mu},\dot{x}_4,p_0)) = 0\end{aligned}$$

holds if and only if $\dot{x}_2 = S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)$. Hence,

$$\begin{aligned}&\hat{F}_{;\dot{x}}(t,x,\dot{x},\ddot{x}) = \\ &\begin{bmatrix} Z_0^T F_{;\dot{x}}(t,x,\dot{x},\ddot{x}) \\ \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}}(t,x_1,x_2,x_3,x_4,\dot{x}_1,S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4),\dot{x}_3,\dot{x}_4,W(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_{3,\mu},\dot{x}_4,p_0)) \\ 0 \end{bmatrix},\end{aligned}$$

provided that $\hat{F}(t,x,\dot{x},\ddot{x}) = 0$, and the kernel of the second block row is given by the span of the columns of T_3T_2. Finally, because of

$$\hat{F}_{;\ddot{x}}(t,x,\dot{x},\ddot{x})T_3T_2 = \begin{bmatrix} Z_0^T F_{;\ddot{x}} T_3 T_2 \\ 0 \\ 0 \end{bmatrix},$$

and since $Z_0^T F_{;\ddot{x}} T_3 T_2$ is nonsingular and of rank $d_\mu^{(2)}$, the reduced system also satisfies Part 7 of Hypothesis 3.26. □

Since the condition $\hat{F}_3(t,x)=0$ is locally equivalent via the implicit function theorem to a relation $x_3=R(t,x_1,x_2,x_4)$ we get from (3.45) the system

$$\begin{aligned}\hat{F}_1(t,x_1,x_2,x_3,x_4,\dot{x}_1,\dot{x}_2,\dot{x}_3,\dot{x}_4,\ddot{x}_1,\ddot{x}_2,\ddot{x}_3,\ddot{x}_4)&=0,\\ \hat{F}_2(t,x_1,x_2,x_3,x_4,\dot{x}_1,\dot{x}_2,\dot{x}_3,\dot{x}_4)&=0,\\ x_3-R(t,x_1,x_2,x_4)&=0.\end{aligned}\tag{3.46}$$

Using the last equation of (3.46) and its derivatives, we can replace every occurrence of x_3, $\dot{x}_3$ and $\ddot{x}_3$, to obtain

$$\hat{F}_2(t,x_1,x_2,R(t,x_1,x_2,x_4),x_4,\dot{x}_1,\dot{x}_2,\tfrac{d}{dt}R(t,x_1,x_2,x_4),\dot{x}_4)=0.\tag{3.47}$$

Since

$$\frac{d}{d\dot{x}}\hat{F}_2=\tilde{Z}_2^T\left[\mathcal{F}_{\mu;\dot{x}_1}+\mathcal{F}_{\mu;\dot{x}_3}R_{;x_1},\mathcal{F}_{\mu;\dot{x}_2}+\mathcal{F}_{\mu;\dot{x}_3}R_{;x_2},\mathcal{F}_{\mu;\dot{x}_4}+\mathcal{F}_{\mu;\dot{x}_3}R_{;x_4}\right]=\tilde{Z}_2^T\mathcal{F}_{\mu;\dot{x}}T_3$$

is nonsingular due to Hypothesis 3.26, we can locally solve (3.47) for $\dot{x}_2$, i.e.,

$$\dot{x}_2=S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4).\tag{3.48}$$

With (3.48) we can also eliminate $\dot{x}_2$ and $\ddot{x}_2$ in $\hat{F}_1(t,x,\dot{x},\ddot{x})$, i.e., we get

$$\begin{aligned}&\hat{F}_1(t,x,\dot{x},\ddot{x})=\\ &Z_0^TF(t,x_1,x_2,R(t,x_1,x_2,x_4),x_4,\dot{x}_1,S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4),R_{;t}(t,x_1,x_2,x_4)\\ &+R_{;x_1}(t,x_1,x_2,x_4)\dot{x}_1+R_{;x_2}(t,x_1,x_2,x_4)S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)+R_{;x_4}(t,x_1,x_2,x_4)\dot{x}_4,\\ &\quad\dot{x}_4,\ddot{x}_1,S_{;t}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)+S_{;x_1}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)\dot{x}_1\\ &+S_{;x_2}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)S(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)+S_{;x4}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)\dot{x}_4\\ &+S_{;\dot{x}_1}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)\ddot{x}_1+S_{;\dot{x}_4}(t,x_1,x_2,x_4,\dot{x}_1,\dot{x}_4)\ddot{x}_4,\tfrac{d^2}{dt^2}R(t,x_1,x_2,x_4),\ddot{x}_4)=0,\end{aligned}\tag{3.49}$$

with

$$\begin{aligned}\tfrac{d^2}{dt^2}R(t,x_1,x_2,x_4)&=R_{;tt}+R_{;tx_1}\dot{x}_1+R_{;tx_2}S+R_{;tx_4}\dot{x}_4\\ &\quad+(R_{;tx_1}+R_{;x_1x_1}\dot{x}_1+R_{;x_1x_2}S+R_{;x_1x_4}\dot{x}_4)\dot{x}_1+R_{;x_1}\ddot{x}_1\\ &\quad+(R_{;tx_2}+R_{;x_1x_2}\dot{x}_1+R_{;x_2x_2}S+R_{;x_2x_4}\dot{x}_4)S\\ &\quad+R_{;x_2}(S_{;t}+S_{;x_1}\dot{x}_1+S_{;x_2}S+S_{;x_4}\dot{x}_4+S_{;\dot{x}_1}\ddot{x}_1+S_{;\dot{x}_4}\ddot{x}_4)\\ &\quad+(R_{;tx_4}+R_{;x_1x_4}\dot{x}_1+R_{;x_4x_2}S+R_{;x_4x_4}\dot{x}_4)\dot{x}_4+R_{;x_4}\ddot{x}_4.\end{aligned}$$

Since

$$\begin{aligned}\tfrac{d}{d\ddot{x}}\hat{F}_1&=Z_0^T\left[F_{;\ddot{x}_1}+F_{;\ddot{x}_2}S_{;\dot{x}_1}+F_{;\ddot{x}_3}(R_{;x_1}+R_{;x_2}S_{;\dot{x}_1}),\ F_{;\ddot{x}_4}+F_{;\ddot{x}_2}S_{;\dot{x}_4}+F_{;\ddot{x}_3}(R_{;x_4}+R_{;x_2}S_{;\dot{x}_4})\right]\\ &=Z_0^TF_{;\ddot{x}}T_3T_2\end{aligned}$$

is nonsingular due to Part 7 of Hypothesis 3.26 the system (3.49) can be locally solved for $\ddot{x}_1$, i.e.,

$$\ddot{x}_1 = T(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4, \ddot{x}_4).$$

In this way, we have obtained a decoupled strangeness-free differential-algebraic system of the form

$$\begin{aligned} \ddot{x}_1 &= T(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4, \ddot{x}_4), \\ \dot{x}_2 &= S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4), \\ x_3 &= R(t, x_1, x_2, x_4), \end{aligned} \tag{3.50}$$

with $d_\mu^{(2)}$ second order differential equations, $d_\mu^{(1)}$ first order differential equations, and a_μ algebraic equations. The variables $x_4 \in C^2(\mathbb{I}, \mathbb{R}^{u_\mu})$ of size $u_\mu = n - d_\mu^{(2)} - d_\mu^{(1)} - a_\mu$ can be chosen arbitrarily, i.e., they can be interpreted as controls. Then, the resulting system has locally a unique solution for x_1, x_2 and x_3, provided that consistent initial values are given.

Theorem 3.30. *Let F as in* (3.34) *be sufficiently smooth and satisfy Hypothesis 3.26 with characteristic values $\mu, r, a_\mu, d_\mu^{(1)}, d_\mu^{(2)}$, v_μ, and $u_\mu = n - d_\mu^{(2)} - d_\mu^{(1)} - a_\mu$. Then every sufficiently smooth solution of* (3.34) *also solves the reduced differential-algebraic equations* (3.45) *and* (3.50) *consisting of $d_\mu^{(2)}$ second order differential equations, $d_\mu^{(1)}$ first order differential equations, and a_μ algebraic equations.*

Proof. If x^* is a sufficiently smooth solution of (3.34), then it must also solve the reduced differential-algebraic equations (3.45) and (3.50), since

$$(t, x^*(t), \dot{x}^*(t), \ldots, (\tfrac{d}{dt})^{\mu+2} x^*(t)) \in \mathbb{L}_\mu$$

for every $t \in \mathbb{I}$. If there are no free solution components then (3.50) fixes a unique solution when we prescribe initial values for x_1, $\dot{x}_1$ and x_2, such that locally there can be only one solution of (3.34) satisfying the given initial conditions. Thus, if there are no free solution components and the initial condition is consistent then the solution exists and is unique. □

Theorem 3.31. *Let F as in* (3.34) *be sufficiently smooth and satisfy Hypothesis 3.26 with characteristic values $\mu, a_\mu, d_\mu^{(2)}, d_\mu^{(1)}, v_\mu$ and with characteristic values $(\mu+1)$ (replacing μ), $a_\mu, d_\mu^{(2)}, d_\mu^{(1)}, v_\mu$. Let $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$ be given and let the parameterization p in* (3.44) *for $\mathcal{F}_{\mu+1}$ include $\ddot{x}_4$. Then, for every function $x_4 \in C^2(\mathbb{I}, \mathbb{R}^{n-a_\mu-d_\mu^{(2)}-d_\mu^{(1)}})$ with $x_4(t_0) = x_{4,0}$, $\dot{x}_4(t_0) = \dot{x}_{4,0}$, and $\ddot{x}_4(t_0) = \ddot{x}_{4,0}$, the reduced differential-algebraic equations* (3.45) *and* (3.50) *have unique solutions x_1, x_2 and x_3 satisfying $x_1(t_0) = x_{1,0}$, $\dot{x}_1(t_0) = \dot{x}_{1,0}$ and $x_2(t_0) = x_{2,0}$. Moreover, the so obtained function $x = (x_1, x_2, x_3, x_4)$ locally solves the original problem* (3.34).

Proof. By assumption, there exists a parameterization $(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)$ locally with respect to $z_{\mu+1,0} \in \mathbb{L}_{\mu+1}$, where p is chosen out of $(\ddot{x}, \dots, x^{(\mu+3)})$, with

$$\begin{aligned}\mathcal{F}_{\mu+1}(t, x_1, x_2, &R(t, x_1, x_2, x_4), x_4, \dot{x}_1,\\ &S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4), \dot{x}_3, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)) \equiv 0.\end{aligned}$$

This includes the equation

$$\begin{aligned}\mathcal{F}_{\mu}(t, x_1, x_2, &R(t, x_1, x_2, x_4), x_4, \dot{x}_1,\\ &S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4), \dot{x}_3, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)) \equiv 0,\end{aligned} \tag{3.51}$$

with trivial dependence on $x^{(\mu+3)}$, as well as the equation

$$\begin{aligned}\tfrac{d}{dt}\mathcal{F}_{\mu}(t, x_1, x_2, &R(t, x_1, x_2, x_4), x_4, \dot{x}_1,\\ &S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4), \dot{x}_3, \dot{x}_4, W(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_{3,\mu}, \dot{x}_4, p)) \equiv 0.\end{aligned} \tag{3.52}$$

Equation (3.51) implies that

$$\mathcal{F}_{\mu;t} + \mathcal{F}_{\mu;x_3} R_{;t} + \mathcal{F}_{\mu;\dot{x}_2} S_{;t} + \mathcal{F}_{\mu;y} W_{;t} \equiv 0, \tag{3.53a}$$
$$\mathcal{F}_{\mu;x_1,x_2,x_4} + \mathcal{F}_{\mu;x_3} R_{;x_1,x_2,x_4} + \mathcal{F}_{\mu;\dot{x}_2} S_{;x_1,x_2,x_4} + \mathcal{F}_{\mu;y} W_{;x_1,x_2,x_4} \equiv 0, \tag{3.53b}$$
$$\mathcal{F}_{\mu;\dot{x}_1,\dot{x}_3,\dot{x}_4} + \mathcal{F}_{\mu;\dot{x}_2} S_{;\dot{x}_1,\dot{x}_4} + \mathcal{F}_{\mu;y} W_{;\dot{x}_1,\dot{x}_{3\mu},\dot{x}_4} \equiv 0, \tag{3.53c}$$
$$\mathcal{F}_{\mu;y} W_{;p} \equiv 0, \tag{3.53d}$$

with $y = (\ddot{x}, \dots, x^{(\mu+3)})$, where we have again omitted the function arguments. The relation $\frac{d}{dt}\mathcal{F}_{\mu} = 0$ has the form

$$\mathcal{F}_{\mu;t} + \mathcal{F}_{\mu;x}\dot{x} + \mathcal{F}_{\mu;\dot{x}}\ddot{x} + \mathcal{F}_{\mu;y} \begin{bmatrix} x^{(3)} \\ \vdots \\ x^{(\mu+3)} \end{bmatrix} = 0,$$

such that inserting the parameterization equation (3.52) can be written as

$$\begin{aligned}&\mathcal{F}_{\mu;t} + \mathcal{F}_{\mu;x_1}\dot{x}_1 + \mathcal{F}_{\mu;x_2}\dot{x}_2 + \mathcal{F}_{\mu;x_3}\dot{x}_3 + \mathcal{F}_{\mu;x_4}\dot{x}_4 +\\ &\quad \mathcal{F}_{\mu;\dot{x}_1} W_1 + \mathcal{F}_{\mu;\dot{x}_2} W_2 + \mathcal{F}_{\mu;\dot{x}_3} W_3 + \mathcal{F}_{\mu;\dot{x}_4} W_4 + \mathcal{F}_{\mu;y} W_5 \equiv 0,\end{aligned}$$

where W_i, $i = 1, \dots, 5$, are the parts of W corresponding to $\ddot{x}_1, \ddot{x}_2, \ddot{x}_3, \ddot{x}_4$ and the remaining variables, respectively. Multiplication with Z_3^T (corresponding to Hypothesis 3.26 with characteristic values $\mu, a_\mu, d_\mu^{(2)}, d_\mu^{(1)}, v_\mu$) gives

$$Z_3^T\mathcal{F}_{\mu;t} + Z_3^T\mathcal{F}_{\mu;x_1}\dot{x}_1 + Z_3^T\mathcal{F}_{\mu;x_2}\dot{x}_2 + Z_3^T\mathcal{F}_{\mu;x_3}\dot{x}_3 + Z_3^T\mathcal{F}_{\mu;x_4}\dot{x}_4 \equiv 0.$$

Inserting the relations (3.53) and observing that $Z_3^T\mathcal{F}_{\mu;x_3}$ is nonsingular, we find that

$$Z_3^T\mathcal{F}_{\mu;x_3}(\dot{x}_3 - R_{;t} - R_{;x_1}\dot{x}_1 - R_{;x_2}\dot{x}_2 - R_{;x_4}\dot{x}_4) \equiv 0,$$

or

$$\dot{x}_3 = R_{;t} + R_{;x_1}\dot{x}_1 + R_{;x_2}\dot{x}_2 + R_{;x_4}\dot{x}_4, \tag{3.54}$$

and multiplication with $\tilde{Z}_2^T = Z_1^T Z_2^T$ (corresponding to Hypothesis 3.26 with characteristic values $\mu, a_\mu, d_\mu^{(2)}, d_\mu^{(1)}, v_\mu$) gives

$$\begin{aligned}\tilde{Z}_2^T \mathcal{F}_{\mu;t} + \tilde{Z}_2^T \mathcal{F}_{\mu;x_1}\dot{x}_1 + \tilde{Z}_2^T \mathcal{F}_{\mu;x_2}\dot{x}_2 + \tilde{Z}_2^T \mathcal{F}_{\mu;x_3}\dot{x}_3 + \tilde{Z}_2^T \mathcal{F}_{\mu;x_4}\dot{x}_4 + \\ \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_1}\ddot{x}_1 + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}\ddot{x}_2 + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_3}\ddot{x}_3 + \tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_4}\ddot{x}_4 \equiv 0.\end{aligned}$$

Further, inserting the relations (3.53) and observing that $\tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}$ is nonsingular, we find that

$$\tilde{Z}_2^T \mathcal{F}_{\mu;\dot{x}_2}(\ddot{x}_2 - S_{;t} - S_{;x_1}\dot{x}_1 - S_{;x_2}\dot{x}_2 - S_{;x_4}\dot{x}_4 - S_{;\dot{x}_1}\ddot{x}_1 - S_{;\dot{x}_4}\ddot{x}_4) \equiv 0$$

using (3.54), or

$$\ddot{x}_2 = S_{;t} + S_{;x_1}\dot{x}_1 + S_{;x_2}\dot{x}_2 + S_{;x_4}\dot{x}_4 + S_{;\dot{x}_1}\ddot{x}_1 + S_{;\dot{x}_4}\ddot{x}_4.$$

In summary, the derivative array equation $\mathcal{F}_{\mu+1} = 0$ implies that

$$Z_0^T F(t, x_1, x_2, x_3, x_4, \dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4, \ddot{x}_1, \ddot{x}_2, \ddot{x}_3, \ddot{x}_4) = 0, \tag{3.55a}$$
$$\dot{x}_2 = S(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_2), \tag{3.55b}$$
$$\ddot{x}_2 = S_{;t} + S_{;x_1}\dot{x}_1 + S_{;x_2}\dot{x}_2 + S_{;x_4}\dot{x}_4 + S_{;\dot{x}_1}\ddot{x}_1 + S_{;\dot{x}_4}\ddot{x}_4, \tag{3.55c}$$
$$x_3 = R(t, x_1, x_2, x_4), \tag{3.55d}$$
$$\dot{x}_3 = R_{;t} + R_{;x_1}\dot{x}_1 + R_{;x_2}\dot{x}_2 + R_{;x_4}\dot{x}_4, \tag{3.55e}$$

and elimination of x_3, $\dot{x}_2$, $\dot{x}_3$, $\ddot{x}_2$ and $\ddot{x}_3$ from (3.55a) gives

$$\ddot{x}_1 = T(t, x_1, x_2, x_4, \dot{x}_1, \dot{x}_4, \ddot{x}_4).$$

In particular, this shows that $\ddot{x}_1$, and $\ddot{x}_2$ are not part of the parameterization. Since $\ddot{x}_4$ is part of the parameterization p, the following construction is possible. Let $x_4 = x_4(t)$, $\dot{x}_4 = \dot{x}_4(t)$ and $\ddot{x}_4 = \ddot{x}_4(t)$, and let $p = p(t)$ be arbitrary but consistent to the choice of $\ddot{x}_4$ and to the initial value $z_{\mu+1,0}$. Further, let $x_1 = x_1(t)$, $x_2 = x_2(t)$ and $x_3 = x_3(t)$ be the solutions of the initial value problem

$$\begin{aligned}&Z_0^T F(t, x_1, x_2, x_3, x_4(t), \dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4(t), \ddot{x}_1, \ddot{x}_2, \ddot{x}_3, \ddot{x}_4(t)) = 0,\\ &\dot{x}_2 = S(t, x_1, x_2, x_4(t), \dot{x}_1, \dot{x}_2),\\ &x_3 = R(t, x_1, x_2, x_4(t)),\\ &x_1(t_0) = x_{1,0}, \quad \dot{x}_1(t_0) = \dot{x}_{1,0}, \quad x_2(t_0) = x_{2,0}.\end{aligned}$$

Although $\ddot{x}_1$ and $\ddot{x}_2$ are not part of the parameterization, we automatically get $\ddot{x}_1 = \ddot{x}_1(t)$ and $\ddot{x}_2 = \ddot{x}_2(t)$. Thus, we have

$$\begin{aligned}\mathcal{F}_{\mu+1}(t, &x_1(t), x_2(t), x_3(t), x_4(t), \dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t), \dot{x}_4(t), \ddot{x}_1(t),\\ &\ddot{x}_2(t), \ddot{x}_3(t), \ddot{x}_4(t), W_5(t, x_1(t), x_2(t), x_4(t), \dot{x}_1(t), \dot{x}_{3,\mu}(t), \dot{x}_4(t), p(t))) \equiv 0\end{aligned}$$

for all t in a neighborhood of t_0, or

$$F(t, x_1(t), x_2(t), x_3(t), x_4(t), \dot{x}_1(t), \dot{x}_2(t), \dot{x}_3(t), \dot{x}_4(t), \ddot{x}_1(t), \ddot{x}_2(t), \ddot{x}_3(t), \ddot{x}_4(t)) \equiv 0$$

for the first block of the derivative array. □

Finally, we give an example to illustrate the index reduction procedure in the case of nonlinear second order DAEs.

Example 3.32. We consider the nonlinear second order differential-algebraic system

$$\begin{aligned} \ddot{x}_1 &= x_1 x_2, \\ x_1 \dot{x}_3 &= x_2 - 1, \\ 0 &= x_3 - 1, \end{aligned} \tag{3.56}$$

with $x = [x_1, x_2, x_3]^T \in C(\mathbb{I}, \mathbb{R}^3)$. We have $x_3(t) = 1$ for all $t \in \mathbb{I}$ such that $\dot{x}_3 = 0$ and $x_2(t) = 1$. Thus, system (3.56) consists of two algebraic equations and one second order differential equations. The nonlinear derivative array of level 0 is given by

$$\mathcal{F}_0(x, \dot{x}, \ddot{x}) = \begin{bmatrix} \ddot{x}_1 - x_1 x_2 \\ x_1 \dot{x}_3 - x_2 + 1 \\ x_3 - 1 \end{bmatrix} = 0,$$

and we have

$$\mathcal{F}_{0;x\dot{x}\ddot{x}} = \left[\begin{array}{ccc|ccc|ccc} -x_2 & -x_1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \dot{x}_3 & -1 & 0 & 0 & 0 & x_1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] = 0.$$

The solution set

$$\mathbb{L}_0 = \{(x, \dot{x}, \ddot{x}) \in \mathbb{R}^{3,3} \mid \ddot{x}_1 = x_1 x_2,\ x_1 \dot{x}_3 = x_2 - 1,\ x_3 = 1\},$$

forms a manifold of dimension $3 = 3n - r = 9 - r$ and further we have

$$\begin{aligned} &\operatorname{rank} \mathcal{F}_{0;x\dot{x}\ddot{x}} = 3 = r, \\ &\operatorname{rank} \mathcal{F}_{0;\dot{x}\ddot{x}} = 2, \\ &\operatorname{rank} \mathcal{F}_{0;\ddot{x}} = 1, \end{aligned}$$

such that the Hypothesis 3.26 is not satisfied for $\mu = 0$. Increasing μ by one yields the derivative array of level 1

$$\mathcal{F}_1(x, \dot{x}, \ddot{x}, x^{(3)}) = \begin{bmatrix} \ddot{x}_1 - x_1 x_2 \\ x_1 \dot{x}_3 - x_2 + 1 \\ x_3 - 1 \\ x_1^{(3)} - \dot{x}_1 x_2 - x_1 \dot{x}_2 \\ \dot{x}_1 \dot{x}_3 + x_1 \ddot{x}_3 - \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = 0,$$

with

$$\mathcal{F}_{1;x\dot{x}\ddot{x}x^{(3)}} = \left[\begin{array}{ccc|ccc|ccc|ccc} -x_2 & -x_1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ \dot{x}_3 & -1 & 0 & 0 & 0 & x_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline -\dot{x}_2 & -\dot{x}_1 & 0 & -x_2 & -x_1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ \ddot{x}_3 & 0 & 0 & \dot{x}_3 & -1 & \dot{x}_1 & 0 & 0 & x_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] = 0.$$

The solution set $\mathbb{L}_1 = \{(x, \dot{x}, \ddot{x}, x^{(3)}) \in \mathbb{R}^{3,4} \mid \ddot{x}_1 = x_1 x_2,\, x_1 \dot{x}_3 = x_2 - 1,\, x_3 = 1,\, x_1^{(3)} = \dot{x}_1 x_2 + x_1 \dot{x}_2,\, \dot{x}_1 \dot{x}_3 + x_1 \ddot{x}_3 = \dot{x}_2,\, \dot{x}_3 = 0\}$ of the derivative array $\mathcal{F}_1$ forms a manifold of dimension $6 = 12 - r$ and we have

$$\begin{aligned} &\operatorname{rank} \mathcal{F}_{1;x\dot{x}\ddot{x}x^{(3)}} = 6 = r, \\ &\operatorname{rank} \mathcal{F}_{1;\dot{x}\ddot{x}x^{(3)}} = 4 = r - a_\mu, \\ &\operatorname{rank} \mathcal{F}_{1;\ddot{x}x^{(3)}} = 3 = r - a_\mu - d_\mu^{(1)} - c_0. \end{aligned}$$

Furthermore, we have $\operatorname{corank} \mathcal{F}_{1;x\dot{x}\ddot{x}x^{(3)}} - \operatorname{corank} \mathcal{F}_{0;x\dot{x}\ddot{x}} = 0 - 0 = 0 = v_\mu$, and choosing $Z_2^T = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$ and $Z_3^T = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & -x_1 \end{bmatrix}$ yields

$$\begin{aligned} &\operatorname{rank} (Z_3^T \mathcal{N}_1 [I_n\ 0]^T) = \operatorname{rank} \left(\begin{bmatrix} 0 & 0 & 1 \\ \dot{x}_3 & -1 & 0 \end{bmatrix} \right) = \operatorname{rank} \left(\begin{bmatrix} 0 & 0 & 1 \\ 0 & -1 & 0 \end{bmatrix} \right) = 2, \\ &\operatorname{rank} (Z_2^T \mathcal{L}_1 [I_n\ 0]^T) = \operatorname{rank} \left(\begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \right) = 1, \end{aligned}$$

on $\mathbb{L}_1$. We can chose $T_3 = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^T$ such that $Z_3^T \mathcal{N}_1 [I_n\ 0]^T T_3 = 0$ and

$$\operatorname{rank} (Z_2^T \mathcal{L}_1 [I_n\ 0]^T T_3) = 0 = d_\mu^{(1)},$$

yielding a matrix $Z_1 = \emptyset_{1,0}$ of size $(1, 0)$ and a matrix $T_2 = 1$ of size $(1, 1)$. Finally, we have

$$\operatorname{rank} (F_{;\ddot{x}} T_3 T_2) = \operatorname{rank} \left(\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \right) = 1 = d_\mu^{(2)},$$

and with $Z_0^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ we have $\operatorname{rank} (Z_0^T F_{;\ddot{x}} T_3 T_2) = \operatorname{rank} ([1]) = 1$. Thus, Hypothesis 3.26 is satisfied for $\mu = 1$, i.e., system (3.56) has strangeness index 1, and we can obtain a reduced second order differentail-algebraic system by

$$\hat{F}(x, \dot{x}, \ddot{x}) = \begin{bmatrix} Z_0^T F(x, \dot{x}, \ddot{x}) \\ Z_3^T \mathcal{F}_1(x, \dot{x}, \ddot{x}, x^{(3)}) \end{bmatrix} = \begin{bmatrix} \ddot{x}_1 - x_1 x_2 \\ x_3 - 1 \\ -x_2 + 1 \end{bmatrix} = 0.$$

3.3 Trimmed First Order Formulation

In the numerical solution of higher order differential-algebraic systems either the direct numerical solution of the higher order system by appropriate numerical methods, or a suitable transformation into a first order system is required. Since most of the numerical methods suited for the solution of DAEs are constructed for first order systems and these methods are well-studied, in general a transformation into a first order system is desired. Furthermore, for a robust solution the numerical methods require differential-algebraic systems of low index, such that on the one hand an index reduction, and on the other hand an order reduction is required. In this section we consider linear second order differential-algebraic equations with variable coefficients of the form (3.6) and discuss different ways to obtain a first order formulation for the second order system that can be solved numerically. Finally, a so-called *trimmed first order formulation* is derived that allows to construct a strangeness-free first order system for linear second order DAEs of arbitrary high index in a numerical feasible way.

The standard way to obtain a strangeness-free first order formulation for a second order system (3.6) is to introduce new variables $v = \dot{x}$ for the derivatives to transform the second order system (3.6) into a first order system, and then apply the usual index reduction procedures to the first order system. The corresponding first order formulation (also called *companion form*) is given by

$$\begin{bmatrix} M(t) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} -C(t) & -K(t) \\ I & 0 \end{bmatrix} \begin{bmatrix} v \\ x \end{bmatrix} + \begin{bmatrix} f(t) \\ 0 \end{bmatrix}. \tag{3.57}$$

This is a linear first order DAE of the form (2.5) with matrices

$$E(t) = \begin{bmatrix} M(t) & 0 \\ 0 & I \end{bmatrix}, \quad A(t) = \begin{bmatrix} -C(t) & -K(t) \\ I & 0 \end{bmatrix}$$

of size $(m+n, 2n)$, and right-hand side $b(t) = \begin{bmatrix} f(t)^T & 0 \end{bmatrix}^T$, with unknowns $y = \begin{bmatrix} v^T & x^T \end{bmatrix}^T$. To obtain a strangeness-free system in a numerical feasible way the derivative arrays defined in (2.22) can be used. Then, by determining suitable projections as defined in Theorem 2.41 an equivalent strangeness-free system

$$\hat{E}(t)\dot{y} = \hat{A}(t)y + \hat{b}(t),$$

of the form (2.24) can be constructed. Many difficulties may arise using this approach as have been described in Chapter 1. The most critical is a possible increase in the index of the DAE that yields higher smoothness requirements for the inhomogeneity $f(t)$ and even can cause the loss of solvability of the system, see e.g. [102]. For k-th order linear DAEs it has been shown in [102] that if μ is the strangeness index of the tuple of matrix-valued functions associated with the k-th order DAE system, then the maximal possible increase in the strangeness index $\tilde{\mu}$ of the first order system obtained by the classical order reduction

procedure is $\tilde{\mu} \leq \mu + k - 1$. Further, the resulting first order system is much larger than the original system, and structures in the coefficient matrices are destroyed. Another drawback of this approach is that the condition number of the first order system (3.57) may increase compared to the original second order system, see e.g. [142].

Example 3.33. We consider again the linear second order system (3.33) given in Example 3.23 of strangeness index $\mu = 2$. The companion form (3.57) of the second order system (3.33) is given by

$$\left[\begin{array}{ccc|ccc} t & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & t & t & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \begin{bmatrix} \dot{v}_1 \\ \dot{v}_2 \\ \dot{v}_3 \\ \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \left[\begin{array}{ccc|ccc} -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & -1-t & -1 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \\ f_3 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \tag{3.58}$$

for $t \in [t_0, t_1]$ with $t_0 > 0$. In comparison to the solution of (3.33) this system has the additional solution components

$$\begin{bmatrix} v_2 \\ v_3 \end{bmatrix} = \begin{bmatrix} \dot{f}_2 - f_3^{(3)} + t f_2^{(3)} + 3\ddot{f}_2 \\ \dot{f}_3 - (t+1)\dot{f}_2 - f_2 + f_3^{(3)} - t f_2^{(3)} - 3\ddot{f}_2 \end{bmatrix},$$

i.e., the third derivative of the inhomogeneity f is required, and thus system (3.58) is of strangeness index $\tilde{\mu} = 3$.

To overcome these problems in the classical order reduction, we can first compute the strangeness-free condensed form (3.14) for the second order system and transform the strangeness-free form (3.14) to a first order system

$$\tilde{E}(t)\dot{\tilde{y}} = \tilde{A}(t)\tilde{y} + \tilde{b}(t)$$

of the form (3.15) afterwards as has been proposed in [102], see also Corollary 3.12. In this way, we can obtain a first order formulation that is strangeness-free without further smoothness requirements or increasing the index.

Example 3.34. For the linear second order DAE (3.33) given in Example 3.23 transformation into global condensed form (3.12) yields

$$\begin{bmatrix} t & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{\tilde{x}}_1 \\ \ddot{\tilde{x}}_2 \\ \ddot{\tilde{x}}_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 \\ -t f_2 + f_3 \end{bmatrix},$$

with $\begin{bmatrix} \tilde{x}_1 & \tilde{x}_2 & \tilde{x}_3 \end{bmatrix}^T = \begin{bmatrix} x_1 & x_2 + x_3 & x_3 \end{bmatrix}^T$. After applying two index reduction steps we obtain a system in the strangeness-free form (3.14) given by

$$\begin{bmatrix} t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{\tilde{x}}_1 \\ \ddot{\tilde{x}}_2 \\ \ddot{\tilde{x}}_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 + t\ddot{f}_2 - 2\dot{f}_2 - \ddot{f}_3 \\ -t f_2 + f_3 \end{bmatrix}.$$

Following Corollary 3.12 the corresponding first order formulation is given by

$$\left[\begin{array}{ccc|c} 1 & 0 & 0 & t \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \\ \dot{\tilde{x}}_4 \end{bmatrix} + \left[\begin{array}{ccc|c} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 0 & -1 \end{array}\right] \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{bmatrix} = \begin{bmatrix} f_1 \\ f_2 + t\ddot{f}_2 - 2\dot{f}_2 - \ddot{f}_3 \\ -tf_2 + f_3 \\ 0 \end{bmatrix},$$

which is strangeness-free and of minimal increased size.

The drawback of this approach is that there is no computational feasible method to compute the condensed form (3.14), except if the structure can be used, since the derivatives of computed transformation matrices are used during the transformations. Further, the strangeness-free system (3.14), and consequently also the first order formulation (3.15), do not have the same solution x as the original second order system (3.6), but a transformed solution $\tilde{x} = Q^{-1}x$. Nevertheless, Corollary 3.12 suggests that first index reduction and then order reduction should be used for a proper treatment of second order systems.
Thus, in the following we will use the index reduction based on derivative arrays derived in Section 3.1.2 to obtain a strangeness-free second order system which can then be used to obtain a trimmed first order formulation in a numerical feasible way. Using the inflated system (3.19) associated with the second order system (3.6) we can determine projections Z_4, Z_3, Z_2, Z_1, and Z_0 as defined in Theorem 3.21 to obtain locally a strangeness-free second order system

$$\hat{M}(t)\ddot{x} + \hat{C}(t)\dot{x} + \hat{K}(t)x = \hat{f}(t), \tag{3.59}$$

with matrix triple of the form

$$(\hat{M}, \hat{C}, \hat{K}) = \left(\begin{bmatrix} \hat{M}_1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{C}_1 \\ \hat{C}_2 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \hat{K}_1 \\ \hat{K}_2 \\ \hat{K}_3 \\ 0 \end{bmatrix} \right).$$

To find a suitable first order formulation, we first have to identify the second order differential variables. As the matrices $\hat{M}_1$, $\hat{C}_2$ and $\hat{K}_3$ have full row rank due to construction (see Theorem 3.21) there exists a pointwise orthogonal matrix-valued function $Q \in C(\mathbb{I}, \mathbb{R}^{n\times n})$ that is sufficiently smooth, such that

$$\begin{bmatrix} \hat{M}_1 \\ \hat{C}_2 \\ \hat{K}_3 \end{bmatrix} Q = \begin{bmatrix} M_{11} & 0 & 0 & 0 \\ C_{21} & C_{22} & 0 & 0 \\ K_{31} & K_{32} & K_{33} & 0 \end{bmatrix}, \tag{3.60}$$

where the matrix-valued functions M_{11} of size $d_\mu^{(2)} \times d_\mu^{(2)}$, C_{22} of size $d_\mu^{(1)} \times d_\mu^{(1)}$ and K_{33} of size $a_\mu \times a_\mu$ are pointwise nonsingular. With the corresponding basis transformation

$$x = Q\hat{x}, \quad \dot{x} = Q\dot{\hat{x}} + \dot{Q}\hat{x}, \quad \ddot{x} = Q\ddot{\hat{x}} + 2\dot{Q}\dot{\hat{x}} + \ddot{Q}\hat{x},$$

we get the equivalent system

$$\begin{aligned}\underbrace{\begin{bmatrix} M_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\hat{M}Q} \begin{bmatrix} \ddot{\hat{x}}_1 \\ \ddot{\hat{x}}_2 \\ \ddot{\hat{x}}_3 \\ \ddot{\hat{x}}_4 \end{bmatrix} + \underbrace{\begin{bmatrix} C_{11} & C_{12} & C_{13} & C_{14} \\ C_{21} & C_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\hat{C}Q+2\hat{M}\dot{Q}} \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{x}}_4 \end{bmatrix} \\ + \underbrace{\begin{bmatrix} K_{11} & K_{12} & K_{13} & K_{14} \\ K_{21} & K_{22} & K_{23} & K_{24} \\ K_{31} & K_{32} & K_{33} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}}_{\hat{K}Q+\hat{C}\dot{Q}+\hat{M}\ddot{Q}} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} = \underbrace{\begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \\ \hat{f}_4 \end{bmatrix}}_{\hat{f}},\end{aligned} \tag{3.61}$$

where the second order differential variables $\hat{x}_1$ are explicitly specified. By introducing the new variable $\hat{v} = \dot{\hat{x}}_1$ we can transform the system (3.61) into first order form

$$\begin{bmatrix} M_{11} & C_{11} & C_{12} & C_{13} & C_{14} \\ 0 & C_{21} & C_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{v}} \\ \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{x}}_4 \end{bmatrix} + \begin{bmatrix} 0 & K_{11} & K_{12} & K_{13} & K_{14} \\ 0 & K_{21} & K_{22} & K_{23} & K_{24} \\ 0 & K_{31} & K_{32} & K_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -I & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{v} \\ \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \\ \hat{f}_4 \\ 0 \end{bmatrix}. \tag{3.62}$$

Here, we have

$$\begin{aligned}&\hat{v} = \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix} \dot{\hat{x}} = \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix} (\dot{Q}^T x + Q^T \dot{x}) = \dot{Q}_1^T x + Q_1^T \dot{x} = \tfrac{d}{dt}(Q_1^T x), \\ &\dot{\hat{v}} = \tfrac{d^2}{dt^2}(Q_1^T x),\end{aligned}$$

with $Q_1 = Q \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix}^T$. Therefore, system (3.62) is equivalent to

$$\begin{bmatrix} \hat{M}Q_1 & \hat{C}Q + 2\hat{M}\dot{Q} \\ 0 & J \end{bmatrix} \begin{bmatrix} \frac{d^2}{dt^2}(Q_1^T x) \\ \frac{d}{dt}(Q^T x) \end{bmatrix} + \begin{bmatrix} 0 & \hat{K}Q + \hat{C}\dot{Q} + \hat{M}\ddot{Q} \\ -I & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{dt}(Q_1^T x) \\ Q^T x \end{bmatrix} = \begin{bmatrix} \hat{f} \\ 0 \end{bmatrix}, \tag{3.63}$$

with $J = \begin{bmatrix} I & 0 & 0 & 0 \end{bmatrix}$. Now, introducing another variable $v = Q_1^T \dot{x} = \frac{d}{dt}(Q_1^T x) - \dot{Q}_1^T x$, the first equation of (3.63) becomes

$$\hat{M}Q_1 \dot{v} + (\hat{C} + \hat{M}(Q_1 \dot{Q}_1^T + 2\dot{Q}Q^T))\dot{x} + (\hat{K} + \hat{M}(Q_1 \ddot{Q}_1^T + 2\dot{Q}\dot{Q}^T + \ddot{Q}Q^T))x = \hat{f},$$

where we have used that $\dot{Q}Q^T + Q\dot{Q}^T = 0$, as Q is orthogonal. Thus, we get a first order system in the original variable x and in v of the form

$$\begin{bmatrix} \hat{M}Q_1 & \tilde{C} \\ 0 & Q_1^T \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{x} \end{bmatrix} + \begin{bmatrix} 0 & \tilde{K} \\ -I & 0 \end{bmatrix} \begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} \hat{f} \\ 0 \end{bmatrix},$$

with

$$\tilde{C} = \hat{C} + \hat{M}(Q_1\dot{Q}_1^T + 2\dot{Q}Q^T) = \hat{C} + \hat{M}(Q_1\dot{Q}_1^T - 2Q\dot{Q}^T),$$
$$\tilde{K} = \hat{K} + \hat{M}(Q_1\ddot{Q}_1^T + 2\dot{Q}\dot{Q}^T + \ddot{Q}Q^T) = \hat{K} + \hat{M}(Q_1\ddot{Q}_1^T - Q\ddot{Q}^T),$$

using that $\ddot{Q}Q^T + 2\dot{Q}\dot{Q}^T + Q\ddot{Q}^T = 0$. In addition, it holds that

$$Q_1\dot{Q}_1^T - 2Q\dot{Q}^T = QJ^TJ\dot{Q}^T - 2Q\dot{Q}^T = Q\begin{bmatrix} -I & 0 & 0 & 0 \\ 0 & -2I & 0 & 0 \\ 0 & 0 & -2I & 0 \\ 0 & 0 & 0 & -2I \end{bmatrix}\dot{Q}^T,$$
$$Q_1\ddot{Q}_1^T - Q\ddot{Q}^T = QJ^TJ\ddot{Q}^T - Q\ddot{Q}^T = Q\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & -I & 0 & 0 \\ 0 & 0 & -I & 0 \\ 0 & 0 & 0 & -I \end{bmatrix}\ddot{Q}^T,$$

such that we get

$$\tilde{C} = \hat{C} - \hat{M}Q\begin{bmatrix} I & 0 & 0 & 0 \\ 0 & 2I & 0 & 0 \\ 0 & 0 & 2I & 0 \\ 0 & 0 & 0 & 2I \end{bmatrix}\dot{Q}^T = \hat{C} - \begin{bmatrix} M_{11} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\dot{Q}^T = \hat{C} - \hat{M}Q\dot{Q}^T,$$
$$\tilde{K} = \hat{K} - \hat{M}Q\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{bmatrix}\ddot{Q}^T = \hat{K}.$$

Altogether, we have derived a first order formulation in the original variable x and v, only using the coefficient matrices of the strangeness-free formulation (3.59) and the orthogonal transformation matrix Q. Due to construction this first order system denoted by

$$\check{E}(t)\dot{y} = \check{A}(t)y + \check{b}(t)$$

is strangeness-free.

Lemma 3.35. *Consider a linear second order differential-algebraic system* (3.59) *that is strangeness-free, with matrix-valued functions* $\hat{M}, \hat{C}, \hat{K} \in C(\mathbb{I}, \mathbb{R}^{m,n})$*, and right-hand side* $\hat{f} \in C(\mathbb{I}, \mathbb{R}^m)$*. Further, let* $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ *be an orthogonal matrix-valued function that decomposes* $\hat{M}, \hat{C}, \hat{K}$ *as in* (3.60)*. Then, the* trimmed first order formulation

$$\begin{bmatrix} \hat{M}Q_1 & \hat{C} + \hat{M}\dot{Q}Q^T \\ 0 & Q_1^T \end{bmatrix}\begin{bmatrix} \dot{v} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} 0 & -\hat{K} \\ I & 0 \end{bmatrix}\begin{bmatrix} v \\ x \end{bmatrix} + \begin{bmatrix} \hat{f} \\ 0 \end{bmatrix}, \tag{3.64}$$

is also strangeness-free, with $Q_1 = Q[I\ 0\ 0\ 0]^T$*, and the characteristic values are given by* $d_\mu = 2d_\mu^{(2)} + d_\mu^{(1)}, a_\mu, v_\mu$ *and* u_μ*.*

Proof. The proof follows directly from the construction of the trimmed first order formulation (3.64). Setting $\hat{x} = Q^T x$ and $\hat{v} = \frac{d}{dt}(Q_1^T x) = v + \dot{Q}_1^T x$, we obtain

$$\begin{bmatrix} M_{11} & C_{11} & C_{12} & C_{13} & C_{14} \\ 0 & C_{21} & C_{22} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & I_{d_\mu^{(2)}} & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{v}} \\ \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \\ \dot{\hat{x}}_4 \end{bmatrix} + \begin{bmatrix} 0 & K_{11} & K_{12} & K_{13} & K_{14} \\ 0 & K_{21} & K_{22} & K_{23} & K_{24} \\ 0 & K_{31} & K_{32} & K_{33} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ -I_{d_\mu^{(2)}} & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \hat{v} \\ \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \\ \hat{x}_4 \end{bmatrix} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \\ \hat{f}_4 \\ 0 \end{bmatrix},$$

which is clearly strangeness-free, since M_{11}, C_{22}, K_{33} are nonsingular. □

Thus, from Lemma 3.35 and the previous discussion we can obtain a strangeness-free first order formulation directly from the coefficients of the strangeness-free second order system (3.59). The trimmed first order formulation (3.64) is of minimal possible size and no further smoothness requirements for the inhomogeneity are required.

Example 3.36. In Example 3.23 we have computed an equivalent strangeness-free formulation

$$\begin{bmatrix} t & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \ddot{x} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \dot{x} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} x = \begin{bmatrix} f_1 \\ -f_2 - 2\dot{f}_2 - t\ddot{f}_2 + \dddot{f}_3 \\ -tf_2 + f_3 \end{bmatrix}$$

for the linear second order system (3.33) using the derivative array approach. Then, following Lemma 3.35 a trimmed first order formulation for system (3.33) is given by

$$\left[\begin{array}{c|ccc} t & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & 0 \end{array}\right] \begin{bmatrix} \dot{v} \\ \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \left[\begin{array}{c|ccc} 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & -1 \\ \hline 1 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} v \\ x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} f_1 \\ -f_2 - 2\dot{f}_2 - t\ddot{f}_2 + \dddot{f}_3 \\ -tf_2 + f_3 \\ 0 \end{bmatrix},$$

which is strangeness-free and has the same solution components x_1, x_2, and x_3 as the original system (3.33).

In the previous discussion we have seen that we can follow different strategies to obtain a strangeness-free first order formulation for a linear second order DAEs of the form (3.6). In general, one should carefully choose the best suited transformation for the given problem. The main difference in the presented approaches is the chronological order of index and order reduction. In the classical approach first a transformation into a first order system is used, and then a strangeness-free formulation is extracted using index reduction techniques. On the other hand, the index reduction can be carried out at first to obtain a strangeness-free second order system, e.g., by transforming to condensed form (3.14), or by using the derivative arrays (3.19), and in a second step this strangeness-free system is transformed into a first order system. In the previous discussion we have seen that for an appropriate treatment of higher order differential-algebraic system the index reduction should be carried

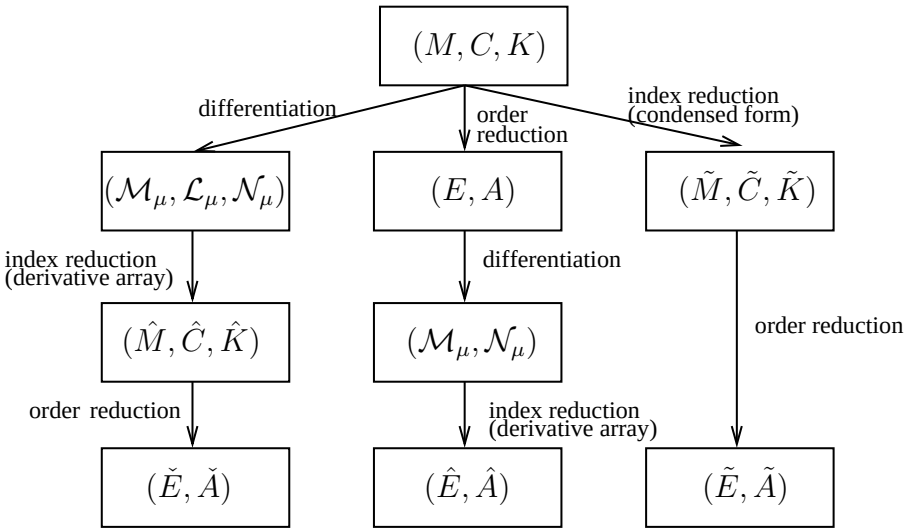

Figure 3.1: Order and index reduction for second order DAE

out at first followed by an order reduction to obtain a strangeness-free first order system. The different strategies are depicted in Figure 3.1.

For constant coefficient higher order differential-algebraic systems the transformation into a first order system corresponds to the linearization of matrix polynomials. In this way higher order differential-algebraic systems are closely related to polynomial eigenvalue problems. The linearization of matrix polynomials is treated in [53, 95]. In [95] large classes of linearizations for matrix polynomials are proposed, which preserve the structure of the matrices as well as the Jordan structure of infinite eigenvalues, corresponding to the index of the DAE. Moreover, different linearizations can have very different condition numbers depending on the magnitude of the eigenvalues, see e.g. [142]. The conditioning of the linearizations introduced in [95] is treated in [63], where it is shown that for any given eigenvalue, we can find a linearization of the matrix polynomial, that will be about as well conditioned as the original problem for that eigenvalue. Analogous to the linearization of matrix polynomials we can find transformations of time-invariant higher order differential-algebraic systems into first order systems, which do not increase the index of the system (corresponding to the preservation of the Jordan structures of infinite eigenvalues), which preserve certain structures in the system, and lead to first order systems that are as well-conditioned as the original problem.

3.4 Explicit Representation of Solutions

For linear second order DAEs with constant coefficients of the form

$$M\ddot{x} + C\dot{x} + Kx = f(t), \tag{3.65}$$

with $M, C, K \in \mathbb{R}^{n,n}$ and $f \in C(\mathbb{I}, \mathbb{R}^n)$, the trimmed first order formulation (3.64) derived in Section 3.3 allows an explicit representation of solutions of the system in terms of the coefficients M, C, K and the inhomogeneity f. For convenience, we restrict ourselves in this section to the square case $m = n$ and assume regularity of the system.
Starting from the corresponding regular strangeness-free system (3.59) we can find an orthogonal matrix $Q \in \mathbb{R}^{n,n}$ as in (3.60) that transforms the system (3.59) into an equivalent strangeness-free system of the form

$$\begin{bmatrix} M_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ddot{\hat{x}}_1 \\ \ddot{\hat{x}}_2 \\ \ddot{\hat{x}}_3 \end{bmatrix} + \begin{bmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \end{bmatrix} + \begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix}, \quad (3.66)$$

with $x = Q\hat{x}$. In the study of matrix polynomials so-called *unimodular transformations* are used as a class of equivalent transformations, such as adding the λa multiple of one row or column to another without increasing the degree of the polynomial. The analogon of these transformations in the context of higher order differential-algebraic systems has been studied in [102]. These unimodular transformations can be reformulated using the concept of differential polynomials, see [66, 92]. Let $\mathbb{R}[D_i]$ be the set of *i-th order differential polynomials with coefficients in* $\mathbb{R}$, i.e.,

$$\mathbb{R}[D_i] := \left\{ a_0 + a_1 \frac{d}{dt} + a_2 \frac{d^2}{dt^2} + \cdots + a_i \frac{d^i}{dt^i} \mid a_k \in \mathbb{R},\ k = 0, 1, \ldots, i \right\}.$$

Since we do not want to increase the order of the differential-algebraic equation, we consider only the following restricted transformations.

Definition 3.37 (Opu-equivalence). Two differential-algebraic systems $M\dot{x}+C\dot{x}+Kx = f$ and $\check{M}\dot{x} + \check{C}\dot{x} + \check{K}x = \check{f}$ are called *order preserving unimodularily equivalent*, or *opu-equivalent*, if there exists a $P \in \mathbb{R}[D_2]^{n,n}$ with constant nonzero determinant such that

$$P(M\dot{x} + C\dot{x} + Kx - f) = \check{M}\dot{x} + \check{C}\dot{x} + \check{K}x - \check{f}.$$

The concept of opu-equivalence transformations requires that the order of differentiation does not increase. Thus, opu-equivalence transformations are nothing else than differentiations of equations and elimination of derivatives in the differential-algebraic system as used in the index reduction procedure described in Section 3.1.1. Now, using an opu-equivalence transformation with

$$P = \begin{bmatrix} I & 0 & -\frac{d}{dt} C_{13} K_{33}^{-1} \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix},$$

we can eliminate the block C_{13} in (3.66) without altering the solution of the system and get the opu-equivalent system

$$\underbrace{\begin{bmatrix} M_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\check{M}=P\hat{M}Q} \begin{bmatrix} \ddot{\hat{x}}_1 \\ \ddot{\hat{x}}_2 \\ \ddot{\hat{x}}_3 \end{bmatrix} + \underbrace{\begin{bmatrix} \check{C}_{11} & \check{C}_{12} & 0 \\ C_{21} & C_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\check{C}=P\hat{C}Q} \begin{bmatrix} \dot{\hat{x}}_1 \\ \dot{\hat{x}}_2 \\ \dot{\hat{x}}_3 \end{bmatrix} + \underbrace{\begin{bmatrix} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ K_{31} & K_{32} & K_{33} \end{bmatrix}}_{\check{K}=P\hat{K}Q} \begin{bmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{bmatrix} = \underbrace{\begin{bmatrix} \check{f}_1 \\ \hat{f}_2 \\ \hat{f}_3 \end{bmatrix}}_{\check{f}=P\hat{f}}, \quad (3.67)$$

where $\check{C}_{11} = C_{11} - C_{13}K_{33}^{-1}K_{31}$, $\check{C}_{12} = C_{12} - C_{13}K_{33}^{-1}K_{32}$ and $\check{f}_1 = \hat{f}_1 - C_{13}K_{33}^{-1}\hat{\dot{f}}_3$. Now, starting from the system

$$\check{M}\ddot{\hat{x}} + \check{C}\dot{\hat{x}} + \check{K}\hat{x} = \check{f},$$

given by (3.67) we can form the trimmed first order formulation (3.64) as described in Section 3.3 to get the first order system

$$\begin{bmatrix} 0 & Q_1^T \\ \hat{M}Q_1 & \check{C}Q^T \end{bmatrix} \begin{bmatrix} \dot{v} \\ \dot{x} \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & -\hat{K} \end{bmatrix} \begin{bmatrix} v \\ x \end{bmatrix} + \begin{bmatrix} 0 \\ \check{f} \end{bmatrix}, \tag{3.68}$$

where in addition we have changed the rows of the system which does not change the solution of the DAE. System (3.68) is a linear first order differential-algebraic system of the form (2.6) with $E, A \in \mathbb{R}^{n+d_\mu^{(2)}, n+d_\mu^{(2)}}$, $b \in C(\mathbb{I}, \mathbb{R}^{n+d_\mu^{(2)}})$ and unknown y given by

$$E = \begin{bmatrix} 0 & Q_1^T \\ \hat{M}Q_1 & \check{C}Q^T \end{bmatrix}, \quad A = \begin{bmatrix} I & 0 \\ 0 & -\hat{K} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \check{f} \end{bmatrix}, \quad y = \begin{bmatrix} v \\ x \end{bmatrix}. \tag{3.69}$$

Since the DAE (3.65) is assumed to be regular, also the matrix pair (E, A) is regular, and the problem (3.68) is uniquely solvable for consistent initial values v_0, x_0. Due to the regularity, we can find a λ such that $(\lambda E - A)$ is nonsingular and by multiplying equation (3.68) with $(\lambda E - A)^{-1}$ we get an equivalent system of the form (2.10) that allows to give an explicit formula for the solution of (3.68) using Theorem 2.31. This solution is of the form

$$y(t) = e^{\hat{E}^D\hat{A}(t-t_0)}\hat{E}^D\hat{E}y_0 + \int_{t_0}^{t} e^{\hat{E}^D\hat{A}(t-s)}\hat{E}^D\hat{b}(s)ds - (I - \hat{E}^D\hat{E})\sum_{i=0}^{\nu-1}(\hat{E}\hat{A}^D)^i\hat{A}^D\hat{b}^{(i)}(t)$$

for some $y_0 \in \mathbb{R}^n$, where $\hat{E} = (\lambda E - A)^{-1}E$, $\hat{A} = (\lambda E - A)^{-1}A$, and $\hat{b} = (\lambda E - A)^{-1}b$. To get an explicit solution representation we first need to determine the index ν of the matrix $\hat{E}$. For linear first order DAEs with constant coefficients of the form (2.6) it is well-known that the index $\nu = \operatorname{ind}(E, A)$ of the matrix pair (E, A) equals the differentiation index ν_d, see e.g. [59]. Further, it holds that for regular DAEs with well-defined strangeness index μ the differentiation index ν_d is also well-defined with

$$\nu_d = \begin{cases} 0 & \text{for } a_\mu = 0, \\ \mu + 1 & \text{for } a_\mu \neq 0, \end{cases}$$

see e.g. [82, Corollary 3.47]. If $a_\mu = 0$, then we have $\nu = \operatorname{ind}(E, A) = \nu_d = 0$, and E is invertible due to the absence of the nilpotent block in the Weierstraß canonical form (2.8). Otherwise, if $a_\mu \neq 0$, then we have $\nu = \operatorname{ind}(E, A) = \nu_d = 1$ for the regular matrix pencil (E, A), as the trimmed first order formulation is strangeness-free due to Lemma 3.35. In this case the matrix pair (E, A) is equivalent to its Weierstrass canonical form (2.8), i.e.,

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

where $N = 0$ as $\nu = 1$. Further, we have $\operatorname{ind}(\hat{E}) = \operatorname{ind}(\hat{E}, I)$ and

$$(\hat{E}, I) = ((\lambda E - A)^{-1}E, I) \sim (E, \lambda E - A) \sim \left(\begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} \lambda I - J & 0 \\ 0 & -I \end{bmatrix} \right),$$

where $\lambda I - J$ is in Jordan canonical form, such that we have $\operatorname{ind}(\hat{E}) = 1$.
With this results the solution of the original second order system (3.65) is either given by

$$x(t) = \begin{bmatrix} 0 & I \end{bmatrix} \left(e^{E^{-1}A(t-t_0)} y_0 + \int_{t_0}^{t} e^{E^{-1}A(t-s)} E^{-1} b(s) ds \right), \tag{3.70}$$

if $a_\mu = 0$ in the first order system (3.68), or by

$$x(t) = \begin{bmatrix} 0 & I \end{bmatrix} \left(e^{\hat{E}^D \hat{A}(t-t_0)} \hat{E}^D \hat{E} y_0 + \int_{t_0}^{t} e^{\hat{E}^D \hat{A}(t-s)} \hat{E}^D \hat{b}(s) ds - (I - \hat{E}^D \hat{E}) \hat{A}^D \hat{b}(t) \right), \tag{3.71}$$

if $a_\mu \neq 0$ in (3.68), for an initial value $y_0 = [\dot{x}_0, x_0]$. In the first case we have

$$E^{-1} = \begin{bmatrix} -M_{11}^{-1}(C_{11} - C_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1} & -M_{11}^{-1}C_{12}C_{22}^{-1} \\ Q_1 - Q_2 C_{22}^{-1} C_{21} & 0 & Q_2 C_{22}^{-1} \end{bmatrix},$$

$$E^{-1}A = \begin{bmatrix} M_{11}^{-1}(C_{12}C_{22}^{-1}C_{21} - C_{11}) & M_{11}^{-1}[(K_{11} - C_{12}C_{22}^{-1}K_{21})Q_1^T + (K_{12} - C_{12}C_{22}^{-1}K_{22})Q_2^T] \\ Q_1 - Q_2 C_{22}^{-1} C_{21} & Q_2 C_{22}^{-1} K_{21} Q_1^T + Q_2 C_{22}^{-1} K_{22} Q_2^T \end{bmatrix},$$

$$E^{-1}b = \begin{bmatrix} M_{11}^{-1}(\hat{f}_1 - C_{12}C_{22}^{-1}\hat{f}_2) \\ 0 \\ Q_2 C_{22}^{-1} \hat{f}_2 \end{bmatrix},$$

such that the solution (3.70) can be formulated only using the coefficient of the original system and the matrix $Q = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}$.
For the second case, i.e., if $a_\mu \neq 0$ in (3.68), we need to describe the products $\hat{E}^D\hat{A}$, $\hat{E}^D\hat{E}$, $\hat{E}^D\hat{b}$, and $\hat{A}^D\hat{b}$ in terms of the coefficients M, C, K of the original second order system (3.65). First of all, we assume that E and A in (3.69) commute, i.e., $EA = AE$. In this case the solution (3.71) can be formulated directly in terms of E and A and we only have to compute the Drazin inverses E^D and A^D. The Drazin inverse of the matrix A is simply given by

$$A^D = \begin{bmatrix} I & 0 \\ 0 & -\hat{K}^D \end{bmatrix}. \tag{3.72}$$

The Drazin inverse of E is given in the following Lemma.

Lemma 3.38. *Consider the matrix $E \in \mathbb{R}^{n+d_\mu^{(2)}, n+d_\mu^{(2)}}$ with*

$$E = \begin{bmatrix} 0 & Q_1^T \\ \hat{M} Q_1 & \check{C} Q^T \end{bmatrix} = \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \hat{Q}^T,$$

as in (3.69) *with* $\nu = \operatorname{ind}(E) = 1$ *coming from the trimmed first order formulation* (3.64), *with* $\hat{Q} = \begin{bmatrix} I & 0 \\ 0 & Q \end{bmatrix} \in \mathbb{R}^{n+d_\mu^{(2)}, n+d_\mu^{(2)}}$ *and orthogonal* $Q = [Q_1\ Q_2\ Q_3] \in \mathbb{R}^{n,n}$. *Then the Drazin inverse of* E *is given by*

$$E^D = \begin{bmatrix} -M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ Q_1 - Q_2C_{22}^{-1}C_{21} & 0 & Q_2C_{22}^{-1} & 0 \end{bmatrix}. \tag{3.73}$$

Proof. We have to verify the axioms (2.2) of the Drazin inverse. First, we have

$$E^DE = \begin{bmatrix} -M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ Q_1 - Q_2C_{22}^{-1}C_{21} & 0 & Q_2C_{22}^{-1} & 0 \end{bmatrix} \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \hat{Q}^T$$

$$= \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & Q_1 & Q_2 & 0 \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Q_1^T \\ 0 & Q_2^T \\ 0 & Q_3^T \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & Q_1Q_1^T + Q_2Q_2^T \end{bmatrix},$$

with $Q_1Q_1^T + Q_2Q_2^T = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{bmatrix}$ and

$$EE^D = \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \hat{Q}^T \begin{bmatrix} -M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ Q_1 - Q_2C_{22}^{-1}C_{21} & 0 & Q_2C_{22}^{-1} & 0 \end{bmatrix}$$

$$= \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} -M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ I & 0 & 0 & 0 \\ -C_{22}^{-1}C_{21} & 0 & C_{22}^{-1} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

As $\nu = \operatorname{ind}(E) = 1$ it follows that $E^DEE = E$ and $E^DEE^D = E^D$. □

With the Drazin inverses (3.72) and (3.73) we have

$$E^DA = \begin{bmatrix} -M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & M_{11}^{-1}(\check{C}_{12}C_{22}^{-1}\hat{K}_2 - \hat{K}_1) \\ Q_1 - Q_2C_{22}^{-1}C_{21} & -Q_2C_{22}^{-1}\hat{K}_2 \end{bmatrix},$$

$$E^Db = \begin{bmatrix} M_{11}^{-1}(\check{f}_1 - \check{C}_{12}C_{22}^{-1}\hat{f}_2) \\ Q_2C_{22}^{-1}\hat{f}_2 \end{bmatrix}, \quad A^Db = \begin{bmatrix} 0 \\ -\hat{K}^D\check{f} \end{bmatrix},$$

such that for the solution representation (3.71) we get

$$x(t) = \begin{bmatrix} 0 & I \end{bmatrix} e^{E^D A(t-t_0)} \begin{bmatrix} v_0 \\ (Q_1 Q_1^T + Q_2 Q_2^T) x_0 \end{bmatrix} + \begin{bmatrix} 0 & I \end{bmatrix} \int_{t_0}^{t} e^{E^D A(t-s)} \begin{bmatrix} M_{11}^{-1}(\check{f}_1 - \check{C}_{12} C_{22}^{-1} \hat{f}_2) \\ Q_2 C_{22}^{-1} \hat{f}_2 \end{bmatrix} ds + Q_3 Q_3^T \hat{K}^D \check{f}.$$

To get a solution representation in the case that E and A do not commute we first consider a regular and homogeneous linear differential-algebraic system

$$E\dot{y}(t) = Ay(t), \quad y(t_0) = y_0, \tag{3.74}$$

with index $\nu = \operatorname{ind}(E, A) = 1$ and E, A as in (3.69). Then, for a consistent initial value y_0, the unique solution of (3.74) is given by

$$y(t) = e^{\hat{E}^D \hat{A}(t-t_0)} \hat{E}^D \hat{E} y_0$$

due to Theorem 2.31. By differentiation we can see that this unique solution also solves the ordinary differential equation

$$\dot{y}(t) = \hat{E}^D \hat{A} y(t), \quad \text{with } y(t_0) = \hat{E}^D \hat{E} y_0. \tag{3.75}$$

This ordinary differential equation is also called the *Drazin ODE*, see [34, 45]. Thus, if we can transform the homogeneous system (3.74) into the corresponding Drazin ODE (3.75), we can get a representation for the Drazin inverse $\hat{E}^D \hat{A}$. Further, consistency conditions for the initial value yield a representation for $\hat{E}^D \hat{E}$. Since the Drazin inverse of a matrix $\hat{E}$ is unique and the products $\hat{E}^D \hat{A}$ and $\hat{E}^D \hat{E}$ are independent of the choice of the parameter λ, see [26], also the representation of $\hat{E}^D \hat{A}$ and $\hat{E}^D \hat{E}$ are unique. Using the basis transformation

$$\hat{Q}^T \begin{bmatrix} v \\ x \end{bmatrix} = \begin{bmatrix} v^T & x_1^T & x_2^T & x_3^T \end{bmatrix}^T, \tag{3.76}$$

where $\hat{Q} = \begin{bmatrix} I & 0 \\ 0 & Q \end{bmatrix}$, and Q chosen as in (3.60), the homogeneous differential-algebraic system (3.74) is equivalent to

$$\begin{aligned} \dot{x}_1 &= v, \\ M_{11}\dot{v} + \check{C}_{11}\dot{x}_1 + \check{C}_{12}\dot{x}_2 + K_{11}x_1 + K_{12}x_2 + K_{13}x_3 &= 0, \\ C_{21}\dot{x}_1 + C_{22}\dot{x}_2 + K_{21}x_1 + K_{22}x_2 + K_{23}x_3 &= 0, \\ K_{31}x_1 + K_{32}x_2 + K_{33}x_3 &= 0. \end{aligned}$$

Since M_{11}, C_{22}, K_{33} are invertible due to the choice of Q we have

$$\begin{aligned}
\dot{x}_1 &= v,\\
\dot{v} &= -M_{11}^{-1}(\check{C}_{11}\dot{x}_1 + \check{C}_{12}\dot{x}_2 + K_{11}x_1 + K_{12}x_2 + K_{13}x_3),\\
\dot{x}_2 &= -C_{22}^{-1}(C_{21}\dot{x}_1 + K_{21}x_1 + K_{22}x_2 + K_{23}x_3),\\
x_3 &= -K_{33}^{-1}(K_{31}x_1 + K_{32}x_2).
\end{aligned} \tag{3.77}$$

Differentiating the last equation in (3.77) once and eliminating all derivatives on the right-hand side yields

$$\begin{aligned}
\dot{x}_1 =& v,\\
\dot{v} =& - M_{11}^{-1}[\check{C}_{11}v - \check{C}_{12}C_{22}^{-1}(C_{21}v + K_{21}x_1 + K_{22}x_2 + K_{23}x_3) + K_{11}x_1 + K_{12}x_2 + K_{13}x_3],\\
\dot{x}_2 =& - C_{22}^{-1}(C_{21}v + K_{21}x_1 + K_{22}x_2 + K_{23}x_3),\\
\dot{x}_3 =& - K_{33}^{-1}(K_{31}v - K_{32}C_{22}^{-1}(C_{21}v + K_{21}x_1 + K_{22}x_2 + K_{23}x_3)).
\end{aligned} \tag{3.78}$$

Further, we can eliminate all occurrences of x_3 using the last equation in (3.77) to get

$$\begin{aligned}
\dot{x}_1 &= v,\\
\dot{v} &= M_{11}^{-1}[(\check{C}_{12}C_{22}^{-1}C_{21} - \check{C}_{11})v + (\check{C}_{12}C_{22}^{-1}Y_{21} - Y_{11})x_1 + (\check{C}_{12}C_{22}^{-1}Y_{22} - Y_{12})x_2],\\
\dot{x}_2 &= -C_{22}^{-1}C_{21}v - C_{22}^{-1}Y_{21}x_1 - C_{22}^{-1}Y_{22}x_2,\\
\dot{x}_3 &= K_{33}^{-1}[(K_{32}C_{22}^{-1}C_{21} - K_{31})v + K_{32}C_{22}^{-1}Y_{21}x_1 + K_{32}C_{22}^{-1}Y_{22}x_2],
\end{aligned}$$

where we have defined

$$\begin{aligned}
Y_{11} &= K_{11} - K_{13}K_{33}^{-1}K_{31}, \qquad & Y_{21} &= K_{21} - K_{23}K_{33}^{-1}K_{31},\\
Y_{12} &= K_{12} - K_{13}K_{33}^{-1}K_{32}, \qquad & Y_{22} &= K_{22} - K_{23}K_{33}^{-1}K_{32}.
\end{aligned}$$

Further, defining

$$\begin{aligned}
V_1 &= \check{C}_{12}C_{22}^{-1}C_{21} - \check{C}_{11}, \qquad & V_4 &= K_{32}C_{22}^{-1}C_{21} - K_{31},\\
V_2 &= \check{C}_{12}C_{22}^{-1}Y_{21} - Y_{11}, \qquad & V_5 &= K_{32}C_{22}^{-1}Y_{21},\\
V_3 &= \check{C}_{12}C_{22}^{-1}Y_{22} - Y_{12}, \qquad & V_6 &= K_{32}C_{22}^{-1}Y_{22},
\end{aligned}$$

we get an ordinary differential equation

$$\begin{bmatrix} \dot{v} \\ \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} M_{11}^{-1}V_1 & M_{11}^{-1}V_2 & M_{11}^{-1}V_3 & 0 \\ I & 0 & 0 & 0 \\ -C_{22}^{-1}C_{21} & -C_{22}^{-1}Y_{21} & -C_{22}^{-1}Y_{22} & 0 \\ K_{33}^{-1}V_4 & K_{33}^{-1}V_5 & K_{33}^{-1}V_6 & 0 \end{bmatrix} \begin{bmatrix} v \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}.$$

Reversing the basis transformation (3.76) yields a system of the form

$$\begin{bmatrix}\dot{v}\\ \dot{x}\end{bmatrix} = \hat{Q}\begin{bmatrix} M_{11}^{-1}V_1 & M_{11}^{-1}V_2 & M_{11}^{-1}V_3 & 0\\ I & 0 & 0 & 0\\ -C_{22}^{-1}C_{21} & -C_{22}^{-1}Y_{21} & -C_{22}^{-1}Y_{22} & 0\\ K_{33}^{-1}V_4 & K_{33}^{-1}V_5 & K_{33}^{-1}V_6 & 0\end{bmatrix}\hat{Q}^T\begin{bmatrix}v\\ x\end{bmatrix} =$$
$$\begin{bmatrix} M_{11}^{-1}V_1 & M_{11}^{-1}V_2Q_1^T + M_{11}^{-1}V_3Q_2^T\\ Q_1 - Q_2C_{22}^{-1}C_{21} + Q_3K_{33}^{-1}V_4 & -Q_2C_{22}^{-1}(Y_{21}Q_1^T + Y_{22}Q_2^T) + Q_3K_{33}^{-1}(V_5Q_1^T + V_6Q_2^T)\end{bmatrix}\begin{bmatrix}v\\ x\end{bmatrix}$$

in the original variables v and x. Due to the algebraic equations, consistent initial values have to satisfy

$$\begin{aligned}\begin{bmatrix}v(t_0)\\ x(t_0)\end{bmatrix} &= \hat{Q}\begin{bmatrix} I & 0 & 0 & 0\\ 0 & I & 0 & 0\\ 0 & 0 & I & 0\\ 0 & -K_{33}^{-1}K_{31} & -K_{33}^{-1}K_{32} & 0\end{bmatrix}\hat{Q}^T\begin{bmatrix}v_0\\ x_0\end{bmatrix},\\ &= \begin{bmatrix} I & 0\\ 0 & I - Q_3K_{33}^{-1}(K_{31}Q_1^T + K_{32}Q_2^T + K_{33}Q_3^T)\end{bmatrix}\begin{bmatrix}v_0\\ x_0\end{bmatrix},\\ &= \begin{bmatrix} I & 0\\ 0 & I - Q_3K_{33}^{-1}\hat{K}_3\end{bmatrix}\begin{bmatrix}v_0\\ x_0\end{bmatrix}.\end{aligned}$$

Altogether, we have derived the Drazin ODE and consistency conditions for initial values for the homogeneous system (3.74).

Lemma 3.39. *Consider a regular linear differential-algebraic system* (3.74) *with index* $\nu = \operatorname{ind}(E, A) = 1$ *and E,A given by* (3.69) *and assume that a consistent initial value* y_0 *is given. Further, consider the corresponding Drazin ODE derived by differentiations*

$$\dot{y} = Sy, \quad y(t_0) = Hy_0,$$

with

$$\begin{aligned}S &= \begin{bmatrix} M_{11}^{-1}V_1 & M_{11}^{-1}V_2Q_1^T + M_{11}^{-1}V_3Q_2^T\\ Q_1 - Q_2C_{22}^{-1}C_{21} + Q_3K_{33}^{-1}V_4 & -Q_2C_{22}^{-1}(Y_{21}Q_1^T + Y_{22}Q_2^T) + Q_3K_{33}^{-1}(V_5Q_1^T + V_6Q_2^T)\end{bmatrix},\\ H &= \begin{bmatrix} I & 0\\ 0 & I - Q_3K_{33}^{-1}(K_{31}Q_1^T + K_{32}Q_2^T + K_{33}Q_3^T)\end{bmatrix}.\end{aligned}$$

Then, it holds that

$$\hat{E}^D\hat{A} = S \quad \text{and} \quad \hat{E}^D\hat{E} = H,$$

and furthermore, $\hat{E}^D\hat{A}$ *and* $\hat{E}^D\hat{E}$ *are unique.*

Proof. As the matrix pair (E, A) is regular there exists a λ such that $(\lambda E - A)$ is nonsingular. Setting $\hat{E} = (\lambda E - A)^{-1}E$ and $\hat{A} = (\lambda E - A)^{-1}A$, the products $\hat{E}^D\hat{A}$ and $\hat{E}^D\hat{E}$ are independent of the choice of the parameter λ, see [26]. Further, the Drazin inverse $\hat{E}^D$ is unique such that also the products $\hat{E}^D\hat{A}$ and $\hat{E}^D\hat{E}$ are unique. We have

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

Thus, to prove that $\hat{E}^D\hat{A} = S$ and $\hat{E}^D\hat{E} = H$ we have to verify the conditions (2.2) of the Drazin inverse for $\hat{E}^D = \lambda H - S$.
In order to get a nonsingular $\lambda E - A$ we can choose $\lambda \in \mathbb{R}$ such that $\lambda C_{22} + (K_{22} - K_{23}K_{33}^{-1}K_{32}) = \lambda C_{22} + Y_{22}$ is invertible. At first, we compute $\hat{E} = (\lambda E - A)^{-1}E$ via block inversion. We have

$$\begin{aligned}
\hat{E} &= \left(\begin{bmatrix} -I & [\lambda I\ 0\ 0] \\ \lambda\hat{M}Q_1 & \lambda\check{C} + \hat{K}Q \end{bmatrix}\hat{Q}^T\right)^{-1}\begin{bmatrix} 0 & [I\ 0\ 0] \\ \hat{M}Q_1 & \check{C} \end{bmatrix}\hat{Q}^T \\
&= \hat{Q}\begin{bmatrix} -I & \lambda I & 0 & 0 \\ \lambda M_{11} & \lambda\check{C}_{11} + K_{11} & \lambda\check{C}_{12} + K_{12} & K_{13} \\ 0 & \lambda C_{21} + K_{21} & \lambda C_{22} + K_{22} & K_{23} \\ 0 & K_{31} & K_{32} & K_{33} \end{bmatrix}^{-1}\begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\hat{Q}^T \\
&= \hat{Q}\left(\begin{bmatrix} I & 0 & 0 & 0 \\ -\lambda M_{11} & I & 0 & K_{13}K_{33}^{-1} \\ 0 & 0 & I & K_{23}K_{33}^{-1} \\ 0 & 0 & 0 & I \end{bmatrix}\begin{bmatrix} -I & 0 & 0 & 0 \\ 0 & \lambda(\lambda M_{11} + \check{C}_{11}) + Y_{11} & \lambda\check{C}_{12} + Y_{12} & 0 \\ 0 & \lambda C_{21} + Y_{21} & \lambda C_{22} + Y_{22} & 0 \\ 0 & 0 & 0 & K_{33} \end{bmatrix}\right. \\
&\quad \left.\begin{bmatrix} I & -\lambda I & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & K_{33}^{-1}K_{31} & K_{33}^{-1}K_{32} & I \end{bmatrix}\right)^{-1}\begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\hat{Q}^T \\
&= \hat{Q}\begin{bmatrix} I & \lambda I & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & -K_{33}^{-1}K_{31} & -K_{33}^{-1}K_{32} & I \end{bmatrix}\begin{bmatrix} -I & 0 & 0 & 0 \\ 0 & A_{11} & A_{12} & 0 \\ 0 & A_{21} & A_{22} & 0 \\ 0 & 0 & 0 & K_{33}^{-1} \end{bmatrix}\begin{bmatrix} I & 0 & 0 & 0 \\ \lambda M_{11} & I & 0 & -K_{13}K_{33}^{-1} \\ 0 & 0 & I & -K_{23}K_{33}^{-1} \\ 0 & 0 & 0 & I \end{bmatrix} \\
&\quad \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}\hat{Q}^T =: \hat{Q}M_1M_2M_3M_4\hat{Q}^T,
\end{aligned}$$

where

$$\begin{aligned}
\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} &= \begin{bmatrix} \lambda^2 M_{11} + \lambda\check{C}_{11} + Y_{11} & \lambda\check{C}_{12} + Y_{12} \\ \lambda C_{21} + Y_{21} & \lambda C_{22} + Y_{22} \end{bmatrix}^{-1} \\
&= \begin{bmatrix} I & 0 \\ -X_1^{-1}(\lambda C_{21} + Y_{21}) & I \end{bmatrix}\begin{bmatrix} X_2^{-1} & -X_2^{-1}(\lambda\check{C}_{12} + Y_{12})X_1^{-1} \\ 0 & X_1^{-1} \end{bmatrix} \\
&= \begin{bmatrix} X_2^{-1} & -X_2^{-1}(\lambda\check{C}_{12} + Y_{12})X_1^{-1} \\ -X_1^{-1}(\lambda C_{21} + Y_{21})X_2^{-1} & X_1^{-1}(\lambda C_{21} + Y_{21})X_2^{-1}(\lambda\check{C}_{12} + Y_{12})X_1^{-1} + X_1^{-1} \end{bmatrix},
\end{aligned}$$

with nonsingular matrices

$$X_1 = \lambda C_{22} + Y_{22},$$

and

$$X_2 = \lambda^2 M_{11} + \lambda \check{C}_{11} + Y_{11} - (\lambda \check{C}_{12} + Y_{12}) X_1^{-1} (\lambda C_{21} + Y_{21})$$

due to construction and due to the choice of λ. To verify the first condition (2.2a) we need to show that $\hat{E}\hat{E}^D = \hat{E}(\lambda H - S) = (\lambda H - S)\hat{E} = \hat{E}^D\hat{E} = H$. If we set

$$(\lambda H - S) =: \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix},$$

with

$$\begin{aligned}
H_{11} =& \lambda I - M_{11}^{-1} V_1, \\
H_{12} =& - M_{11}^{-1} [V_2 Q_1^T + V_3 Q_2^T], \\
H_{21} =& - Q_1 + Q_2 C_{22}^{-1} C_{21} - Q_3 K_{33}^{-1} V_4, \\
H_{22} =& \lambda I + Q_2 C_{22}^{-1} (Y_{21} Q_1^T + Y_{22} Q_2^T) - Q_3 K_{33}^{-1} [(V_5 + \lambda K_{31}) Q_1^T \\
& + (V_6 + \lambda K_{32}) Q_2^T + \lambda K_{33} Q_3^T],
\end{aligned}$$

then we get

$$\hat{E}(\lambda H - S) = \hat{Q} M_1 M_2 M_3 M_4 \hat{Q}^T \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} = \hat{Q} M_1 M_2 M_3 M_4 \begin{bmatrix} H_{11} & H_{12} \\ Q_1^T H_{21} & Q_1^T H_{22} \\ Q_2^T H_{21} & Q_2^T H_{22} \\ Q_3^T H_{21} & Q_3^T H_{22} \end{bmatrix}$$

$$= \hat{Q} M_1 M_2 M_3 \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} H_{11} & H_{12} \\ -I & \lambda Q_1^T \\ C_{22}^{-1} C_{21} & \lambda Q_2^T + C_{22}^{-1} (Y_{21} Q_1^T + Y_{22} Q_2^T) \\ -K_{33}^{-1} V_4 & \lambda Q_3^T - K_{33}^{-1} [(V_5 + \lambda K_{31}) Q_1^T + (V_6 + \lambda K_{32}) Q_2^T + \lambda K_{33} Q_3^T] \end{bmatrix}$$

$$= \hat{Q} M_1 M_2 \begin{bmatrix} I & 0 & 0 & 0 \\ \lambda M_{11} & I & 0 & -K_{13} K_{33}^{-1} \\ 0 & 0 & I & -K_{23} K_{33}^{-1} \\ 0 & 0 & 0 & I \end{bmatrix} \begin{bmatrix} -I & \lambda Q_1^T \\ \lambda M_{11} & (\lambda \check{C}_{11} + Y_{11}) Q_1^T + (\lambda \check{C}_{12} + Y_{12}) Q_2^T \\ 0 & (\lambda C_{21} + Y_{21}) Q_1^T + (\lambda C_{22} + Y_{22}) Q_2^T \\ 0 & 0 \end{bmatrix}$$

$$= \hat{Q} M_1 \begin{bmatrix} -I & 0 & 0 & 0 \\ 0 & A_{11} & A_{12} & 0 \\ 0 & A_{21} & A_{22} & 0 \\ 0 & 0 & 0 & K_{33}^{-1} \end{bmatrix} \begin{bmatrix} -I & \lambda Q_1^T \\ 0 & (\lambda^2 M_{11} + \lambda \check{C}_{11} + Y_{11}) Q_1^T + (\lambda \check{C}_{12} + Y_{12}) Q_2^T \\ 0 & (\lambda C_{21} + Y_{21}) Q_1^T + (\lambda C_{22} + Y_{22}) Q_2^T \\ 0 & 0 \end{bmatrix}$$

$$= \hat{Q} \begin{bmatrix} I & \lambda I & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & -K_{33}^{-1} K_{31} & -K_{33}^{-1} K_{32} & I \end{bmatrix} \begin{bmatrix} I & -\lambda Q_1^T \\ 0 & A_{11} X_4 + A_{12} X_5 \\ 0 & A_{21} X_4 + A_{22} X_5 \\ 0 & 0 \end{bmatrix}$$

$$
\begin{aligned}
&= \hat{Q}\begin{bmatrix} I & \lambda(-Q_1^T + Q_1^T) \\ 0 & Q_1^T \\ 0 & Q_2^T \\ 0 & -K_{33}^{-1}[(K_{31}A_{11} + K_{32}A_{21})X_4 + (K_{31}A_{12} + K_{32}A_{22})X_5] \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ 0 & Q_1Q_1^T + Q_2Q_2^T - Q_3K_{33}^{-1}(K_{31}Q_1^T + K_{32}Q_2^T) \end{bmatrix} \\
&= H,
\end{aligned}
$$

where

$$
\begin{aligned}
X_4 &= (\lambda^2 M_{11} + \lambda\check{C}_{11} + Y_{11})Q_1^T + (\lambda\check{C}_{12} + Y_{12})Q_2^T, \\
X_5 &= (\lambda C_{21} + Y_{21})Q_1^T + (\lambda C_{22} + Y_{22})Q_2^T,
\end{aligned}
$$

and using that

$$
\begin{aligned}
A_{11}X_4 + A_{12}X_5 &= Q_1^T, \\
A_{21}X_4 + A_{22}X_5 &= Q_2^T.
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
(\lambda H - S)\hat{E} &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \hat{Q}M_1M_2M_3M_4\hat{Q}^T \\
&= \begin{bmatrix} H_{11} & H_{12}Q_1 & H_{12}Q_2 & 0 \\ H_{21} & H_{22}Q_1 & H_{22}Q_2 & 0 \end{bmatrix} \begin{bmatrix} I & \lambda I & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & -K_{33}^{-1}K_{31} & -K_{33}^{-1}K_{32} & I \end{bmatrix} M_2M_3M_4\hat{Q}^T \\
&= \begin{bmatrix} H_{11} & \lambda H_{11} + H_{12}Q_1 & H_{12}Q_2 & 0 \\ H_{21} & \lambda H_{21} + H_{22}Q_1 & H_{22}Q_2 & 0 \end{bmatrix} \begin{bmatrix} -I & 0 & 0 & 0 \\ 0 & A_{11} & A_{12} & 0 \\ 0 & A_{21} & A_{22} & 0 \\ 0 & 0 & 0 & K_{33}^{-1} \end{bmatrix} M_3M_4\hat{Q}^T \\
&= \begin{bmatrix} -H_{11} & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ -H_{21} & 0 & (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1} & 0 \end{bmatrix} M_3M_4\hat{Q}^T \\
&= \begin{bmatrix} -H_{11} & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & 0 \\ -H_{21} & 0 & (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1} & 0 \end{bmatrix} \begin{bmatrix} I & 0 & 0 & 0 \\ \lambda M_{11} & I & 0 & -K_{13}K_{33}^{-1} \\ 0 & 0 & I & -K_{23}K_{33}^{-1} \\ 0 & 0 & 0 & I \end{bmatrix} M_4\hat{Q}^T \\
&= \begin{bmatrix} -H_{11} + \lambda I & M_{11}^{-1} & -M_{11}^{-1}\check{C}_{12}C_{22}^{-1} & M_{11}^{-1}(\check{C}_{12}C_{22}^{-1}K_{23} - K_{13})K_{33}^{-1}) \\ -H_{21} & 0 & (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1} & -(Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}K_{23}K_{33}^{-1} \end{bmatrix} \\
&\quad \begin{bmatrix} 0 & I & 0 & 0 \\ M_{11} & \check{C}_{11} & \check{C}_{12} & 0 \\ 0 & C_{21} & C_{22} & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \hat{Q}^T
\end{aligned}
$$

$$
\begin{aligned}
&= \begin{bmatrix} I & -H_{11} + \lambda I + M_{11}^{-1}(\check{C}_{11} - \check{C}_{12}C_{22}^{-1}C_{21}) & 0 & 0 \\ 0 & -H_{21} + (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}C_{21} & Q_2 - Q_3K_{33}^{-1}K_{32} & 0 \end{bmatrix} \hat{Q}^T \\
&= \begin{bmatrix} I & 0 \\ 0 & -H_{21}Q_1^T + (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}C_{21}Q_1^T + (Q_2 - Q_3K_{33}^{-1}K_{32})Q_2^T \end{bmatrix} \\
&= \begin{bmatrix} I & 0 \\ 0 & Q_1Q_1^T + Q_2Q_2^T - Q_3K_{33}^{-1}(K_{31}Q_1^T + K_{32}Q_2^T) \end{bmatrix} = H,
\end{aligned}
$$

where we have used that

$$
\begin{aligned}
&(\lambda H_{11} + H_{12}Q_1)A_{11} + H_{12}Q_2A_{21} = M_{11}^{-1}, \\
&(\lambda H_{11} + H_{12}Q_1)A_{12} + H_{12}Q_2A_{22} = -M_{11}^{-1}\check{C}_{12}C_{22}^{-1}, \\
&(\lambda H_{21} + H_{22}Q_1)A_{11} + H_{22}Q_2A_{21} = 0, \\
&(\lambda H_{21} + H_{22}Q_1)A_{12} + H_{22}Q_2A_{22} = (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}.
\end{aligned}
$$

Further, the structure of H implies that $SH = HS = S$ and $H^2 = H$. Thus, for the second condition (2.2b) we have

$$
\hat{E}^D\hat{E}\hat{E}^D = (\lambda H - S)\hat{E}(\lambda H - S) = (\lambda H - S)H = \lambda H - S = \hat{E}^D.
$$

Finally, for the third condition (2.2c) we have

$$
\begin{aligned}
(\lambda H - S)\hat{E}^2 &= \hat{E}(\lambda H - S)\hat{E} = \hat{E}H = (\lambda E - A)^{-1}EH \\
&= (\lambda E - A)^{-1} \begin{bmatrix} 0 & Q_1^T \\ \hat{M}Q_1 & \check{C}Q^T(I - Q_3K_{33}^{-1}(K_{31}Q_1^T + K_{32}Q_2^T + K_{33}Q_3^T)) \end{bmatrix} \\
&= (\lambda E - A)^{-1} \begin{bmatrix} 0 & Q_1^T \\ \hat{M}Q_1 & \check{C}Q^T - \check{C}Q^TQ_3K_{33}^{-1}\hat{K}_3 \end{bmatrix} = \hat{E}.
\end{aligned}
$$

□

Thus, for the linear homogeneous DAE (3.74) we have the explicit solution representation

$$
y(t) = e^{S(t-t_0)}Hy_0 = e^{S(t-t_0)} \begin{bmatrix} v_0 \\ (I - Q_3K_{33}^{-1}\hat{K}_3)x_0 \end{bmatrix},
$$

where S and H are given as in Lemma 3.39, and $y_0 = \begin{bmatrix} v_0^T & x_0^T \end{bmatrix}^T$, and for the solution $x(t)$ of the corresponding homogeneous second order system (3.65) we therefore have

$$
x(t) = \begin{bmatrix} 0 & I \end{bmatrix} e^{S(t-t_0)} \begin{bmatrix} v_0 \\ (I - Q_3K_{33}^{-1}\hat{K}_3)x_0 \end{bmatrix}.
$$

From the well-known principle that two solutions of a linear inhomogeneous problem differ only by a solution of the corresponding homogeneous problem, we only need to append a particular solution of the corresponding inhomogeneous problem to describe all solutions

of the inhomogeneous problem. Thus, for an inhomogeneous problem (3.68) of index $\nu = 1$ we get the solution representation

$$y(t) = e^{S(t-t_0)}Hy_0 + \int_{t_0}^{t} e^{S(t-s)}(\lambda H - S)\hat{b}(s)ds - (I-H)\hat{A}^D\hat{b}(t),$$

with

$$I - H = \begin{bmatrix} 0 & 0 \\ 0 & Q_3K_{33}^{-1}\hat{K}_3 \end{bmatrix}, \quad Hy_0 = \begin{bmatrix} v_0 \\ (I - Q_3K_{33}^{-1}\hat{K}_3)x_0 \end{bmatrix},$$

as well as

$$\begin{aligned}(\lambda H - S)(\lambda E - A)^{-1}b(s) &= \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix} \hat{Q}M_1M_2M_3b(s) \\ &= \begin{bmatrix} M_{11}^{-1}(\check{f}_1 - \check{C}_{12}C_{22}^{-1}\hat{f}_2 + (\check{C}_{12}C_{22}^{-1}K_{23} + K_{13})K_{33}^{-1}\hat{f}_3) \\ (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}(\hat{f}_2 - K_{23}K_{33}^{-1}\hat{f}_3) \end{bmatrix},\end{aligned}$$

following from the proof of Lemma 3.39. Further, we have

$$\begin{aligned}(I-H)\hat{A}^D\hat{b}(t) = \hat{A}^D(I-H)\hat{b}(t) &= \hat{A}^D \begin{bmatrix} 0 & 0 \\ 0 & Q_3K_{33}^{-1}\hat{K}_3 \end{bmatrix} (\lambda E - A)^{-1}b(t) \\ &= \hat{A}^D \begin{bmatrix} 0 & 0 \\ 0 & Q_3K_{33}^{-1}\hat{K}_3 \end{bmatrix} \hat{Q}M_1M_2M_3b(t) \\ &= \hat{A}^D \begin{bmatrix} 0 \\ Q_3K_{33}^{-1}\hat{f}_3(t) \end{bmatrix}.\end{aligned}$$

Therefore, for the solution $x(t)$ of the original second order system (3.65) we have

$$\begin{aligned}x(t) = \begin{bmatrix} 0 & I \end{bmatrix} \Bigg\{ & e^{S(t-t_0)} \begin{bmatrix} v_0 \\ (I - Q_3K_{33}^{-1}\hat{K}_3)x_0 \end{bmatrix} - \hat{A}^D \begin{bmatrix} 0 \\ Q_3K_{33}^{-1}\hat{f}_3 \end{bmatrix} \\ & + \int_{t_0}^{t} e^{S(t-s)} \begin{bmatrix} M_{11}^{-1}(\check{f}_1 - \check{C}_{12}C_{22}^{-1}\hat{f}_2 + (\check{C}_{12}C_{22}^{-1}K_{23} + K_{13})K_{33}^{-1}\hat{f}_3) \\ (Q_2 - Q_3K_{33}^{-1}K_{32})C_{22}^{-1}(\hat{f}_2 - K_{23}K_{33}^{-1}\hat{f}_3) \end{bmatrix} ds \Bigg\}.\end{aligned} \tag{3.79}$$

Remark 3.40. *For the solution representation* (3.79) *the Drazin inverse* $\hat{A}^D$ *is required. In most applications the matrix* $\hat{K}$ *might be nonsingular or* $\hat{K} = I$ *meaning that* E *and* A *in* (3.69) *commute such that* $\hat{A}^D$ *is simply given by* (3.72).

3.5 Future Work

In the previous discussion we have restricted to second order differential-algebraic systems. In general, all presented concepts can also be extended to arbitrary high order systems of the form (3.2). The theoretical analysis and the condensed forms given in Section 3.1.1

have been generalized to linear k-th order DAEs (3.2) in [102, 135], where in particular a strangeness-free canonical form as in Theorem 3.11 and a first order formulation as in Corollary 3.12 are derived. Also the derivative array approach presented in Section 3.1.2 can be generalized to linear k-th order DAEs, see also Remark 3.24, and by linearization along solution trajectories also nonlinear k-th order systems of the form (3.1) can be handled in the same way as in Section 3.2. Further, given a strangeness-free linear k-th order system, the trimmed order reduction formalism derived in Section 3.3 can be applied successively to the k-th order system to reduce the order by one in each reduction step. In this process the derivative of order $(k-1)$ of the transformation matrix Q, chosen similar as in (3.60), will occur. In the constant coefficient case that corresponds to the theory of matrix polynomials, structure preserving staircase forms for matrix tuples are given in [20], that allow trimmed linearizations for arbitrary high order systems in the context of matrix polynomials. For the variable coefficient case it is not clear if such structure preserving staircase forms exist and how trimmed first order formulations can be derived in this case.
Furthermore, it remains to prove Theorem 3.19, and consequently also Corollary 3.20, Theorem 3.21 and Theorem 3.22, for arbitrary strangeness index $\mu > 2$. To do this another global canonical form analogous to the form given in [82, Theorem 3.21] might be helpful. Here, we only state the following conjecture.

Conjecture. *Let the strangeness-index of (M, C, K) with $M, C, K \in C(\mathbb{I}, \mathbb{R}^{m,n})$ be well-defined. Then (M, C, K) is globally equivalent to a matrix triple of the form*

$$\left(\begin{bmatrix} I_{d_\mu^{(2)}} & 0 & 0 & 0 \\ 0 & C & 0 & F \\ 0 & D & 0 & G \\ 0 & E & 0 & H \end{bmatrix}, \begin{bmatrix} \star & 0 & \star & \star \\ 0 & I_{d_\mu^{(1)}} & 0 & J \\ 0 & 0 & 0 & K \\ 0 & 0 & 0 & L \end{bmatrix}, \begin{bmatrix} \star & \star & \star & 0 \\ \star & \star & \star & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I_{a_\mu} \end{bmatrix} \right),$$

with

$$\mathcal{X} = \begin{bmatrix} 0 & \mathcal{X}_\mu & & \star \\ & \ddots & \ddots & \\ & & \ddots & \mathcal{X}_1 \\ & & & 0 \end{bmatrix}, \quad \mathcal{Y} = \begin{bmatrix} 0 & \mathcal{Y}_{\mu,\mu} & & \star \\ \vdots & \vdots & \ddots & \\ 0 & \mathcal{Y}_{1,\mu} & \dots & \mathcal{Y}_{1,1} \\ 0 & 0 & \dots & 0 \end{bmatrix},$$

for $\mathcal{X} \in \{C, D, E, J, K, L\}$, $\mathcal{Y} \in \{F, G, H\}$, where the blocks K_i, L_i, C_i, D_i, E_i have sizes $w_i \times c_{i-1}$, $c_i \times c_{i-1}$, $q_i \times q_{i-1}$, $w_i \times q_{i-1}$, and $c_i \times q_{i-1}$, respectively, and $J = \begin{bmatrix} 0 & \star & \dots & \star \end{bmatrix}$ is partitioned accordingly. Further, the blocks $G_{i,j}$, $H_{i,j}$ and $F_{i,j}$ have sizes $w_i \times c_{j-1}$, $c_i \times c_{j-1}$ and $q_i \times c_{j-1}$, respectively. In particular, we have the full row rank condition

$$\operatorname{rank} \begin{bmatrix} C_\mu & F_{\mu,\mu} & J_\mu \\ D_\mu & G_{\mu,\mu} & K_\mu \\ E_\mu & H_{\mu,\mu} & L_\mu \end{bmatrix} = 2s_{\mu-1}^{(MCK)} + s_{\mu-1}^{(MC)} + s_{\mu-1}^{(MK)} + s_{\mu-1}^{(CK)} = c_\mu + w_\mu + q_\mu.$$

Chapter 4

Structured Differential-Algebraic Systems

In many technical applications the arising differential-algebraic equations exhibit certain structures as e.g. the equations of motions of multibody systems (1.1), the circuit equations (1.4), or linear systems as in (1.2) or (1.3), where the coefficient matrices are structured, see also Chapter 1. In the numerical solution of these systems the structural information can be used to develop efficient index reduction and solution methods. The equations of motion of multibody systems (1.1) have been an important research topic for many years and efficient methods for the index reduction and for the numerical solution have been developed, see e.g. [14, 34, 50, 137]. Also index reduction methods for electrical circuit equations (1.4) have been studied [6, 7, 8, 36]. But, the development of structure preserving index reduction methods for linear time-variant systems with symmetries in the coefficient matrices has remained open.
In general, the structure of a system reflects a physical property of the system that should be preserved during the numerical solution. In the case of linear DAEs with constant coefficient of the form (2.6), for example, the algebraic structure of the problem forces the eigenvalues of the corresponding eigenvalue problem to lie in certain regions in the complex plane (e.g., on the unit circle or the real axis) or to occur in different kind of pairings. If such a system is solved numerically without considering the structure then these physical properties are obscured and we might get physically meaningless results as rounding errors can cause eigenvalues to wander out of their required region, see e.g. [37]. In this field, mainly from the point of view of generalized eigenvalue problems, structure preserving canonical forms as well as structure preserving solution methods for matrix pairs have been investigated, see e.g. [20, 70, 132]. These results can be applied for differential-algebraic systems with constant coefficients, but they do not allow the treatment of time-variant differential-algebraic systems. Another important aspect in the numerical solution of structured differential-algebraic systems is that the structure of the system can be used for an efficient the solution of the linear systems arising in each integration step, which usually has the highest computational effort during the numerical integration.

In this chapter we consider linear differential-algebraic systems with variable coefficients of the form (2.5) where the coefficient matrices $E(t)$ and $A(t)$ are symmetric, e.g. as in the linearized equations of motions of mechanical systems (1.2), see also [53, 155, 156], or in the semidiscretization of the Stokes equation and the linearized Navier-Stokes equation [149]. On the other hand, we consider linear systems of the form (2.5) that have a self-adjoint structure as in linear-quadratic optimal control problems (1.3), see also [10, 83], or in gyroscopic mechanical systems [65, 89]. In Section 4.1 we review structured condensed

forms for symmetric matrix pairs that are extended to the case of pairs of symmetric matrix-valued functions in Section 4.2. Analogous structure preserving condensed forms for pairs of Hermitian matrix-valued function have also been derived in [153]. In Section 4.3 we derive a structure preserving condensed form and a strangeness-free formulation for self-adjoint pairs of matrix-valued functions. Finally, in Section 4.4, we present a structure preserving index reduction method for self-adjoint systems based on index reduction by minimal extension.

4.1 Condensed Forms for Symmetric Matrix Pairs

To derive structure preserving condensed forms for linear DAEs we start with linear time-invariant systems of the form

$$E\dot{x} = Ax + b(t), \quad t \in \mathbb{I}, \tag{4.1}$$

where $E, A \in \mathbb{R}^{n\times n}$ are symmetric, i.e., $E = E^T$ and $A = A^T$ and $b \in C(\mathbb{I}, \mathbb{R}^n)$. In order to obtain a structure preserving condensed form for the symmetric matrix pair (E, A) we cannot use general equivalence transformations but have to restrict to congruence transformations.

Definition 4.1 (Strong congruence). Two pairs of matrices (E_i, A_i), $i = 1, 2$, with $E_i, A_i \in \mathbb{R}^{n,n}$ are called *strongly congruent* if there exists a nonsingular matrix $P \in \mathbb{R}^{n,n}$ such that

$$E_2 = P^T E_1 P, \quad A_2 = P^T A_1 P. \tag{4.2}$$

This congruence transformation defines an equivalence relation. The canonical form for matrix pairs under general equivalence transformations is the well-known Kronecker canonical form, see e.g. [47]. In the symmetric case there also exists a symmetric version of the Kronecker canonical form under congruence transformations (4.2), see [140], but the numerical computation of this canonical form is an ill-conditioned problem as small rounding errors can radically change the kind and number of Kronecker blocks. A numerically computable structured staircase form for symmetric matrix pairs that displays the invariants of the structured Kronecker form is given e.g. in [20]. For the analysis of existence and uniqueness of solutions of DAEs, however, we do not need the complete information of generalized eigenvalues and eigenspaces provided by the invariants of the Kronecker canonical form, but only the infomation about the eigenvalues at infinity. Therefore, it is sufficient to consider condensed forms for pairs of symmetric matrices that can be computed numerically using rank decisions based on orthogonal transformations and allow to analyze the index of a DAE as well as existence and uniqueness of solutions. To derive such a condensed form we use the symmetric Schur decomposition of a symmetric matrix, see e.g. [54, 90].

Lemma 4.2. *Let $A \in \mathbb{R}^{n,n}$ be symmetric with* $\operatorname{rank} A = r$. *Then there exists an orthogonal matrix $P \in \mathbb{R}^{n,n}$ such that*

$$P^T A P = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \tag{4.3}$$

with nonsingular and diagonal $\Sigma_r \in \mathbb{R}^{r,r}$.

Proof. See e.g. [54, Theorem 8.1.1]. □

Now, we can derive a condensed form for pairs of symmetric matrices using orthogonal congruence transformations.

Theorem 4.3. *Let* $E, A \in \mathbb{R}^{n,n}$ *be symmetric and let*

$$\begin{aligned} T &\quad \text{be a basis of } \operatorname{kernel} E, \\ T' &\quad \text{be a basis of } \operatorname{cokernel} E = \operatorname{range} E, \\ V &\quad \text{be a basis of } \operatorname{corange}(T^T A T). \end{aligned}$$

Then there exists an orthogonal matrix $P \in \mathbb{R}^{n,n}$ *such that the matrix pair* (E, A) *is strongly congruent to a symmetric matrix pair of the form*

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} & \Sigma_s & 0 \\ A_{12}^T & A_{22} & A_{23} & 0 & 0 \\ A_{13}^T & A_{23}^T & \Sigma_a & 0 & 0 \\ \Sigma_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right), \quad \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} \tag{4.4}$$

where the matrices $\Sigma_a \in \mathbb{R}^{a,a}$, *and* $\Sigma_s \in \mathbb{R}^{s,s}$ *are nonsingular and diagonal, the block* $\begin{bmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{bmatrix} \in \mathbb{R}^{r,r}$ *is nonsingular, and the last block rows and block columns are of dimension* u. *Further, the quantities*

(a)	$r = \operatorname{rank} E$,	*(rank)*
(b)	$a = \operatorname{rank}(T^T A T)$,	*(algebraic part)*
(c)	$s = \operatorname{rank}(V^T T^T A T')$,	*(strangeness)*
(d)	$d = r - s$,	*(differential part)*
(e)	$u = n - r - a - s$	*(undetermined unknowns/vanishing equations)*

are invariant under the congruence relation (4.2).

Proof. To derive the condensed form (4.4) we use the following sequence of congruence transformations with orthogonal transformation matrices

$$\begin{aligned} (E, A) &\sim \left(\begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix} \right) \sim \left(\begin{bmatrix} \Sigma_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & \Sigma_a & 0 \\ A_{13}^T & 0 & 0 \end{bmatrix} \right), \\ &\sim \left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} & \Sigma_s & 0 \\ A_{12}^T & A_{22} & A_{23} & 0 & 0 \\ A_{13}^T & A_{23}^T & \Sigma_a & 0 & 0 \\ \Sigma_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right). \end{aligned}$$

To show the invariance of the quantities r, a, s, d, u under congruence transformations (4.2), we consider two matrix pairs (E_i, A_i), $i = 1, 2$ that are congruent, i.e., there exits a nonsingular matrix P such that

$$E_2 = P^T E_1 P, \quad A_2 = P^T A_1 P.$$

Since

$$\operatorname{rank} E_2 = \operatorname{rank}(P^T E_1 P) = \operatorname{rank} E_1,$$

it follows that r is invariant under congruence transformation. The quantities a and s are well-defined as they do not depend on the choice of the bases. Let T_2, T_2' and V_2 be the bases associated with (E_2, A_2), i.e.,

$$\begin{array}{lll} \operatorname{rank}(E_2 T_2) = 0, & T_2^T T_2 \text{ nonsingular}, & \operatorname{rank}(T_2^T T_2) = n - r, \\ \operatorname{rank}(E_2 T_2') = r, & T_2'^T T_2' \text{ nonsingular}, & \operatorname{rank}(T_2'^T T_2') = r, \\ \operatorname{rank}(V_2^T T_2^T A_2 T_2) = 0, & V_2^T V_2 \text{ nonsingular}, & \operatorname{rank}(V_2^T V_2) = k, \end{array}$$

with $k = \dim \operatorname{corange}(T_2^T A_2 T_2)$. Inserting the congruence relation (4.2) and defining

$$T_1 = P T_2, \quad T_1' = P T_2', \quad V_1^T = V_2^T,$$

we obtain the same relations for (E_1, A_1) with the matrices T_1 and T_1'. Hence, T_1 is a basis of kernel E_1 and T_1' is a basis of range E_1. Because of

$$k = \dim \operatorname{corange}(T_2^T A_2 T_2) = \dim \operatorname{corange}(T_2^T P^T A_1 P T_2) = \dim \operatorname{corange}(T_1^T A_1 T_1),$$

this also applies to V_1. With

$$\operatorname{rank}(T_2^T A_2 T_2) = \operatorname{rank}(T_2^T P^T A_1 P T_2) = \operatorname{rank}(T_1^T A_1 T_1),$$

and

$$\operatorname{rank}(V_2^T T_2^T A_2 T_2') = \operatorname{rank}(V_2^T T_2^T P^T A_1 P T_2') = \operatorname{rank}(V_1^T T_1^T A_1 T_1'),$$

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

Note, that the matrix pair (E, A) can be reduced further if we also allow non-orthogonal transformations, see e.g. [73].

4.2 Condensed Forms for Pairs of Symmetric Matrix-Valued Functions

After the introduction of a condensed form for symmetric matrix pairs in Section 4.1 we now consider condensed forms for pairs of symmetric matrix-valued functions. In [73] it has been posed as an open problem to derive a condensed form for linear differential-algebraic systems of the form (2.5), where $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ are symmetric, i.e., $E(t) = E^T(t)$, and $A(t) = A^T(t)$ for all $t \in \mathbb{I}$. Here, we will show that it is possible to derive a structure

preserving condensed form similar as in Theorem 4.3 for pairs of symmetric matrix-valued functions

$$(E(t), A(t)) \tag{4.5}$$

under certain additional assumptions. This form allows to characterize existence and uniqueness of solutions of the DAE as well as consistency of initial values similar as in Theorem 2.34 and Theorem 2.36. Since the concept of strong congruence transformations (4.2) is not adequate to treat time-varying systems, we need the concept of global congruence transformations.

Definition 4.4 (Global congruence). Two pairs of matrix-valued functions (E_i, A_i) with $E_i, A_i \in C(\mathbb{I}, \mathbb{R}^{n,n})$, $i = 1, 2$ are called *globally congruent* if there exists a pointwise nonsingular matrix-valued function $P \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$E_2 = P^T E_1 P, \quad A_2 = P^T A_1 P - P^T E_1 \dot{P}. \tag{4.6}$$

We can see that for a symmetric pair (E_1, A_1) the matrix pair (E_2, A_2) in (4.6) is only symmetric again if

$$P^T(t) E_1(t) \dot{P}(t) = \dot{P}^T(t) E_1(t) P(t), \quad \text{for all } t \in \mathbb{I}. \tag{4.7}$$

This condition holds for example in the special case where $E_1(t)\dot{P}(t) = 0$ for all $t \in \mathbb{I}$. Similar as in Section 3.1.1 we first consider the action of the congruence relation (4.6) locally at a fixed point, since in the numerical solution of differential-algebraic equations it is usually important to consider local quantities that are numerically computable and give information on the global behavior of the solution in the neighborhood of a fixed point $\hat{t} \in \mathbb{I}$. At a fixed point $\hat{t}$ we can choose $P(\hat{t}) = \tilde{P}$ and $\dot{P}(\hat{t}) = \tilde{R}$ independently, such that we get the following local version of the congruence relation (4.6).

Definition 4.5 (Local congruence). Two pairs of matrices (E_i, A_i), $i = 1, 2$ with $E_i, A_i \in \mathbb{R}^{n,n}$, are called *locally congruent* if there exist matrices $P, R \in \mathbb{R}^{n,n}$, with P nonsingular such that

$$E_2 = P^T E_1 P, \quad A_2 = P^T A_1 P - P^T E_1 R. \tag{4.8}$$

Again, for a symmetric pair (E_1, A_1) not every matrix R in (4.8) will lead again to a symmetric pair (E_2, A_2). To obtain a symmetric matrix pair we have to require that $P^T E_1 R = R^T E_1 P$, e.g., we can choose R such that $R = P$ or such that $E_1 R$ vanishes. Now, we can derive a local condensed form under the congruence transformation (4.8) for symmetric matrix pairs similar as in Theorem 4.3.

Theorem 4.6. *Let $E, A \in \mathbb{R}^{n,n}$ be symmetric and let*

$$\begin{aligned} T &\quad \textit{be a basis of } \operatorname{kernel} E, \\ T' &\quad \textit{be a basis of } \operatorname{cokernel} E = \operatorname{range} E, \\ V &\quad \textit{be a basis of } \operatorname{corange}(T^T A T). \end{aligned}$$

Then there exist an orthogonal matrix $P \in \mathbb{R}^{n,n}$ and a matrix $R \in \mathbb{R}^{n,n}$ such that the matrix pair (E, A) is locally congruent to a symmetric matrix pair of the form

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & A_{13} & \Sigma_s & 0 \\ 0 & 0 & A_{23} & 0 & 0 \\ A_{13}^T & A_{23}^T & \Sigma_a & 0 & 0 \\ \Sigma_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right), \quad \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} \tag{4.9}$$

where the matrices $\Sigma_a \in \mathbb{R}^{a,a}$, and $\Sigma_s \in \mathbb{R}^{s,s}$ are nonsingular and diagonal, the matrix $\begin{bmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{bmatrix} \in \mathbb{R}^{r,r}$ is nonsingular, and the last block rows and block columns are of dimension u. Further, the quantities

(a)	$r = \operatorname{rank} E$,	*(rank)*
(b)	$a = \operatorname{rank}(T^T A T)$,	*(algebraic part)*
(c)	$s = \operatorname{rank}(V^T T^T A T')$,	*(strangeness)*
(d)	$d = r - s$,	*(differential part)*
(e)	$u = n - r - a - s$	*(undetermined unknowns/vanishing equations)*

are invariant under the congruence relation (4.8).

Proof. Following the proof of Theorem 4.3 we have the following sequence of congruence transformations

$$\begin{aligned} (E, A) &\sim \left(\begin{bmatrix} \Sigma_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & \Sigma_a & 0 \\ A_{13}^T & 0 & 0 \end{bmatrix} \right), \quad \text{with } R = 0, \\ &\sim \left(\begin{bmatrix} \Sigma_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & A_{12} & A_{13} \\ A_{12}^T & \Sigma_a & 0 \\ A_{13}^T & 0 & 0 \end{bmatrix} \right), \quad \text{with } R = \begin{bmatrix} \Sigma_r^{-1} A_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \end{aligned}$$

which is congruent to the matrix pair in (4.9). The invariance of the local quantities can be shown in the same way as in the proof of Theorem 4.3. For two matrix pairs (E_i, A_i), $i = 1, 2$ that are local congruent, i.e., there exist a nonsingular matrix P and a matrix R such that

$$E_2 = P^T E_1 P, \quad A_2 = P^T A_1 P - P^T E_1 R,$$

we consider the corresponding bases T_2, T_2' and V_2 associated with (E_2, A_2). Then, similar as in the proof of Theorem 4.3, the matrices

$$T_1 = P T_2, \quad T_1' = P T_2', \quad V_1^T = V_2^T$$

form the corresponding bases associated with (E_1, A_1), since

$$\begin{aligned} k &= \dim \operatorname{corange}(T_2^T A_2 T_2) = \dim \operatorname{corange}(T_2^T P^T A_1 P T_2 - T_2^T P^T E_1 R T_2) \\ &= \dim \operatorname{corange}(T_1^T A_1 T_1). \end{aligned}$$

With

$$\operatorname{rank}(T_2^T A_2 T_2) = \operatorname{rank}(T_2^T P^T A_1 P T_2 - T_2^T P^T E_1 R T_2) = \operatorname{rank}(T_1^T A_1 T_1)$$

and

$$\operatorname{rank}(V_2^T T_2^T A_2 T_2') = \operatorname{rank}(V_2^T T_2^T P^T A_1 P T_2' - V_2^T T_2^T P^T E_1 R T_2') = \operatorname{rank}(V_1^T T_1^T A_1 T_1'),$$

we get the invariance of a and s and therefore also of d and u. □

If we also allow non-orthogonal transformations, then we can reduce the matrix pair further, see e.g. [73]. For the pair of matrix-valued functions (4.5) the local condensed form (4.9) can be computed for each fixed value $\hat{t} \in \mathbb{I}$. Then, we obtain integer-valued functions $d, a, s, u : \mathbb{I} \to \mathbb{N}_0$, and we assume that the regularity assumptions

$$d(t) \equiv d, \quad a(t) \equiv a, \quad s(t) \equiv s, \quad u(t) \equiv u, \quad \text{for all } t \in \mathbb{I} \tag{4.10}$$

hold, i.e., the ranks of the matrices and the sizes of the blocks in the local condensed form (4.9) do not depend on $t \in \mathbb{I}$. This restriction then allows for the application of the following property of a symmetric matrix-valued function of constant rank.

Lemma 4.7. *Let $E \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$, $k \in \mathbb{N}_0 \cup \{\infty\}$ be symmetric, with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. Then there exists a pointwise orthogonal matrix-valued function $P \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$ such that*

$$P^T(t) E(t) P(t) = \begin{bmatrix} \Delta_r(t) & 0 \\ 0 & 0 \end{bmatrix}$$

where $\Delta_r \in C^k(\mathbb{I}, \mathbb{R}^{r,r})$ is pointwise nonsingular and symmetric for all $t \in \mathbb{I}$.

Proof. The general non-symmetric version of the Theorem is proved e.g., in [82, Theorem 3.9] for complex matrix-valued functions or in [137, Theorem 2.1.4] for the real case. For symmetric matrix-valued functions the result follows from [110]. □

In order to preserve the symmetry of a pair of matrix-valued functions under global congruence transformation (4.6) we need to ensure that condition (4.7) holds. Therefore, additionally we need the following Assumption and Lemma.

Assumption 4.8. *Let $E \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be symmetric, with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. There exists a constant matrix $Q \in \mathbb{R}^{n,r}$ such that the columns of Q form an orthogonal basis of $\operatorname{range} E$ for all $t \in \mathbb{I}$.*

Lemma 4.9. *Let $E \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ be symmetric, with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$ and assume that Assumption 4.8 holds. Then there exists an orthogonal matrix $P \in \mathbb{R}^{n,n}$ such that*

$$P^T E(t) P = \begin{bmatrix} \Delta_r(t) & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{for all } t \in \mathbb{I},$$

with pointwise nonsingular and symmetric $\Delta_r \in C^1(\mathbb{I}, \mathbb{R}^{r,r})$.

Proof. Due to Assumption 4.8, there exists a matrix $Q \in \mathbb{R}^{n,r}$ such that the columns of Q form an orthogonal basis of range E for all $t \in \mathbb{I}$. Then, we can find a matrix $Q' \in \mathbb{R}^{n,n-r}$ such that $\begin{bmatrix} Q & Q' \end{bmatrix}$ is orthogonal and Q' is a basis of corange $E(\hat{t}) = \operatorname{kernel} E(\hat{t})$ for each $\hat{t} \in \mathbb{I}$. Thus,

$$\begin{bmatrix} Q^T \\ Q'^T \end{bmatrix} E(\hat{t}) \begin{bmatrix} Q & Q' \end{bmatrix} = \begin{bmatrix} \Delta_r(\hat{t}) & 0 \\ 0 & 0 \end{bmatrix}$$

for each $\hat{t} \in \mathbb{I}$, with pointwise nonsingular $\Delta_r \in C^1(\mathbb{I}, \mathbb{R}^{r,r})$. □

Using Lemma 4.9 we can derive the following global condensed form for pairs of symmetric matrix-valued functions.

Theorem 4.10. *Let the pair $(E(t), A(t))$ of matrix-valued function $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and symmetric with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. Suppose that the regularity assumptions (4.10) hold and that E fulfills Assumption 4.8. Then the pair $(E(t), A(t))$ is globally congruent to a pair of symmetric matrix-valued functions of the form*

$$\left(\begin{bmatrix} \Delta_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & 0 & S_{13} & 0 \\ 0 & \Delta_a & 0 & 0 \\ S_{13}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right), \quad \begin{matrix} r \\ a \\ s \\ u \end{matrix} \tag{4.11}$$

where the matrix-valued functions $\Delta_r \in C(\mathbb{I}, \mathbb{R}^{r,r})$ and $\Delta_a \in C(\mathbb{I}, \mathbb{R}^{a,a})$ are pointwise nonsingular and $S_{13} \in C(\mathbb{I}, \mathbb{R}^{r,s})$ has pointwise full column rank.

Proof. We give a constructive proof using Lemma 4.7 and Lemma 4.9. First, we can determine an orthogonal matrix $P_1 \in \mathbb{R}^{n,n}$ such that

$$E_1 := P_1^T E P_1 = \begin{bmatrix} \Delta_r & 0 \\ 0 & 0 \end{bmatrix}, \quad A_1 := P_1^T A P_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix},$$

where $\Delta_r \in C(\mathbb{I}, \mathbb{R}^{r,r})$ is symmetric and pointwise nonsingular. As $\operatorname{rank} A_{22} = a$ is constant in $\mathbb{I}$, we can determine a pointwise orthogonal matrix-valued function $P_2 \in C(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$E_2 := P_2^T E_1 P_2 = \begin{bmatrix} \Delta_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_2 := P_2^T A_1 P_2 - P_2^T E_1 \dot{P}_2 = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{12}^T & \Delta_a & 0 \\ A_{13}^T & 0 & 0 \end{bmatrix},$$

with pointwise nonsingular $\Delta_a \in C(\mathbb{I}, \mathbb{R}^{a,a})$ and $E_1 \dot{P}_2 = 0$ for all $t \in \mathbb{I}$. Next, we can eliminate the blocks A_{12} and A_{12}^T with a transformation $P_3 \in C(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$E_3 := P_3^T E_2 P_3 = \begin{bmatrix} \Delta_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad A_3 := P_3^T A_2 P_3 - P_3^T E_2 \dot{P}_3 = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & \Delta_a & 0 \\ A_{13}^T & 0 & 0 \end{bmatrix},$$

where again $E_2\dot{P}_3 = 0$ for all $t \in \mathbb{I}$. Further, as $\operatorname{rank} A_{13} = s$ for all $t \in \mathbb{I}$ we can find a pointwise orthogonal matrix-valued function $\hat{P}_4$ such that $A_{13}\hat{P}_4 = \begin{bmatrix} S_{13} & 0 \end{bmatrix}$, with $S_{13} \in C(\mathbb{I}, \mathbb{R}^{r,s})$ of pointwise full column rank. Choosing P_4 accordingly, we get

$$E_4 := P_4^T E_3 P_4 = \begin{bmatrix} \Delta_r & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_4 := P_4^T A_3 P_4 - P_4^T E_3 \dot{P}_4 = \begin{bmatrix} A_{11} & 0 & S_{13} & 0 \\ 0 & \Delta_a & 0 & 0 \\ S_{13}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

with $E_3\dot{P}_4 = 0$ for all $t \in \mathbb{I}$. □

Under the assumptions of Theorem 4.10 we can now transform the symmetric pair (4.5) to the global condensed form (4.11). The restrictions due to the regularity assumptions (4.10) imply that the previous considerations can be applied only on a dense subset of the given closed interval as has been discussed in Section 2.2.4.

In order to obtain a strangeness-free formulation we have to eliminate the 'strange' coupled parts by differentiating and eliminating certain parts of the system until we have $s = 0$ in the global condensed form (4.11). Let $(\tilde{E}, \tilde{A})$ be the pair of matrix-valued functions in global condensed form (4.11). As $\operatorname{rank} S_{13}(t) = s$ for all $t \in \mathbb{I}$ we can find a pointwise orthogonal matrix function $\hat{Q} \in C(\mathbb{I}, \mathbb{R}^{r,r})$ such that we can decompose S_{13} into $S_{13}^T\hat{Q}^T = \begin{bmatrix} \Delta_s^T & 0 \end{bmatrix}$ with pointwise nonsingular $\Delta_s \in C(\mathbb{I}, \mathbb{R}^{s,s})$. Setting $P^T = \begin{bmatrix} \hat{Q} & 0 & 0 & 0 \\ 0 & I_a & 0 & 0 \\ 0 & 0 & I_s & 0 \\ 0 & 0 & 0 & I_u \end{bmatrix}$, we get the congruent matrix pair

$$\begin{aligned} &(P^T\tilde{E}P, P^T\tilde{A}P - P^T\tilde{E}\dot{P}) \\ &= \left(\begin{bmatrix} \hat{Q}\Delta_r\hat{Q}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} \hat{Q}A_{11}\hat{Q}^T & 0 & \hat{Q}S_{13} & 0 \\ 0 & \Delta_a & 0 & 0 \\ S_{13}^T\hat{Q}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} - \begin{bmatrix} \hat{Q}\Delta_r\dot{\hat{Q}}^T & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \right). \end{aligned}$$

Again, symmetry of the matrix pair is only preserved if $\hat{Q}\Delta_r\dot{\hat{Q}}^T = \dot{\hat{Q}}\Delta_r\hat{Q}^T$. If Assumption 4.8 holds also for S_{13}, then $\hat{Q}$ can be chosen as a constant matrix and we have $\dot{\hat{Q}} = 0$ such that the resulting pair denoted by

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & 0 & \Delta_s & 0 \\ A_{12}^T & A_{22} & 0 & 0 & 0 \\ 0 & 0 & \Delta_a & 0 & 0 \\ \Delta_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right),$$

where $\begin{bmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{bmatrix} = \hat{Q}\Delta_r\hat{Q}^T$ is again symmetric. Further, we can eliminate certain blocks via block Gaussian elimination to get the equivalent pair

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & 0 & \Delta_s & 0 \\ 0 & A_{22} & 0 & 0 & 0 \\ 0 & 0 & \Delta_a & 0 & 0 \\ \Delta_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \right). \tag{4.12}$$

Note that for this transformations no requirement to keep symmetry is needed. The DAE associated with the pair (4.12) can be written as

$$\begin{aligned} E_{11}\dot{x}_1 + E_{12}\dot{x}_2 &= \Delta_s x_4 + b_1, \\ E_{12}^T\dot{x}_1 + E_{22}\dot{x}_2 &= A_{22}x_2 + b_2, \\ 0 &= \Delta_a x_3 + b_3, \\ 0 &= \Delta_s^T x_1 + b_4, \\ 0 &= b_5. \end{aligned} \tag{4.13}$$

Now, we can use the derivatives of the fourth equation in (4.13) to eliminate the terms with $\dot{x}_1$ in the first two equations to get

$$\begin{aligned} E_{12}\dot{x}_2 &= \Delta_s x_4 + \tilde{b}_1, \\ E_{22}\dot{x}_2 &= A_{22}x_2 + \tilde{b}_2, \\ 0 &= \Delta_a x_3 + b_3, \\ 0 &= \Delta_s^T x_1 + b_4, \\ 0 &= b_5, \end{aligned} \tag{4.14}$$

where $\tilde{b}_1 = b_1 + E_{11}\frac{d}{dt}\left(\Delta_s^{-T}b_4\right)$, $\tilde{b}_2 = b_2 + E_{12}^T\frac{d}{dt}\left(\Delta_s^{-T}b_4\right)$. This step only affects the right hand side such that the symmetry of each block matrix is preserved. But, the symmetry of the overall system cannot be preserved after this elimination step due to the occurrence of the block E_{12}. In general, the existence and uniqueness of solutions of the system (4.14) depend on the rank of $\begin{bmatrix} E_{12} \\ E_{22} \end{bmatrix}$. If $r = s$, then the blocks E_{12} and E_{22} do not occur and the system reduces after one differentiation and elimination step to a purely algebraic equation such that the system is of strangeness index $\mu = 1$. If E_{22} is invertible, then all information concerning existence and uniqueness is available and no further steps are needed. Thus, if $r - s > 0$ and E_{22} in (4.12) is nonsingular, the system (4.12) has strangeness index $\mu \leq 1$. Assuming that E_{22} is nonsingular we can eliminate the block E_{12} and get the following

system

$$\begin{aligned} 0 &= \Delta_s x_4 - E_{12}E_{22}^{-1}A_{22}x_2 + \tilde{b}_1 - E_{12}E_{22}^{-1}\tilde{b}_2, \\ E_{22}\dot{x}_2 &= A_{22}x_2 + \tilde{b}_2, \\ 0 &= \Delta_a x_3 + b_3, \\ 0 &= \Delta_s^T x_1 + b_4, \\ 0 &= b_5. \end{aligned} \tag{4.15}$$

Further, we can eliminate the term $E_{12}E_{22}^{-1}A_{22}$ with a block Gaussian elimination using the invertible block Δ_s to get a strangeness-free system which is again symmetric. Rearanging and renaming the matrices and vector-valued functions finally yields the strangeness-free symmetric system

$$\begin{aligned} \hat{E}_{11}(t)\dot{\hat{x}}_1 &= \hat{A}_{11}(t)\hat{x}_1 + \hat{b}_1(t), \\ 0 &= \hat{A}_{22}(t)\hat{x}_2 + \hat{b}_2(t), \\ 0 &= \hat{b}_3(t), \end{aligned} \tag{4.16}$$

consisting of $d_\mu = r - s$ differential equations, $a_\mu = a + 2s$ algebraic equations, and $u_\mu = u$ vanishing equations, where $\hat{E}_{11}$ and $\hat{A}_{22}$ are nonsingular. Here, we have

$$\hat{E}_{11} = E_{22}, \quad \hat{A}_{11} = A_{22}, \quad \hat{A}_{22} = \begin{bmatrix} \Delta_a & 0 & 0 \\ 0 & 0 & \Delta_s \\ 0 & \Delta_s^T & 0 \end{bmatrix},$$

$$\hat{x}_1 = x_2, \quad \hat{x}_2 = \begin{bmatrix} x_3 \\ x_1 \\ x_4 - \Delta_s^{-1}E_{12}E_{22}^{-1}A_{22}x_2 \end{bmatrix}, \quad \hat{x}_3 = x_5,$$

$$\hat{b}_1 = \tilde{b}_2, \quad \hat{b}_2 = \begin{bmatrix} b_3 \\ \tilde{b}_1 - E_{12}E_{22}^{-1}\tilde{b}_2 \\ b_4 \end{bmatrix}, \quad \hat{b}_3 = b_5.$$

Example 4.11. We consider the symmetric linear differential-algebraic system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & t & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} 0 & 0 & t \\ 0 & 1 & 0 \\ t & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix},$$

in an interval $\mathbb{I} = [t_0, t_1]$ with $t_0 > 0$. This system has strangeness index $\mu = 1$ and is already given in the form (4.12), where $E_{22} = t$ is nonsingular for all $t \neq 0$. The differentiation-and-elimination step yields the strangeness-free system

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & t & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \end{bmatrix} = \begin{bmatrix} \frac{1}{t} & 0 & t \\ 0 & 1 & 0 \\ t & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} b_1 + \frac{1}{t}\dot{b}_3 \\ b_2 \\ b_3 \end{bmatrix},$$

which is again symmetric.

From the previous discussion it follows that for linear differential-algebraic systems with symmetric coefficients, even under Assumption 4.8, a structure preserving strangeness-free formulation (4.16) in general only exists if the strangeness index is $\mu \leq 1$. For systems with strangeness index $\mu > 1$ the inverse E_{22}^{-1} does not exist and usually we cannot preserve the symmetry of the strangeness-free system in this case.

Example 4.12. Consider the linear symmetric differential-algebraic system

$$\left[\begin{array}{cc|cc|cc} 0 & 0 & t & 0 & 0 & 0 \\ 0 & 0 & 0 & t & 0 & 0 \\ \hline t & 0 & 0 & 0 & 0 & 0 \\ 0 & t & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \\ \dot{x}_6 \end{bmatrix} = \left[\begin{array}{cc|cc|cc} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right] \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix},$$

in an interval $\mathbb{I} = [t_0, t_1]$ with $t_0 > 0$ and $b \in C^2([t_0, t_1], \mathbb{R}^6)$. This system has strangeness index $\mu = 2$ and Assumption 4.8 is fulfilled. The corresponding matrix pair is already in the form (4.12) and $E_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ is clearly singular. Differentiating the last two equations of the system and inserting the derivative into the third and fourth equation yields the corresponding matrix pair

$$\left(\left[\begin{array}{cc|cc|cc} 0 & 0 & t & 0 & 0 & 0 \\ 0 & 0 & 0 & t & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc|cc} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right] \right).$$

In a second step differentiating the third equation and eliminating the corresponding derivative in the first equation yields the matrix pair

$$\left(\left[\begin{array}{cc|cc|cc} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & t & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right], \left[\begin{array}{cc|cc|cc} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \end{array}\right] \right).$$

The second matrix is still symmetric, but the symmetry of the first matrix cannot be preserved.

From the previous discussion we get another result.

Corollary 4.13. *Consider a linear differential-algebraic equation* (2.5) *with symmetric coefficient matrices* $(E(t), A(t))$*, where* $E(t)$ *is positive semidefinite for all* $t \in \mathbb{I}$*. Then the system has strangeness index* $\mu \leq 1$*.*

Proof. The proof follows directly from the previous discussion. If E is positive semidefinite then the matrix $\begin{bmatrix} E_{11} & E_{12} \\ E_{12}^T & E_{22} \end{bmatrix}$ in (4.12) is positive definite and thus E_{22} is positive definite and therefore nonsingular. □

In conclusion, we can see that we can obtain a structure preserving condensed form (4.11) for symmetric systems only under the strong Assumption 4.8 and a strangeness-free system (4.16) only for symmetric differential-algebraic systems of s-index $\mu \leq 1$, also under Assumption 4.8. This is not very convenient, since too many restrictions and assumptions have to be made in order to preserve the structure using global congruence transformations (4.6). Other index reduction techniques can be applied that allow to preserve the symmetry of the system as will be seen in Section 4.4.

4.3 Condensed Forms for Self-Adjoint Differential-Algebraic Equations

We have seen in Section 4.2 that a structure preserving condensed form for symmetric pairs of matrix-valued functions under global congruence transformations only exists under the strong Assumption 4.8. It turns out that for linear time-dependent differential-algebraic systems a better structure to consider is self-adjointness of the system, as arises e.g. in linear-quadratic optimal control problems (1.3) or in gyroscopic mechanical systems, since global congruence transformations preserve the self-adjointness of the system.
In this section we consider linear differential-algebraic systems of the form (2.5), with sufficiently smooth matrix-valued functions $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ on an interval $\mathbb{I} = [t_0, t_1]$. In order to define the adjoint equation of a linear DAE we consider the *differential-algebraic operator*

$$D : \mathbb{X} \to \mathbb{Y}, \quad Dx(t) = E(t)\dot{x}(t) - A(t)x(t) \tag{4.17}$$

according to (2.5), with the function spaces

$$\begin{aligned} \mathbb{X} &= \{x \in C^0(\mathbb{I}, \mathbb{R}^n) \mid E^+Ex \in C^1(\mathbb{I}, \mathbb{R}^n),\ E^+Ex(t_0) = 0\}, \\ \mathbb{Y} &= C^0(\mathbb{I}, \mathbb{R}^n), \end{aligned} \tag{4.18}$$

where E^+ denotes the Moore-Penrose pseudo-inverse of E (see Definition 2.17), see also [82, Section 3.4]. At first, we assume that the pair (E, A) is regular and strangeness-free. Note that the function space $\mathbb{X}$ corresponds to the solution space of the DAE using the concept of strong solutions (see Definition 5.14). Further, we assume without loss of generality that we have homogeneous initial conditions. This can always be achieved by shifting $x(t)$ to $x(t) - x_0$ and changing the inhomogeneity $b(t)$ to $b(t) + A(t)x_0$. Furthermore, a sesquilinear form on $\mathbb{Y} \times \mathbb{Y}$ is defined by

$$(f, g) = \int_{\mathbb{I}} f^T(t)g(t)dt \tag{4.19}$$

for $f, g \in C^0(\mathbb{I}, \mathbb{R}^n)$.

Definition 4.14 (Sesquilinear form). Consider two real vector spaces $\mathbb{X}$, $\mathbb{X}^\star$. A mapping $(.,.) : \mathbb{X} \times \mathbb{X}^\star \to \mathbb{R}$ is called a *sesquilinear form* if

$$\begin{aligned} &(a) \qquad (x, x^\star + y^\star) = (x, x^\star) + (x, y^\star), \\ &(b) \qquad (x, \alpha x^\star) = \alpha(x, x^\star), \\ &(c) \qquad (x + y, x^\star) = (x, x^\star) + (y, x^\star), \\ &(d) \qquad (\alpha x, x^\star) = \alpha(x, x^\star), \end{aligned}$$

for all $x, y \in \mathbb{X}$, $x^\star, y^\star \in \mathbb{X}^\star$ and $\alpha \in \mathbb{R}$.

Then, we can define a conjugate operator corresponding to the differential-algebraic operator D, see also [82].

Definition 4.15 (Conjugate operator). Given a linear differential-algebraic operator $D : \mathbb{X} \to \mathbb{Y}$ as in (4.17). Then a *conjugate operator* is defined as $D^\star : \mathbb{Y}^\star \to \mathbb{X}^\star$ such that

$$(Dx, y) = (x, D^\star y) \quad \text{for all } x \in \mathbb{X},\ y \in \mathbb{Y}^\star$$

with sesquilinear form $(.,.)$ as in (4.19) and function spaces

$$\begin{aligned} &\mathbb{X}^\star = C^0(\mathbb{I}, \mathbb{R}^n), \\ &\mathbb{Y}^\star = \{y \in C^0(\mathbb{I}, \mathbb{R}^n) \,|\, EE^+ y \in C^1(\mathbb{I}, \mathbb{R}^n),\ EE^+ y(t_1) = 0\}. \end{aligned} \tag{4.20}$$

Theorem 4.16. *The differential-algebraic operator $D : \mathbb{X} \to \mathbb{Y}$ defined in* (4.17) *with regular and strangeness-free pair (E, A) has a unique conjugate operator $D^\star : \mathbb{Y}^\star \to \mathbb{X}^\star$ with function spaces defined as in* (4.20) *that is given by*

$$D^\star y = -\frac{d}{dt}(E^T y) - A^T y.$$

Proof. We have

$$\begin{aligned} (Dx, y) &= \int_{\mathbb{I}} (E\dot{x} - Ax)^T y \, dt \\ &= \int_{\mathbb{I}} (\dot{x}^T E^T y - x^T A^T y) \, dt \\ &= \int_{\mathbb{I}} \left(\frac{d}{dt}(x^T E^T y) - x^T \dot{E}^T y - x^T E^T \dot{y} - x^T A^T y \right) dt \\ &= \left[x^T E^T y \right]_{t_0}^{t_1} + \int_{\mathbb{I}} x^T \left(-\frac{d}{dt}(E^T y) - A^T y \right) dt \\ &= (x, D^\star y), \end{aligned}$$

where $\left[x^T E^T y \right]_{t_0}^{t_1} = 0$, since

$$x^T E^T y = x^T (EE^+ E)^T y = x^T E^+ E E^T y = x^T E^T E E^+ y$$

due to the properties (2.1) of the Moore-Penrose pseudo-inverse (see Definition 2.17) and

$$x(t_1)^T E(t_1)^T y(t_1) = x^T E^T E E^+ y(t_1) = 0,$$
$$x(t_0)^T E(t_0)^T y(t_0) = x^T(t_0) E^+ E E^T y(t_0) = 0,$$

due to the homogenous initial conditions in $\mathbb{X}$ and $\mathbb{Y}^\star$. To show uniqueness we assume that there exists another conjugate $\tilde{D}^\star$ for D. Then it holds that

$$(Dx, y) = (x, D^\star y) \quad \text{for all } x \in \mathbb{X},\ y \in \mathbb{Y}^\star,$$

as well as

$$(Dx, y) = (x, \tilde{D}^\star y) \quad \text{for all } x \in \mathbb{X},\ y \in \mathbb{Y}^\star.$$

Thus,

$$0 = (x, \tilde{D}^\star y) - (x, D^\star y) \quad \text{for all } x \in \mathbb{X},\ y \in \mathbb{Y}^\star,$$

and therefore

$$\tilde{D}^\star y - D^\star y = 0 \quad \text{for all } y \in \mathbb{Y}^\star.$$

□

Due to Theorem 4.16, the differential-algebraic operator D belonging to a regular and strangeness-free pair of matrix-valued functions (E, A) has a unique conjugate operator $D^\star$ that can be described by the pair $(-E^T, (A + \dot{E})^T)$. For arbitrary pairs of matrix-valued functions we can now introduce the following terminology.

Definition 4.17 (Adjoint). For a pair of matrix-valued functions (E, A) with $E \in C^1(\mathbb{I}, \mathbb{R}^{m,n})$ and $A \in C^0(\mathbb{I}, \mathbb{R}^{m,n})$, the pair $(-E^T, (A + \dot{E})^T)$ is called the *adjoint* of (E, A).

Due to Definition 4.17, the linear differential-algebraic equation

$$-E^T \dot{y} = (A + \dot{E})^T y + g(t), \tag{4.21}$$

with $g \in C(\mathbb{I}, \mathbb{R}^m)$ is also called the *adjoint differential-algebraic equation* of (2.5), see also [11, 83].

Lemma 4.18. *The adjoint of a pair of matrix-valued functions (E, A) with $E \in C^1(\mathbb{I}, \mathbb{R}^{m,n})$ and $A \in C^0(\mathbb{I}, \mathbb{R}^{m,n})$ has itself an adjoint which corresponds to (E, A).*

Proof. As $-E^T \in C^1(\mathbb{I}, \mathbb{R}^{n,m})$ and $(A + \dot{E})^T \in C^0(\mathbb{I}, \mathbb{R}^{n,m})$ the adjoint of (E, A) has itself an adjoint, which is given by

$$(-(-E^T)^T, ((A + \dot{E})^T + (-\dot{E}^T))^T) = (E, A).$$

□

Global equivalence transformations of the form (2.11) and the transfer to the adjoint are commutative. In particular, the adjoints of equivalent pairs of matrix-valued functions are again equivalent.

Theorem 4.19. *Let (E, A) be a pair of matrix-valued functions with $E \in C^1(\mathbb{I}, \mathbb{R}^{m,n})$ and $A \in C^0(\mathbb{I}, \mathbb{R}^{m,n})$ and consider the globally equivalent pair*

$$(\tilde{E}, \tilde{A}) = (PEQ, PAQ - PE\dot{Q}),$$

where $P \in C^1(\mathbb{I}, \mathbb{R}^{m,m})$ and $Q \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ are pointwise nonsingular. Then, the adjoints of (E, A) and $(\tilde{E}, \tilde{A})$ are again globally equivalent.

Proof. The adjoint of $(\tilde{E}, \tilde{A})$ is given by

$$\begin{aligned}(-\tilde{E}^T, (\tilde{A} + \dot{\tilde{E}})^T) &= (-(PEQ)^T, (PAQ - PE\dot{Q} + \tfrac{d}{dt}(PEQ))^T) \\ &= (-Q^T E^T P^T, Q^T A^T P^T + Q^T \dot{E}^T P^T + Q^T E^T \dot{P}^T).\end{aligned}$$

On the other hand, equivalence transformation of the adjoint $(-E^T, (A + \dot{E})^T)$ of (E, A) with Q^T and P^T yields

$$\begin{aligned}(-E^T, (A + \dot{E})^T) &\sim (-Q^T E^T P^T, Q^T (A + \dot{E})^T P^T + Q^T E^T \dot{P}^T) \\ &= (-Q^T E^T P^T, Q^T A^T P^T + Q^T \dot{E}^T P^T + Q^T E^T \dot{P}^T).\end{aligned}$$

□

Now, we can define self-adjointness for pairs of matrix-valued functions.

Definition 4.20 (Self-adjointness). A pair of matrix-valued functions (E, A) with $E \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ and $A \in C^0(\mathbb{I}, \mathbb{R}^{n,n})$ is called *self-adjoint* if it holds that

$$E = -E^T, \quad A = (A + \dot{E})^T \quad \text{for all } t \in \mathbb{I}.$$

A linear differential-algebraic system (2.5) is called *self-adjoint* if the corresponding matrix pair (E, A) is self-adjoint.

Self-adjoint matrix pairs arise for example in linear-quadratic optimal control problems (1.3), or in gyroscopic mechanical systems.

Example 4.21. Consider a constraint gyroscopic mechanical system

$$\begin{aligned}M\ddot{p} + C\dot{p} + Kp &= f(t) + G^T\lambda, \\ Gp &= 0,\end{aligned}$$

with $M, C, K \in \mathbb{R}^{n,n}$ and $M = M^T$ positive definite, $C = -C^T$ and $K = K^T$ positive definite. Then, a structure preserving first order formulation introducing the new variable $v = \dot{p}$ is given by the second companion form

$$\begin{bmatrix} C & M & 0 \\ -M & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{p} \\ \dot{v} \\ \dot{\lambda} \end{bmatrix} = \begin{bmatrix} -K & 0 & G^T \\ 0 & -M & 0 \\ G & 0 & 0 \end{bmatrix} \begin{bmatrix} p \\ v \\ \lambda \end{bmatrix} + \begin{bmatrix} f(t) \\ 0 \\ 0 \end{bmatrix},$$

and this system is self-adjoint.

In contrast to the symmetric structure global congruence transformations of the form (4.6) preserve the self-adjoint structure of a pair of matrix-valued functions.

Lemma 4.22. *Let a pair of matrix-valued functions (E, A) with $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and self-adjoint. Then for each pointwise nonsingular matrix-valued function $P \in C^1(\mathbb{I}, \mathbb{R}^{n,n})$ the global congruent pair*

$$(\tilde{E}, \tilde{A}) = (P^T E P, P^T A P - P^T E \dot{P})$$

is also self-adjoint.

Proof. It holds that

$$\tilde{E}^T = P^T E^T P = -P^T E P = -\tilde{E},$$

as well as

$$\begin{aligned}(\tilde{A} + \dot{\tilde{E}})^T &= (P^T A P - P^T E \dot{P})^T + \tfrac{d}{dt}(P^T E P)^T \\ &= P^T A^T P - \dot{P}^T E^T P + \dot{P}^T E^T P + P^T \dot{E}^T P + P^T E^T \dot{P} \\ &= P^T (A + \dot{E})^T P + P^T E^T \dot{P} \\ &= P^T A P - P^T E \dot{P} = \tilde{A}.\end{aligned}$$

□

To derive a global condensed form for self-adjoint pairs of matrix-valued functions we use the following factorization for skew-symmetric matrix-valued functions.

Lemma 4.23. *Let $A \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$, $k \in \mathbb{N}_0 \cup \{\infty\}$ be skew-symmetric, i.e., $A = -A^T$, with $\operatorname{rank} A(t) = r$ for all $t \in \mathbb{I}$. Then there exists a pointwise orthogonal matrix-valued function $P \in C^k(\mathbb{I}, \mathbb{R}^{n,n})$ such that*

$$P^T(t) A(t) P(t) = \begin{bmatrix} \Sigma(t) & 0 \\ 0 & 0 \end{bmatrix},$$

with pointwise nonsingular and skew-symmetric $\Sigma \in C^k(\mathbb{I}, \mathbb{R}^{r,r})$.

Proof. The Schur decomposition for skew-symmetric matrices is given in [54]. Then, the Lemma follows in the same way as Theorem 2.25 and Lemma 4.9. □

We can derive a local condensed form for self-adjoint pairs of matrix-valued functions similar as in Theorem 4.6 with the corresponding invariant characteristic quantities.

Theorem 4.24. *Let $E, A \in \mathbb{R}^{n,n}$ and (E, A) be self-adjoint. Further, let*

$$\begin{aligned} T &\quad \text{be a basis of } \operatorname{kernel} E, \\ T' &\quad \text{be a basis of } \operatorname{cokernel} E = \operatorname{range} E, \\ V &\quad \text{be a basis of } \operatorname{corange}(T^T A T). \end{aligned}$$

Then there exists an orthogonal matrix $P \in \mathbb{R}^{n,n}$ and a matrix $R \in \mathbb{R}^{n,n}$ such that the matrix pair (E, A) is locally congruent to a self-adjoint matrix pair of the form

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ -E_{12}^T & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 & A_{13} & \Sigma_s & 0 \\ 0 & 0 & A_{23} & 0 & 0 \\ A_{13}^T & A_{23}^T & \Sigma_a & 0 & 0 \\ \Sigma_s & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} \tag{4.22}$$

where the matrices $\Sigma_a \in \mathbb{R}^{a,a}$ and $\Sigma_s \in \mathbb{R}^{s,s}$ are nonsingular and diagonal, the matrix $\begin{bmatrix} E_{11} & E_{12} \\ -E_{12}^T & E_{22} \end{bmatrix} \in \mathbb{R}^{r,r}$ is nonsingular, and the last block rows and block columns are of dimension u. Further, the quantities

(a)	$r = \operatorname{rank} E$,	*(rank)*
(b)	$a = \operatorname{rank}(T^T A T)$,	*(algebraic part)*
(c)	$s = \operatorname{rank}(V^T T^T A T')$,	*(strangeness)*
(d)	$d = r - s$,	*(differential part)*
(e)	$u = n - r - a - s$	*(undetermined unknowns/vanishing equations)*

are invariant under the congruence relation (4.8).

Proof. The proof is analogous to the proof of Theorem 4.6. □

We can also derive a global condensed form for self-adjoint pairs of matrix-valued functions under the regularity assumptions (4.10).

Theorem 4.25. *Let the pair (E, A) of matrix-valued functions $E, A \in C(\mathbb{I}, \mathbb{R}^{n,n})$ be sufficiently smooth and self-adjoint with $\operatorname{rank} E(t) = r$ for all $t \in \mathbb{I}$. Suppose that a regularity assumption as in* (4.10) *holds. Then the pair (E, A) is globally congruent to a self-adjoint pair of matrix-valued functions of the form*

$$\left(\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \begin{bmatrix} A_{11} & A_{12} & 0 & \Sigma_s & 0 \\ 0 & A_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}\right) \begin{matrix} s \\ d \\ a \\ s \\ u \end{matrix} \tag{4.23}$$

where all blocks are again matrix-valued functions, the matrices $\Sigma_a = \Sigma_a^T$ and Σ_s are pointwise nonsingular, and $\begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$ is pointwise nonsingular and skew-symmetric.

Proof. Again, we give a constructive proof. First, we determine a pointwise orthogonal matrix-valued function $P_1 \in C(\mathbb{I}, \mathbb{R}^{n,n})$ such that

$$E_1 := P_1^T E P_1 = \begin{bmatrix} \Sigma_r & 0 \\ 0 & 0 \end{bmatrix}, \qquad A_1 := P_1^T A P_1 - P_1^T E \dot{P}_1 = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

where $\Sigma_r \in C(\mathbb{I}, \mathbb{R}^{r,r})$ is skew-symmetric and pointwise nonsingular and the pair (E_1, A_1) is again self-adjoint. As $\operatorname{rank} A_{22} = a$ is constant in $\mathbb{I}$, there exists a pointwise orthogonal matrix-valued function $Q \in C(\mathbb{I}, \mathbb{R}^{n-r,n-r})$ such that

$$Q^T A_{22} Q = \begin{bmatrix} \Sigma_a & 0 \\ 0 & 0 \end{bmatrix},$$

where $\Sigma_a \in C(\mathbb{I}, \mathbb{R}^{a,a})$ is pointwise nonsingular and symmetric. Defining P_2 accordingly, we get

$$E_2 := P_2^T E_1 P_2 = \begin{bmatrix} \Sigma_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad A_2 := P_2^T A_1 P_2 - P_2^T E_1 \dot{P}_2 = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & \Sigma_a & 0 \\ A_{31} & 0 & 0 \end{bmatrix}.$$

We can now eliminate the blocks A_{12} and A_{21} with a nonsingular transformation P_3 using the block Σ_a such that

$$E_3 := P_3^T E_2 P_3 = \begin{bmatrix} \Sigma_r & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \qquad A_3 := P_3^T A_2 P_3 - P_3^T E_2 \dot{P}_3 = \begin{bmatrix} A_{11} & 0 & A_{13} \\ 0 & \Sigma_a & 0 \\ A_{31} & 0 & 0 \end{bmatrix}.$$

As $\operatorname{rank} A_{13} = s$ for all $t \in \mathbb{I}$ and $A_{13} = A_{31}^T$, we can find pointwise orthogonal matrix-valued functions $\hat{P}_4 \in C(\mathbb{I}, \mathbb{R}^{n-r-a,n-r-a})$ and $\hat{Q}_4 \in C(\mathbb{I}, \mathbb{R}^{r,r})$ such that $\hat{Q}_4 A_{13} \hat{P}_4^T = \begin{bmatrix} \Sigma_s & 0 \\ 0 & 0 \end{bmatrix}$ with $\Sigma_s \in C(\mathbb{I}, \mathbb{R}^{s,s})$ pointwise nonsingular. Setting P_4 accordingly, we get

$$E_4 := P_4^T E_3 P_4 = \begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_4 := P_4^T A_3 P_4 - P_4^T E_3 \dot{P}_4 = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \Sigma_s & 0 \\ \tilde{A}_{21} & \tilde{A}_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\hat{Q}_4 \Sigma_r \hat{Q}_4^T = \begin{bmatrix} E_{11} & E_{12} \\ E_{21} & E_{22} \end{bmatrix}$ is pointwise nonsingular and skew-symmetric and $\hat{Q}_4 A_{11} \hat{Q}_4^T - \hat{Q}_4 \Sigma_r \dot{\hat{Q}}_4^T = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix}$. Further, we can eliminate the block $\tilde{A}_{21}$ with a nonsingular transformation P_5 to get

$$E_5 := P_5^T E_4 P_5 = \begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$A_5 := P_5^T A_4 P_5 - P_5^T E_4 \dot{P}_5 = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} + \tilde{A}_{21}^T & 0 & \Sigma_s & 0 \\ 0 & \tilde{A}_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where $\tilde{A}_{12} + \tilde{A}_{21}^T = -\dot{E}_{12}$. The pair (E_5, A_5) is still self-adjoint as all congruence transformations preserve the structure. □

Under the assumptions of Theorem 4.25, we can now transform a self-adjoint pair of matrix-valued functions (E, A) into the global condensed form (4.23). The DAE associated with the pair in global condensed form (4.23) can be written as

$$\begin{aligned} E_{11}\dot{x}_1 + E_{12}\dot{x}_2 &= A_{11}x_1 + A_{12}x_2 + \Sigma_s x_4 + b_1, \\ E_{21}\dot{x}_1 + E_{22}\dot{x}_2 &= A_{22}x_2 + b_2, \\ 0 &= \Sigma_a x_3 + b_3, \\ 0 &= \Sigma_s^T x_1 + b_4, \\ 0 &= b_5. \end{aligned}$$

In order to obtain a strangeness-free formulation we have to eliminate the strangeness parts. To do this, we use again the derivative of the fourth equation to eliminate the terms with $\dot{x}_1$ in the first two equations and get

$$\begin{aligned} E_{12}\dot{x}_2 &= A_{11}x_1 + A_{12}x_2 + \Sigma_s x_4 + \tilde{b}_1, \\ E_{22}\dot{x}_2 &= A_{22}x_2 + \tilde{b}_2, \\ 0 &= \Sigma_a x_3 + b_3, \\ 0 &= \Sigma_s^T x_1 + b_4, \\ 0 &= b_5, \end{aligned} \tag{4.24}$$

where $\tilde{b}_1 = b_1 + E_{11}\frac{d}{dt}\left((\Sigma_s^T)^{-1}b_4\right)$, and $\tilde{b}_2 = b_2 + E_{21}\frac{d}{dt}\left((\Sigma_s^T)^{-1}b_4\right)$. Due to the occurrence of the block E_{12}, the self-adjoint structure of the system is destroyed. In the same way as in Section 4.2, if E_{22} is nonsingular, i.e., if the strangeness index is $\mu \leq 1$, we can eliminate the block E_{12} and get the equivalent system

$$\begin{aligned} 0 &= A_{11}x_1 + (A_{12} - E_{12}E_{22}^{-1}A_{22})x_2 + \Sigma_s x_4 + \tilde{b}_1 - E_{12}E_{22}^{-1}\tilde{b}_2, \\ E_{22}\dot{x}_2 &= A_{22}x_2 + \tilde{b}_2, \\ 0 &= \Sigma_a x_3 + b_3, \\ 0 &= \Sigma_s^T x_1 + b_4, \\ 0 &= b_5. \end{aligned} \tag{4.25}$$

We can further eliminate the term $A_{12} - E_{12}E_{22}^{-1}A_{22}$ via block Gaussian elimination using the invertible block Σ_s and get a strangeness-free system which is again self-adjoint. Rear-

anging and renaming the matrices and vector-valued functions finally yields the strangeness-free self-adjoint differential-algebraic system

$$\begin{aligned} \hat{E}_{11}(t)\dot{\hat{x}}_1 &= \hat{A}_{11}(t)\hat{x}_1 + \hat{b}_1(t), \\ 0 &= \hat{A}_{22}(t)\hat{x}_2 + \hat{b}_2(t), \\ 0 &= \hat{b}_3(t), \end{aligned} \tag{4.26}$$

consisting of d_μ differential equations, a_μ algebraic equations, and u_μ vanishing equations, with

$$\hat{E}_{11} = E_{22}, \quad \hat{A}_{11} = A_{22}, \quad \hat{A}_{22} = \begin{bmatrix} \Sigma_a & 0 & 0 \\ 0 & 0 & \Sigma_s \\ 0 & \Sigma_s^T & 0 \end{bmatrix},$$

$$\hat{x}_1 = x_2, \quad \hat{x}_2 = \begin{bmatrix} x_3 \\ x_1 \\ x_4 - \Sigma_s^{-1} E_{12} E_{22}^{-1} A_{22} x_2 \end{bmatrix}, \quad \hat{x}_3 = x_5,$$

$$\hat{b}_1 = \tilde{b}_2, \quad \hat{b}_2 = \begin{bmatrix} b_3 \\ \tilde{b}_1 - E_{12} E_{22}^{-1} \tilde{b}_2 \\ b_4 \end{bmatrix}, \quad \hat{b}_3 = b_5.$$

In the same way as in the case of symmetric matrix pairs, a structure preserving strangeness-free formulation (4.26) for self-adjoint differential-algebraic systems in general only exists if the strangeness index is $\mu \leq 1$. For systems with strangeness index $\mu > 1$ in general we cannot preserve the self-adjointness of the strangeness-free system. Counterexamples similar to Example 4.12 can be found for self-adjoint matrix pairs.

Remark 4.26. *For constant coefficient systems of the form* (2.6) *a matrix pair* (E, A) *is self-adjoint if it is skew-symmetric/symmetric, i.e.,* $E = -E^T$ *and* $A = A^T$*. Similar results as obtained in Section 4.1 can be derived analogously for skew-symmetric/symmetric matrix pairs as strong congruence* (4.2) *preserve this structure. A local condensed form as given in Theorem 4.6 can be obtained in the same way. Structured staircase forms for skew-symmetric/symmetric matrix pairs have also been considered in [20].*

Remark 4.27. *Linear-quadratic optimal control problems as in* (1.3) *have been the motivation for considering self-adjoint differential-algebraic systems. The solution strategies for these systems actually lead to boundary value problems of the form* (1.3c) *with two-point boundary conditions. The classical approach to solve these boundary value problems is the use of Riccati differential-algebraic equations, see e.g. [77, 83]. First, index reduction and feedback regularization are used to transform the system to a regular, strangeness-free control problem and then the Riccati approach is used on the reduced system. If the reduced problem can be obtained in a structure preserving way, then the solution of the Riccati equations can be adapted to the self-adjoint structure.*

4.4 STRUCTURE PRESERVING INDEX REDUCTION BY MINIMAL EXTENSION

In the numerical solution of differential-algebraic equations it was suggested in [22, 75, 76, 82] to transform the system into an equivalent strangeness-free differential-algebraic system. The basic idea of the general approach described in Section 2.2.2 is to consider the original system together with a sufficient number of its derivatives as a derivative array and to derive locally at every integration step an equivalent system of strangeness index $\mu = 0$ that has the same solution set as the original system, but contains all the information on the manifold in which the dynamics of the system take place. In principle, this general approach provides a uniform framework for the numerical solution of differential-algebraic systems, but it has high computational complexity since from the derivative array certain nullspaces of the Jacobians and associated projections onto these nullspaces have to be computed at every integration step. Further, a particular structure of the system is not reflected in the general approach. Another index reduction technique, the so-called *index reduction by minimal extension* described in [80], can be applied that uses the introduction of new variables to reduce the index of the differential-algebraic system. In the following, we will see that this can be done in a structure preserving way. This approach has also been used for the index reduction of electrical circuit equations where it allows to preserve certain symmetry structures in the reduced equations, see [6]. In the following, we will use these ideas to obtain a structure preserving index reduction method for self-adjoint differential-algebraic systems.

In this section, we consider a self-adjoint initial value problem

$$E(t)\dot{x} = A(t)x + b(t), \quad x(t_0) = x_0, \tag{4.27}$$

with $E, A \in C([t_0, t_f], \mathbb{R}^{n,n})$, $E = -E^T$, $A = (A + \dot{E})^T$ of strangeness index $\mu = 1$. Further, we assume that the system is already given in global condensed form (4.23)

$$\begin{bmatrix} E_{11} & E_{12} & 0 & 0 & 0 \\ E_{21} & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \dot{x}_3 \\ \dot{x}_4 \\ \dot{x}_5 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & 0 & \Sigma_s & 0 \\ 0 & A_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{bmatrix},$$

where E_{22} is nonsingular since $\mu = 1$. Via a global congruence transformation with a matrix P we can eliminate the blocks E_{12} and E_{21} and get the equivalent system

$$\begin{bmatrix} \tilde{E}_{11} & 0 & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \\ \dot{\tilde{x}}_4 \\ \dot{\tilde{x}}_5 \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \Sigma_s & 0 \\ \tilde{A}_{21} & A_{22} & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \\ \tilde{x}_5 \end{bmatrix} + \begin{bmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \tilde{b}_3 \\ \tilde{b}_4 \\ \tilde{b}_5 \end{bmatrix}, \tag{4.28}$$

with

$$\begin{aligned}
\tilde{E}_{11} &= E_{11} - E_{12}E_{22}^{-1}E_{21},\\
\tilde{A}_{11} &= A_{11} - (A_{12} - E_{12}E_{22}^{-1}A_{22})E_{22}^{-1}E_{21},\\
\tilde{A}_{12} &= A_{12} - E_{12}E_{22}^{-1}A_{22},\\
\tilde{A}_{21} &= -A_{22}E_{22}^{-1}E_{21} + E_{22}\tfrac{d}{dt}(E_{22}^{-1}E_{21}),
\end{aligned}$$

and corresponding transformed $\tilde{x} = Px$ and $\tilde{b} = Pb$, which is again self-adjoint, i.e.,

$$\tilde{E}_{11} = -\tilde{E}_{11}^T, \quad \tilde{A}_{11} = \tilde{A}_{11}^T + \dot{\tilde{E}}_{11}^T, \quad \tilde{A}_{12} = \tilde{A}_{21}^T, \quad \tilde{A}_{21} = \tilde{A}_{12}^T. \tag{4.29}$$

The equations that have to be differentiated to reduce the index of the system are given by the fourth block row of system (4.28). Differentiating these equations and adding the derivatives to the system we obtain the reduced derivative array

$$\begin{bmatrix} \tilde{E}_{11} & 0 & 0 & 0 & 0\\ 0 & E_{22} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ I_s & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1\\ \dot{\tilde{x}}_2\\ \dot{\tilde{x}}_3\\ \dot{\tilde{x}}_4\\ \dot{\tilde{x}}_5 \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \Sigma_s & 0\\ \tilde{A}_{21} & A_{22} & 0 & 0 & 0\\ 0 & 0 & \Sigma_a & 0 & 0\\ \Sigma_s^T & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \tilde{x}_1\\ \tilde{x}_2\\ \tilde{x}_3\\ \tilde{x}_4\\ \tilde{x}_5 \end{bmatrix} + \begin{bmatrix} \tilde{b}_1\\ \tilde{b}_2\\ \tilde{b}_3\\ \tilde{b}_4\\ \tilde{b}_5\\ -\tilde{b}_6 \end{bmatrix}, \tag{4.30}$$

with $\tilde{b}_6 = \frac{d}{dt}(\Sigma_s^{-T}\tilde{b}_4)$. Now, we introduce a new variable x_6 to replace every occurrence of $\dot{\tilde{x}}_1$ in (4.30) and thus reduces the index of the system. We therefore obtain the extended system

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0\\ 0 & E_{22} & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1\\ \dot{\tilde{x}}_2\\ \dot{\tilde{x}}_3\\ \dot{\tilde{x}}_4\\ \dot{\tilde{x}}_5\\ \dot{x}_6 \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \Sigma_s & 0 & -\tilde{E}_{11}\\ \tilde{A}_{21} & A_{22} & 0 & 0 & 0 & 0\\ 0 & 0 & \Sigma_a & 0 & 0 & 0\\ \Sigma_s^T & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & I_s \end{bmatrix} \begin{bmatrix} \tilde{x}_1\\ \tilde{x}_2\\ \tilde{x}_3\\ \tilde{x}_4\\ \tilde{x}_5\\ x_6 \end{bmatrix} + \begin{bmatrix} \tilde{b}_1\\ \tilde{b}_2\\ \tilde{b}_3\\ \tilde{b}_4\\ \tilde{b}_5\\ \tilde{b}_6 \end{bmatrix}. \tag{4.31}$$

In order to preserve the self-adjoint structure we have to eliminate the block $\tilde{E}_{11}$ in (4.31) using the last equation to get

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0\\ 0 & E_{22} & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1\\ \dot{\tilde{x}}_2\\ \dot{\tilde{x}}_3\\ \dot{\tilde{x}}_4\\ \dot{\tilde{x}}_5\\ \dot{x}_6 \end{bmatrix} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} & 0 & \Sigma_s & 0 & 0\\ \tilde{A}_{21} & A_{22} & 0 & 0 & 0 & 0\\ 0 & 0 & \Sigma_a & 0 & 0 & 0\\ \Sigma_s^T & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & I_s \end{bmatrix} \begin{bmatrix} \tilde{x}_1\\ \tilde{x}_2\\ \tilde{x}_3\\ \tilde{x}_4\\ \tilde{x}_5\\ x_6 \end{bmatrix} + \begin{bmatrix} \tilde{b}_1\\ \tilde{b}_2\\ \tilde{b}_3\\ \tilde{b}_4\\ \tilde{b}_5\\ \tilde{b}_6 \end{bmatrix},$$

with $\bar{b}_1 = \tilde{b}_1 + \tilde{E}_{11}\tilde{b}_6$. We have $\tilde{A}_{11} = \tilde{A}_{11}^T + \dot{\tilde{E}}_{11}^T$ from (4.29), but since the block $\tilde{E}_{11}$ has been eliminated, we also have to eliminate the block $\tilde{A}_{11}$ using the nonsingular block Σ_s^T in order to preserve the self-adjoint structure. We finally get

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & E_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \\ \dot{\tilde{x}}_4 \\ \dot{\tilde{x}}_5 \\ \dot{x}_6 \end{bmatrix} = \begin{bmatrix} 0 & \tilde{A}_{12} & 0 & \Sigma_s & 0 & 0 \\ \tilde{A}_{21} & A_{22} & 0 & 0 & 0 & 0 \\ 0 & 0 & \Sigma_a & 0 & 0 & 0 \\ \Sigma_s^T & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & I_s \end{bmatrix} \begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \\ \tilde{x}_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} \hat{b}_1 \\ \tilde{b}_2 \\ \tilde{b}_3 \\ \tilde{b}_4 \\ \tilde{b}_5 \\ \tilde{b}_6 \end{bmatrix}, \tag{4.32}$$

with $\hat{b}_1 = \tilde{b}_1 + \tilde{E}_{11}\tilde{b}_6 - \tilde{A}_{11}\Sigma_s^{-T}\tilde{b}_4$, and this system is strangeness-free and self-adjoint.

Lemma 4.28. *Let the differential-algebraic system in global condensed form* (4.23) *have strangeness index* $\mu = 1$*. Then the extended system* (4.32) *has strangeness index* $\mu = 0$*.*

Proof. Following Theorem 4.24, we can compute matrices T and V whose columns span the nullspaces of $\tilde{E}$ and $T^T\tilde{A}T$ where $\tilde{E}$ and $\tilde{A}$ are the extended matrices in (4.32). Possible choices for T and V are

$$T = \begin{bmatrix} I_s & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 & 0 & 0 \\ 0 & 0 & 0 & I_s & 0 & 0 \\ 0 & 0 & 0 & 0 & I_u & 0 \\ 0 & 0 & 0 & 0 & 0 & I_s \end{bmatrix}, \quad V = \begin{bmatrix} 0 & 0 & 0 \\ I_d & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I_u & 0 \\ 0 & 0 & I_s \end{bmatrix}.$$

Further, let $T' = \begin{bmatrix} 0 & I_d & 0 & 0 & 0 & 0 \end{bmatrix}^T$ complete T to a nonsingular matrix. It follows that $V^T T^T \tilde{A} T' = 0$, i.e., the extended system (4.32) has strangeness index $\mu = 0$. □

Note that this approach can also be applied in the case of linear symmetric differential-algebraic equations of strangeness-index $\mu = 1$ as considered in Section 4.2 in an analogous way.

4.5 Future Work

In the previous section we have seen that a structure preserving strangeness-free form for symmetric as well as self-adjoint linear DAEs only exists if the strangeness-index of the system is lower or equal to 1. For symmetric systems we need in addition strong assumptions on the coefficient matrices, in order to be able to obtain a structure preserving condensed form. It does not seem to be possible to lessen these assumptions while preserving the symmetric structure using global congruence transformations. Nevertheless, the index reduction by minimal extension that allows the formulation of a structure preserving strangeness-free system for self-adjoint systems of index $\mu = 1$, can also be applied to linear symmetric systems of index $\mu = 1$ without the need of Assumption 4.8. All obtained results can also be extended to linear DAEs with Hermitian coefficient matrices, see [153].

Chapter 5

Switched Differential-Algebraic Systems

A particular feature of many complex dynamical systems modeled by DAEs is that they are *switched systems or hybrid systems*, i.e., the mathematical model itself may change with time, depending on certain indicators. This in often an artifact from the modeling, as fast nonlinear phenomena of physical systems are often approximated resulting in piecewise continuous systems with discrete transitions. Switched systems also arise naturally in control systems where the value of a control switches.
Typical examples for switched systems are electronic circuits, where different device models are used for different frequency ranges or switching elements like diodes or electric switches are used.

Example 5.1 (Boost converter)**.** We consider the boost converter given in Figure 5.1 consisting of a capacitor with capacitance C, an ideal diode D, a voltage source V_i, an inductor with inductance L, a resistor with resistance R, and an ideal switch S. Depending

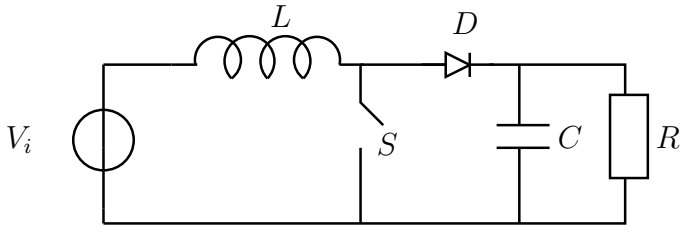

Figure 5.1: The boost converter

on the states of the diode and of the switch we can distinguish four cases:

1. the switch S is open and the diode D is conducting,
2. the switch S is closed and the diode D is blocking,
3. the switch S is open and the diode D is blocking,
4. the switch S is closed and the diode D is conducting.

Let i_D, i_S, i_L, i_C, and i_R denote the currents through the diode, switch, inductance, capacitance, and resistance respectively, and v_D, v_S, v_L, v_C, and v_R the voltages across the corresponding elements. If the switch is open we have $i_S = 0$, and if the switch S is closed

then $v_S = 0$. In the same way, if the diode D is conducting we have $v_D = 0$, and if the diode is blocking then $i_D = 0$. Using the relations

$$v_R(t) = Ri_R(t), \quad v_L(t) = L\frac{d}{dt}i_L(t), \quad i_C(t) = C\frac{d}{dt}v_C(t),$$

and Kirchhoff's current and voltage laws the circuit equations are given by

$$\begin{aligned} C\dot{v}_C &= i_D - i_R, \\ L\dot{i}_L &= v_S - V_i, \\ 0 &= v_D + v_C + v_S, \\ 0 &= Ri_R - v_C, \\ 0 &= i_S + i_L - i_D, \end{aligned} \tag{5.1}$$

together with the algebraic constraints

$$\begin{cases} i_S = 0,\ v_D = 0 & \text{in mode 1,} \\ v_S = 0,\ i_D = 0 & \text{in mode 2,} \\ i_S = 0,\ i_D = 0 & \text{in mode 3,} \\ v_S = 0,\ v_D = 0 & \text{in mode 4.} \end{cases} \tag{5.2}$$

The system switches between the different modes based on the states of the diode and the switch. That means, starting e.g. in mode 1, the system switches to another mode either if the current through the diode i_D becomes negative (switch to mode 3), or if the switch is closed (switch to mode 4).

Another important class of applications that display switching or hybrid behavior are mechanical systems with dry friction [34, 88], impact phenomena, or structure varying systems with changing number of degrees of freedom, e.g., robot manipulators, or automatic gear-boxes [60].

Example 5.2. (See [34]) Multibody systems with dry friction between the bodies in contact are usually modeled by the Coulomb friction law. Here, the friction force F is assumed to be proportional to the normal force F_N on the surface between the bodies in the contact point, i.e., $\|F\| = \mu_F\|F_N\|$, where $\mu_F > 0$ is the coefficient of friction which depends on material properties. The friction force is directed tangential to the friction surface and is opposite to the direction of motion of the body, i.e.,

$$F = -\mu_F\|F_N\|c(p)\,\mathrm{sign}(c(p)^T\dot{p}),$$

where $c(p)$ is a unit vector parallel to the friction surface and $c(p)^T\dot{p}$ describes the relative tangential velocity at the contact point, see Figure 5.2. If the friction surface is modeled by an algebraic equation $g(p) = 0$ (this is also called the *contact condition*), then the normal

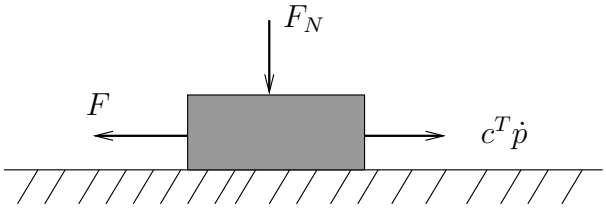

Figure 5.2: Mechanical systems with dry friction

force equals the constraint force, i.e., $F_N = G(p)^T\lambda$ with $G(p) = \frac{\partial g(p)}{\partial p}$, and the equations of motion are given by

$$\begin{aligned} M\ddot{p} &= f_a(p,\dot{p}) - G(p)^T\lambda - \mu_F\|F_N\|c(p)\,\mathrm{sign}(c(p)^T\dot{p}), \\ 0 &= g(p). \end{aligned}$$

Dry friction models usually exhibit two distinct modes: *stiction* or *static friction* and *sliding* or *kinetic friction*. During static friction the tangential contact force F maintains zero relative velocity between the contacting bodies. The bodies stick together and the friction force F is bounded by the equation $\|F\| \leq \mu_s\|F_N\|$, where $\mu_s > 0$ is now the coefficient of static friction. On the other hand, in kinetic friction the relative contact velocity is non-zero and the kinetic friction force F has constant magnitude and is directly opposing the sliding velocity with $\|F\| = \mu_k\|F_N\|$, with coefficient of kinetic friction μ_k. In this case the bodies in contact are in relative motion, i.e., one body is slipping across the other. In general $\mu_s \geq \mu_k$ and the discrepancy is of order $10 - 20\%$. The transition between the two modes is known as *stick-slip transition*, see also [34, 88].

Further examples of switched systems are biological or chemical systems which act different in different day cycles or depending on certain nutritions, or traffic systems which operate different depending on delays, see e.g. [28].
Typically, the continuous dynamics of the system in the different operation modes are described by sets of ordinary differential equations or differential-algebraic equations. The changing between different operation modes is modeled by discrete transitions resulting in switching between sets of equations describing each operation mode. In the following, systems of differential-algebraic equations that switch between several modes of operation are called *switched differential-algebraic equations* or *hybrid differential-algebraic equations*, see [61]. Hybrid systems are loosely defined as dynamical systems whose state has two components, one of which evolves in a continuous set, while the other evolves in a discrete set according to some transition rule, i.e., the term *hybrid* refers to the combination of discrete event and continuous time dynamics which interact and define the behavior of the system. The terminology *switched system* emphasizes the switching between different system representations and refers to the behavior of the continuous state. In the literature, a switched system often implicitly assumes a hybrid system model in which the discrete dynamics are "simple". In the following and in the literature, both terms are often used synonymously.

As the discrete and continuous dynamics interact they must be analyzed simultaneously. The mathematical theory of switched differential-algebraic systems, the control theory for such systems as well as the development of efficient and accurate numerical methods is still in an early stage. For an overview of modeling, analysis, simulation and control of hybrid systems, see e.g. [12, 94]. Further works concerning hybrid systems diagnosis and stabilization of hybrid systems are [28, 101, 116, 117]. One of the basic difficulties in switched differential-algebraic systems is that after a mode switch takes place, the model dimension and the structure of the system as well as its properties such as the index, the number of algebraic or differential equations or redundancies may change. Thus, mode switching may lead to a DAE with a different index or a different number of degrees of freedom resulting in a discontinuity in the solution manifold. From this point of view, DAEs with discontinuities or singular points are also included in the hybrid system approach. In [60, 61] it was shown how the theory for general over- and underdetermined DAEs can be applied to hybrid differential-algebraic systems. Besides the already existing problems in the numerical integration of DAEs there are new difficulties in the numerical simulation of switched systems. First of all, the reduction to strangeness-free form has to be done in the same way as for standard DAEs and appropriate numerical methods for DAEs have to be used for the numerical integration. However, in switched systems the integration is often done over small intervals and in addition the states at the switch points have to be determined exactly, as they are the basis for the consistent initialization in the successor mode. Further, a special phenomena that can occur during the simulation of hybrid systems is a cyclic change between different modes of operation, called *chattering* or *sliding*, for example if nearly equal thresholds for the transition conditions of different modes are given and the system starts to oscillate around these. These oscillations may be real in the physical model as hysteresis, delays and other dynamic nonidealities lead to fast oscillations, but also may arise due to errors in the numerical method. Chattering behavior has to be treated in an appropriate way to ensure that the numerical integration terminates in reasonable time.

In this chapter we consider the analysis and numerical solution of general nonlinear switched differential-algebraic equations. For the formulation of switched systems we follow the ideas proposed in [12, 60, 61] and define hybrid differential-algebraic systems. In Section 5.2 we extend the general theory of over- and underdetermined systems of differential-algebraic equations to switched systems of DAEs and show how index reduction can be done for switched systems. In Section 5.3 we study existence and uniqueness of solutions of hybrid DAE systems. In general, in order to guarantee existence and uniqueness of solutions of a hybrid system after mode switching, the current state has to be transfered to the new mode in a consistent way. Further, we have to deal with non-uniqueness of solutions after a switch. In Section 5.4 we will investigate how the numerical methods for the consistent initialization of DAEs that were derived in [85, 87] and allow to fix certain state components and change others can be extended to switched systems and develop methods for correct initialization at switch points. In Section 5.5 we develop mathematical methods to detect (numerical) chattering and show how the chattering behavior can be approximated by so-called *sliding modes*. Finally, in Section 5.6 we consider the control of switched systems

and show how the principle concepts of control theory for linear descriptor systems can be extended to the case of hybrid systems.

5.1 Formulation of Switched Differential-Algebraic Systems

In the following, switched systems of differential-algebraic equations are described using a hybrid system formulation. Hybrid systems are generally described by a collection of discrete subsystems, a collection of continuous subsystems and the possible interaction between these subsystems. The continuous subsystems in general can consist of ODEs, DAEs, PDEs or Integro-DAEs. For the modeling of hybrid systems there are many different approaches [94, 101] coming from different areas that were developed for specific tasks, e.g., the hybrid automaton model [2, 46] that combines continuous state space models for the continuous dynamics with finite automata for the discrete dynamics, hybrid Petri nets [31], general abstract dynamical models [16], state-transition network representations [5], or bond graph representations [107]. For example, in hybrid automata dynamic components are added to a discrete state automaton. This formulation is suited for the examination of reachability of states but not for the examination of the dynamics of the system. In this section, we choose a formulation of hybrid differential-algebraic systems following an approach given in [12, 60, 61]. In particular, we consider hybrid systems that are composed of several different constrained nonlinear dynamical systems described by differential-algebraic equations for the different operation modes and transition conditions between these DAEs. Further, we assume that the discrete and continuous subsystems only interact via instantaneous discrete transitions at distinct points in time called *events*.

Definition 5.3 (Hybrid differential-algebraic system). Let $\mathbb{I} = [t_0, t_f] \subset \mathbb{R}$ be an integration interval that is decomposed into subintervals $\mathbb{I}_i = [\tau_i, \tau_i')$ for $i = 1, \ldots, N_\mathbb{I} - 1$ and $\mathbb{I}_{N_\mathbb{I}} = [\tau_{N_\mathbb{I}}, \tau'_{N_\mathbb{I}}]$, $N_\mathbb{I} \in \mathbb{N}$ such that $\mathbb{I} = \bigcup_{i=1}^{N_\mathbb{I}} \mathbb{I}_i$, with $\tau_1 = t_0$, $\tau'_{N_\mathbb{I}} = t_f$ and $\tau_i' = \tau_{i+1}$ for all $i = 1, \ldots, N_\mathbb{I} - 1$ and $\tau_i < \tau_i'$ for all $i = 1, \ldots, N_\mathbb{I}$. Further, let $\mathbb{M} := \{1, \ldots, N_F\}$, $N_F \in \mathbb{N}$ be the *set of modes* and for each $l \in \mathbb{M}$ let D_l be the union of certain integration intervals $\mathbb{I}_i$, such that $\bigcup_{l \in \mathbb{M}} D_l = \mathbb{I}$ and $D_l \cap D_k = \emptyset$ for $l, k \in \mathbb{M}$ with $l \neq k$. Then, a *hybrid system of differential-algebraic equations* $\mathcal{H}$ is defined as the collection of

- a set of N_F systems of nonlinear differential-algebraic equations

$$F^l(t, x^l, \dot{x}^l) = 0, \quad l \in \mathbb{M}, \tag{5.3}$$

  with sufficiently smooth functions $F^l : D_l \times \mathbb{R}^{n_l} \times \mathbb{R}^{n_l} \to \mathbb{R}^{m_l}$,

- an index set of autonomous transitions $J^l = \{1, 2, \ldots, n_T^l\}$ for each mode $l \in \mathbb{M}$, where $n_T^l \in \mathbb{N}$ is the number of possible transitions of mode l,

- *transition conditions* $L_j^l(t, x^l, \dot{x}^l)$ for all transitions $j \in J^l$, and all modes $l \in \mathbb{M}$ with

$$L_j^l : D_l \times \mathbb{R}^{n_l} \times \mathbb{R}^{n_l} \to \{TRUE, FALSE\}, \tag{5.4}$$

- *switching functions* of the form

$$g^l_{j,i} : D_l \times \mathbb{R}^{n_l} \times \mathbb{R}^{n_l} \to \mathbb{R}, \quad \text{for all } i = 1, \ldots, n^l_j, \quad j \in J^l,$$

with $g^l_{j,i}(t, x^l, \dot{x}^l) > 0$ in mode l,

- *mode allocation functions* for all $l \in \mathbb{M}$ of the form

$$S^l : J^l \to \mathbb{M}, \text{ with } S^l(j) = k, \tag{5.5}$$

that determine the successor mode k after a mode change, and

- *transition functions* $T^k_l : \mathbb{R}^{n_l} \times \mathbb{R}^{n_l} \to \mathbb{R}^{n_k \times 2}$ of the form

$$T^k_l(x^l(\tau'_i), \dot{x}^l(\tau'_i)) = [x^k(\tau_{i+1}), \dot{x}^k(\tau_{i+1})], \tag{5.6}$$

for all $l \in \mathbb{M}$ with successor mode $k \in \mathbb{M}$ that map the final values of the variables in mode l to the initial values in mode k at event time $\tau'_i = \tau_{i+1} \in D_k$.

Definition 5.4 (Linear hybrid differential-algebraic system). A hybrid system $\mathcal{H}$ as in Definition 5.3 is called *linear* if the DAE in each mode is a linear DAE of the form

$$E^l(t)\dot{x}^l = A^l(t)x^l + b^l(t), \quad l = 1, \ldots, N_F, \tag{5.7}$$

with sufficiently smooth functions $b^l : D_l \to \mathbb{R}^{m_l}$ and $E^l, A^l : D_l \to \mathbb{R}^{m_l \times n_l}$.

If in addition an initial value

$$x^{l_1}(t_0) = x^{l_1}_0 \in \mathbb{R}^{n_{l_1}} \tag{5.8}$$

is given in some initial mode l_1, then a hybrid system $\mathcal{H}$ as in Definition 5.3 together with the initial condition (5.8) and initial mode $l_1 \in \mathbb{M}$ is called a *hybrid initial value problem*. In this setting, (5.3) or (5.7) are the DAEs that describe the dynamics of the hybrid system in mode $l \in \mathbb{M}$ and in each subinterval the dynamics of the system are governed by only one DAE. The hybrid system is said to be *in mode* $l \in \mathbb{M}$ if $t \in D_l$. Further, the piecewise continuous functions $x^l : D_l \to \mathbb{R}^{n_l}$ describe the continuous state of the hybrid system in mode l and $x^l(\tau'_i)$ is the smooth extension of x^l to the interval boundary $\tau'_i = \tau_{i+1}$ of an integration interval $\mathbb{I}_i \in D_l$. We further define the *hybrid time trajectory* $T_\tau = \{\mathbb{I}_i\}_{i=1,\ldots,N_\mathbb{I}}$ as a sequence of intervals and the *hybrid mode trajectory* $T_m = \{l_i\}_{i=1,\ldots,N_\mathbb{I}}$ as the corresponding sequence of modes, where $l_i \in \mathbb{M}$ is the mode in interval $\mathbb{I}_i$. The hybrid time trajectory and the hybrid mode trajectory depend on the initial mode and initial conditions as well as on the defined switching conditions. The *set of event times* corresponding to a hybrid time trajectory is given by $\mathcal{E}(T_\tau) = \{\tau_i \mid i = 1, \ldots, N_\mathbb{I}\}$, and the state of a hybrid system $\mathcal{H}$ is described by the *hybrid solution trajectory* $\{(x^{l_i}(t), l_i)\}$ consisting of a sequence of the continuous states $x^{l_i}(t)$ with corresponding modes l_i.
The hybrid system $\mathcal{H}$ changes between different modes on the basis of the transition conditions. If $L^l_j(\hat{t}, x^l(\hat{t}), \dot{x}^l(\hat{t})) = FALSE$ for all $j \in J^l$ at a time $\hat{t} \in D_l$, then the system

stays in the current mode. On the other hand, if there exists an integer $j \in J^l$ such that $L_j^l(\hat{t}, x^l(\hat{t}), \dot{x}^l(\hat{t})) = TRUE$ at time $\hat{t}$, then the system switches to another mode. The switch points are defined as the roots of the switching functions $g_{j,i}^l(t, x^l, \dot{x}^l)$ that are given as threshold functions, i.e., if $g_{j,i}^l(t, x^l, \dot{x}^l) > 0$ for all $i = 1, \ldots, n_j^l, j \in J^l$, then the system stays in the current mode l, but if $g_{j,i}^l(t, x^l, \dot{x}^l) \leq 0$ for a $j \in J^l$ and some i then the system may switch to a new mode. Note, that a transition condition L_j^l is described by n_j^l separated switching functions $g_{j,i}^l$, $i = 1, \ldots, n_j^l$, which logical combination determines if the transition condition L_j^l is satisfied. In this way, the switching functions $g_{j,i}^l$ can be chosen as simple as possible, e.g. linear, allowing an efficient and reliable computation of the switch points. Thus, each time a switching function crosses zero the associated transition condition may switch its logical value. Each switching function can be seen as a *switching surface* in the state space given by

$$\Gamma_{j,i}^l = \left\{(t, x^l, \dot{x}^l) \in D_l \times \mathbb{R}^{n_l} \times \mathbb{R}^{n_l} \mid g_{j,i}^l(t, x^l, \dot{x}^l) = 0\right\}, \quad j \in J^l, \quad l \in \mathbb{M}, \tag{5.9}$$

along which discontinuous changes in the system may occur, i.e., mode switching occurs at points on these switching surfaces. For convenience of expression, in the following we assume that each transition condition L_j^l is described by exactly one switching function g_j^l (i.e., $n_j^l = 1$ for all $j \in J^l, l \in \mathbb{M}$) and the transition condition L_j^l is satisfied if and only if $g_j^l \leq 0$. Then, the satisfaction of a transition condition corresponds to the crossing of the switching surfaces Γ_j^l. The union of the switching surfaces for all $j \in J^l$ in mode l is given by

$$\Gamma^l = \bigcup_{j \in J^l} \Gamma_j^l, \quad l \in \mathbb{M}. \tag{5.10}$$

Further, we assume deterministic models, i.e., only one transition condition is becoming true at a time. The solutions of a nonlinear DAE lie in the constraint manifold given by the algebraic constraints, see Section 2.2.2. In hybrid systems the solution trajectory stays in the constraint manifold $\mathbb{L}^l$ of the DAE in the current mode l as long as no transition condition is satisfied. Thus, the *constraint manifold of the hybrid system in mode l* is given by

$$\Lambda^l = \left\{(t, x) \in D_l \times \mathbb{R}^{n_l} \mid (t, x) \in \mathbb{L}^l \text{ and } g_j^l(t, x, \dot{x}) > 0 \text{ for all } j \in J^l\right\}, \tag{5.11}$$

and Γ^l describes the boundary of Λ^l in mode l. Finally, the transition function T_l^k transfers the state at the mode change from mode l to mode k according to the jth transition. This transfer can result in jumps in the state vector of the hybrid system. Further, in order to obtain a solution in the new mode k, the initial value obtained by the transition function has to be consistent with the DAE in mode k.

In Figure 5.3 an example of a typical path for a hybrid solution trajectory is plotted. The solution trajectory of the hybrid system starts in initial mode l_1 at a consistent initial point $(t_0, x_0^{l_1}) \in \Lambda^{l_1}$ and evolves continuously within the constraint manifold Λ^{l_1} of mode l_1, as

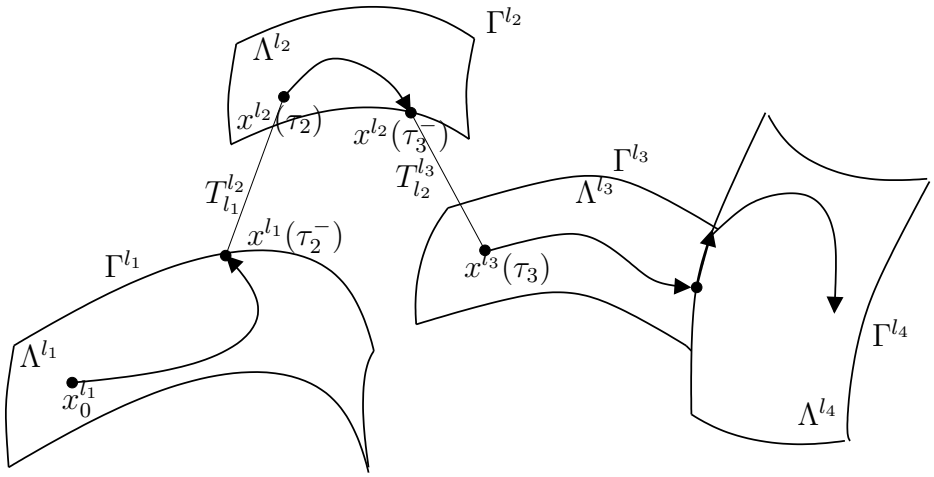

Figure 5.3: Evolution of a hybrid system trajectory

specified by the DAE F^{l_1}, until the minimal time τ_2 at which $L_j^{l_1}(\tau_2, x^{l_1}, \dot{x}^{l_1}) = TRUE$ for some $j \in J^{l_1}$, i.e., the switching surface $\Gamma_j^{l_1}$ is reached. The transition function $T_{l_1}^{l_2}$ moves the trajectory from $x^{l_1}(\tau_2^-)$ to $x^{l_2}(\tau_2)$ with limit value $x^{l_1}(\tau_2^-) = \lim_{t \to \tau_2} x^{l_1}(t)$. The new initial point $x^{l_2}(\tau_2)$ in mode l_2 has to lie in the constraint manifold Λ^{l_2}, i.e., it has to be consistent for the DAE in mode l_2 and $L_j^{l_2}(\tau_2, x^{l_2}(\tau_2), \dot{x}^{l_2}(\tau_2)) = FALSE$ for all $j \in J^{l_2}$. Then, the system trajectory continues in mode l_2 within the constraint manifold Λ^{l_2} until the switching surface Γ^{l_2} is crossed. It may happen that the constraint manifolds Λ^{l_3} and Λ^{l_4} corresponding to mode l_3 and mode l_4 are separated by a common switching surface $\Gamma_j^{l_3} = \Gamma_i^{l_4}$, $j \in J^{l_3}$, $i \in J^{l_4}$ from one another. If the solution trajectory reaches this switching surface and some sliding condition is fulfilled, the system enters into a sliding state along the switching surface Γ^{l_3} resulting in repeated switching between the two modes and the solution trajectory evolves along the switching surface. When the existence condition for sliding is no longer fulfilled, then the system evolves in the solution manifold Λ^{l_4} until the next event.

The possible transition behaviors at a mode change from mode l to mode k are depicted in Figure 5.4. In the first case, the transition function T_l^k moves the state vector from the switching surface Γ^l into the interior of the constraint manifold Λ^k of the new mode and the further solution behavior is governed by the DAE in mode k (case (a) in Figure 5.4). In the following, this will be called *regular switching*. In the second case, the transition function T_l^k transfers the state vector from the switching surface Γ^l to the boundary Γ^k of the new mode, which causes an immediate further mode transition (case (b) in Figure 5.4). In this case oscillations between the two modes can occur if this transition moves the trajectory back to the first mode. In the third case, the transition function T_l^k moves the state vector beyond the region and boundary of the new mode (case (c) in Figure 5.4). This can happen for example if the state vector is not consistent with the DAE in the new mode. This case can be handled by another immediate transition, e.g., a projection onto the constraint

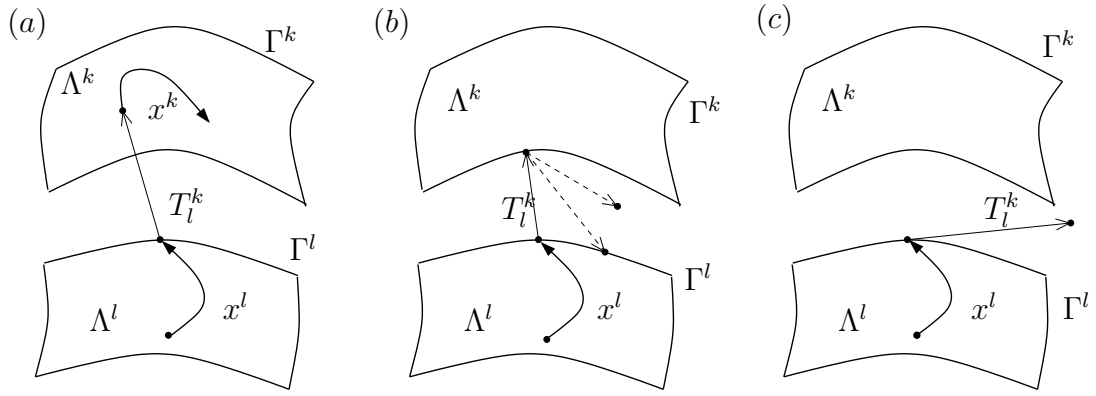

Figure 5.4: Transition behavior of a hybrid system

manifold Λ^k. Modes that do not affect the state vector but are immediately followed by another transition are sometimes called *mythical modes*, see [105]. If the behavior of the system in a mode is governed only by algebraic equations then no continuous evolution takes place in this mode but jumps in the state vector can occur. These modes are also called *pinnacles*, see also [105].

Remark 5.5. *Definition 5.3 does not allow multiple events at the same time, but so-called Zeno behavior is possible, i.e., infinite number of transitions at almost the same time leading to accumulation points of events times τ_i (this arises e.g. in the simple example of a bouncing ball). In the following, we assume that no accumulation of event times occur in order to make the system well-defined.*

5.2 Index Reduction

For hybrid differential-algebraic systems a reduction to strangeness-free form must be realized just as for standard DAEs to be able to apply numerical methods suited for DAEs in the integration process. In hybrid systems this must be possibly done very often on possibly short intervals and for different modes. The index reduction procedure described in Section 2.2.2 leading to the reduced systems (2.19) or (2.21) can be applied to switched systems as has been shown in [60, 61]. In this way, a strangeness-free formulation can be determined independently for each mode. As described in Section 2.2.2, we need the information about several derivatives of the given DAE, so we consider the derivative array $\mathcal{F}_i^l$ of level i in mode $l \in \mathbb{M}$, defined as in (2.15), which stacks the original equations of the DAE in mode $l \in \mathbb{M}$ and all its derivatives up to level i into one large system

$$\mathcal{F}_i^l(t, x^l, \dot{x}^l, \ldots, \frac{d^{i+1}}{dt^{i+1}} x^l) = 0. \tag{5.12}$$

In order to be able to derive a reduced system for each mode we need to assume that Hypothesis 2.37 holds in each mode.

Hypothesis 5.6. *In each mode $l \in \mathbb{M}$ and each domain D_l the strangeness index μ^l is well-defined, i.e., the DAE* (5.3) *in mode l satisfies Hypothesis 2.37 with constant characteristic values $\mu^l, r^l, a^l_\mu, d^l_\mu,$ and v^l_μ.*

If Hypothesis 5.6 is violated at a finite number of points we can introduce new modes and further switching points to satisfy Hypothesis 5.6. In the following, we consider hybrid systems that satisfy Hypothesis 5.6, i.e., the integers $m_l, n_l, \mu^l, r^l, d^l_\mu, a^l_\mu, u^l_\mu, v^l_\mu$ are constant within each mode but may be different for different modes.

Definition 5.7 (Regularity). A hybrid system $\mathcal{H}$ is called *regular* if for each mode $l \in \mathbb{M}$ the corresponding DAE F^l is regular.

Definition 5.8 (Maximal Strangeness Index). For a hybrid system $\mathcal{H}$ that satisfies Hypothesis 5.6 the *maximal strangeness index* μ_{max} is defined as

$$\mu_{max} = \max_{l \in \mathbb{M}} \{\mu^l\}.$$

A hybrid system is called *strangeness-free* if $\mu_{max} = 0$.

We can locally obtain a reduced system of the form (2.19) independently in each mode $l \in \mathbb{M}$ as described in Section 2.2.2, which is denoted by

$$\begin{aligned} \hat{F}^l_1(t, x^l, \dot{x}^l) &= 0, \\ \hat{F}^l_2(t, x^l) &= 0, \end{aligned} \tag{5.13}$$

with $\hat{F}^l_1(t, x^l, \dot{x}^l) = (Z^l_1)^T F^l(t, x^l, \dot{x}^l)$ and $\hat{F}^l_2(t, x^l) = (Z^l_2)^T \mathcal{F}^l_{\mu_l}(t, x^l, H(t, x^l))$, where Z^l_1 and Z^l_2 are the matrices defined in Hypothesis 2.37 for the DAE in mode l. Analogously, a decoupled differential-algebraic system of the form

$$\begin{aligned} \dot{x}^l_1 &= L^l(t, x^l_1, x^l_2, \dot{x}^l_2), \\ x^l_3 &= R^l(t, x^l_1, x^l_2), \end{aligned} \tag{5.14}$$

with vanishing strangeness index and d^l_μ differential and a^l_μ algebraic equations can be derived. Thus, a reduced system as in (5.13) or (5.14) with the same solution as the original DAE in mode l (5.3) can be extracted independently in each mode and therefore also for the complete hybrid system. Connecting all the reduced systems together we locally obtain an equivalent reduced hybrid system denoted by $\hat{\mathcal{H}}$, which is strangeness-free, i.e., $\mu_{max} = 0$. By solving the corresponding transformed systems, the solution of the complete hybrid system can be computed for every time t. From Theorem 2.40 it follows that every sufficiently smooth solution x^l of the DAE in mode l (5.3) that satisfies Hypothesis 2.37 also solves the reduced problems (5.13) and (5.14). For a hybrid systems $\mathcal{H}$ satisfying Hypothesis 5.6 we can show that every sufficiently smooth solution of $\mathcal{H}$ is also a solution of the equivalent strangeness-free hybrid system $\hat{\mathcal{H}}$.

Theorem 5.9. *Let $\mathcal{H}$ be a hybrid system as in Definition 5.3 with sufficiently smooth functions F^l in each mode $l \in \mathbb{M}$ that satisfy the Hypothesis 5.6 with characteristic values $\mu^l, a^l_\mu, d^l_\mu, v^l_\mu$ and $u^l_\mu = n_l - a^l_\mu - d^l_\mu$ in each mode $l \in \mathbb{M}$. Then, every sufficiently smooth solution of $\mathcal{H}$ is also a solution of $\hat{\mathcal{H}}$.*

Proof. Let $T_\tau = \{\mathbb{I}_i\}_{i=1,\dots,N_\mathbb{I}}$ and $T_m = \{l_i\}_{i=1,\dots,N_\mathbb{I}}$ be the hybrid time and mode trajectory corresponding to the hybrid system $\mathcal{H}$. If the sequence $\{x^*_{l_i}\}_{l_i \in T_m}$, corresponding to T_m is a sufficiently smooth solution of $\mathcal{H}$ determined from the DAEs F^{l_i} in each mode $l_i \in \mathbb{M}$, then the functions $x^*_{l_i}$ also solve the reduced strangeness-free systems of DAEs (5.13) and (5.14), since for every mode l_i, and all $t \in D_{l_i}$

$$(t, x^*_{l_i}(t), \dot{x}^*_{l_i}(t), \dots, (\tfrac{d}{dt})^{\mu^{l_i}+1} x^*_{l_i}(t)) \in \mathbb{L}^{l_i}_{\mu^l}.$$

Since the transition functions yield consistent initial values after each mode change as $\{x^*_{l_i}\}_{l_i \in T_m}$ is a solution, these values are also consistent for the reduced systems. Thus, the sequence $\{x^*_{l_i}\}_{l_i \in T_m}$ is also a solution of the reduced hybrid system $\hat{\mathcal{H}}$. □

Note that the reduced hybrid system $\hat{\mathcal{H}}$ depends strongly on the choice and the consistency of the initial values in each mode. Therefore, in each interval $\mathbb{I}_i$ and for every mode l, the initial values must be chosen in a consistent way, so that the solution in each mode exists and is unique if $u^l_\mu = 0$.

Remark 5.10. *A special case of hybrid systems are* hybrid multibody systems*, where the DAEs in each mode $l \in \mathbb{M}$ are given by the equations of motion of a multibody system of the form*

$$\begin{aligned} M^l \ddot{p}^l &= f^l(t, p^l, \dot{p}^l) - G^l(p^l)^T \lambda^l, \\ 0 &= g^l(p^l). \end{aligned}$$

The strangeness index for multibody systems resulting from the constraints $g^l(p^l) = 0$ is given by $\mu^l = 2$, see e.g. [82]. In a mode l without constraints the strangeness index is $\mu^l = 0$. Therefore, the maximal strangeness index μ_{max} of a hybrid multibody system is at most 2, and it is $\mu_{max} = 0$ if the hybrid system is completely unconstrained. Further, the transition functions can be simplified, as the position variable p^l potentially stays the same in all modes and only changes in the algebraic variables λ^l occur, i.e., the transition functions only have to determine initial values for eventually new Lagrange multipliers λ^l.

5.3 Existence and Uniqueness of Solutions

Solution concepts for hybrid systems have to deal with nonsmooth solutions and with changes in the number of equations or unknowns after mode changes. Therefore, an investigation of solution concepts and sufficient conditions for the existence and uniqueness of solutions of hybrid differential-algebraic systems is required. In general, if we want overall continuous solutions of the hybrid system, then the transition functions must guarantee

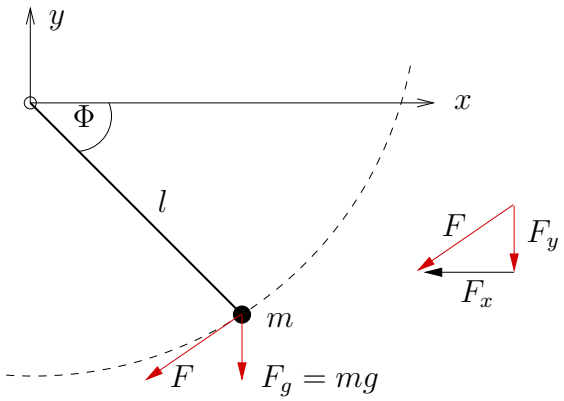

Figure 5.5: The accelerated pendulum

this. However, if the number of equations or the number of free variables changes at a mode change, then this condition may be difficult to realize. In particular, we may face the situation that the solution is not unique after a mode change. In any case, we need consistency of the initial values with the DAE in the new mode for the existence of a solution. In this section, we derive necessary and sufficient conditions for existence and uniqueness of solutions of hybrid differential-algebraic systems. Since different solution concepts for DAEs, see e.g. [81, 82, 86], can be applied for the DAEs in each mode $l \in \mathbb{M}$, in the following we consider the classical solvability concept resulting in continuously differentiable functions as solution of the DAEs in each mode and the concept of strong solutions allowing weaker smoothness requirements. The concept of generalized solutions defined in a distributional setting allowing also nonsmooth behavior is considered in Section 5.3.2.

A major difficulty in defining a solution for the overall hybrid system $\mathcal{H}$ is that at a switch point not only the index of the DAE, but also the number of unknowns of the system and the number of differential, algebraic and undetermined variables may change.

Example 5.11. [61] Consider a pendulum of mass m and length l under the influence of gravity $F_g = -mg$ that is tangentially accelerated by a linearly increasing force $F(x, y) = F_x(x) + F_y(y)$ in the position coordinates x and y, see Figure 5.5. The classical constrained motion of the system is described by the following DAE

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

This system is of strangeness index $\mu = 2$ with two differential variables x, y and one algebraic variable λ. If we suppose that the rope is cut when we reach a certain centrifugal force $F_{c_{max}}$, i.e., if $\dot{x}^2 + \dot{y}^2 > F_{c_{max}}$, the system changes from a pendulum to a flying mass point. In this case, the system is not constrained anymore and the equations of motions

are given by

$$\begin{aligned} m\ddot{x} &= 0, \\ m\ddot{y} &= -mg. \end{aligned}$$

This system is an ordinary differential equation, which is strangeness-free, with two differential variables x, y. In general, it is not clear how the algebraic variable λ should be continued after the mode change such that the solution of the overall system (x, y, λ) is not unique anymore after the mode change. But clearly in this simple case we can just choose λ constant as the last value.

In the following, let

$$n := \max_{l \in \mathbb{M}} n_l$$

be the maximal size of all solution vectors x^l. If $n > n_l$ in some mode $l \in \mathbb{M}$, then it is not clear how some solution components can be continued in this mode after a mode change. In this case, we extend the system in mode l by solution components $\hat{x}^l$ of size $n - n_l$, i.e., we consider the extended system

$$\begin{bmatrix} E^l(t) & 0 \end{bmatrix} \begin{bmatrix} \dot{x}^l \\ \dot{\hat{x}}^l \end{bmatrix} = \begin{bmatrix} A^l(t) & 0 \end{bmatrix} \begin{bmatrix} x^l \\ \hat{x}^l \end{bmatrix} + b^l(t). \tag{5.15}$$

For the first n_l components of the solution vector the extended system (5.15) has the same solution as the original system. Furthermore, (5.15) has the same strangeness index as the original system in mode l. But, now we have to deal with nonuniqueness of solutions, while the original system in mode l may have been uniquely solvable. With regard to the solution of the overall hybrid system this nonuniqueness was already present in the hybrid system before the extension (see Example 5.11), such that considering the extended systems (5.15) does not interfere with the solvability of the overall hybrid system. The nonuniqueness of solutions can be overcome by embedding the DAE (5.15) into a minimization problem

$$\frac{1}{2}\|\tilde{x}^l - \tilde{x}_0^l\|^2 = \min! \quad \text{such that} \quad \frac{1}{2}\|\tilde{E}^l(t)\dot{\tilde{x}}^l - \tilde{A}^l(t)\tilde{x}^l - b^l\|^2 = \min!, \tag{5.16}$$

where $\tilde{E}^l(t) = \begin{bmatrix} E^l(t) & 0 \end{bmatrix}$, $\tilde{A}^l(t) = \begin{bmatrix} A^l(t) & 0 \end{bmatrix}$, $\tilde{x}^l = \begin{bmatrix} x^l \\ \hat{x}^l \end{bmatrix}$, and $\tilde{x}_0^l \in \mathbb{R}^n$ is a given initial value. In this way, the undetermined state components are continued constantly with the last value of the previous mode described by the initial value $\tilde{x}_0^l$. For systems with well-defined strangeness index we may assume without loss of generality that the pair (E^l, A^l) and therefore also the pair $(\tilde{E}^l, \tilde{A}^l)$ are strangeness-free, since we can always transform the system to the corresponding strangeness-free form. Then, the minimization problem (5.16) has a unique solution which is called the *least squares solution* of the DAE, see [82].

5.3.1 Continuous Solutions of Linear Switched Systems

The first way to formulate necessary and sufficient conditions for the existence and uniqueness of solutions of linear hybrid systems is to demand continuous functions as solution of the hybrid system. In this section, we consider linear hybrid systems $\mathcal{H}$ and we assume without loss of generality that $n_l = n$ in all modes $l \in \mathbb{M}$. Locally in each mode $l \in \mathbb{M}$ and for every interval $\mathbb{I}_i \subseteq D_l$, we consider the classical solution concept introduced in Chapter 2, see Definition 2.26. If the strangeness index μ^l of (E^l, A^l) in mode $l \in \mathbb{M}$ is well-defined, i.e., Hypothesis 5.6 holds, and if b^l is sufficiently smooth, then by Theorem 2.36 the linear DAEs (5.7) in each mode are equivalent to strangeness-free systems of the form

$$\begin{aligned} \dot{x}_1^l &= A_{13}^l(t)x_3^l + b_1^l(t), \\ 0 &= x_2^l + b_2^l(t), \\ 0 &= b_3^l(t). \end{aligned} \tag{5.17}$$

with $A_{13}^l \in C(\mathbb{I}_i, \mathbb{R}^{d_\mu^l, u_\mu^l})$, $\mathbb{I}_i \subseteq D_l$ and the inhomogeneities b_i^l are determined by the derivatives of b^l. This equivalent strangeness-free formulation allows us to read off existence and uniqueness of solutions of the linear DAE (5.7) in mode $l \in \mathbb{M}$.

Corollary 5.12. *Let the strangeness index μ^l of the linear DAE (5.7) in mode $l \in \mathbb{M}$ be well-defined and let $b^l \in C^{\mu^l+1}(\mathbb{I}_i, \mathbb{R}^{m_l})$, with $\mathbb{I}_i = [\tau_i, \tau_i') \subseteq D_l$. Then we have:*

1. *The DAE (5.7) in mode l is solvable if and only if the v_μ^l functional consistency conditions*

$$b_3^l = 0$$

   are fulfilled.

2. *An initial condition $x^l(\tau_i) = x_{\tau_i}^l \in \mathbb{R}^{n_l}$ is consistent if and only if in addition the a_μ^l conditions*

$$x_2^l(\tau_i) = -b_2^l(\tau_i)$$

   are implied by $x^l(\tau_i) = x_{\tau_i}^l$.

3. *The corresponding initial value problem is uniquely solvable if and only if in addition it holds that*

$$u_\mu^l = 0.$$

Remark 5.13. *The smoothness assumption for the inhomogeneity b^l in Corollary 5.12 is used to guarantee that the solution is continuously differentiable, in particular with regard to the algebraic solution component x_2^l. To obtain a continuous solution the assumption that $b^l \in C^{\mu^l}(\mathbb{I}_i, \mathbb{R}^{m_l})$ is sufficient.*

The strangeness-free form (5.17) also allows to identify the minimal smoothness requirements for a solution, since only the variable x_1^l has to be differentiated. Thus, the concept of classical solutions can be weakend as the derivative $\dot{x}^l$ does not occur in the kernel of the coefficient matrix E^l, see also [81, 86].

Definition 5.14 (Strong solution). Consider a linear DAE (5.7) in mode l with sufficiently smooth coefficient functions E^l, A^l, and b^l. A function $x^l : \mathbb{I}_i \to \mathbb{R}^{n_l}$, $\mathbb{I}_i \subseteq D_l$ is called a *strong solution* of the DAE (5.7) in mode l if $\dot{x}^l$ exists in the cokernel of E^l, x^l is continuous and satisfies (5.7) pointwise.

For linear DAEs of the form (5.7) we can consider the projector function $(E^l)^+E^l$ which projects onto the cokernel of E^l (see Lemma 2.18), where $(E^l)^+$ denotes the pointwise Moore-Penrose pseudo-inverse of E^l (see Definition 2.17). With this, a weaker solution space for strangeness-free linear systems of the form (5.7) has been defined in [74, 82] as

$$C^1_{(E^l)^+E^l}(\mathbb{I}_i, \mathbb{R}^{n_l}) = \{x^l \in C(\mathbb{I}_i, \mathbb{R}^{n_l}) \mid (E^l)^+E^lx^l \in C^1(\mathbb{I}_i, \mathbb{R}^{n_l})\}.$$

The solution of a hybrid system $\mathcal{H}$ depends on the initial mode, initial conditions, mode switching conditions and on the transition functions. For a given time trajectory T_τ, with corresponding mode trajectory T_m, the initial mode as well as the mode switching sequence due to the transition conditions are fixed and the solution of the overall hybrid system is a sequence of continuous functions $x^l : \mathbb{I}_i \to \mathbb{R}^{n_l}$, with $\mathbb{I}_i \subseteq D_l$.

Definition 5.15 (Continuous solution of a linear hybrid system). A function

$$x \in C(\mathbb{I}, \mathbb{R}^n),$$

with $\mathbb{I} = [t_0, t_f] = \bigcup_{i=1}^{N_\mathbb{I}} \mathbb{I}_i$ is called a *continuous solution of a linear hybrid system* $\mathcal{H}$ with hybrid time trajectory $T_\tau = \{\mathbb{I}_i\}_{i=1,\dots,N_\mathbb{I}}$ and corresponding hybrid mode trajectory $T_m = \{l_i\}_{i=1,\dots,N_\mathbb{I}}$ if

$$x|_{\mathbb{I}_i} \in C^1_{(E^{l_i})^+E^{l_i}}(\mathbb{I}_i, \mathbb{R}^n) \text{ for all } \mathbb{I}_i \in T_\tau,\ l_i \in T_m,$$

and $x|_{\mathbb{I}_i}$ is a strong solution of the DAE (5.7) in the corresponding mode l_i. The function x is called a *continuous solution of the hybrid initial value problem* with initial condition $x_0 \in \mathbb{R}^n$ at t_0 if it is a continuous solution and satisfies the initial condition $x(t_0) = x_0$.

Definition 5.16 (Consistency of initial conditions). An initial condition $x(t_0) = x_0$ is called *consistent* with the hybrid system $\mathcal{H}$ if the corresponding hybrid initial value problem has at least one solution.

From Corollary 5.12 we get conditions for the existence and uniqueness of solutions locally in each mode. If we assume that the DAEs in each mode are solvable, we can give conditions for the existence and uniqueness of a continuous solution of a hybrid system $\mathcal{H}$.

Assumption 5.17. *For a linear hybrid system $\mathcal{H}$ let the strangeness-index μ^l be well-defined for all modes $l \in \mathbb{M}$ and assume that the linear DAEs (5.7) in each mode are solvable, i.e., $b_3^l = 0$, provided that consistent initial conditions are given.*

Under Assumption 5.17 there exists a solution if at each mode change the transition function is such that the resulting initial condition is consistent and if moreover $u_\mu^l = 0$ for all $l \in \mathbb{M}$, then the solution is unique as well.

Theorem 5.18. *Consider a linear hybrid system $\mathcal{H}$ that satisfies Hypothesis 5.6 with hybrid time trajectory T_τ, corresponding hybrid mode trajectory T_m, and a initial value $x_0 \in \mathbb{R}^n$. Let $\mathcal{E}(T_\tau)$ be the set of event times. Further, assume that Assumption 5.17 holds. Then there exists a continuous solution x of the linear hybrid system $\mathcal{H}$ in the sense of Definition 5.15 if and only if*

1. *the initial value x_0 is consistent for the DAE in the initial mode $l_1 \in T_m$,*
2. *the transition functions $T_{l_{i-1}}^{l_i}$ are the identity mappings, i.e.,*

$$T_{l_{i-1}}^{l_i}(x(\tau_{i-1}'), \dot{x}(\tau_{i-1}')) = [x(\tau_{i-1}'), \dot{x}(\tau_{i-1}')] = [x(\tau_i), \dot{x}(\tau_i)],$$

   and for every $\tau_i \in \mathcal{E}(T_\tau)$ the states $x(\tau_i)$ are consistent with the DAE in mode $l_i \in T_m$.

The continuous solution x is unique if and only if in addition $u_\mu^l = 0$ for all modes $l \in T_m$.

Proof. Due to Assumption 5.17, the DAE (5.7) in each mode is solvable provided that consistent initial values are given. For the solvability of the hybrid system we therefore need consistency of the initial value x_0 for the DAE in the initial mode l_1 and in addition consistency of the values given by the transition functions after each mode change. Further, if $T_{l_i}^{l_{i+1}}$ is the identity mapping we have $x^{l_{i+1}}(\tau_{i+1}) = x^{l_i}(\tau_i')$, which ensures continuity of the solution. Obviously, if there are no free solution components, i.e., $u_\mu^l = 0$ for all modes $l \in \mathbb{M}$, and the initial conditions are consistent, then in each mode the solution exists and is unique. □

5.3.2 Generalized Solutions of Linear Switched Systems

In the solutions of hybrid systems discontinuities at switch points can occur. In order to deal with these discontinuities we can consider impulsive smooth solutions which can be treated within a standard distributional framework as introduced in Section 2.2.3. In this way, solutions can be defined over discontinuities at switch points, but this approach requires infinitely often differentiable matrix functions E^l, A^l and right hand sides, which is not always fulfilled (see also the Remarks in Section 2.2.4).
In a hybrid system $\mathcal{H}$ discontinuities or jumps in the solution can occur at the switch points $\tau_j \in \mathcal{E}(T_\tau)$. In the following, we consider impulsive smooth functions on the set of event times $\mathcal{E}(T_\tau)$ as solutions of $\mathcal{H}$, i.e., we consider the set of impulsive smooth distributions $\mathcal{C}_{imp}^n(\mathcal{E}(T_\tau))$. The transfer of the states after mode changes by the transition functions may cause a nontrivial impulsive part resulting in an instantaneous jump in the solution, but nevertheless the state has to be transfered in a consistent way. Note, that also inconsistency of initial values can be treated with the distributional approach as presented in Section 2.2.3, but this will not be considered here. Now, we can define generalized solutions of a linear hybrid system.

Definition 5.19 (Generalized solution of linear hybrid systems). A function

$$x : \mathbb{I} \to \mathbb{R}^n,$$

with $\mathbb{I} = [t_0, t_f] = \bigcup_{i=1}^{N_\mathbb{I}} \mathbb{I}_i$ is called a *generalized solution of a linear hybrid system* $\mathcal{H}$ with hybrid time trajectory $T_\tau = \{\mathbb{I}_i\}_{i=1,\dots,N_\mathbb{I}}$, corresponding hybrid mode trajectory $T_m = \{l_i\}_{i=1,\dots,N_\mathbb{I}}$, and set of event times $\mathcal{E}(T_\tau)$, if

$$x \in \mathcal{C}^n_{imp}(\mathcal{E}(T_\tau)),$$

and x satisfies the linear DAE (5.7) in mode $l_i \in T_m$ for every $t \in \mathbb{I}_i$, $i = 1, \dots, N_\mathbb{I}$. The function x is called a *generalized solution of the linear hybrid initial value problem* with initial condition $x_0 \in \mathbb{R}^n$ at t_0, if it is a generalized solution of the hybrid system and satisfies one of the initial condition

$$x(t_0) = x_0, \quad x(t_0^+) = x_0.$$

In the following, we specify conditions to guarantee the existence and uniqueness of solutions for a linear hybrid system $\mathcal{H}$ with impulsive smooth solutions assuming solvability of the DAE in each mode.

Assumption 5.20. *For a linear hybrid system $\mathcal{H}$ as in Definition 5.3 let the matrix-valued functions $E^l, A^l \in C^\infty(\mathbb{I}, \mathbb{R}^{m,n})$ and let the strangeness index μ^l be well-defined for all modes $l \in \mathbb{M}$. Further, assume that the linear DAE (5.7) in each mode is solvable.*

Theorem 5.21. *Consider a linear hybrid system $\mathcal{H}$, with hybrid time trajectory T_τ, corresponding hybrid mode trajectory T_m, and an initial value $x_0 \in \mathbb{R}^n$ and assume that Assumption 5.20 holds. Then there exists a generalized solution of the linear hybrid system $\mathcal{H}$ in the sense of Definition 5.19 if and only if*

1. *the initial value x_0 is consistent for the initial mode $l_1 \in T_m$, and*
2. *for every $\tau_i \in \mathcal{E}(T_\tau)$ the values $x(\tau_i^+)$ obtained from the transition functions*

$$T^{l_i}_{l_{i-1}}(x(\tau_i^-), \dot{x}(\tau_i^-)) = [x(\tau_i^+), \dot{x}(\tau_i^+)]$$

are consistent with the DAE in mode $l_i \in T_m$.

The generalized solution is unique if and only if in addition $u^l_\mu = 0$ for all modes $l \in T_m$.

Proof. Due to Assumption 5.20, all individual modes have a solution for consistent initial conditions. Thus, for a consistent initial state $x_0 \in \mathbb{R}^n$ and initial mode l_1 there exists a smooth solution for the DAE in mode l_1. The system stays in the initial mode l_1 until the next time event $\tau_2 \in \mathcal{E}(T_\tau)$. If the initial condition $x(\tau_2^+)$ obtained from $T^{l_2}_{l_1}(x(\tau_2^-), \dot{x}(\tau_2^-)) = [x(\tau_2^+), \dot{x}(\tau_2^+)]$ is consistent with the DAE in the new mode, then there exists a smooth solution for the DAE in the new mode and only a state jump at τ_2 occurs. The same holds for all following mode switches at $\tau_i \in \mathcal{E}(T_\tau)$. Obviously, the solution is unique if the solution in each mode is unique, i.e., if $u^l_\mu = 0$ for all modes. □

5.3.3 Solutions of Nonlinear Switched Systems

In this section we consider solvability conditions for hybrid systems $\mathcal{H}$ with nonlinear DAEs (5.3) in each mode. To analyze nonlinear problems one usually uses the implicit function theorem (Theorem 2.9) to show that a solution is locally unique. To be able to apply the implicit function theorem we must require that for a given solution the derivative of F^l has a continuous inverse. Therefore, we need to assume that the differential-algebraic system in each mode is sufficiently smooth in a small interval following the switch point such that the implicit function theorem can be applied at all switch points τ_j.

Assumption 5.22. *Consider a hybrid system $\mathcal{H}$ as in Definition 5.3 with nonlinear DAEs of the form (5.3) and $n_l = n$ in each mode $l \in \mathbb{M}$. Assume that F^l is sufficiently smooth in $[\tau_i, \tau_i' + \epsilon]$ for small $\epsilon > 0$ for each interval $\mathbb{I}_i = [\tau_i, \tau_i') \in D_l$.*

Remark 5.23. *Also in the nonlinear case we assume that $n_l = n$ for all $l \in \mathbb{M}$. Otherwise, undetermined components are inserted into the system. This causes nonuniqueness of the solution but a definition of an overall solution is possible.*

Linearization of the nonlinear DAE (5.3) in mode $l \in \mathbb{M}$ along a solution trajectory x^l yields a linear DAE with variable coefficients in the form (5.7) with

$$E^l(t) = F^l_{;\dot{x}}(t, x^l, \dot{x}^l), \quad A^l(t) = -F^l_{;x}(t, x^l, \dot{x}^l), \quad b^l(t) = -F^l(t, x^l, \dot{x}^l) = 0, \tag{5.18}$$

such that locally similar results as for the linear case can be expected. Further, it can be shown that differentiation and linearization commute, i.e., linearization of the nonlinear derivative array (5.12) along a solution yields the same results as the derivative array based on the linearization of the nonlinear DAE along a solution, see e.g. [23, 82].
We can locally transform the nonlinear DAEs (5.3) to the reduced systems (5.13) or (5.14) and obtain existence conditions for a continuous solution. Sufficient conditions such that the reduced system (5.13) in mode l locally reflects the solvability properties of the original system (5.3) in mode l are given in the following Theorem.

Theorem 5.24. *Consider a hybrid system $\mathcal{H}$ as in Definition 5.3 with sufficiently smooth function F^l as in (5.3) in each mode $l \in \mathbb{M}$ that satisfies the Hypothesis 5.6 with characteristic values $\mu^l, a^l_\mu, d^l_\mu, v^l_\mu$ and with characteristic values $\mu^l + 1$ (replacing μ^l), $a^l_\mu, d^l_\mu, v^l_\mu$ in each mode $l \in \mathbb{M}$. For each $l \in \mathbb{M}$ let $z_{\mu^l+1,i} \in \mathbb{L}^l_{\mu^l+1}$ be given and let the parameterization p in (2.18) for $\mathcal{F}^l_{\mu^l+1}$ include $\dot{x}^l_2$. Then, for every function $x^l_2 \in C^1(\mathbb{I}_i, \mathbb{R}^{n_l - a^l_\mu - d^l_\mu})$, $\mathbb{I}_i \subseteq D_l$ with $x^l_2(\tau_i) = x^l_{2,i}$, $\dot{x}^l_2(\tau_i) = \dot{x}^l_{2,i}$, the reduced differential-algebraic systems (5.13) and (5.14) have unique solutions x^l_1 and x^l_3 satisfying $x^l_1(\tau_i) = x^l_{1,i}$. Moreover, the so obtained function $x^l = (x^l_1, x^l_2, x^l_3)$ locally solves the original problem (5.3) in mode l.*

Proof. The Theorem follows directly from [82, Theorem 4.34]. □

With this we can give conditions for the solvability of the overall hybrid system $\mathcal{H}$ similar to Theorem 5.18.

Theorem 5.25. *Consider a hybrid system $\mathcal{H}$ that satisfies Hypothesis 5.6 with hybrid time trajectory T_τ, corresponding hybrid mode trajectory T_m, and a initial value $x_0 \in \mathbb{R}^n$. Let $\mathcal{E}(T_\tau)$ be the set of event times and let the assumptions of Theorem 5.25 be fulfilled. Further, let Assumption 5.22 hold. Then there exists a continuous solution x of the hybrid system $\mathcal{H}$ in the sense of Definition 5.15 (with E^l as in* (5.18)*) if and only if*

1. *the initial value x_0 is consistent for the DAE in the initial mode $l_1 \in T_m$,*
2. *the transition functions $T^{l_i}_{l_{i-1}}$ are the identity mappings, i.e.,*

$$T^{l_i}_{l_{i-1}}(x(\tau'_{i-1}), \dot{x}(\tau'_{i-1})) = [x(\tau'_{i-1}), \dot{x}(\tau'_{i-1})] = [x(\tau_i), \dot{x}(\tau_i)],$$

and for every $\tau_i \in \mathcal{E}(T_\tau)$ the states $x(\tau_i)$ are consistent with the DAE in mode $l_i \in T_m$.

The continuous solution x is unique if and only if in addition $u^l_\mu = 0$ for all modes $l \in T_m$.

Proof. The proof is analogous to the proof of Theorem 5.18. □

5.4 Consistent Reinitialization

One of the difficulties in the numerical integration of differential-algebraic equations is to compute consistent initial values before starting the integration, i.e., calculating values at the initial time t_0 that satisfy the given algebraic constraints as well as the hidden constraints for higher index problems. For switched differential-algebraic equations consistent initial values are needed in addition at all switch points and therefore may have to be computed frequently during the simulation. Thus, an efficient and accurate reinitialization routine is required as the computation of consistent initial values influence the transition conditions and thus the mode switching of the system. Reinitialization of DAEs after discontinuities has been discussed in [12, 124] for regular linear time-invariant systems and in [96] for quasi-linear d-index 1 DAEs by solving nonlinear consistency equations. In this section, we discuss the consistent reinitialization for general nonlinear hybrid differential-algebraic systems after mode switching. Here, we use the ideas of the consistent initialization for nonlinear DAEs as described in [82], which allows over- and underdetermined solutions and also allows to fix certain state components and change others.

In the following, we consider a general nonlinear hybrid system $\mathcal{H}$ as in Definition 5.3 with an initial value x_0 at some initial or switching time $\tau_i \in \mathcal{E}$ and we assume that Hypothesis 2.37 holds in a neighborhood of a path $(\tau_i, x^\star(\tau_i), \mathcal{P}(\tau_i))$ belonging to a solution $x^\star \in C^1(\mathbb{I}, \mathbb{R}^{n_l})$ of the nonlinear DAE (5.3) in mode $l \in \mathbb{M}$. Here, the function $\mathcal{P} \in C(\mathbb{I}, \mathbb{R}^{(\mu^l+1)n_l})$ is a parameterization of the solution set $\mathbb{L}^l_{\mu^l}$ that coincides with $\dot{x}^\star$ in the first n_l components, i.e., $\mathcal{P}(t)[I_{n_l}\ 0 \dots\ 0]^T = \dot{x}^\star(t)$, such that $\mathcal{F}^l_{\mu^l}(t, x^\star(t), \mathcal{P}(t)) \equiv 0$. It has been shown in ([82], Remark 4.15) that every (x_0, y_0) with $y_0 = (\dot{x}_0, \dots, x_0^{(\mu^l+1)})$ in a neighborhood of $(x^\star(\tau_i), \mathcal{P}(\tau_i))$ can be locally extended to a solution of (5.3). Thus, consistency of an initial value x_0 at time τ_i means that (τ_i, x_0) is part of some $(\tau_i, x_0, y_0) \in$

$\mathbb{L}^l_{\mu^l}$. To determine a consistent initial value or to check it for consistency we must therefore solve the underdetermined system

$$\mathcal{F}^l_{\mu^l}(\tau_i, x_0, y_0) = 0 \tag{5.19}$$

for (x_0, y_0). We use the Gauss-Newton method [33, 113] started with a sufficiently good initial guess $(\tilde{x}_0, \tilde{y}_0)$ to solve this underdetermined systems of nonlinear equations in a least squares sense. The *Gauss-Newton method* for a nonlinear system of the form

$$G(z) = 0,$$

where $G : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth function, generates a sequence $\{z^k\}$ of approximations of the form

$$z^{k+1} = z^k - G^+_{;z}(z^k)G(z^k), \tag{5.20}$$

starting with an initial guess $z^0 \in \mathbb{R}^n$. Here, $G^+_{;z}(z)$ is the Moore-Penrose pseudo-inverse of the Jacobian $G_{;z}(z) = [\frac{\partial G_i}{\partial z_j}(z)]_{i,j}$ (see Definition 2.17).

Theorem 5.26. *Let $G : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$, $m < n$, with D open, convex denote a continuously differentiable mapping, and assume a full row rank of the Jacobian. Consider the Gauss-Newton method* (5.20) *and assume that a starting point $z^0 \in D$, and constants $\alpha, \omega \geq 0$ exist such that $\|G^+_{;z}(z^0)G(z^0)\| \leq \alpha$, and $\|G^+_{;z}(u)(G_{;z}(v) - G_{;z}(u))(v-u)\| \leq \omega\|v-u\|^2$ for all $u, v \in D$, $v - u \in \operatorname{range}(G^+_{;z}(u))$. Moreover, let*

$$h := \alpha\omega < 2, \quad \overline{S}(z^0, r) \subset D \text{ with } r := 2\alpha/(2-h).$$

Then:

1. *The sequence $\{z^k\}$ of Gauss-Newton iterates is well-defined, remains in $\overline{S}(z^0, r)$ and converges to some $z^\star \in \overline{S}(z^0, r)$ with $G^+_{;z}(z^\star)G(z^\star) = 0$.*
2. *Quadratic convergence can be estimated according to*

$$\|z^{k+1} - z^k\| \leq \frac{1}{2}\omega\|z^k - z^{k-1}\|^2.$$

Proof. See [33, Theorem 4.19]. □

Thus, if the Jacobian $G_{;z}(z)$ has full row rank and fulfills a Lipschitz condition in an open convex set, and if $G^+_{;z}(z^k)$ is bounded in this set, then we have quadratic convergence of the iterates $\{z^k\}$ to a least squares solution $z^\star$ of $G(z) = 0$ satisfying

$$G^+_{;z}(z^\star)G(z^\star) = 0.$$

Since the Jacobian of (5.19) at a solution (x_0, y_0) has full row rank at a solution and thus in a whole neighborhood by Hypothesis 2.37, we have local quadratic convergence of

the Gauss-Newton method, provided that the starting point z^0 is sufficiently close to the solution.

During the numerical integration of a hybrid system, events cause switching from mode l to mode k at a switch point τ_i. The transition function T_l^k as defined in (5.6) maps the state at the switch point in mode l to the state at the switch point in the new mode k via

$$T_l^k(x^l(\tau_i), \dot{x}^l(\tau_i)) = [x^\star, \dot{x}^\star].$$

But, the transfered state $x^\star$ is not necessarily consistent with the DAE in mode k at time τ_i, and to continue the integration a consistent initial value $x^k(\tau_i)$ at τ_i has to be computed. In order to find a reasonable continuation of the solution of the hybrid system, we try to find a consistent initial value $x^k(\tau_i)$ at τ_i from among all consistent values in the constraint manifold $\mathbb{L}^k$, on the basis of the given but inconsistent initial state $x^\star$, in such a way that the solution x^k extends the past solution x^l in a physically reasonable way. Since algebraic variables need to be chosen consistently with the DAE in the current mode they have to be computed as the solution of a nonlinear system describing the algebraic constraints. On the other hand, initial values for differential variables and possibly undetermined variables can be chosen freely, such that these components of the initial value vector should be kept fixed during the computation of consistent initial values in order to find a continuation of the hybrid system solution that is as smooth as possible. Thus, even if the transition function T_l^k provides continuity of the state variables over a switch point the consistent reinitialization can cause discontinuities in the solution. If possible, these discontinuities should only occur in the algebraic variables, which have to be consistent, while the differential variables and undetermined variables should proceed continuously over the switch point.
By a slight modification of the above approach it is possible to prescribe initial values for the differential variables and controls, i.e., fix certain components of an initial guess $\tilde{x}_0$ during the computation of consistent initial values, and only compute consistent values for the algebraic variables. This requires the classification of a component of $\tilde{x}_0$ to be a differential variable or a control, such that eliminating the associated columns from the Jacobian of the nonlinear system (5.19) does not lead to a rank deficiency, since we must guarantee that the remaining columns of the Jacobian still have full row rank to ensure quadratic convergence of the Gauss-Newton method. Due to Hypothesis 2.37, there exist continuous matrix functions

$$Z_2^l \in C(\mathbb{I}, \mathbb{R}^{(\mu^l+1)n_l, a_\mu^l}), \quad T_2^l \in C(\mathbb{I}, \mathbb{R}^{n_l, n_l - a_\mu^l}), \quad Z_1^l \in C(\mathbb{I}, \mathbb{R}^{n_l, d_\mu^l}),$$

with the properties described in Hypothesis 2.37 and pointwise orthonormal columns. Let the matrix functions

$$Z_2^{l'} \in C(\mathbb{I}, \mathbb{R}^{(\mu^l+1)n_l, (\mu^l+1)n_l - a_\mu^l}), \quad T_2^{l'} \in C(\mathbb{I}, \mathbb{R}^{n_l, a_\mu^l}), \quad Z_1^{l'} \in C(\mathbb{I}, \mathbb{R}^{n_l, n_l - d_\mu^l}),$$

be chosen such that

$$\begin{bmatrix} Z_2^{l'} & Z_2^l \end{bmatrix}, \quad \begin{bmatrix} T_2^{l'} & T_2^l \end{bmatrix}, \quad \begin{bmatrix} Z_1^{l'} & Z_1^l \end{bmatrix}$$

are pointwise orthogonal, i.e., in particular nonsingular. Then, Hypothesis 2.37 yields

$$\hat{F}^l_{2;x} T^l_2 = 0, \quad \operatorname{rank} T^l_2 = n_l - a^l_\mu, \quad \operatorname{rank} \hat{F}^l_{1;\dot{x}} T^l_2 = d^l_\mu,$$

and multiplication with the nonsingular matrix $\begin{bmatrix} T^{l'}_2 & T_2 \end{bmatrix}$ yields the separation

$$\operatorname{rank} \begin{bmatrix} \hat{F}^l_{1;\dot{x}} \\ \hat{F}^l_{2;x} \end{bmatrix} = \operatorname{rank} \begin{bmatrix} \hat{F}^l_{1;\dot{x}} T^{l'}_2 & \hat{F}^l_{1;\dot{x}} T^l_2 \\ \hat{F}^l_{2;x} T^{l'}_2 & 0 \end{bmatrix} = \operatorname{rank} \hat{F}^l_{1;\dot{x}} T^l_2 + \operatorname{rank} \hat{F}^l_{2;x} T^{l'}_2 = d^l_\mu + a^l_\mu.$$

Now, let $\tilde{T}^l_2$ be a fixed approximation with orthonormal columns to T^l_2 that spans the nullspace of $\hat{F}^l_{2;x}$ at the desired solution. Then, we can solve

$$\mathcal{F}^l_{\mu^l}(t_0, \tilde{T}^l_2 \tilde{T}^{lT}_2 \tilde{x}_0 + (I - \tilde{T}^l_2 \tilde{T}^{lT}_2) x, \dot{x}, \ldots, x^{(\mu^l+1)}) = 0 \tag{5.21}$$

for $(x, \dot{x}, \ldots x^{(\mu^l+1)})$, with initial guess $\tilde{x}_0$, where $\tilde{T}^l_2 \tilde{T}^{lT}_2$ is an orthogonal projection of rank d^l_μ onto kernel $\hat{F}^l_{2;x}$, while $I - \tilde{T}^l_2 \tilde{T}^{lT}_2$ is a projection onto cokernel $\hat{F}^l_{2;x}$. The d^l_μ differential components of the initial guess $\tilde{x}_0$ are kept fixed during the Gauss-Newton iterations as the corresponding columns of the Jacobian are set to zero. Note that this approach will lead to a rank drop in the Jacobian if any of the algebraic variables are fixed. A drawback of this approach to solve the nonlinear systems (5.19) or (5.21) with the Gauss-Newton method is the limited region of convergence. This means that the Gauss-Newton method may not converge if the initial guess is not sufficiently close to the solution. Therefore, after mode switching, the starting value $(\tau_i, x^\star)$ for the Gauss-Newton iteration given by the transition function T^k_l should be sufficiently close to a solution in the new mode to guarantee convergence.

After a mode change from mode l to mode k, differential variables in the predecessor mode l may change to algebraic or undetermined variables in the successor mode k or vice versa. Assuming that $n^l = n^k$ (otherwise undetermined variables can be inserted into the system to meet this requirement, see Section 5.3) the different possibilities are summarized in Table 5.1. Whenever algebraic variables or undetermined parts change into differential variables, no problems with consistency occur as the initial conditions fits into the differential equation (cases 9,10,11,12 in Table 5.1). Thus, if $a^k_\mu \leq a^l_\mu$ then it is possible to obtain a continuous solution provided that the constraint manifold has not changed. On the other hand, if differential or undetermined variables change into algebraic variables, then inconsistency can occur and reinitialization results in discontinuities in the solution (cases 2,4,6,8 in Table 5.1). If $u^k_\mu > 0$, then the solution is not unique and the DAE can only be solved in a least squares sense. If variables change into undetermined variables (cases 3,5,6,7,12 in Table 5.1) and the least squares solution is obtained in such a way that $\|x\|_2$ is minimized, then the continuity condition and the minimum norm condition can contradict each other. Therefore, the minimization problem for the least squares solution should be chosen as in (5.16). In addition, the index of the differential-algebraic system might have changed form μ^l to μ^k. If $\mu^l \geq \mu^k$, then no problems occur, but if the index increases after a mode change, then higher smoothness requirements are needed to guarantee the

	Differential part	Algebraic part	Undetermined part	Changes in char. val.
1	$d_\mu^l = d_\mu^k$	$a_\mu^l = a_\mu^k$	$u_\mu^l = u_\mu^k$	—
2	$d_\mu^l = d_\mu^k$	$a_\mu^l < a_\mu^k$	$u_\mu^l > u_\mu^k$	$u \curvearrowright a$
3	$d_\mu^l = d_\mu^k$	$a_\mu^l > a_\mu^k$	$u_\mu^l < u_\mu^k$	$a \curvearrowright u$
4	$d_\mu^l > d_\mu^k$	$a_\mu^l < a_\mu^k$	$u_\mu^l = u_\mu^k$	$d \curvearrowright a$
5	$d_\mu^l > d_\mu^k$	$a_\mu^l = a_\mu^k$	$u_\mu^l < u_\mu^k$	$d \curvearrowright u$
6	$d_\mu^l > d_\mu^k$	$a_\mu^l < a_\mu^k$	$u_\mu^l < u_\mu^k$	$d \curvearrowright a + u$
7	$d_\mu^l > d_\mu^k$	$a_\mu^l > a_\mu^k$	$u_\mu^l < u_\mu^k$	$d + a \curvearrowright u$
8	$d_\mu^l > d_\mu^k$	$a_\mu^l < a_\mu^k$	$u_\mu^l > u_\mu^k$	$d + a \curvearrowright a$
9	$d_\mu^l < d_\mu^k$	$a_\mu^l = a_\mu^k$	$u_\mu^l > u_\mu^k$	$u \curvearrowright d$
10	$d_\mu^l < d_\mu^k$	$a_\mu^l > a_\mu^k$	$u_\mu^l = u_\mu^k$	$a \curvearrowright d$
11	$d_\mu^l < d_\mu^k$	$a_\mu^l > a_\mu^k$	$u_\mu^l > u_\mu^k$	$a + u \curvearrowright d$
12	$d_\mu^l < d_\mu^k$	$a_\mu^l > a_\mu^k$	$u_\mu^l < u_\mu^k$	$a \curvearrowright d + u$
13	$d_\mu^l < d_\mu^k$	$a_\mu^l < a_\mu^k$	$u_\mu^l > u_\mu^k$	$u \curvearrowright d + a$

Table 5.1: Changes in the characteristic quantities after a mode change from mode l to mode k

existence of a solution and more effort is needed to obtain the reduced system in the new mode which might alter the convergence of numerical methods.

Remark 5.27. *For the computation of consistent initial value for a linear DAE* (2.5) *(see also [82, p. 308ff]) we can consider the reduced system of the form* (2.24)*, where the algebraic equations are displayed directly. The condition for a given* $\tilde{x}_0$ *at* t_0 *to be consistent is given by*

$$\hat{A}_2(t_0)\tilde{x}_0 + \hat{b}_2(t_0) = 0. \tag{5.22}$$

In the case that the given $\tilde{x}_0$ *is not consistent, we can use* (5.22) *to determine a related consistent* x_0*. Setting* $\tilde{x}_0 = x_0 + \delta$ *we determine the correction* δ *by solving the minimization problem*

$$\|\delta\|_2 = min!$$

subject to the constraint

$$\|\hat{A}_2(t_0)\delta - \hat{b}_2(t_0) - \hat{A}_2(t_0)\tilde{x}_0\|_2 = min!.$$

The solution of this least squares problem is given by

$$\delta = \hat{A}_2^+(t_0)(\hat{A}_2(t_0)\tilde{x}_0 + \hat{b}_2(t_0)),$$

where $\hat{A}_2^+(t_0)$ is the Moore-Penrose pseudo-inverse of $\hat{A}_2(t_0)$. Since $\hat{A}_2(t_0)$ has full row rank a_μ, due to Theorem 2.41, it follows that $\hat{A}_2(t_0)\hat{A}_2^+(t_0) = I_a$, and therefore

$$\hat{A}_2(t_0)x_0 + \hat{b}_2(t_0) = \hat{A}_2(t_0)(\tilde{x}_0 - \delta) + \hat{b}_2(t_0) = \hat{A}_2(t_0)\tilde{x}_0 - (\hat{A}_2(t_0)\tilde{x}_0 + \hat{b}_2(t_0)) + \hat{b}_2(t_0) = 0.$$

Also in this case, we can prescribe initial values for the differential variables, whereas initial values for the algebraic variables are not known. This requires a separation of the unknown x into differential, algebraic and unknown parts. For a system in the form (2.24) we can compute an orthogonal matrix $U = [U_1, U_2]$ of size (n, n) such that

$$\hat{E}_1(t_0)\begin{bmatrix} U_1 & U_2 \end{bmatrix} = \begin{bmatrix} E_{11} & 0 \end{bmatrix},$$

where E_{11} has size (d_μ, d_μ) and is nonsingular. Then, we determine an orthogonal matrix $V = [V_1, V_2]$ of size $(n - d_\mu, n - d_\mu)$ such that

$$\hat{A}_2(t_0)U_2V = \begin{bmatrix} A_{22} & 0 \end{bmatrix},$$

where A_{22} is of size (a_μ, a_μ) and nonsingular. This allows a reinterpretation of variables as differential, algebraic or undetermined variables using the basis transformation

$$x = Q\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix}, \quad Q = \begin{bmatrix} U_1 & U_2V_1 & U_2V_2 \end{bmatrix},$$

with orthogonal matrix Q and corresponding DAE

$$\begin{bmatrix} E_{11} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & 0 \\ 0 & 0 & 0 \end{bmatrix}\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \end{bmatrix} + \begin{bmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{b}_3 \end{bmatrix}.$$

From the second block row we get a partitioning of the consistency condition (5.22) into

$$0 = [A_{21}(t_0)\ A_{22}(t_0)]\begin{bmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{bmatrix} + \hat{b}_2(t_0).$$

Now, let an estimate $\tilde{x}_0 = (\tilde{x}_{1,0}, \tilde{x}_{2,0})$ for a consistent initial value be given. Keeping $\tilde{x}_{1,0}$ fixed, we can determine a correction δ_2 for the estimate $\tilde{x}_{2,0} = x_{2,0} + \delta_2$ by solving the minimization problem

$$\|\delta_2\|_2 = min!$$

subject to the constraint

$$\|\hat{A}_{22}(t_0)\delta_2 - \hat{A}_2(t_0)\tilde{x}_0 - \hat{b}_2(t_0)\|_2 = min!,$$

i.e., $\delta_2 = \hat{A}_{22}^+(t_0)(\hat{A}_2(t_0)\tilde{x}_0 + \hat{b}_2(t_0))$. The corrected consistent initial condition is then given by $x_{1,0} = \tilde{x}_{1,0}$ and $x_{2,0} = \tilde{x}_{2,0} - \delta_2$, and thus $x_0 = Q\begin{bmatrix} \tilde{x}_{1,0} \\ \tilde{x}_{2,0} - \delta_2 \end{bmatrix}$.

5.5 Sliding Motion

A special phenomena that can occur during the simulation of hybrid systems is a cyclic changing between different modes of continuous operation, called *chattering* or *sliding*, for example if nearly equal thresholds for the transition conditions of different modes are given and the system starts to oscillate around these. These oscillations may be real in the physical model since hysteresis, delays and other dynamic nonidealities lead to fast oscillations. An example for such a system with physical chattering is the anti-lock braking system in automobiles, for a simple model of an anti blocking system in a truck see [34]. On the other hand, numerical errors may lead to numerical chattering as switching conditions may be satisfied due to local errors. The numerical solution of a hybrid system exhibiting chattering behavior requires high computational costs as small stepsizes are required to restart the integration after each mode change. In the worst case, the numerical integration breaks down, as it does not proceed in time, but chatters between modes. As chattering causes severe problems in the numerical simulation it has to be treated in an appropriate way. One possibility to prevent numerical chattering is the introduction of hysteresis such that the integration in each mode is done in an interval of a length bounded from below. Another way to avoid oscillations around switching surfaces and to reduce the computational costs is to detect regions in which chattering can occur and to approximate the system dynamics along the switching surface in this region. An additional mode, the so-called *sliding mode*, can be inserted into the system that represents the dynamics during sliding, and thus replaces the chattering. In the following, we will first consider sliding motion for ODEs, extend the ideas to DAEs and finally apply the results to switched differential-algebraic systems. Furthermore, in Section 5.5.4 we present the basic ideas of introducing hysteresis to prevent chattering behavior.

5.5.1 Sliding Motion for Ordinary Differential Equations

Sliding motion is well understood for ordinary differential equations, see e.g. [39, 40, 145, 147]. To explain the basic ideas, we consider the following autonomous ODE system with discontinuous right-hand side

$$\dot{x} = f(x), \tag{5.23}$$

where the function $f : \mathbb{D}_x \subset \mathbb{R}^n \to \mathbb{R}^n$ is piecewise continuous. This is no restriction of the general case as each non-autonomous systems can be transformed into an autonomous system by adding t as new variable. In particular, we restrict to the case where the function f is discontinuous on a smooth switching surface Γ given by

$$\Gamma := \{x \in \mathbb{D}_x \mid g(x) = 0\},$$

where $g : \mathbb{D}_x \to \mathbb{R}$ is continuously differentiable. Then, Γ separates the phase space $\mathbb{D}_x$ into two domains

$$\Gamma^+ = \{x \in \mathbb{D}_x \mid g(x) > 0\} \quad \text{and} \quad \Gamma^- = \{x \in \mathbb{D}_x \mid g(x) < 0\},$$

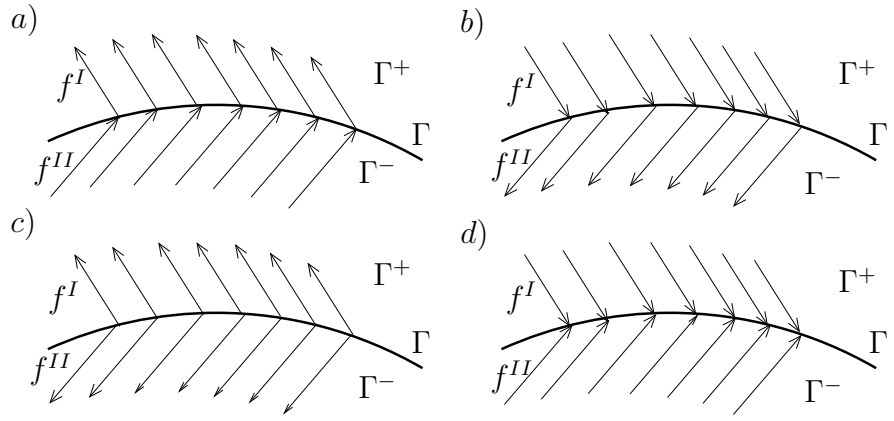

Figure 5.6: Phase space behavior at a switching surface

and we can consider the following system of differential equations

$$\dot{x} = \begin{cases} f^I(x) & \text{for } x \in \Gamma^+, \\ f^{II}(x) & \text{for } x \in \Gamma^-, \end{cases} \tag{5.24}$$

where $f^I = f|_{\Gamma^+}$ and $f^{II} = f|_{\Gamma^-}$. The system (5.23) is completely described by (5.24) in the domains Γ^+ and Γ^-, but on the switching surface Γ the standard definition of solution for ODEs may not be applicable, as the behavior of the solution of (5.23) on the switching surface is not defined. To cope with this, the discontinuous right-hand side of (5.23) can be replaced by a *differential inclusion*, see e.g. [40], i.e.,

$$\dot{x}(t) \in \eta(t, x). \tag{5.25}$$

If at a point (t, x) the function f is continuous, then the set $\eta(t, x)$ consist only of one point which is the value of the function f at this point. If (t, x) is a point of discontinuity of f, then the set $\eta(t, x)$ is to be defined in some other way.
In general, there are four types of solution behavior in the neighborhood of a switching surface characterized by the directions of the vector fields f^I and f^{II} as depicted in Figure 5.6. If the vector fields point towards the surface from one side and away from the surface from the other side as in cases *a*) and *b*) in Figure 5.6, the solution trajectory crosses the discontinuity and the system has a classical solution. On the other hand, if both vector fields point towards the switching surface Γ as in *d*) in Figure 5.6, then the solution cannot leave this manifold, but sticks to the manifold, and the solution can be defined via the differential inclusion (5.25). In this case chattering behavior during the numerical integration can occur as depicted in Figure 5.7. In the last case, where the vector fields on both sides point away from the surface as in *c*) in Figure 5.6, the switching surface cannot be crossed and there exists a point beyond which no classical solution exists.

In reality, small parameters in the system prevent the system from chattering and induce a smooth motion along the surface. In sliding motion the system dynamics are approximated

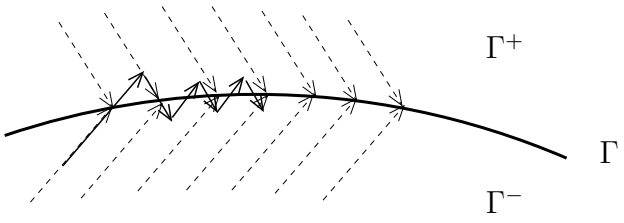

Figure 5.7: Chattering behavior along a switching surface

in such a way that the state trajectory moves along the switching surface. There are two main approaches to describe the dynamics of the system on a switching surface. The first approach is called *equivalence in dynamics* or *Filippov regularization*, see e.g. [39, 40, 145]. Here, approximations of the solution trajectories on both sides in a small neighborhood around the surface are used to determine the average velocity on the surface. Another approach called *equivalence in control* is presented in [145]. Here, free solution components, i.e., controls, in the system are chosen such that the solution trajectory moves along the switching surface. It has been stated in [145] that if the true system behavior near the switching surface can be attributed to hysteresis phenomena, then the method of equivalent dynamics derives sliding behavior closer to the true system behavior than the method of equivalent control. On the other hand, if there are no hysteresis effects the equivalent control method may generate better approximation. In the following, we will describe the approach of equivalence in dynamics in detail and afterwards shortly present the main ideas of the method of equivalence in control.
In the Filippov regularization, for each point $x \in \mathbb{D}_x$, the differential inclusion $\eta(t,x)$ is defined to be the smallest closed convex set containing all the limit values of $f(x^\star)$ for $x^\star \notin \Gamma$, $x^\star \to x$. Then, a function $x(t)$ is said to be a solution of (5.23) if it is absolutely continuous and satisfies (5.25) almost everywhere. For $x^\star$ approaching the point $x \in \Gamma$ from Γ^- and Γ^+, let the function $f(x^\star)$ have the limit values

$$\lim_{\substack{x^\star \in \Gamma^- \\ x^\star \to x}} f^{II}(x^\star) = f_\Gamma^{II}(x) \quad \text{and} \quad \lim_{\substack{x^\star \in \Gamma^+ \\ x^\star \to x}} f^{I}(x^\star) = f_\Gamma^{I}(x).$$

Then, the set $\eta(t,x)$ is the line segment joining the end points of the vectors $f_\Gamma^{II}(x)$ and $f_\Gamma^{I}(x)$ for $x \in \Gamma$. If this line is on one side of the tangent plane to the switching surface Γ, then the solution passes from one side of the surface to the other side, see Figure 5.8. On the other hand, if the line segment intersects the tangent plane, then the solutions approach Γ from both sides, see Figure 5.9. In this case, the standard notion of solution is not suitable as there is no indication of how a solution can be continued. Nevertheless, if the line segment intersects the tangent plane, the intersection point is the endpoint of the vector $f_\Gamma(x)$ which determines the velocity of the motion $\dot{x} = f_\Gamma(x)$ along the surface Γ at x. From (5.25) the solution $x(t)$ of the differential equation satisfies

$$\dot{x} = f_\Gamma(x), \tag{5.26}$$

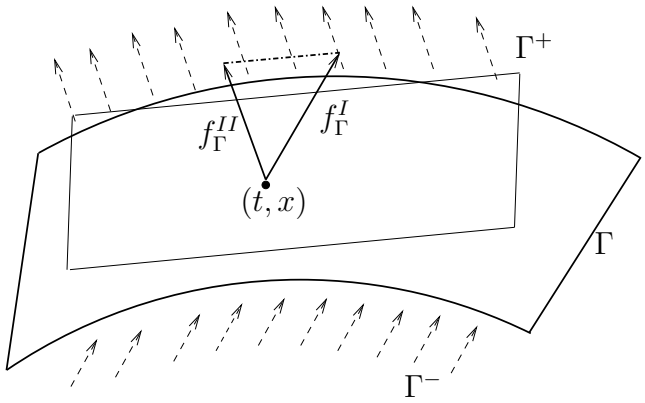

Figure 5.8: Regular switching at a switching surface

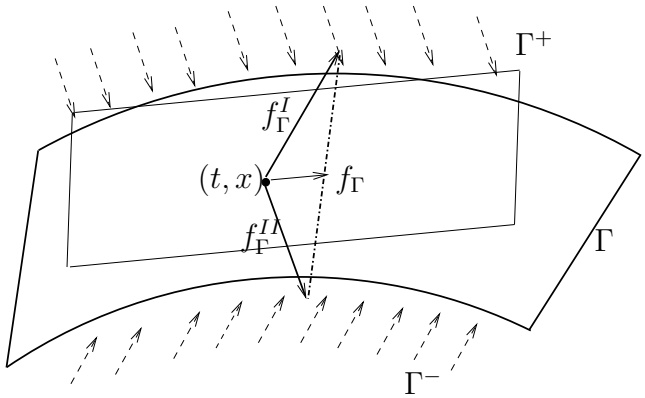

Figure 5.9: Filippov's construction of equivalent dynamics

where f_Γ is a linear combination of f_Γ^I and f_Γ^{II} and therefore it is also a solution of (5.23). If $x \in \Gamma^+$, then f_Γ equals f^I and if $x \in \Gamma^-$, then f_Γ equals f^{II}. Note that f_Γ is a particular selection from the set $\eta(t,x)$. It is also possible to define other differential inclusions as we will see below. The velocity vector f_Γ of sliding motion in (5.26) lies on a plane tangential to the surface, and therefore its end point is the intersection point of the tangential plane and the straight line connecting the end points of f_Γ^I and f_Γ^{II}. This line segment can be written as a convex combination of f_Γ^I and f_Γ^{II}, such that the equation for sliding motion is given by

$$\dot{x} = f_\Gamma(x) = \alpha f_\Gamma^I(x) + (1-\alpha) f_\Gamma^{II}(x), \quad (0 \le \alpha \le 1). \tag{5.27}$$

In the following, we assume that $g_{;x}(x) \neq 0$ in a neighborhood of the switching surface Γ. The parameter α should be selected such that the velocity vector is tangential to the switching surface , i.e., $g_{;x}(x) f_\Gamma(x) = 0$ and therefore α is given by

$$\alpha = [g_{;x}(f_\Gamma^{II} - f_\Gamma^I)]^{-1} g_{;x} f_\Gamma^{II}.$$

The Filippov construction of equivalent dynamics is depicted in Figure 5.9.

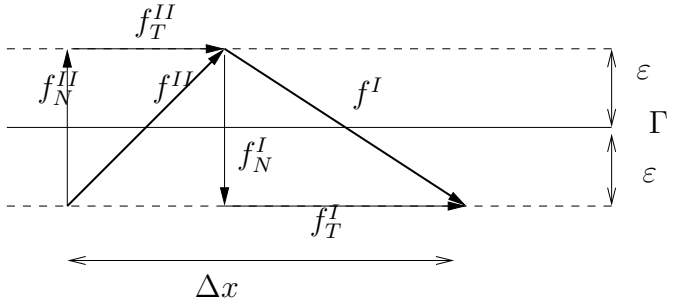

Figure 5.10: Equivalent dynamics via hysteresis effects

In real systems delays, hysteresis and other nonidealities result in real sliding. The sliding equations (5.27) derived by equivalent dynamics on the surface can be considered as the motion of a limiting process. If we consider an infinitesimal hysteresis band of width ε around the switching surface, then the dynamics on the surface are defined as the behavior in the limit as $\varepsilon \to 0$. Once the system hits the surface, oscillation in a neighborhood of width 2ε occur. If ε is small, then the velocity vectors f_Γ^{II} and f_Γ^I in the neighborhood of the discontinuity surface can be represented by their normal components f_N^{II}, f_N^I and tangential components f_T^{II}, f_T^I, i.e., we neglect curvature and the gradient of the surface is assumed to be constant. To determine the direction of the motion along the surface, we calculate the average velocity on the surface. The time to cross the ε band is $\Delta t_1 = \frac{2\varepsilon}{f_N^I}$ for f_Γ^I and $\Delta t_2 = -\frac{2\varepsilon}{f_N^{II}}$ for f_Γ^{II}, where $f_N^I = g_{;x} f_\Gamma^I$ and $f_N^{II} = g_{;x} f_\Gamma^{II}$ are the normal projections of f_Γ^I and f_Γ^{II} onto the switching surface, see also Figure 5.10. The time to move back and forth over the band is therefore given by

$$\Delta t = \Delta t_1 + \Delta t_2$$

and the tangential distance the system travels over the time interval Δt is

$$\Delta x = f_\Gamma^I \Delta t_1 + f_\Gamma^{II} \Delta t_2.$$

Then the average state velocity of the motion on the surface is given by

$$\dot{x}_{av} = \frac{\Delta x}{\Delta t} = \frac{\Delta t_1}{\Delta t} f_\Gamma^I + \left(1 - \frac{\Delta t_1}{\Delta t}\right) f_\Gamma^{II}, \tag{5.28}$$

with

$$\frac{\Delta t_1}{\Delta t} = \frac{f_N^{II}}{f_N^{II} - f_N^I}.$$

Thus, the average velocity (5.28) equals f_Γ in (5.27) and the equivalent dynamics on a sliding surface corresponds to the limiting behavior when switching tends to be infinitely fast.

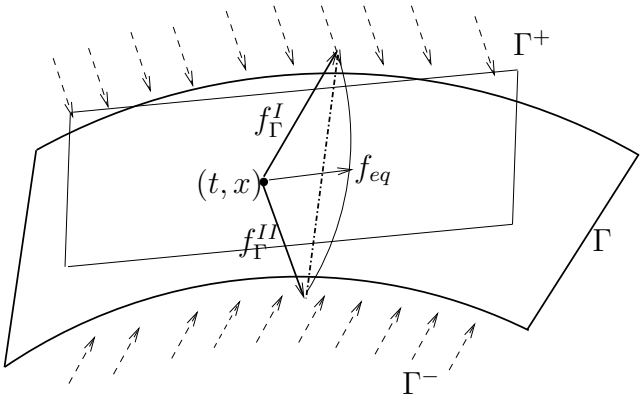

Figure 5.11: Equivalence in control vs. equivalence in dynamics

Another way to construct the set $\eta(t,x)$ in (5.25) is given by the equivalence in control method [145, 147]. In this case, we consider a system

$$\dot{x} = f(x, u(x)), \tag{5.29}$$

where $f : \mathbb{D}_x \times \mathbb{D}_u \to \mathbb{R}^n$ is a continuous function and the function $u : \mathbb{D}_x \to \mathbb{R}$ is discontinuous on a smooth switching surface $\Gamma = \{x \in \mathbb{D}_x \mid g(x) = 0\}$. At points belonging to the surface Γ we assume that the equation of sliding motion is given by

$$\dot{x} = f(x, u_{eq}(x)), \tag{5.30}$$

where the equivalent control u_{eq} is defined such that the vector f lies tangentially to the surface Γ and the value u_{eq} is contained in an interval $[u^-, u^+]$, where $u^\pm$ are limiting values of u on both sides of the surface Γ. In contrast to the Filippov construction, the endpoint of the vector $f(x, u_{eq}(x))$ lies on the intersection of the tangential plane to Γ at the point x with the arc that is spanned by the endpoint of the vector $f(x,u)$ when u varies from u^- to u^+. Thus, in this case the set $\eta(t,x)$ is an arc while in the Filippov construction $\eta(t,x)$ is the straight line connecting $f(x,u^+)$ and $f(x,u^-)$, see Figure 5.11. If the function f is linear in u, then near the switching surface Γ the equation (5.29) can be written in the form

$$\dot{x} = f_0(x) + B(x)u(x). \tag{5.31}$$

To obtain the motion along the surface Γ the equivalent control u_{eq} must be chosen such that $\dot{x}$ is tangential to the surface Γ, i.e.,

$$g_{;x}(x)\dot{x} = g_{;x}(x)f_0(x) + g_{;x}(x)B(x)u_{eq} = 0,$$

and thus

$$u_{eq} = -[g_{;x}(x)B(x)]^{-1}g_{;x}(x)f_0(x), \tag{5.32}$$

if $g_{;x}(x)B(x)$ is nonsingular. The regularity of $g_{;x}(x)B(x)$ is also known as the *transversality condition* and establishes that the control vector field $B(x)$ is not tangential to the switching surface Γ at any point $x \in \mathbb{D}_x$. If u_{eq} from (5.32) satisfies

$$u^- \leq u_{eq} \leq u^+ \quad \text{or} \quad u^+ \leq u_{eq} \leq u^-,$$

then by substituting the vector u_{eq} into (5.31), we obtain the velocity vector of sliding motion along Γ as

$$\dot{x} = f_0(x) - B(x)[g_{;x}(x)B(x)]^{-1} g_{;x}(x) f_0(x).$$

The equivalence in control method is also applicable if discontinuities occur along the intersection of several switching surfaces Γ_i and the control u is a vector with components u_i that are discontinuous on Γ_i, see [40, 145]. For systems that are linear with respect to the control the equivalence in control approach coincides with the Filippov construction.

In the case of sliding motion we pursue the solution along the switching manifold Γ. The Filippov construction (5.27) and also the equivalent control method approximate this motion as a motion tangential to the switching surface to construct an ordinary differential equation for sliding motion. From a DAE point of view a better way to define the system behavior during sliding is to append the condition that the solution should stay on the manifold Γ as an algebraic constraint and define the *differential-algebraic system in sliding motion* by

$$\begin{aligned} \dot{x} &= \alpha f^I(x) + (1-\alpha) f^{II}(x), \\ 0 &= g(x), \end{aligned} \tag{5.33}$$

where the algebraic variable α is chosen such that the solution remains in Γ.

Theorem 5.28. *Consider an ordinary differential system* (5.23) *where the right-hand side* $f(x)$ *is discontinuous on a smooth switching surface* $\Gamma = \{x \in \mathbb{D}_x \,|\, g(x) = 0\}$ *such that* (5.23) *can be separated into* $f^I(x)$ *and* $f^{II}(x)$ *as in* (5.24)*. If*

$$g_{;x}(x)(f^I(x) - f^{II}(x))$$

is nonsingular for all $x \in \mathbb{D}_x$*, then the equivalent dynamics of the system* (5.23) *during sliding motion are described by the DAE in sliding motion* (5.33)*, and the DAE* (5.33) *is of strangeness index* $\mu = 1$*.*

Proof. Every solution of (5.23) is also a solution of (5.33) for $\alpha = 0$ and $\alpha = 1$, respectively. Differentiation of the algebraic constraints yields

$$\begin{aligned} 0 = \frac{d}{dt} g(x) &= g_{;x}(x)\dot{x} \\ &= g_{;x}(x)(\alpha f^I(x) + (1-\alpha) f^{II}(x)) \\ &= \alpha g_{;x}(x)(f^I(x) - f^{II}(x)) + g_{;x}(x) f^{II}(x), \end{aligned}$$

which can be solved for α, if $g_{;x}(x)(f^I(x) - f^{II}(x))$ is nonsingular. Thus, we get a strangeness-free system after one differentiation and therefore $\mu = 1$. □

5.5.2 Sliding Motion for Differential-Algebraic Equations

The ideas of sliding motion for ODEs as described in Section 5.5.1 can be used to describe sliding motion for discontinuous DAEs. For a first approach in this direction in the case of semi-explicit d-index 1 DAEs in chemical engineering see [1]. Sliding motion for constrained multibody systems is also treated in [34, 88].
In this section, we consider a general nonlinear DAE of the form

$$F(t, x, \dot{x}) = 0, \tag{5.34}$$

with piecewise continuous function $F : \mathbb{I} \times \mathbb{D}_x \times \mathbb{D}_{\dot{x}} \to \mathbb{R}^n$, $\mathbb{D}_x, \mathbb{D}_{\dot{x}} \subset \mathbb{R}^n$ that is discontinuous on a smooth switching surface $\Gamma = \{(t, x) \in \mathbb{I} \times \mathbb{D}_x \mid g(t, x) = 0\}$ described by a switching function $g : \mathbb{I} \times \mathbb{D}_x \to \mathbb{R}$. Again, Γ separates the phase space into two domains

$$\begin{aligned} \Gamma^+ &= \{(t, x) \in \mathbb{I} \times \mathbb{D}_x \mid g(t, x) > 0\}, \\ \Gamma^- &= \{(t, x) \in \mathbb{I} \times \mathbb{D}_x \mid g(t, x) < 0\}. \end{aligned}$$

Therefore, we can rewrite the differential-algebraic equations as

$$\begin{cases} F^I(t, x, \dot{x}) = 0 & \text{for } (t, x) \in \Gamma^+, \\ F^{II}(t, x, \dot{x}) = 0 & \text{for } (t, x) \in \Gamma^-, \end{cases} \tag{5.35}$$

where $F^I = F|_{\Gamma^+}$ and $F^{II} = F|_{\Gamma^-}$ are smooth on $\Gamma^{\pm}$, and we assume that the strangeness-index is well-defined for F^I and F^{II} on $\Gamma^{\pm}$, respectively. By index reduction as described in Section 2.2.2, we can transform both systems in (5.35) to the corresponding reduced systems provided that Hypothesis 2.37 holds for F^I and F^{II} in Γ^+ and Γ^+, respectively. This means that without loss of generality we can consider the reduced systems

$$\begin{aligned} \hat{F}_1^I(t, x, \dot{x}) &= 0, \\ \hat{F}_2^I(t, x) &= 0, \end{aligned} \tag{5.36a}$$

for $(t, x) \in \Gamma^+$ and

$$\begin{aligned} \hat{F}_1^{II}(t, x, \dot{x}) &= 0, \\ \hat{F}_2^{II}(t, x) &= 0, \end{aligned} \tag{5.36b}$$

for $(t, x) \in \Gamma^-$. The reduced systems consist of decoupled ordinary differential equations and algebraic equations, where the equations $\hat{F}_1^I(t, x, \dot{x}) = 0$ and $\hat{F}_1^{II}(t, x, \dot{x}) = 0$ describe the dynamics of the system, while $\hat{F}_2^I(t, x) = 0$ and $\hat{F}_2^{II}(t, x) = 0$ are algebraic constraints that force the solution onto a specific manifold. Further, we assume that the solution of the DAE (5.35) within each region exists and is unique, i.e., there are no undetermined parts in the system. Otherwise, if there are undetermined parts in the system we can consider sliding mode control as described in Section 5.6.4. Further, we assume that changes in the characteristic values only occur on the switching surface Γ and the number of differential

and algebraic variables are the same for (5.36a) and (5.36b). Then, the reduced systems (5.36) can be further transformed to the systems

$$\begin{aligned} \dot{x}_1 &= L^I(t, x_1), \\ x_2 &= R^I(t, x_1), \end{aligned} \tag{5.37a}$$

for $(x_1, x_2) \in \Gamma^+$, and

$$\begin{aligned} \dot{x}_1 &= L^{II}(t, x_1), \\ x_2 &= R^{II}(t, x_1), \end{aligned} \tag{5.37b}$$

for $(x_1, x_2) \in \Gamma^-$. Now, we can use the Filippov construction to describe the equivalent dynamics of sliding motion along the switching surface by

$$\dot{x}_1 = \alpha L^I_\Gamma(t, x_1) + (1 - \alpha) L^{II}_\Gamma(t, x_1),$$

where for $t = const.$ and $(t, x_1, x_2) \in \Gamma$

$$L^I_\Gamma(t, x_1) = \lim_{(t,x^\star)\in\Gamma^+, x^\star \to x} L^I(t, x_1^\star), \quad L^{II}_\Gamma(t, x_1) = \lim_{(t,x^\star)\in\Gamma^-, x^\star \to x} L^{II}(t, x_1^\star),$$

and α is chosen such that the solution remains on the switching surface Γ given by the algebraic constraint $g(t, x_1, x_2) = 0$. To obtain the corresponding DAE for the system in sliding motion, the algebraic constraints have to be considered in an appropriate way to force the solution onto a specific manifold suited for both systems (5.37a) and (5.37b). In a similar way, this constraint manifold can be defined by rotation of the constraint manifolds $R^I(t, x_1)$ and $R^{II}(t, x_1)$, such that $R^I(t, x_1)$ is turned into $R^{II}(t, x_1)$ across the discontinuity or vice versa depending on the direction of the discontinuity crossing. This means that the constraint manifold during sliding motion is given by

$$x_2 = \alpha R^I_\Gamma(t, x_1) + (1 - \alpha) R^{II}_\Gamma(t, x_1),$$

where for $t = const.$ and $(t, x_1, x_2) \in \Gamma$

$$R^I_\Gamma(t, x_1) = \lim_{(t,x^\star)\in\Gamma^+, x^\star \to x} R^I(t, x_1^\star), \quad R^{II}_\Gamma(t, x_1) = \lim_{(t,x^\star)\in\Gamma^-, x^\star \to x} R^{II}(t, x_1^\star).$$

Altogether, the *DAE in sliding motion* is given by

$$\begin{aligned} \dot{x}_1 &= \alpha L^I(t, x_1) + (1 - \alpha) L^{II}(t, x_1), \\ x_2 &= \alpha R^I(t, x_1) + (1 - \alpha) R^{II}(t, x_1), \\ 0 &= g(t, x_1, x_2). \end{aligned} \tag{5.38}$$

Theorem 5.29. *Consider a regular DAE* (5.34) *that is discontinuous on a smooth switching surface given by* $\Gamma = \{(t, x) \in \mathbb{I} \times \mathbb{D}_x \,|\, g(t, x) = 0\}$ *such that* (5.34) *can be separated into the reduced systems* (5.37a) *and* (5.37b)*. If*

$$g_{;x_2}(t, x_1, x_2) \left(R^{II}(t, x_1) - R^I(t, x_1) \right)$$

is nonsingular for all $(t, x_1, x_2) \in \mathbb{I} \times \mathbb{D}_x$*, then the equivalent dynamics during sliding motion are described by the DAE* (5.38)*, and the system* (5.38) *is regular and of strangeness index* $\mu = 1$*.*

Proof. As system (5.37) is regular also system (5.38) is regular due to construction. Differentiating the two algebraic constraints yields

$$\dot{x}_2 = \dot{\alpha}(R^I - R^{II}) + \alpha(R^I_{;t} + R^I_{;x_1}\dot{x}_1) + (1-\alpha)(R^{II}_{;t} + R^{II}_{;x_1}\dot{x}_1),$$

and

$$0 = g_{;t} + g_{;x_1}\dot{x}_1 + g_{;x_2}\dot{x}_2, \tag{5.39}$$

omitting the function arguments. Replacing now the derivatives $\dot{x}_1$ and $\dot{x}_2$ in (5.39) yields

$$\begin{aligned}\dot{\alpha} g_{;x_2}(R^{II} - R^I) =& g_{;t} + g_{;x_1}(\alpha L^I + (1-\alpha)L^{II}) \\ &+ g_{;x_2}\left[\alpha R^I_{;t} + (1-\alpha)R^{II}_{;t} + (\alpha R^I_{;x_1} + (1-\alpha)R^{II}_{;x_1})(\alpha L^I + (1-\alpha)L^{II})\right],\end{aligned}$$

such that under the assumption that g is differentiable and

$$g_{;x_2}(R^{II} - R^I)$$

is nonsingular we get an explicit differential equation for the variable $\dot{\alpha}$. □

5.5.3 Sliding Motion for Switched Differential-Algebraic Systems

Sliding motion for discontinuous DAEs as described in the previous section can be used to handle chattering behavior in switched differential-algebraic systems. In this section, we consider the following scenario. During the integration of a hybrid system $\mathcal{H}$ a mode switch from mode l to mode k is detected. Let the two modes be separated by the j-th switching surface $\Gamma^l_j = \{(t,x) \in D_l \times \mathbb{R}^{n_l} \mid g^l_j(t,x) = 0\}$, $j \in J^l$, and assume that there exists a mode transition $\tilde{j} \in J^k$ such that $\Gamma^l_j = \Gamma^k_{\tilde{j}} = \{(t,x) \in D_k \times \mathbb{R}^{n_k} \mid g^k_{\tilde{j}}(t,x) = 0\}$, i.e., $g^l_j(t,x) = -g^k_{\tilde{j}}(t,x)$. Note that here, we restrict to switching functions independent of the state derivative $\dot{x}$. Similar as in the previous section, under the assumption of regularity and well-definedness of the strangeness index in each mode, the differential-algebraic systems in the adjacent modes l and k can be transformed to the strangeness-free forms

$$\begin{aligned}\dot{x}^l_1 &= L^l(t, x^l_1), \\ x^l_2 &= R^l(t, x^l_1),\end{aligned} \tag{5.40}$$

and

$$\begin{aligned}\dot{x}^k_1 &= L^k(t, x^k_1), \\ x^k_2 &= R^k(t, x^k_1).\end{aligned} \tag{5.41}$$

In the following, we assume that $g^l_{j;x}(x) \neq 0$ in a neighborhood of the switching surface Γ^l_j. The hybrid differential-algebraic system exhibit sliding motion or chattering behavior if the dynamical parts of the differential-algebraic systems (5.40) and (5.41) fulfill some

sliding condition. In Section 5.5.1 we have seen that sliding motion occurs if all solutions near the surface Γ_j^l approach it from both sides, i.e., if the projections of the vectors L^l and L^k onto the surface gradient are of opposite signs and are directed towards the surface from both sides in a neighborhood of the switching surface. Thus, sliding occurs at a point $(t,x) \in \Gamma_j^l$ if the *sliding condition*

$$L_N^l = g_{j;x_1^l}^l(t,x_1^l,x_2^l)L_\Gamma^l(t,x_1^l) < 0 \quad \text{and} \quad L_N^k = g_{\tilde{j};x_1^k}^k(t,x_1^k,x_2^k)L_\Gamma^k(t,x_1^k) > 0, \tag{5.42}$$

is satisfied, where for $t = const.$ and $(t,x) \in \Gamma_j^l = \Gamma_{\tilde{j}}^k$

$$L_\Gamma^l(t,x_1^l) = \lim_{(t,x^\star)\in\Lambda^l, x_1^\star \to x_1} L^l(t,x_1^\star), \quad L_\Gamma^k(t,x_1^k) = \lim_{(t,x^\star)\in\Lambda^k, x_1^\star \to x_1} L^k(t,x_1^\star).$$

This means that we consider the directional derivatives of $g_j^l = -g_{\tilde{j}}^k$ along L_Γ^l, L_Γ^k, respectively (see Definition 2.3) which correspond to the projections L_N^l and L_N^k of the vectors L_Γ^l and L_Γ^k onto the gradient of the switching surface Γ_j^l. These directional derivatives can be approximated numerically by

$$L_N^l \approx \frac{1}{\delta} g_j^l(t, x_1^l + \delta L^l(t,x_1^l), x_2^l), \quad L_N^k \approx \frac{1}{\delta} g_{\tilde{j}}^k(t, x_1^k + \delta L^k(t,x_1^k), x_2^k), \tag{5.43}$$

for small enough δ. The four different cases of phase space behavior near the switching surface can then be characterized in terms of the projections L_N^l and L_N^k as follows:

1. If $L_N^l > 0$ and $L_N^k > 0$, then the system switches from mode l to mode k.
2. If $L_N^l < 0$ and $L_N^k < 0$, then the system switches from mode k to mode l.
3. If $L_N^l > 0$ and $L_N^k < 0$, then both flows are directed away from the surface.
4. If $L_N^l < 0$ and $L_N^k > 0$, then the sliding condition is satisfied.

The *sliding surface* $\Gamma_S^l \subseteq \Gamma$ can then be defined by

$$\Gamma_S^l := \{(t,x) \in \Gamma_j^l \mid L_N^l(t,x) < 0 \text{ and } L_N^k(t,x) > 0\},$$

as that part of the switching surface, where sliding occurs. This means that trajectories in this region stay within this region until the boundary is reached, since trajectories leaving the sliding surface will immediately return to it. If the solution trajectory of a hybrid system directly traverse the discontinuity, i.e., the sliding condition is not fulfilled, the solution continues in mode k after the mode change. If the solution trajectories in both modes l and k are directed away from the surface, then the solution cannot be continued uniquely after the mode change. Except in the case of an unhappily chosen initial value, this normally should not occur.
In the numerical simulation of hybrid systems an immediate switch back to mode l after one or a few integration steps in the numerical solution in mode k would result if the

sliding condition is satisfied. To avoid this we can add an additional mode for the sliding motion by defining the DAE during sliding and switch to the sliding mode instead. The system should stay in sliding mode as long as the solution trajectory stays in the sliding region, and resume in mode l or k, depending on the sign of the directional derivatives. An additional difficulty in defining the DAE in sliding motion for hybrid systems is that it can also happen that the characteristic values or the index change at a mode switch. Let d_μ^l, d_μ^k and a_μ^l, a_μ^k denote the number of differential and algebraic equations in mode l and mode k, i.e., the dimension of x_1^l, x_1^k and x_2^l, x_2^k in (5.40) and (5.41), respectively. If $d_\mu^l = d_\mu^k$ and $a_\mu^l = a_\mu^k$, then the system during sliding can be defined as in (5.38). But, it may also happen that $d_\mu^l \neq d_\mu^k$ and differential variables change to algebraic variables or vice verse after the discontinuity. Let $d_\mu^l + a_\mu^l = d_\mu^k + a_\mu^k = n$ and without loss of generality assume that $d_\mu^l > d_\mu^k$ and $a_\mu^l < a_\mu^k$. Then x_1^l and x_2^k can be further partitioned into

$$x_1^l = \begin{bmatrix} x_{1,1}^l \\ x_{1,2}^l \end{bmatrix}, \quad x_2^k = \begin{bmatrix} x_{2,1}^k \\ x_{2,2}^k \end{bmatrix},$$

with $x_{1,1}^l \in \mathbb{R}^{d_\mu^k}$, $x_{1,2}^l \in \mathbb{R}^{d_\mu^l - d_\mu^k}$ and $x_{2,1}^k \in \mathbb{R}^{a_\mu^l}$, $x_{2,2}^k \in \mathbb{R}^{a_\mu^k - a_\mu^l}$. Furthermore, let the reduced systems (5.40) and (5.41) be partitioned accordingly into

$$\begin{aligned} \begin{bmatrix} \dot{x}_{1,1}^l \\ \dot{x}_{1,2}^l \end{bmatrix} &= \begin{bmatrix} L_1^l(t, x_{1,1}^l, x_{1,2}^l) \\ L_2^l(t, x_{1,1}^l, x_{1,2}^l) \end{bmatrix}, \\ x_2^l &= R^l(t, x_{1,1}^l, x_{1,2}^l), \end{aligned} \tag{5.44}$$

and

$$\begin{aligned} \dot{x}_1^k &= L^k(t, x_1^k), \\ \begin{bmatrix} x_{2,1}^k \\ x_{2,2}^k \end{bmatrix} &= \begin{bmatrix} R_1^k(t, x_1^k) \\ R_2^k(t, x_1^k) \end{bmatrix}. \end{aligned} \tag{5.45}$$

Then, the differential-algebraic system during sliding motion can be defined as

$$\begin{aligned} \dot{x}_1 &= \alpha \begin{bmatrix} L_1^l(t, x_1) \\ L_2^l(t, x_1) \end{bmatrix} + (1-\alpha) \begin{bmatrix} L^k(t, x_1) \\ 0 \end{bmatrix}, \\ x_2 &= \alpha \begin{bmatrix} R^l(t, x_1) \\ 0 \end{bmatrix} + (1-\alpha) \begin{bmatrix} R_1^k(t, x_1) \\ R_2^k(t, x_1) \end{bmatrix}, \\ 0 &= g_j^l(t, x_1, x_2), \end{aligned} \tag{5.46}$$

similar as in (5.38). In the same way as in Theorem 5.29, the DAE in sliding motion (5.46) is regular and of strangeness index $\mu = 1$.

Example 5.30. [34] Consider a multibody system with dry friction as given in Example 5.2 with equation of motion of the form

$$\begin{aligned} M\ddot{p} &= f_a(p, \dot{p}) - G(p)^T \lambda - \mu_F \|F_N\| c(p) \operatorname{sign}(c(p)^T \dot{p}), \\ 0 &= g(p), \end{aligned}$$

where M is positive definite and the switching function is given by $q(p,\dot{p}) = c(p)^T\dot{p}$ describing the relative tangential velocity between the bodies. Here, c is a unit vector parallel to the friction surface modeled by $g(p) = 0$ at the contact point. The sliding condition is fulfilled if

$$\|c^T M^{-1}(f_a - G^T\lambda) + \dot{c}^T\dot{p}\| < \mu_F\|F_N\|c^T M^{-1}c,$$

i.e., if the force in the direction of c is smaller than the maximal friction force. The Filippov construction results in

$$\begin{aligned} M\ddot{p} &= f_a(p,\dot{p}) - G(p)^T\lambda + (1-2\alpha)\mu_F\|F_N\|c(p), \\ 0 &= g(p), \\ 0 &= c(p)^T\dot{p}, \end{aligned} \tag{5.47}$$

with additional algebraic variable α. When α is chosen such that $\dot{q} = c^T\ddot{p} + \dot{c}^T\dot{p} = 0$, the DAE (5.47) can be transformed to

$$\begin{aligned} M\ddot{p} &= f_a(p,\dot{p}) - G(p)^T\lambda - \frac{c^T M^{-1}(f_a - G^T\lambda) + \dot{c}^T\dot{p}}{c^T M^{-1}c}c, \\ 0 &= g(p). \end{aligned} \tag{5.48}$$

Alternatively, the equations of motion for stiction can be obtained by adding the algebraic equation $q = c^T\dot{p} = 0$ and the Lagrange parameter λ_S to the system

$$\begin{aligned} M\ddot{p} &= f_a(p,\dot{p}) - G(p)^T\lambda + c(p)\lambda_S, \\ 0 &= g(p), \\ 0 &= c(p)^T\dot{p}. \end{aligned}$$

Differentiation of the equation $c(p)^T\dot{p} = 0$ gives

$$\lambda_S = -\frac{c^T M^{-1}(f_a - G^T\lambda) + \dot{c}^T\dot{p}}{c^T M^{-1}c},$$

and elimination of λ_S yields the above equation (5.48), i.e., $\lambda_S = (1-2\alpha)\mu_F\|F_N\|$.

5.5.4 Hysteresis Switching

Another possibility to prevent a hybrid system from chattering is to built in hysteresis that prevents the system from changing modes too quickly and thereby precluding the possibility of unbounded chattering. The introduction of hysteresis can be properly applied if numerical chattering between two modes l and k occurs that have transition conditions that only differ in sign, i.e., there are some switching functions such that

$$g_j^l(t,x^l,\dot{x}^l) = -g_i^k(t,x^k,\dot{x}^k), \text{ with } j \in J^l,\ i \in J^k.$$

In this case, a hysteresis can be realized by adding a term $\epsilon > 0$ to the transition conditions, i.e., by defining the switching functions

$$\hat{g}_i^k = g_i^k + \epsilon \quad \text{and} \quad \hat{g}_j^l = g_j^l + \epsilon.$$

For independent transition conditions between two modes l and k or if numerical chattering between more than two modes occur, the integration of a hysteresis is not so easy to realize. In addition, if the determination of the exact switch point is essential for the system behavior, then an artificial hysteresis cannot be inserted into the system. In this cases only a different but possibly complex modeling can be used to obtain a system without numerical chattering.
A further possibility to suppress chattering is so-called *dwell-time switching*, see e.g. [103]. Here, the basic idea is to have some fixed time $\tau > 0$, called the *dwell-time*, such that, once a mode l is chosen the system will remain in this mode for at least a time τ before another mode transition can occur.
In general, during the numerical integration of hybrid systems all mode transitions have to be observed closely to detect numerical chattering. If numerical chattering occurs during the numerical simulation appropriate measures should be taken.

5.6 Control of Switched Systems

In this section we consider hybrid control problems, i.e., hybrid systems $\mathcal{H}$ as in Definition 5.3 consisting of DAEs (5.3) with specified undetermined parts u^l describing the *controls*. In control problems the system inputs u^l are used to steer the solution of the system so that a given property is satisfied. In general, classical control concepts for DAEs can be applied to hybrid systems locally in every mode in the same way as the index reduction described in Section 5.2, but some attention has to be paid to the transition of the system state between modes. Choosing a control u^l in some mode l influences the transition conditions and mode changes of the hybrid system as well as the points in time at with switching occurs. Thus, changes in the controls lead to a huge number of possible hybrid mode trajectories and hybrid time trajectories. In addition, transitions between modes often cause nonsmoothness of the solution which complicates the minimization problem used in the optimal control theory.
In the following, we will restrict ourselves to hybrid systems $\mathcal{H}$ with linear time-invariant DAEs of the form

$$E^l \dot{x}^l = A^l x^l + B^l u^l, \tag{5.49a}$$
$$y^l = C^l x^l, \tag{5.49b}$$

in each mode $l \in \mathbb{M}$, where $E^l, A^l \in \mathbb{R}^{m,n}$, $B^l \in \mathbb{R}^{m,k}$ and $C^l \in \mathbb{R}^{p,n}$ are constant matrices. In the control context, systems of the form (5.49) are also known as *descriptor systems*. Here, $x^l : D_l \to \mathbb{R}^n$ represents the state of the system in mode l, $u^l : D_l \to \mathbb{R}^k$ is the input or control and $y^l : D_l \to \mathbb{R}^p$ is the output of the system in mode $l \in \mathbb{M}$. As the

output equation (5.49b) does not contribute to the analysis of the system behavior, it is often omitted in theoretical considerations. For a particular input u^l the system (5.49) represents a differential-algebraic equation, such that the solvability theory for control problems is related to that of DAEs. Using a *behavior approach*, see e.g. [67], by setting

$$z^l = \begin{bmatrix} x^l \\ u^l \\ y^l \end{bmatrix}$$

the system (5.49) corresponds to a linear DAE and the general theory of hybrid systems can be applied, especially the reduction to strangeness-free form described in Section 5.2. In the following, we will describe the main ideas for controlling hybrid systems $\mathcal{H}$ with descriptor systems of the form (5.49) in each mode. At first, we consider open loop control problems in Section 5.6.1, then feedback control in Section 5.6.2, and hybrid optimal control problems in Section 5.6.3. Finally, in Section 5.6.4 we present an approach allowing sliding mode control for linear hybrid descriptor systems.

5.6.1 Open Loop Control

In open loop control the question whether a system can be steered from an initial state x_0 at time t_0 to another state x_f at time t_f is examined. Thus, we have to analyze if a hybrid system $\mathcal{H}$ can be transfered from every possible state to every other state by choosing suitable input functions $u^l(t)$ in every mode $l \in \mathbb{M}$. We start with recalling some important definitions in control theory. The first term concerns the solvability of the descriptor system in mode $l \in \mathbb{M}$ for every input function and every initial value that is consistent with this input.

Definition 5.31 (Consistency and regularity of control problems). The control problem (5.49) in mode $l \in \mathbb{M}$ is called *consistent* if there exists an input function u^l such that the DAE (5.49) is solvable. It is called *regular* if for every sufficiently smooth input function u^l the DAE (5.49) is solvable and the solution in mode l is unique for consistent initial values.

Then the following Corollary can be formulated characterizing the solvability of the descriptor system (5.49) in mode l.

Corollary 5.32. *If the pair (E^l, A^l) of square matrices is regular (see Definition 2.28), then the control problem* (5.49) *in mode $l \in \mathbb{M}$ is consistent and regular. If (E^l, A^l) with $E^l, A^l \in \mathbb{R}^{m,n}$ is a singular matrix pair, then the control problem* (5.49) *in mode $l \in \mathbb{M}$ is not regular.*

Proof. See [82, Corollarys 2.54 and 2.55]. □

Further, we define the terms controllability and observability for the descriptor systems (5.49) in each mode.

Definition 5.33 (Controllability). The descriptor system (5.49) in mode $l \in \mathbb{M}$ is called *completely controllable* if for any given initial state $x^l(t_0) = x^0 \in \mathbb{R}^n$ at $t_0 \in \mathbb{I}_i \subseteq D_l$ and any final state $x^1 \in \mathbb{R}^n$ there exists a control input u^l such that the solution of (5.49) with this control input fulfills $x^l(t_1) = x^1$ after finite time t_1 with $t_0 < t_1 < \infty$, and $t_1 \in \mathbb{I}_i$.

Definition 5.34 (Observability). The descriptor system (5.49) in mode $l \in \mathbb{M}$ is called *completely observable* if the zero output of the descriptor system with $u^l = 0$ implies that this system has the trivial solution $x^l = 0$ only.

Note that Definition 5.33 implies that, if the descriptor system in mode l is completely controllable from (t_0, x_0) to (t_1, x_1), then the hybrid system $\mathcal{H}$ stays in mode l at least until x_1 is reached. In general, descriptor systems of the form (5.49) are not completely controllable or completely observable, since the algebraic constraints fix the solution and the output onto the constraint manifold. For this reason we also consider the following definitions.

Definition 5.35 (Reachability, R-controllability). For the descriptor system (5.49) in mode $l \in \mathbb{M}$, a set $\mathcal{R}^l \subseteq \mathbb{R}^n$ is called *reachable from* x_0^l if there exists a control input u^l that transfers the system from x_0^l to some $x_1^l \in \mathcal{R}^l$ in finite time, while staying in mode l. System (5.49) is called *controllable within the reachable set* $\mathcal{R}^l$ *(R-controllable)* if any state in $\mathcal{R}^l$ can be reached from any consistent initial state x_0^l.

Definition 5.36 (R-Observability). The descriptor system (5.49) in mode $l \in \mathbb{M}$ is called *observable within the reachable set (R-observable)* if the zero output of the descriptor system with $u^l = 0$ implies that all solutions of the system satisfy $P_r^l x^l = 0$, where P_r^l is the projection onto the right deflating subspace corresponding to the finite eigenvalues of (E^l, A^l) (see the definition in (2.9)). A hybrid system $\mathcal{H}$ is called *observable within the reachable set (R-observable)* if it is R-observable within each mode $l \in \mathbb{M}$.

In the following, we assume R-controllability and R-observability of the hybrid system $\mathcal{H}$, as well as unique solvability locally in each mode $l \in \mathbb{M}$.

Assumption 5.37. *For a hybrid system $\mathcal{H}$ as in Definition 5.3 with linear descriptor systems* (5.49) *in each mode $l \in \mathbb{M}$, let (E^l, A^l) be square and regular with $m = n$ for all $l \in \mathbb{M}$. Further, assume that* (5.49) *is R-controllable as well as R-observable for each mode.*

For linear time-invariant descriptor systems (5.49), these controllability and observability concepts can be characterized algebraically in terms of the matrices E^l, A^l, B^l and C^l.

Theorem 5.38. *Consider the quadruple (E^l, A^l, B^l, C^l) as in* (5.49).

1. *The descriptor system* (5.49) *in mode $l \in \mathbb{M}$ is completely controllable if and only if*

$$\operatorname{rank}[\alpha E^l - \beta A^l, B^l] = n$$

for all $(\alpha, \beta) \in \mathbb{C}^2 \backslash \{(0,0)\}$.

2. *The descriptor system* (5.49) *in mode* $l \in \mathbb{M}$ *is completely observable if and only if*

$$\operatorname{rank} \begin{bmatrix} \alpha E^l - \beta A^l \\ C^l \end{bmatrix} = n$$

for all $(\alpha, \beta) \in \mathbb{C}^2 \backslash \{(0,0)\}$.

3. *The descriptor system* (5.49) *in mode* $l \in \mathbb{M}$ *is R-controllable if and only if*

$$\operatorname{rank}[\lambda E^l - A^l, B^l] = n$$

for all $\lambda \in \mathbb{C}$.

4. *The descriptor system* (5.49) *in mode* $l \in \mathbb{M}$ *is R-observable if and only if*

$$\operatorname{rank} \begin{bmatrix} \lambda E^l - A^l \\ C^l \end{bmatrix} = n$$

for all $\lambda \in \mathbb{C}$.

Proof. See [19, 30]. □

Before we define the terms reachability and R-controllability for a hybrid system $\mathcal{H}$ we introduce the term *execution of a hybrid system* for convenience.

Definition 5.39 (Execution). An *execution* of a hybrid system $\mathcal{H}$ as in Definition 5.3 is given by $(T_\tau, \{(x^{l_i}(t), l_i)\})$ with hybrid time trajectory $T_\tau = \{[\tau_i, \tau_i')\}$ and hybrid solution trajectory $\{(x^{l_i}(t), l_i)\}$, where for each interval $[\tau_i, \tau_i')$ we have that $x^{l_i}(t)$ is a solution of the DAE in mode l_i for all $t \in [\tau_i, \tau_i')$, and $L_j^{l_i}(\tau_i', x^{l_i}(\tau_i'), \dot{x}^{l_i}(\tau_i')) = TRUE$ for some $j \in J^{l_i}$. Further, we have $l_{i+1} = S^{l_i}(j)$ and $T_{l_i}^{l_{i+1}}(x^{l_i}(\tau_i'), \dot{x}^{l_i}(\tau_i')) = [x^{l_{i+1}}(\tau_{i+1}), \dot{x}^{l_{i+1}}(\tau_{i+1})]$.

Definition 5.40 (Reachable state). A hybrid state $(\hat{x}, \hat{l}) \in \mathbb{R}^n \times \mathbb{M}$ is called *reachable* if there exists a finite execution $(T_\tau, \{(x^{l_i}(t), l_i)\}_{i=1}^N)$ with $T_\tau = \{[\tau_i, \tau_i')\}_{i=1}^N$ and $(x^{l_N}(\tau_N), l_N) = (\hat{x}, \hat{l})$. The set of all reachable states of a hybrid system $\mathcal{H}$ is denoted by $\mathcal{R}_\mathcal{H}$.

Definition 5.41 (Reachability and R-controllability of a hybrid system). For a hybrid system $\mathcal{H}$ as in Definition 5.3 with descriptor systems (5.49) in each mode a set $\mathcal{R} \subseteq \mathbb{R}^n \times \mathbb{M}$ is called *reachable from the initial state* $(x_0^{l_1}, l_1)$ if there exists a sequence of control inputs $u^{l_1}, u^{l_2}, \ldots, u^{l_N}$ and a corresponding mode trajectory $T_m = \{l_i\}_{i=1}^N$, such that the system state is transfered from $(x_0^{l_1}, l_1)$ to some $(x_1^{l_k}, l_k) \in \mathcal{R}$ in finite time. The hybrid system $\mathcal{H}$ is called *controllable within the reachable set (R-controllable)* if any hybrid state in $\mathcal{R}_\mathcal{H}$ can be reached from any consistent initial state $(x_0^{l_1}, l_1)$.

Under Assumption 5.37, we can give necessary conditions such that the hybrid system $\mathcal{H}$ is R-controllable, i.e., for every consistent initial hybrid state we can find a sequence of controls that steers the system to any state in $\mathcal{R}_\mathcal{H}$.

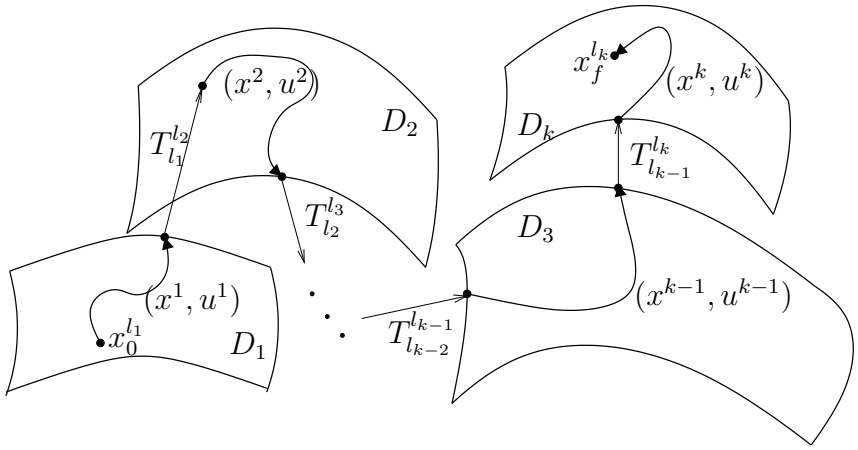

Figure 5.12: Controllability of a hybrid system

Theorem 5.42. *Consider a hybrid system $\mathcal{H}$ as in Definition 5.3 with descriptor systems* (5.49) *in each mode $l \in \mathbb{M}$ and assume that Assumption 5.37 holds. Further, let $(x_0^{l_1}, l_1) \in \mathbb{R}^n \times \mathbb{M}$ be a consistent initial hybrid state, i.e., $x_0^{l_1}$ is consistent for the DAE in mode $l_1 \in \mathbb{M}$. Then, the hybrid system $\mathcal{H}$ is R-controllable and R-observable if and only if*

1. *for any mode $l_k \in \mathbb{M}$ there exits a hybrid mode trajectory $T_m = \{l_i\}_{i=1}^k$ with finite number of transitions $k < \infty$, and*

2. *$x^{l_i}(t)$ given by $T_{l_{i-1}}^{l_i}(x^{l_{i-1}}(t), \dot{x}^{l_{i-1}}(t)) = [x^{l_i}(t), \dot{x}^{l_i}(t)]$ is consistent for every $t \in \mathbb{I}$ with the DAE in mode $l_i \in T_m$.*

Proof. Under Assumption 5.37 we have unique solvability for consistent initial values and given controls u^l in each mode $l \in \mathbb{M}$, due to Corollary 5.32. Assume that $\mathcal{H}$ is R-controllable and R-observable. Then, for any consistent initial state $(x_0^{l_1}, l_1)$ any hybrid state $(x_f^{l_k}, l_k) \in \mathcal{R}_\mathcal{H}$ can be reached. This means that $x_f^{l_k} \in \mathcal{R}^{l_k}$ and there exists a sequence of controls $\{u^{l_i}\}_{i=1}^k$ and a corresponding mode trajectory $T_m = \{l_i\}_{i=1}^k$ such that the state $x_0^{l_1}$ is transfered to $x_f^{l_k}$. Further, the initial conditions after mode changes have to be consistent in the new mode to ensure the existence of a solution. On the other hand, we consider an arbitrary $x_f^{l_k} \in \mathcal{R}^{l_k}$ in the reachable set $\mathcal{R}^{l_k}$ of some mode $l_k \in \mathbb{M}$. Then, due to the assumptions, there exists a mode trajectory $T_m = \{l_i\}_{i=0}^k$. Further, due to Assumption 5.37, in each mode $l_i \in T_m$ we can find a corresponding control u^{l_i} that steers the solution to some $x^\star$ for which the next transition condition is satisfied, and thus the next switch point τ_i is defined. As the initial conditions after each mode transitions are consistent in the new mode for every $t \in \mathbb{I}$, the mode transition can be performed and the solution evolves in the new mode. After a finite number of transitions the system reaches mode l_k and a control u^{l_k} can be chosen that transfers the state to $x_f^{l_k}$. See also Figure 5.12. □

5.6.2 Feedback Control

In the control context a common approach is to modify the system properties using so-called *feedbacks*, i.e., the input is chosen on the basis of observations from the state or the output that can be measured. In a hybrid system $\mathcal{H}$ possible feedbacks for the input u^l in mode $l \in \mathbb{M}$ are given by

$$u^l = F^l x^l + w^l, \tag{5.50}$$

$$u^l = F^l y^l + w^l, \tag{5.51}$$

i.e., *proportional state feedback* (5.50) with $F^l \in \mathbb{R}^{k,n}$, or *proportional output feedback* (5.51) with $F^l \in \mathbb{R}^{k,p}$, respectively. If we apply these feedbacks as controls to the system in mode $l \in \mathbb{M}$, we obtain the so-called *closed-loop systems*

$$E^l \dot{x}^l = (A^l + B^l F^l) x^l + B^l w^l, \tag{5.52}$$

or

$$E^l \dot{x}^l = (A^l + B^l F^l C^l) x^l + B^l w^l. \tag{5.53}$$

Under certain conditions we can find a feedback in mode l such that the closed-loop systems (5.52) or (5.53) in mode l are regular and of nilpotence index ν^l at most one.

Theorem 5.43. *Given a matrix quadruple (E^l, A^l, B^l, C^l) as in* (5.49).

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

$$\operatorname{rank} [E^l,\ A^l T^l,\ B^l] = n, \tag{5.54}$$

   where T^l is a matrix whose columns span $\operatorname{kernel} E^l$.

2. *There exists a matrix $F^l \in \mathbb{R}^{k,p}$ such that the matrix pair $(E^l, A^l + B^l F^l C^l)$ is regular and of index $\nu^l = \operatorname{ind}(E^l, A^l + B^l F^l C^l)$ at most one if and only if E^l and A^l are square and* (5.54) *as well as*

$$\operatorname{rank} \begin{bmatrix} E^l \\ (Z^l)^T A^l \\ C^l \end{bmatrix} = n, \tag{5.55}$$

   hold, where Z^l is a matrix whose columns span $\operatorname{kernel} (E^l)^T$.

Proof. See, [82, Theorem 2.56]. □

In a similar way as before, if feedbacks in each mode $l \in \mathbb{M}$ can be chosen such that the closed-loop systems in each mode l are regular and of index at most one, then the complete hybrid system $\mathcal{H}$ is regular and of maximal index at most one.

Theorem 5.44. *Consider a hybrid system $\mathcal{H}$ as in Definition 5.3 with descriptor systems* (5.49) *in each mode $l \in \mathbb{M}$. Suppose that E^l, A^l are square and the rank conditions* (5.54)*, and* (5.55) *hold for every $l \in \mathbb{M}$. Then, there exist feedback controls* (5.50) *or* (5.51) *in each mode such that the hybrid system $\mathcal{H}$ is regular and of maximal index at most one.*

Proof. Due to Theorem 5.43, there exist matrices $F_1^l \in \mathbb{R}^{k,n}$ and $F_2^l \in \mathbb{R}^{k,p}$ in each mode $l \in \mathbb{M}$ such that the matrix pairs $(E^l, A^l + B^l F_1^l)$ and $(E^l, A^l + B^l F_2^l C^l)$ are regular and of index $\nu^l \leq 1$. Thus, the hybrid system $\mathcal{H}$ is regular and $\max_{l \in \mathbb{M}} \nu^l \leq 1$. □

Note, that Theorem 5.44 does not guarantee the existence of a solution of the overall hybrid control problem.

Example 5.45. Consider a hybrid differential-algebraic control problem $\mathcal{H}$ consisting of the following two closed-loop control systems

$$\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} F_1 & F_2 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{in mode 1,}$$

and

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} \begin{bmatrix} F_1 & F_2 \end{bmatrix} \right) \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{in mode 2.}$$

Condition (5.54) holds for both systems such that we can find a feedback for which the closed-loop systems in both modes are regular and of index at most 1. Choosing $\begin{bmatrix} F_1 & F_2 \end{bmatrix} = \begin{bmatrix} 0 & 1 \end{bmatrix}$, for the system in mode 1 we get the matrix pair

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

which is clearly regular and of index $\nu^1 = 0$ and the unique solution is given by

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

for some initial values $x_{1,0}, x_{2,0}$. For the system in mode 2 we get the matrix pair

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

which is also regular and of index $\nu^2 = 1$ and the unique solution is given by

$$\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Thus, if the hybrid system switches from mode 1 to mode 2, then a solution of the hybrid control system only exists if the initial values after the switching are consistent with the DAE in mode 2.

As emphasized in Example 5.45, in order to guarantee that a feedback control exists such that the hybrid system $\mathcal{H}$ is regular and strangeness-free and possesses a solution, additionally consistency conditions for the transition functions need to hold in the same way as in Theorem 5.42.

5.6.3 Hybrid Optimal Control

Classical control applications such as stabilization of a system or path following can often be formulated in terms of optimal control problems. In hybrid optimal control problems the aim is to find an optimal hybrid solution trajectory such that a cost functional is minimized subject to the systems dynamics, as well as further constraints defining the transition conditions. For hybrid systems the optimal control problem has been introduced in [12, 94]. The linear-quadratic optimal control problem for hybrid systems consisting of ordinary differential equations has been studied in [126].
In contrast to the standard optimal control problem, the parameters that can be optimized in hybrid control systems are the control inputs u^l, the number of switchings and the switching times, as well as the mode sequence, i.e., the order in which the transition conditions are satisfied. Here, we will assume that the system switches between modes a fixed number of times. In [154] a two stage optimization method for switched systems has been proposed which first computes optimal control inputs u^l and a hybrid time trajectory for a given number of switchings and a given sequence of active modes, and in a second stage updates the number of switchings and the mode sequence to optimize the solution obtained in the first stage.

In the following, we assume that the initial time t_0 and a consistent initial state x_0 are given and that the final time t_f is fixed. The final state $x(t_f)$ is assumed to be free. Further, we restrict ourselves to autonomous switching at times τ_i, $i = 2, \ldots, N_{\mathbb{I}}$ induced by sign changes in the switching functions given by $g_j^l(t, x)$ for all $l \in \mathbb{M}$, $j \in J^l = \{1, \ldots, n_T^l\}$. Further, we set $\tau_1 = t_0$ and $\tau_{N_{\mathbb{I}}+1} = t_f$. Then, the linear-quadratic optimal control problem for a hybrid system $\mathcal{H}$ as defined in Definition 5.3, with linear descriptor systems of the form (5.49a) in each mode $l \in \mathbb{M}$ (omitting the output equations) and with a finite number $N_{\mathbb{I}} < \infty$ of subintervals, can be defined as follows. For given t_0, t_f, with $t_0 < t_f < \infty$, initial condition

$$x(t_0) = x^0, \tag{5.56a}$$

and initial mode $l_1 \in \mathbb{M}$, the hybrid optimal control problem is to determine a sequence of control input $u(t) = \{u^{l_i}(t)\}_{i=1}^{N_{\mathbb{I}}}$, a corresponding sequence of switch points $\tau = \{\tau_i\}_{i=2}^{N_{\mathbb{I}}}$, and a corresponding mode sequence $\{l_i\}_{i=1}^{N_{\mathbb{I}}}$, where $l_i \in \mathbb{M}$ is the active mode in the interval $[\tau_i, \tau_{i+1})$, which drives the state $x(t) = \{x^{l_i}(t)\}_{i=1}^{N_{\mathbb{I}}}$, starting from the initial value x^0 at t_0 in initial mode l_1, while minimizing the following quadratic cost functional

$$\mathcal{S}(x, u, \tau) = \frac{1}{2}\left\{ x(t_f)^T M x(t_f) + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i}^{\tau_{i+1}} \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} Q_{l_i} & S_{l_i} \\ S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} dt \right\}, \tag{5.56b}$$

where $M \in \mathbb{R}^{n,n}$ and $Q_{l_i} \in \mathbb{R}^{n,n}$ are symmetric positive semi-definite matrices, $R_{l_i} \in \mathbb{R}^{k,k}$ are symmetric positive definite matrices, and $S_{l_i} \in \mathbb{R}^{n,k}$, for all $l_i \in \mathbb{M}$, subject to the system dynamics

$$E^{l_i}\dot{x} = A^{l_i}x + B^{l_i}u, \quad \text{for } t \in [\tau_i, \tau_{i+1}), \quad i = 1, \ldots, N_{\mathbb{I}}, \tag{5.56c}$$

with $E^{l_i}, A^{l_i} \in \mathbb{R}^{n,n}$, $B^{l_i} \in \mathbb{R}^{n,k}$ in each mode l_i, and subject to the constraints

$$g_j^{l_i}(\tau_{i+1}, x(\tau_{i+1})) = 0, \quad \text{for } i = 1, \ldots, N_{\mathbb{I}} - 1, \quad j \in J^{l_i}. \tag{5.56d}$$

In the following, we set

$$g^{l_i}(t, x(t)) = \begin{bmatrix} g_1^{l_i}(t, x(t)) \\ \vdots \\ g_{n_T^l}^{l_i}(t, x(t)) \end{bmatrix}.$$

Note that u and x in (5.56b), (5.56c), (5.56d) always denote the control u^{l_i} and the state x^{l_i} in the corresponding mode l_i for $t \in [\tau_i, \tau_{i+1})$. We omit the subscript l_i for ease of representation. Further, we require that the solution $x(t)$ of the hybrid system is continuous, i.e., at the switch points we have

$$x(\tau_{i+1}) = x(\tau_{i+1}^-), \quad \text{for } i = 1, \ldots, N_{\mathbb{I}} - 1, \tag{5.56e}$$

where $x(\tau_{i+1}^-) = \lim_{t \to \tau_{i+1}^-} x(t)$ is the value of the solution in the previous mode expanded to the switch point. The sequence of switch points and the sequence of modes depend on the state $x(t)$ as the switch points are the roots of the switching functions, and the next mode l_{i+1} is determined by the mode allocation function via

$$l_{i+1} = S^{l_i}(j). \tag{5.56f}$$

In the following, we assume without loss of generality that the DAE (5.56c) in each mode l_i is regular and strangeness-free as a free system without control, i.e., when $u = 0$. This is not a restriction, since we can always use index reduction for the behavior formulation and feedback regularization to obtain a reduced system with these properties in each mode, see e.g. [83]. Further, we assume that the initial conditions $x(\tau_i^-)$ are consistent for all successor modes l_i in $[\tau_i, \tau_{i+1})$ and that $Mx(t_f)$ lies in the cokernel of $E^{l_{N_{\mathbb{I}}}}$. Again, the requirement of consistent initial conditions is not a restriction, since they can be obtained from the reduced system. Then, by using calculus of variations we can derive necessary conditions for optimality of the hybrid system that lead to a sequence of two-point boundary value problems with additional transversality conditions at the switching times.

Theorem 5.46. *Consider the optimal control problem* (5.56). *Let* $u_\star$, $x_\star$, *and* $\tau_\star = \{\tau_i^\star\}_{i=2}^{N_{\mathbb{I}}}$ *be the optimal solution of the optimal control problem, with* $u_\star \in U_k(\tau_\star) := \{u(t) \in \mathbb{R}^k \mid u(t)$ *piecewise continuous on* $[\tau_i^\star, \tau_{i+1}^\star)$ *for all* $\tau_i^\star \in \tau_\star\}$ *and* $x_\star \in C^0([t_0, t_f], \mathbb{R}^n)$ *the corresponding solution of*

$$E^{l_i}\dot{x} = A^{l_i}x + B^{l_i}u_\star, \quad \textit{for } t \in [\tau_i^\star, \tau_{i+1}^\star).$$

Suppose that (5.56c) *is regular and strangeness-free as a behavior system and that* $Mx(t_f) \in$ cokernel $(E^{l_{N_{\mathbb{I}}}})$. *Further, let* $g^{l_i}(t, x(t))$ *be differentiable and assume that* $\frac{d}{dt}g^{l_i}(t, x(t))$ *is*

nonsingular for all $t \in [\tau_i^\star, \tau_{i+1}^\star]$, $l_i \in \mathbb{M}$. *Then, there exist a piecewise continuous co-state* $\lambda(t) \in \mathbb{R}^n$ *and Lagrange multipliers* $\eta^{l_i} \in \mathbb{R}^{n_T^{l_i}}$, *with* $\eta^{l_i} \neq 0$, $l_i \in \mathbb{M}$, *such that* $x_\star(t), \lambda(t), u_\star(t), \tau_\star$ *and* η^{l_i} *solve the linear boundary value problem*

$$\begin{bmatrix} 0 & E^{l_i} & 0 \\ -(E^{l_i})^T & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \dot\lambda \\ \dot x \\ \dot u \end{bmatrix} = \begin{bmatrix} 0 & A^{l_i} & B^{l_i} \\ (A^{l_i})^T & Q_{l_i} & S_{l_i} \\ (B^{l_i})^T & S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} \lambda \\ x \\ u \end{bmatrix}, \quad \text{for } t \in [\tau_i, \tau_{i+1}), \tag{5.57a}$$

with boundary conditions

$$x(t_0) = x^0, \quad (E^{l_{N_\mathbb{I}}})^T \lambda(t_f) = M x(t_f), \tag{5.57b}$$

and transversality conditions at the switch points of the form

$$(E^{l_i})^T \lambda(\tau_{i+1}^-) = \frac{1}{2} g_{;x}^{l_i}(\tau_{i+1}, x(\tau_{i+1}))^T \eta^{l_i}, \tag{5.57c}$$

$$(E^{l_{i+1}})^T \lambda(\tau_{i+1}^+) = (E^{l_i})^T \lambda(\tau_{i+1}^-) - \frac{1}{2} g_{;x}^{l_i}(\tau_{i+1}, x(\tau_{i+1}))^T \eta^{l_i}, \tag{5.57d}$$

$$H^{l_{i+1}}(x,u,\lambda)\big|_{\tau_{i+1}^+} = H^{l_i}(x,u,\lambda)\big|_{\tau_{i+1}^-} + g_{;t}^{l_i}(\tau_{i+1}, x(\tau_{i+1}))^T \eta^{l_i}, \tag{5.57e}$$

for $i = 1, \ldots, N_\mathbb{I} - 1$, *where the Hamiltonians* $H^l(x,u,\lambda)$ *in each mode* $l \in \mathbb{M}$ *are defined by*

$$H^l(x,u,\lambda) = x^T Q_l x + x^T S_l u + u^T S_l^T x + u^T R_l u + \lambda^T (A^l x + B^l u) + (A^l x + B^l u)^T \lambda.$$

Note that $\tau_1^\star = \tau_1 = t_0$ and $\tau_{N_\mathbb{I}+1}^\star = \tau_{N_\mathbb{I}+1} = t_f$ are fixed and we have discontinuities in λ and in the Hamiltonians at the switching times τ_{i+1}.

Proof. We use a variation of the Pontryagin maximum principle, see e.g. [69]. Let $u_\star$ be an optimal control. We consider the first order variation

$$u(t) = u_\star(t) + \varepsilon v(t), \tag{5.58}$$

with $v(t)$ chosen such that $u(t) \in U_k(\tau)$. Then, for each mode l_i we have

$$E^{l_i} \dot x = A^{l_i} x + B^{l_i}(u_\star(t) + \varepsilon v(t)), \quad t \in [\tau_i, \tau_{i+1}),$$

and the local solution is given by

$$\begin{aligned} x(t) =& e^{\hat E^D \hat A (t-\tau_i)} \hat E^D \hat E x(\tau_i^-) + \int_{\tau_i}^t e^{\hat E^D \hat A (t-s)} \hat E^D \hat B (u_\star(s) + \varepsilon v(s)) ds \\ & - (I - \hat E^D \hat E) \sum_{i=0}^{\nu^{l_i}-1} (\hat E \hat A^D)^i \hat A^D \hat B \frac{d^i}{dt^i}(u_\star(t) + \varepsilon v(t)), \end{aligned}$$

where $x(\tau_i^-) = \lim_{t\to\tau_i^-} x(t)$ is the value of the solution in the previous mode expanded to the switch point and $\hat{E} = (\hat{\lambda}E^{l_i} - A^{l_i})^{-1}E^{l_i}$, $\hat{A} = (\hat{\lambda}E^{l_i} - A^{l_i})^{-1}A^{l_i}$, $\hat{B} = (\hat{\lambda}E^{l_i} - A^{l_i})^{-1}B^{l_i}$ for some $\hat{\lambda} \in \mathbb{R}$, and $\nu^{l_i} = \operatorname{ind}(E^{l_i}, A^{l_i})$, see also Theorem 2.31. This gives

$$\begin{aligned} x(t) =& x_\star(t) + \varepsilon \left[\int_{\tau_i}^{t} e^{\hat{E}^D \hat{A}(t-s)} \hat{E}^D \hat{B} v(s) ds - (I - \hat{E}^D \hat{E}) \sum_{i=0}^{\nu^{l_i}-1} (\hat{E}\hat{A}^D)^i \hat{A}^D \hat{B} v^{(i)}(t) \right], \\ =& x_\star(t) + \varepsilon\varphi(t), \end{aligned}$$

i.e., the corresponding variation of x, where $\varphi(t)$ solves the DAE

$$E^{l_i}\dot{\varphi} = A^{l_i}\varphi + B^{l_i}v(t), \quad \varphi(\tau_i) = 0, \quad \text{for } t \in [\tau_i, \tau_{i+1}).$$

Further, we consider the corresponding variation of the switch points

$$\tau_i = \tau_i^\star + \delta\tau_i, \quad i = 1, \ldots, N_{\mathbb{I}} + 1,$$

that depends on the variation of x in such a way, that if $\tau_{i+1}^\star$ is a root of $g_j^{l_i}(t, x_\star(t))$, then τ_{i+1} is a root of $g_j^{l_i}(t, x(t))$. Thus, from

$$\begin{aligned} 0 = g_j^{l_i}(\tau_{i+1}, x(\tau_{i+1})) =& g_j^{l_i}(\tau_{i+1}^\star + \delta\tau_{i+1}, x_\star(\tau_{i+1}^\star + \delta\tau_{i+1}) + \varepsilon\varphi(\tau_{i+1}^\star + \delta\tau_{i+1})) \\ =& \delta\tau_{i+1}[\dot{x}_\star(\tau_{i+1}^\star) g_{j;x}^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star)) + g_{j;t}^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star))] \\ &+ \varepsilon\varphi(\tau_{i+1}^\star) g_{j;x}^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star)) + h.o.t \end{aligned}$$

we get

$$\delta\tau_{i+1} = -\varepsilon\varphi(t) \left[\frac{d}{dt} g_j^{l_i}(t, x_\star(t)) \right]^{-1} \frac{\partial}{\partial x} g_j^{l_i}(t, x_\star(t)) \Big|_{t=\tau_{i+1}^\star}. \tag{5.59}$$

We assume that all $\delta\tau_i$ are sufficiently small, i.e., $\tau_{i-1}^\star + \delta\tau_{i-1} < \tau_i^\star + \delta\tau_i < \tau_{i+1}^\star + \delta\tau_{i+1}$, and the successor mode is still given by $l_{i+1} = S^{l_i}(j)$, i.e., no other root in $g_k^{l_i}$, $k \in J^{l_i}$, $k \neq j$ occurs before. Next, we introduce Lagrange multipliers $\eta^{l_i} \neq 0 \in \mathbb{R}^{n_T^{l_i}}$ for all $l_i \in \mathbb{M}$ and form an augmented cost functional

$$\begin{aligned} \mathcal{S}_a(x, u, \tau) = \frac{1}{2} \Bigg\{ & x(t_f)^T M x(t_f) + \sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T g^{l_i}(\tau_{i+1}, x(\tau_{i+1})) \\ & + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i}^{\tau_{i+1}} \begin{bmatrix} x \\ u \end{bmatrix}^T \begin{bmatrix} Q_{l_i} & S_{l_i} \\ S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} dt \Bigg\}. \end{aligned}$$

Introducing another Lagrange multiplier function $\lambda(t)$ and using the Hamiltonians H^{l_i} locally in each mode l_i, the cost functional $\mathcal{S}_a(x,u,\tau)$ can be written as

$$\mathcal{S}_a(x,u,\tau) = \frac{1}{2}\Bigg\{ x(t_f)^T M x(t_f) + \sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T g^{l_i}(\tau_{i+1}, x(\tau_{i+1})) + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i}^{\tau_{i+1}} (H^{l_i}(x,u,\lambda) - \lambda^T(E^{l_i}\dot{x}) - (E^{l_i}\dot{x})^T\lambda)dt \Bigg\},$$

and analogously for $u_\star$, $x_\star$, $\tau_\star$

$$\mathcal{S}_a(x_\star,u_\star,\tau_\star) = \frac{1}{2}\Bigg\{ x_\star(t_f)^T M x_\star(t_f) + \sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T g^{l_i}(\tau^\star_{i+1}, x_\star(\tau^\star_{i+1})) + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau^\star_{i+1}} (H^{l_i}(x_\star,u_\star,\lambda) - \lambda^T(E^{l_i}\dot{x}_\star) - (E^{l_i}\dot{x}_\star)^T\lambda)dt \Bigg\}.$$

Combining these formulas, we get

$$\begin{aligned}
\mathcal{S}_a(x,u,\tau) - \mathcal{S}_a(x_\star,u_\star,\tau_\star) = \frac{1}{2}\Bigg\{ & \quad x(t_f)^T M x(t_f) - x_\star(t_f)^T M x_\star(t_f) \\
& + \sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T [g^{l_i}(\tau_{i+1}, x(\tau_{i+1})) - g^{l_i}(\tau^\star_{i+1}, x_\star(\tau^\star_{i+1}))] \\
& + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau^\star_{i+1}} (H^{l_i}(x,u,\lambda) - H^{l_i}(x_\star,u_\star,\lambda))dt \\
& + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau^\star_{i+1}} \underbrace{(\lambda^T \underbrace{(E^{l_i}\dot{x}_\star - E^{l_i}\dot{x})}_{-\varepsilon E^{l_i}\dot{\varphi}} + \underbrace{(E^{l_i}\dot{x}_\star - E^{l_i}\dot{x})^T}_{-\varepsilon (E^{l_i}\dot{\varphi})^T}\lambda)}_{-2\varepsilon\lambda^T E^{l_i}\dot{\varphi}} dt \\
& + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau^\star_{i+1}}^{\tau^\star_{i+1}+\delta\tau_{i+1}} (H^{l_i}(x,u,\lambda) - \lambda^T(E^{l_i}\dot{x}) - (E^{l_i}\dot{x})^T\lambda)dt \\
& - \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau_i^\star+\delta\tau_i} (H^{l_i}(x,u,\lambda) - \lambda^T(E^{l_i}\dot{x}) - (E^{l_i}\dot{x})^T\lambda)dt \Bigg\},
\end{aligned}$$

using that $\tau_i = \tau_i^\star + \delta\tau_i$ and $\delta\tau_i$ small enough. For $t \in [\tau_i^\star, \tau^\star_{i+1})$ we have

$$\begin{aligned}
\frac{1}{2}\left\{H^{l_i}(x,u,\lambda) - H^{l_i}(x_\star,u_\star,\lambda)\right\} = \frac{1}{2}\Big\{ & x^T Q_{l_i} x + x^T S_{l_i} u + u^T S_{l_i}^T x + u^T R_{l_i} u \\
& + \lambda^T(A^{l_i}x + B^{l_i}u) + (A^{l_i}x + B^{l_i}u)^T\lambda \\
& - x_\star^T Q_{l_i} x_\star - x_\star^T S_{l_i} u_\star - u_\star^T S_{l_i}^T x_\star - u_\star^T R_{l_i} u_\star \\
& - \lambda^T(A^{l_i}x_\star + B^{l_i}u_\star) - (A^{l_i}x_\star + B^{l_i}u_\star)^T\lambda \Big\},
\end{aligned}$$

and using that

$$u = u_\star + \varepsilon v, \quad x = x_\star + \varepsilon\varphi,$$

we get

$$\begin{aligned}\frac{1}{2}\left\{H^{l_i}(x,u,\lambda) - H^{l_i}(x_\star,u_\star,\lambda)\right\} =& \frac{1}{2}\varepsilon\left\{x_\star^T Q_{l_i}\varphi + \varphi^T Q_{l_i} x_\star + x_\star^T S_{l_i} v + \varphi^T S_{l_i} u_\star\right.\\ &+ u_\star^T S_{l_i}^T \varphi + v^T S_{l_i}^T x_\star + u_\star^T R_{l_i} v + v^T R_{l_i} u_\star\\ &\left.+\lambda^T(A^{l_i}\varphi + B^{l_i} v) + (A^{l_i}\varphi + B^{l_i} v)^T\lambda\right\} + O(\varepsilon^2)\\ =& \varepsilon\left\{[x_\star^T Q_{l_i} + u_\star^T S_{l_i}^T + \lambda^T A^{l_i}]\varphi\right.\\ &\left.+[x_\star^T S_{l_i} + u_\star^T R_{l_i} + \lambda^T B^{l_i}]v\right\} + O(\varepsilon^2).\end{aligned}$$

Further, we have

$$\begin{aligned}&g^{l_i}(\tau_{i+1}, x(\tau_{i+1})) - g^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star))\\ &= g^{l_i}(\tau_{i+1}^\star + \delta\tau_{i+1}, x_\star(\tau_{i+1}^\star + \delta\tau_{i+1}) + \varepsilon\varphi(\tau_{i+1}^\star + \delta\tau_{i+1})) - g^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star))\\ &= \delta\tau_{i+1} g_{;t}^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star)) + (\varepsilon\varphi(\tau_{i+1}^\star) + \delta\tau_{i+1}\dot{x}_\star(\tau_{i+1}^\star)) g_{;x}^{l_i}(\tau_{i+1}^\star, x_\star(\tau_{i+1}^\star)) + R(\delta\tau_{i+1}, \varepsilon),\end{aligned}$$

where the remainder term $R(\delta\tau_{i+1}, \varepsilon)$ contains higher order terms of the variations $\delta\tau_{i+1}$ and ε, and

$$x^T M x - x_\star^T M x_\star = 2\varepsilon x_\star^T M\varphi + O(\varepsilon^2).$$

In addition, via partial integration we get

$$\begin{aligned}-\int_{\tau_i^\star}^{\tau_{i+1}^\star} \varepsilon\lambda^T E^{l_i}\dot{\varphi} dt =& -\varepsilon\lambda^T E^{l_i}\varphi\big|_{\tau_i^\star}^{\tau_{i+1}^\star} + \varepsilon\int_{\tau_i^\star}^{\tau_{i+1}^\star} \dot{\lambda}^T E^{l_i}\varphi dt\\ =& -\varepsilon\lambda^T(\tau_{i+1}^\star) E^{l_i}\varphi(\tau_{i+1}^\star) + \varepsilon\int_{\tau_i^\star}^{\tau_{i+1}^\star} \dot{\lambda}^T E^{l_i}\varphi dt,\end{aligned}$$

such that, altogether, we have

$$\begin{aligned}
&\mathcal{S}_a(x,u,\tau) - \mathcal{S}_a(x_\star,u_\star,\tau_\star) = O(\varepsilon^2) + \varepsilon x_\star^T(t_f) M \varphi(t_f) \\
&+ \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T \left[\delta\tau_{i+1} g_{;t}^{l_i}(t,x_\star) + (\varepsilon\varphi(t) + \delta\tau_{i+1}\dot{x}_\star(t)) g_{;x}^{l_i}(t,x_\star)\right]\Big|_{\tau_{i+1}^\star} + R(\delta\tau_{i+1},\varepsilon) \\
&+ \varepsilon \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau_{i+1}^\star} ([x_\star^T Q_{l_i} + u_\star^T S_{l_i}^T + \lambda^T A^{l_i} + \dot{\lambda}^T E^{l_i}]\varphi + [x_\star^T S_{l_i} + u_\star^T R_{l_i} + \lambda^T B^{l_i}]v) dt \\
&- \varepsilon \sum_{i=1}^{N_{\mathbb{I}}} \lambda^T(\tau_{i+1}^\star) E^{l_i} \varphi(\tau_{i+1}^\star) \\
&+ \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} \int_{\tau_{i+1}^\star}^{\tau_{i+1}^\star+\delta\tau_{i+1}} (H^{l_i}(x,u,\lambda) - H^{l_{i+1}}(x,u,\lambda) - 2\lambda^T(E^{l_i} - E^{l_{i+1}})\dot{x}) dt \\
&+ \frac{1}{2} \int_{\tau_{N_{\mathbb{I}}+1}^\star}^{\tau_{N_{\mathbb{I}}+1}^\star+\delta\tau_{N_{\mathbb{I}}+1}} (H^{l_{N_{\mathbb{I}}}}(x,u,\lambda) - 2\lambda^T E^{l_{N_{\mathbb{I}}}}\dot{x}) dt - \frac{1}{2}\int_{\tau_1^\star}^{\tau_1^\star+\delta\tau_1} (H^{l_1}(x,u,\lambda) - 2\lambda^T E^{l_1}\dot{x}) dt.
\end{aligned}$$

Since $\delta\tau_1 = \delta\tau_{N_{\mathbb{I}}+1} = 0$, the last two integrals vanish and from the midpoint theorem of integral calculus (see e.g. [41, §18, Theorem 7]) it follows that there exist a $\xi_{i+1} \in [\tau_{i+1}^\star, \tau_{i+1}^\star + \delta\tau_{i+1}]$, $i = 1,\dots,N_{\mathbb{I}}-1$, such that

$$\begin{aligned}
&\int_{\tau_{i+1}^\star}^{\tau_{i+1}^\star+\delta\tau_{i+1}} (H^{l_i}(x,u,\lambda) - H^{l_{i+1}}(x,u,\lambda) - 2\lambda^T((E^{l_i} - E^{l_{i+1}})\dot{x})) dt \\
&= \delta\tau_{i+1}\left(H^{l_i}(x,u,\lambda) - H^{l_{i+1}}(x,u,\lambda) - 2\lambda^T(E^{l_i} - E^{l_{i+1}})\dot{x}\right)\Big|_{\xi_{i+1}}.
\end{aligned}$$

Since $\delta\tau_{i+1}$ is small, we can assume that $\xi_{i+1} = \tau_{i+1}^\star$, and with (5.59) and defining

$$G(t,x) = \left[\frac{d}{dt} g_j^{l_i}(t,x)\right]^{-1} \frac{\partial}{\partial x} g_j^{l_i}(t,x)$$

we get

$$\begin{aligned}
&\mathcal{S}_a(x,u,\tau) - \mathcal{S}_a(x_\star,u_\star,\tau_\star) = O(\varepsilon^2) + \varepsilon x_\star^T(t_f) M \varphi(t_f) \\
&- \frac{1}{2}\varepsilon \sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T \varphi [G(t,x_\star)(g_{;t}^{l_i}(t,x_\star) + \dot{x}_\star g_{;x}^{l_i}(t,x_\star)) - g_{;x}^{l_i}(t,x_\star)]\Big|_{\tau_{i+1}^\star} + R(\delta\tau_{i+1},\varepsilon) \\
&+ \varepsilon \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau_{i+1}^\star} ([x_\star^T Q_{l_i} + u_\star^T S_{l_i}^T + \lambda^T A^{l_i} + \dot{\lambda}^T E^{l_i}]\varphi + [x_\star^T S_{l_i} + u_\star^T R_{l_i} + \lambda^T B^{l_i}]v) dt \\
&- \varepsilon \sum_{i=1}^{N_{\mathbb{I}}} \lambda^T(\tau_{i+1}^\star) E^{l_i} \varphi(\tau_{i+1}^\star)
\end{aligned}$$

$$-\frac{1}{2}\varepsilon\sum_{i=1}^{N_{\mathbb{I}}-1}\varphi G(t,x)\,[H^{l_i}(x,u,\lambda)-H^{l_{i+1}}(x,u,\lambda)-2\lambda^T(E^{l_i}-E^{l_{i+1}})\dot{x}]\big|_{\tau^\star_{i+1}}.$$

Since $\mathcal{S}_a(x,u,\tau)-\mathcal{S}_a(x_\star,u_\star,\tau_\star)\geq 0$ for all ε that are sufficiently small, it follows that the factor of ε must vanish for all v and corresponding φ. Choosing λ as solution of

$$-(E^{l_i})^T\dot{\lambda}(t)=(A^{l_i})^T\lambda(t)+Q_{l_i}x_\star+S_{l_i}u_\star,\quad \text{for } t\in[\tau_i^\star,\tau_{i+1}^\star) \tag{5.60}$$

with terminal conditions

$$(E^{l_i})^T\lambda(\tau_{i+1}^\star)=\frac{1}{2}g_{;x}^{l_i}(\tau_{i+1}^\star,x_\star(\tau_{i+1}^\star))^T\eta^{l_i},\quad \text{for } i=1,\ldots,N_{\mathbb{I}}-1, \tag{5.61}$$

and

$$(E^{l_{N_{\mathbb{I}}}})^T\lambda(\tau_{N_{\mathbb{I}}+1}^\star)=Mx_\star(t_f), \tag{5.62}$$

the first part of the integrand vanishes and we get as second condition that

$$\int_{\tau_i^\star}^{\tau_{i+1}^\star}[x_\star^TS_{l_i}+u_\star^TR_{l_i}+\lambda^TB^{l_i}]v\;dt=0,\quad \text{for all } v\in U_k(\tau).$$

Thus, it follows immediately that

$$x_\star^TS_{l_i}+u_\star^TR_{l_i}+\lambda^TB^{l_i}\equiv 0,\quad \text{for all } t\in[\tau_i^\star,\tau_{i+1}^\star). \tag{5.63}$$

The conditions for the remaining terms in the first and last sum to vanish are given by

$$\begin{aligned}H^{l_{i+1}}(x,u,\lambda)\big|_{\tau_{i+1}^{\star+}}&=H^{l_i}(x,u,\lambda)\big|_{\tau_{i+1}^{\star-}}+g_{;t}^{l_i}(\tau_{i+1}^\star,x_\star(\tau_{i+1}^\star))^T\eta^{l_i},\\(E^{l_{i+1}})^T\lambda(\tau_{i+1}^+)&=(E^{l_i})^T\lambda(\tau_{i+1}^-)-\frac{1}{2}g_{;x}^{l_i}(\tau_{i+1}^\star,x_\star(\tau_{i+1}^\star))^T\eta^{l_i}.\end{aligned} \tag{5.64}$$

Taking together equations (5.56a), (5.56c), (5.56d), (5.60), (5.61), (5.62), (5.63), and (5.64) we get a sequence of two-point boundary value problems (5.57). □

To prove the opposite statement, i.e., that the solution of the boundary value problem (5.57) yields a solution of the optimal control problem (5.56) we need some further assumptions on the cost functional and on the switching functions.

Theorem 5.47. *Let $x_\star,\lambda,u_\star,\tau_\star,\eta^l$ be chosen such that they solve the boundary value problem* (5.57)*. Further, let the matrices $\begin{bmatrix}Q_{l_i} & S_{l_i}\\ S_{l_i}^T & R_{l_i}\end{bmatrix}$ be positive semi-definite and assume that $g^{l_i}\in C^2([\tau_i,\tau_{i+1}],\mathbb{R}^{n_T^{l_i}})$ for all modes $l_i\in\mathbb{M}$ with*

$$(\eta^{l_i})^Tg_{;xx}^{l_i}(\tau_{i+1},x(\tau_{i+1}))\geq 0. \tag{5.65}$$

Then,

$$\mathcal{S}_a(x,u,\tau)\geq\mathcal{S}_a(x_\star,u_\star,\tau_\star)$$

for all x,u,τ satisfying (5.56a)*,* (5.56c)*,* (5.56e)*, and* (5.56d)*.*

Proof. We define

$$\phi(s) = \mathcal{S}_a(sx_\star + (1-s)x, su_\star + (1-s)u, s\tau_\star + (1-s)\tau).$$

Then, the assertion is equivalent to the statement that $\phi(s)$ has its minimum at $s = 1$ for all $x_\star, u_\star, \tau_\star$ satisfying the hybrid system $\mathcal{H}$. As $\phi(s)$ is quadratic in s it has a minimum for $s = 1$ if and only if

$$\left.\frac{d\phi}{ds}\right|_{s=1} = 0, \quad \text{and} \quad \left.\frac{d^2\phi}{ds^2}\right|_{s=1} \geq 0.$$

We have

$$\left.\frac{d\phi}{ds}\right|_{s=1} = \mathcal{S}_{a;x}(x_\star, u_\star, \tau_\star)(x_\star - x) + \mathcal{S}_{a;u}(x_\star, u_\star, \tau_\star)(u_\star - u) + \mathcal{S}_{a;\tau}(x_\star, u_\star, \tau_\star)(\tau_\star - \tau),$$

and

$$\begin{aligned}
\mathcal{S}_{a;x}(x_\star, u_\star, \tau_\star) &= x_\star^T M|_{t_f} + \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T g_{;x}^{l_i}(t, x_\star)\big|_{\tau_{i+1}^\star} + \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau_{i+1}^\star} (x_\star^T Q_{l_i} + u_\star^T S_{l_i}^T)dt, \\
\mathcal{S}_{a;u}(x_\star, u_\star, \tau_\star) &= \sum_{i=1}^{N_{\mathbb{I}}} \int_{\tau_i^\star}^{\tau_{i+1}^\star} (x_\star^T S_{l_i} + u_\star^T R_{l_i})dt, \\
\mathcal{S}_{a;\tau}(x_\star, u_\star, \tau_\star) &= \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T (g_{;t}^{l_i}(t, x_\star) + g_{;x}^{l_i}(t, x_\star)\dot{x}_\star)\big|_{\tau_{i+1}^\star} \\
&\quad + \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}} \begin{bmatrix} x_\star \\ u_\star \end{bmatrix}^T \begin{bmatrix} Q_{l_i} & S_{l_i} \\ S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} x_\star \\ u_\star \end{bmatrix}\Bigg|_{\tau_i^\star}^{\tau_{i+1}^\star}.
\end{aligned}$$

Further, we have

$$\begin{aligned}
&\sum_{i=1}^{N_{\mathbb{I}}} \begin{bmatrix} x_\star \\ u_\star \end{bmatrix}^T \begin{bmatrix} Q_{l_i} & S_{l_i} \\ S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} x_\star \\ u_\star \end{bmatrix}\Bigg|_{\tau_i^\star}^{\tau_{i+1}^\star} = \sum_{i=1}^{N_{\mathbb{I}}} \left[H^{l_i}(x_\star, u_\star, \lambda) - \lambda^T(E^{l_i}\dot{x}_\star) - (E^{l_i}\dot{x}_\star)^T\lambda\right]\Big|_{\tau_i^\star}^{\tau_{i+1}^\star} \\
&= -\sum_{i=1}^{N_{\mathbb{I}}-1} \left[H^{l_{i+1}}(x_\star, u_\star, \lambda) - H^{l_i}(x_\star, u_\star, \lambda) - 2\lambda^T(E^{l_{i+1}} - E^{l_i})\dot{x}_\star\right]\big|_{\tau_{i+1}^\star} \\
&\quad - \left[H^{l_1}(x_\star, u_\star, \lambda) - 2\lambda^T E^{l_1}\dot{x}_\star\right]\big|_{\tau_1^\star} + \left[H^{l_{N_{\mathbb{I}}}}(x_\star, u_\star, \lambda) - 2\lambda^T E^{l_{N_{\mathbb{I}}}}\dot{x}_\star\right]\big|_{\tau_{N_{\mathbb{I}}+1}^\star},
\end{aligned}$$

and thus, we get

$$\begin{aligned}
\left.\frac{d\phi}{ds}\right|_{s=1} &= x_\star^T M(x_\star - x)|_{t_f} \\
&\quad + \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T \left[g_{;x}^{l_i}(t, x_\star)(x_\star - x) + (g_{;t}^{l_i}(t, x_\star) + g_{;x}^{l_i}(t, x_\star)\dot{x}_\star)(\tau_{i+1}^\star - \tau_{i+1})\right]\big|_{\tau_{i+1}^\star}
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{i=1}^{N_{\mathbb{I}}}\int_{\tau_i^\star}^{\tau_{i+1}^\star}(x_\star^T Q_{l_i}(x_\star - x) + u_\star^T S_{l_i}^T(x_\star - x))dt\\
&+\sum_{i=1}^{N_{\mathbb{I}}}\int_{\tau_i^\star}^{\tau_{i+1}^\star}(x_\star^T S_{l_i}(u_\star - u) + u_\star^T R_{l_i}(u_\star - u))dt\\
&-\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}[H^{l_{i+1}}(x_\star,u_\star,\lambda) - H^{l_i}(x_\star,u_\star,\lambda) - 2\lambda^T(E^{l_{i+1}} - E^{l_i})\dot{x}_\star]_{\tau_{i+1}^\star}(\tau_{i+1}^\star - \tau_{i+1})\\
&-\frac{1}{2}[H^{l_1}(x_\star,u_\star,\lambda) - 2\lambda^T E^{l_1}\dot{x}_\star]_{\tau_1^\star}(\tau_1^\star - \tau_1)\\
&+\frac{1}{2}[H^{l_{N_{\mathbb{I}}}}(x_\star,u_\star,\lambda) - 2\lambda^T E^{l_{N_{\mathbb{I}}}}\dot{x}_\star]_{\tau_{N_{\mathbb{I}}+1}^\star}(\tau_{N_{\mathbb{I}}+1}^\star - \tau_{N_{\mathbb{I}}+1}).
\end{aligned}
$$

Since $\tau_1 = \tau_1^\star$ and $\tau_{N_{\mathbb{I}}+1} = \tau_{N_{\mathbb{I}}+1}^\star$, the last two terms vanish. Multiplying the second equation of (5.57a) once with $x_\star^T$ and once with x^T and inserting the other two equations of (5.57a) yields

$$
\begin{aligned}
x_\star^T Q_{l_i} x_\star &= -x_\star^T (E^{l_i})^T\dot{\lambda} - x_\star^T (A^{l_i})^T\lambda - x_\star^T S_{l_i} u_\star\\
&= -x_\star^T (E^{l_i})^T\dot{\lambda} - \dot{x}_\star^T (E^{l_i})^T\lambda + u_\star^T (B^{l_i})^T\lambda - x_\star^T S_{l_i} u_\star\\
&= -x_\star^T (E^{l_i})^T\dot{\lambda} - \dot{x}_\star^T (E^{l_i})^T\lambda - u_\star^T S_{l_i}^T x_\star - u_\star^T R_{l_i} u_\star - x_\star^T S_{l_i} u_\star, \qquad (5.66)\\
x^T Q_{l_i} x_\star &= -x^T (E^{l_i})^T\dot{\lambda} - \dot{x}^T (E^{l_i})^T\lambda - u^T S_{l_i}^T x_\star - u^T R_{l_i} u_\star - x^T S_{l_i} u_\star, \qquad (5.67)
\end{aligned}
$$

and therefore

$$
\begin{aligned}
x_\star^T Q_{l_i}(x_\star - x) = &-\dot{\lambda}^T E^{l_i}(x_\star - x) - \lambda^T E^{l_i}(\dot{x}_\star - \dot{x}) - x_\star^T S_{l_i}(u_\star - u)\\
&- u_\star^T R_{l_i}(u_\star - u) - u_\star^T S_{l_i}^T(x_\star - x).
\end{aligned} \qquad (5.68)
$$

With this we have

$$
\begin{aligned}
\left.\frac{d\phi(s)}{ds}\right|_{s=1} &= x_\star^T M(x_\star - x)|_{t_f} - \sum_{i=1}^{N_{\mathbb{I}}}\int_{\tau_i^\star}^{\tau_{i+1}^\star}\dot{\lambda}^T E^{l_i}(x_\star - x) + \lambda^T E^{l_i}(\dot{x}_\star - \dot{x})dt\\
&+\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}(\eta^{l_i})^T\left[g_{;x}^{l_i}(t,x_\star)(x_\star - x) + (g_{;t}^{l_i}(t,x_\star) + g_{;x}^{l_i}(t,x_\star)\dot{x}_\star)(\tau_{i+1}^\star - \tau_{i+1})\right]\Big|_{\tau_{i+1}^\star}\\
&-\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}[H^{l_{i+1}}(x_\star,u_\star,\lambda) - H^{l_i}(x_\star,u_\star,\lambda) - 2\lambda^T(E^{l_{i+1}} - E^{l_i})\dot{x}_\star]\Big|_{\tau_{i+1}^\star}(\tau_{i+1}^\star - \tau_{i+1}).
\end{aligned}
$$

Using partial integration, we get

$$
\int_{\tau_i^\star}^{\tau_{i+1}^\star}\dot{\lambda}^T E^{l_i}(x_\star - x)dt = \lambda^T E^{l_i}(x_\star - x)\Big|_{\tau_i^\star}^{\tau_{i+1}^\star} - \int_{\tau_i^\star}^{\tau_{i+1}^\star}\lambda^T E^{l_i}(\dot{x}_\star - \dot{x})dt,
$$

and thus, we have

$$\begin{aligned}\left.\frac{d\phi(s)}{ds}\right|_{s=1} &= x_\star^T M(x_\star - x)|_{t_f} - \sum_{i=1}^{N_{\mathbb{I}}} \lambda^T E^{l_i}(x_\star - x)\big|_{\tau_{i+1}^\star} + \sum_{i=1}^{N_{\mathbb{I}}} \lambda^T E^{l_i}(x_\star - x)\big|_{\tau_i^\star} \\ &+ \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T \left[g_{;x}^{l_i}(t,x_\star)(x_\star - x) + (g_{;t}^{l_i}(t,x_\star) + g_{;x}^{l_i}(t,x_\star)\dot{x}_\star)(\tau_{i+1}^\star - \tau_{i+1})\right]\big|_{\tau_{i+1}^\star} \\ &- \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} \left[H^{l_{i+1}}(x_\star,u_\star,\lambda) - H^{l_i}(x_\star,u_\star,\lambda) - 2\lambda^T(E^{l_{i+1}} - E^{l_i})\dot{x}_\star\right]\big|_{\tau_{i+1}^\star} (\tau_{i+1}^\star - \tau_{i+1}).\end{aligned}$$

With the conditions (5.57b), (5.57c), (5.57d) and (5.57e) all of the above terms vanish and we have $\frac{d\phi}{ds}\big|_{s=1} = 0$. For the second derivative of ϕ with respect to s we get

$$\begin{aligned}\left.\frac{d^2\phi(s)}{ds^2}\right|_{s=1} &= (x_\star - x)^T \mathcal{S}_{a;xx}(x_\star,u_\star,\tau_\star)(x_\star - x) + (u_\star - u)^T \mathcal{S}_{a;uu}(x_\star,u_\star,\tau_\star)(u_\star - u) \\ &\quad + \mathcal{S}_{a;\tau\tau}(x_\star,u_\star,\tau_\star)(\tau_\star - \tau)^2 + 2\mathcal{S}_{a;xu}(x_\star,u_\star,\tau_\star)(x_\star - x)(u_\star - u) \\ &\quad + 2\mathcal{S}_{a;\tau x}(x_\star,u_\star,\tau_\star)(\tau_\star - \tau)(x_\star - x) + 2\mathcal{S}_{a;\tau u}(x_\star,u_\star,\tau_\star)(\tau_\star - \tau)(u_\star - u).\end{aligned}$$

We have

$$\begin{aligned}\mathcal{S}_{a;xx}(x_\star,u_\star,\tau_\star) &= M + \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T g_{;xx}^{l_i}(t,x_\star)\Big|_{\tau_{i+1}^\star} + \sum_{i=1}^{N_{\mathbb{I}}} Q_{l_i}(\tau_{i+1}^\star - \tau_i^\star), \\ \mathcal{S}_{a;xu}(x_\star,u_\star,\tau_\star) &= \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}} (S_{l_i} + S_{l_i}^T)(\tau_{i+1}^\star - \tau_i^\star), \\ \mathcal{S}_{a;x\tau}(x_\star,u_\star,\tau_\star) &= \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T (g_{;xt}^{l_i}(t,x_\star) + g_{;xx}^{l_i}(t,x_\star)\dot{x}_\star)\Big|_{\tau_{i+1}^\star} + \sum_{i=1}^{N_{\mathbb{I}}} (x_\star^T Q_{l_i} + u_\star^T S_{l_i}^T)\big|_{\tau_i^\star}^{\tau_{i+1}^\star}, \\ \mathcal{S}_{a;uu}(x_\star,u_\star,\tau_\star) &= \sum_{i=1}^{N_{\mathbb{I}}} R_{l_i}(\tau_{i+1}^\star - \tau_i^\star), \\ \mathcal{S}_{a;u\tau}(x_\star,u_\star,\tau_\star) &= \sum_{i=1}^{N_{\mathbb{I}}} (x_\star^T S_{l_i} + u_\star^T R_{l_i})\big|_{\tau_i^\star}^{\tau_{i+1}^\star}, \\ \mathcal{S}_{a;\tau\tau}(x_\star,u_\star,\tau_\star) &= \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1} (\eta^{l_i})^T (g_{;tt}^{l_i}(t,x_\star) + 2g_{;tx}^{l_i}(t,x_\star)\dot{x}_\star + g_{;xx}^{l_i}(t,x_\star)\dot{x}_\star^2 + g_{;x}^{l_i}(t,x_\star)\ddot{x}_\star)\Big|_{\tau_{i+1}^\star} \\ &\quad + \frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}} \frac{\partial}{\partial t}\left[H^{l_i}(x_\star,u_\star,\lambda) - \lambda^T(E^{l_i}\dot{x}_\star) - (E^{l_i}\dot{x}_\star)^T\lambda\right]_{\tau_i^\star}^{\tau_{i+1}^\star},\end{aligned}$$

using the symmetry of second derivatives of g^{l_i}. Altogether, we have

$$\left.\frac{d^2\phi}{ds^2}\right|_{s=1} = (x_\star - x)^T M(x_\star - x) + \sum_{i=1}^{N_{\mathbb{I}}} \begin{bmatrix} x_\star - x \\ u_\star - u \end{bmatrix}^T \begin{bmatrix} Q_{l_i} & S_{l_i} \\ S_{l_i}^T & R_{l_i} \end{bmatrix} \begin{bmatrix} x_\star - x \\ u_\star - u \end{bmatrix} (\tau_{i+1}^\star - \tau_i^\star) \tag{5.69a}$$

$$+\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}(x_\star-x)^T(\eta^{l_i})^T g^{l_i}_{;xx}(t,x_\star)(x_\star-x)\Big|_{\tau^\star_{i+1}} \tag{5.69b}$$

$$+\sum_{i=1}^{N_{\mathbb{I}}-1}(\eta^{l_i})^T(g^{l_i}_{;xt}(t,x_\star)+g^{l_i}_{;xx}(t,x_\star)\dot{x}_\star)\Big|_{\tau^\star_{i+1}}(\tau_\star-\tau)(x_\star-x) \tag{5.69c}$$

$$+\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}(\eta^{l_i})^T(g^{l_i}_{;tt}(t,x_\star)+2g^{l_i}_{;tx}(t,x_\star)\dot{x}_\star+g^{l_i}_{;xx}(t,x_\star)\dot{x}_\star^2+g^{l_i}_{;x}(t,x_\star)\ddot{x}_\star)\Big|_{\tau^\star_{i+1}}(\tau_\star-\tau)^2 \tag{5.69d}$$

$$-2\sum_{i=1}^{N_{\mathbb{I}}-1}\dot{\lambda}^T(E^{l_i}-E^{l_{i+1}})(x_\star-x)|_{\tau^\star_{i+1}}(\tau_\star-\tau) \tag{5.69e}$$

$$+\sum_{i=1}^{N_{\mathbb{I}}-1}\left[\frac{1}{2}(H^{l_i}_{;t}-H^{l_{i+1}}_{;t})-\dot{\lambda}^T(E^{l_i}-E^{l_{i+1}})\dot{x}_\star-\lambda^T(E^{l_i}-E^{l_{i+1}})\ddot{x}_\star\right]_{\tau^\star_{i+1}}(\tau_\star-\tau)^2, \tag{5.69f}$$

using (5.68). From (5.57d) and (5.57e) we get that

$$\begin{aligned}
\dot{\lambda}^T E^{l_i}\Big|_{\tau_{i+1}} &= \frac{1}{2}(\eta^{l_i})^T\left[g^{l_i}_{;xt}(t,x)+g^{l_i}_{;xx}(t,x)\dot{x}\right]\Big|_{\tau_{i+1}},\\
\dot{\lambda}^T(E^{l_i}-E^{l_{i+1}})\Big|_{\tau_{i+1}} &= \frac{1}{2}(\eta^{l_i})^T\left[g^{l_i}_{;xt}(t,x)+g^{l_i}_{;xx}(t,x)\dot{x}\right]\Big|_{\tau_{i+1}},\\
H^{l_i}_{;t}-H^{l_{i+1}}_{;t}\Big|_{\tau_{i+1}} &= -(\eta^{l_i})^T\left[g^{l_i}_{;tt}(t,x)+g^{l_i}_{;tx}(t,x)\dot{x}\right]\Big|_{\tau_{i+1}},
\end{aligned}$$

such that the terms (5.69c), (5.69d), (5.69e) and (5.69f) vanish and we obtain

$$\begin{aligned}
\frac{d^2\phi}{ds^2}\Big|_{s=1} =&(x_\star-x)^T M(x_\star-x)+\sum_{i=1}^{N_{\mathbb{I}}}\begin{bmatrix}x_\star-x\\u_\star-u\end{bmatrix}^T\begin{bmatrix}Q_{l_i} & S_{l_i}\\ S_{l_i}^T & R_{l_i}\end{bmatrix}\begin{bmatrix}x_\star-x\\u_\star-u\end{bmatrix}(\tau^\star_{i+1}-\tau^\star_i)\\
&+\frac{1}{2}\sum_{i=1}^{N_{\mathbb{I}}-1}(x_\star-x)^T(\eta^{l_i})^T g^{l_i}_{;xx}(t,x_\star)(x_\star-x)\Big|_{\tau^\star_{i+1}}.
\end{aligned}$$

Since M and $\begin{bmatrix}Q_{l_i} & S_{l_i}\\ S_{l_i}^T & R_{l_i}\end{bmatrix}$ are positive semidefinite for all modes and

$$(\eta^{l_i})^T g^{l_i}_{;xx}(\tau^\star_{i+1},x^\star(\tau^\star_{i+1}))\geq 0$$

for all $l_i\in\mathbb{M}$ we have that $\frac{d^2\phi}{ds^2}\Big|_{s=1}\geq 0$. □

Thus, if the matrices $\begin{bmatrix}Q_{l_i} & S_{l_i}\\ S_{l_i}^T & R_{l_i}\end{bmatrix}$ are positive definite for each mode l_i, the condition (5.65) holds for the switching functions, and if $x_\star$, $\lambda_\star$, $u_\star$, $\tau_\star$, and η satisfy the boundary value problem (5.57) with the given transversality conditions, then $x_\star$, $u_\star$ and $\tau_\star$ form the optimal solution of the hybrid optimal control problem (5.56). Unfortunately, these conditions are, in general, not necessary, since the solution of the boundary value problem (5.57) may

not exist or may not be unique. The solution of the boundary value problem (5.57) for each mode $l_i \in \mathbb{M}$ can be obtained by solving an initial value problem for a matrix Riccati equation in the same way as in the standard linear-quadratic optimal control problem for DAEs, see e.g. [83].

5.6.4 Sliding Mode Control

The freedom in the choice of the controls u^l in each mode can also be used to steer the system dynamics during sliding motion in such a way that the solution trajectory of the hybrid system stays on the switching surface and in addition evolves as smooth as possible over the switch point. Thus, the principle of *sliding mode control* is to choose a control law, such that the solution of the descriptor system is forced to stay on a certain switching surface. Sliding mode control is widely used in the control theory for ordinary differential equations, for an introduction to sliding mode control see e.g. [145, 146]. To force the system state to stay on a switching surface, one must ensure that the system is able to reach the switching surface from any initial condition and, having reached the switching surface, that the control action is able to maintain the system state on the switching surface. In the following, we consider a linear hybrid system $\mathcal{H}$ that switches between two modes l and k described by the linear time-invariant descriptor systems

$$E^l \dot{x}^l = A^l x^l + B^l u^l + b^l(t), \tag{5.70}$$

and

$$E^k \dot{x}^k = A^k x^k + B^k u^k + b^k(t), \tag{5.71}$$

with $x^l(t), x^k(t) \in \mathbb{R}^n$. Here, we have again omitted the output equations as they do not contribute to the following analysis. We assume that switching occurs along the switching surface $g^l_{j_1}(t, x^l) = -g^k_{j_2}(t, x^k) = 0$ with $j_1 \in J^l$, $j_2 \in J^k$ and that a sliding condition is satisfied. Further, we assume that the descriptor systems (5.70) and (5.71) are regular and R-controllable as well as R-observable, i.e., the Assumption 5.37 should be satisfied, and further the transition function T^k_l should provide a smooth solution. Using a behavior approach, i.e., by setting

$$z^l = \begin{bmatrix} x^l \\ u^l \end{bmatrix},$$

the two systems can be transformed to strangeness-free form, i.e., we can consider equivalent strangeness-free descriptor systems similar as in Theorem 2.34.

Theorem 5.48. *Let $E, A \in \mathbb{R}^{n,n}$, $B \in \mathbb{R}^{n,k}$ and (E, A) be regular (i.e. the corresponding descriptor system is consistent and regular due to Corollary 5.32). Then, the corresponding descriptor system is equivalent (in the sense that the solution sets are in one-to one correspondence via a scaling by nonsingular matrices) to a descriptor system of the form*

$$\begin{aligned} \dot{x}_1 &= B_{12} u_2 + b_1(t), \\ 0 &= x_2 + B_{22} u_2 + b_2(t). \end{aligned} \tag{5.72}$$

Here, d_μ and a_μ are again the number of differential and algebraic equations.

Proof. See [84, Theorem 7]. □

Thus, using Theorem 5.48, without loss of generality we can consider the following two reduced descriptor systems

$$\begin{aligned} \dot{x}_1^l &= B_{12}^l u_2^l + b_1^l(t), \\ 0 &= x_2^l + B_{22}^l u_2^l + b_2^l(t), \end{aligned} \tag{5.73}$$

and

$$\begin{aligned} \dot{x}_1^k &= B_{12}^k u_2^k + b_1^k(t), \\ 0 &= x_2^k + B_{22}^k u_2^k + b_2^k(t). \end{aligned} \tag{5.74}$$

Next, we want to determine an equivalent control u_{eq} that forces the solutions of the two descriptor systems (5.73) and (5.74) onto the switching surface. For convenience, we assume that the switching function $g_{j_1}^l$ is linear in x and already partitioned according to the reduced form (5.73), i.e.,

$$g_{j_1}^l(t, x^l) = C_1^l x_1^l + C_2^l x_2^l, \tag{5.75}$$

with $C_1^l \in \mathbb{R}^{1,d_\mu^l}, C_2^l \in \mathbb{R}^{1,a_\mu^l}$. From the equivalent control method introduced in Section 5.5.1 for ordinary differential equations we know that an equivalent control $u_{2,eq}^l$ for system (5.73) can be obtained from the sliding condition via the solution of the system

$$\frac{d}{dt} g_{j_1}^l(t, x^l) = C_1^l \dot{x}_1^l + C_2^l \dot{x}_2^l = 0.$$

Using the differential equation in (5.73) and the derivative of the algebraic equation in (5.73) we get

$$u_{2,eq}^l = (C_1^l B_{12}^l)^{-1} [C_2^l \dot{b}_2^l(t) - C_1^l b_1^l(t)],$$

assuming that u_2^l is a piecewise constant function, i.e., $\dot{u}_2^l = 0$, and assuming that $C_1^l B_{12}^l$is invertible (transversality condition). Then, substituting the equivalent control $u_{2,eq}^l$ into (5.73) yields the differential-algebraic system during sliding motion of the form

$$\begin{aligned} \dot{x}_1^l &= B_{12}^l (C_1^l B_{12}^l)^{-1} [C_2^l \dot{b}_2^l(t) - C_1^l b_1^l(t)] + b_1^l(t), \\ 0 &= x_2^l + B_{22}^l (C_1^l B_{12}^l)^{-1} [C_2^l \dot{b}_2^l(t) - C_1^l b_1^l(t)] + b_2^l(t). \end{aligned} \tag{5.76}$$

Analogously, an equivalent control $u_{2,eq}^k$ and a differential-algebraic system in sliding motion can be derived for the descriptor system (5.74).

Theorem 5.49. *Consider the reduced regular descriptor system* (5.73)*. Assume that $g_{j_1}^l(t, x^l)$ is given by* (5.75) *and $C_1^l B_{12}^l$ is nonsingular. Then, the differential-algebraic system* (5.76) *during sliding is regular and strangeness-free.*

Proof. The differential-algebraic system (5.76) consists of an ordinary differential equation for x_1^l and a decoupled algebraic equation for x_2^l. Thus, the system is regular and strangeness-free. □

5.7 Future Work

A number of open questions remain in the analysis and in particular in the control of general nonlinear hybrid differential-algebraic systems. In the modeling of hybrid systems an important point is to allow that multiple transition conditions can be satisfied at the same point in time. Hybrid systems with this property arise frequently in practical applications, due to the use of modeling tools that implicitly define switching functions for each model component. Coupling together a number of similar components each having its own switching function often leads to hybrid models where a number of switching functions change their values at the same time. Further, we have assumed that the successor mode is uniquely determined by the mode allocation function. In practical applications this might be difficult to realize since the successor mode can depend on the system state. Another problem that arises is that the transition condition that causes the mode switching might not be fulfilled anymore after the consistent reinitialization. This problem of so-called *discontinuity sticking* as well as consistent event location is treated e.g. in [13].
In Section 5.6 we have extended the basic concepts of control theory for linear time-invariant descriptor systems to the case of hybrid systems. A more detailed investigation of control theoretical concepts is required considerig time-variant hybrid systems as well as nonlinear systems. Again, the control theory for linear DAEs with variable coefficients as presented in [9, 51, 77] and the theory for nonlinear DAEs as presented in [79, 83] can be used to describe the control theoretical concepts for the system in each mode, such that the results obtained in Section 5.6 in principle can also be extended to linear time-variant and nonlinear hybrid systems. Again, the number of characteristic values and therefore also the number of controls can change at a switch point, such that an investigation of the transitions at a mode change is required. For hybrid optimal control problems necessary conditions for optimality have been derived that lead to a sequence of boundary value problems with additional transversality conditions at the switch points, under the assumption that a fixed number of switch points occur. The problem to determine an optimal number of switch points and a corresponding optimal sequence of modes has to be considered for hybrid optimal control problems. Further, a more detailed investigation of sliding mode control for differential-algebraic systems with variable coefficients as well as nonlinear differential-algebraic systems is required considering also more general switching functions.

Chapter 6

Numerical Methods for Switched Differential-Algebraic Systems

The numerical simulation of hybrid differential-algebraic systems requires the efficient treatment of certain aspects in the hybrid system behavior. Besides the robust numerical integration of the DAEs in each operation mode, the points in time at which a mode change occur have to be detected accurately and in strict temporal sequence as they influence the mode switching and the future behavior of the hybrid systems. Thus, the points in time at which a transition condition changes its logical value have to be detected, which correspond to roots of the switching functions. These switch points have to be located presicely, as they are the initial points for the further integration. Further, the system state at the switch point has to be determined to restart the integration method in the new mode. Since the switch points will in general not coincide with the points chosen by the stepsize control, an interpolation method is needed that interpolates the computed solution at the detected switch point. In the following, we describe how these aspects of hybrid system simulation can be treated efficiently. At first, in Section 6.1 we review some basic aspects of polynomial interpolation. Next, in Section 6.2 we present numerical integration methods for DAEs, namely BDF methods and implicit Runge-Kutta methods, and show how interpolation between gridpoints can be realized using these methods. In Section 6.3 we consider the state event location and detection and present a root finding procedure based on a modified secant method that is used to determine the roots of the switching functions. Finally, in Section 6.4 we design a mode controller that combines the previously discussed methods for the numerical solution of hybrid differential-algebraic systems.

6.1 Polynomial Interpolation

In this section we consider the problem to approximate a function $f(t)$ for which function values f_i at discrete points t_i are given. The most frequently used method to approximate the function $f(t)$ is polynomial interpolation. The proofs of the following results are given e.g. in [134, Chapter 3].
Let $(n+1)$ discrete, pairwise disjoint nodes $t_0, \ldots, t_n$ and corresponding values $f_0, \ldots, f_n$ be given. Further, let Π_n denote the space of polynomials of maximal degree n, i.e.,

$$\Pi_n := \{p : \mathbb{R} \to \mathbb{R} \mid p(t) = \sum_{i=0}^{k} a_i t^i,\ a_i \in \mathbb{R},\ k \leq n\}. \tag{6.1}$$

The aim of polynomial interpolation is to find a polynomial $p_n \in \Pi_n$ of degree n which satisfies the interpolation condition

$$p_n(t_i) = f_i, \text{ for } i = 0, \ldots, n. \tag{6.2}$$

The existence and uniqueness of polynomial interpolation is the basis for the following observations.

Theorem 6.1. *Consider $(n+1)$ arbitrary points (t_i, f_i), $i = 0, \ldots, n$ with pairwise disjoint nodes $t_i \neq t_j$ for all $i \neq j$. Then, there exists a unique interpolation polynomial $p_n(t) \in \Pi_n$ such that $p_n(t_i) = f_i$ for all $i = 0, \ldots, n$.*

There are different possibilities to represent the interpolation polynomial. On the one hand, the Lagrange interpolation formula can be used. For given nodes $t_0, \ldots, t_n$ we consider the *Lagrange interpolation polynomials*

$$L_i(t) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{t - t_j}{t_i - t_j}, \quad i = 0, \ldots, n. \tag{6.3}$$

These polynomials of degree n have the property

$$L_i(t_k) = \begin{cases} 1, & \text{if } i = k, \\ 0, & \text{if } i \neq k \end{cases},$$

and the polynomial defined by

$$p_n(t) := \sum_{i=0}^{n} f_i L_i(t) \tag{6.4}$$

fulfills the interpolation condition (6.2). The explicit computation of the Lagrange polynomials is too expensive for the calculate of an interpolation polynomial, but the Lagrange interpolation formula (6.4) allows to derive formulas for numerical differentiations.

Theorem 6.2. *Let $f \in C^n([t_0, t_n], \mathbb{R})$. Then, there exists an $\xi \in (t_0, t_n)$ such that*

$$f^{(n)} = n! \sum_{i=0}^{n} f_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{1}{t_i - t_j}.$$

The error in the polynomial interpolation depends on the degree of the interpolation polynomial.

Theorem 6.3. *Let $f \in C^{n+1}([t_0, t_n], \mathbb{R})$ and consider an interpolation polynomial $p_n \in \Pi_n$ with pairwise disjoint nodes (t_0, f_0), $\ldots$, (t_n, f_n). Then for each $\bar{t} \in [t_0, t_n]$ we have*

$$f(\bar{t}) - p_n(\bar{t}) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (\bar{t} - t_i), \quad \xi \in (t_0, t_n).$$

On the other hand, the Newton interpolation formula can be used for the formulation of the interpolation polynomial.

Definition 6.4 (Divided difference). For a given number of nodes (t_j, f_j), $j = 0, \ldots, n$ the recursion

$$\begin{aligned} f[t_j] &:= f_j, \\ f[t_j, \ldots, t_{j+k}] &:= \frac{f[t_{j+1}, \ldots, t_{j+k}] - f[t_j, \ldots, t_{j+k-1}]}{t_{j+k} - t_j}, \end{aligned}$$

for $j = 0, \ldots, n, k = 0, \ldots, n - j$, defines the *k-th divided difference* of $(t_0, f_0), \ldots, (t_n, f_n)$.

Using divided differences, the polynomial $p_n(t)$ of degree n that interpolates (t_i, f_i) for $i = 0, \ldots, n$ can be expressed by the Newton interpolation formula

$$p_n(t) = \sum_{j=0}^{n} \prod_{l=0}^{j-1} (t - t_l) f[t_0, \ldots t_j]. \tag{6.5}$$

The Newton interpolation formula (6.5) allows to simply add further nodes and thus increase the degree of the interpolation polynomial successively.

6.2 DAE Integration Methods

In the numerical integration of DAEs, it is known for some time that applying standard discretization schemes for ordinary differential equations directly to differential-algebraic equations may lead to many difficulties due to the algebraic and hidden constraints, see e.g. [17, 59]. It may happen that the solution of the discretized equation is not uniquely solvable, while the original problem has a unique solution, or the numerical solution may drift-off from the analytical solution due to discretization errors. Further, many differential-algebraic systems behave like stiff differential equations which forces one to use methods with good stability properties. Therefore, not all numerical methods that are suitable for ordinary differential equations are also suitable for the numerical treatment of differential-algebraic equations. Further, the consistency of the approximate solution plays an important role in the accuracy and stability of numerical algorithms. In the following, we will use implicit Runge-Kutta methods and BDF methods for the numerical integration of differential-algebraic equations. We restrict to BDF and collocation Runge-Kutta methods since they have the great advantage to provide continuous solution representations which enable an efficient interpolation at switch points and since they are well suited for the numerical integration of DAEs.

In the following, we consider an initial value problem for nonlinear DAEs of the form (2.3) with initial value (2.4) in an interval $\mathbb{I} = [t_0, t_f]$, and we assume that the system has a unique solution provided that the initial value is consistent, i.e., the system is regular. As described in Section 2.2.2, we can transform a higher index problem (2.3) to a strangeness-free system

with the same solution. Therefore, in the following, we restrict our considerations to regular strangeness-free DAEs of the form

$$\begin{aligned} \hat{F}_1(t, x, \dot{x}) &= 0, \\ \hat{F}_2(t, x) &= 0, \end{aligned} \tag{6.6}$$

i.e., instead of the direct discretization of a higher index differential-algebraic equation (2.3), we discretize the equivalent strangeness-free formulation (6.6). The advantage of this equivalent strangeness-free formulation is that a parameterization of the constraint manifold is explicitly available, such that the numerical solution can be forced to lie on this manifold. In the following, let $t_0 < t_1 < t_2 < \cdots < t_N = t_f$ be the gridpoints in the interval $\mathbb{I}$ with $t_i = t_{i-1} + h$ and stepsize h, and we denote by x_i the approximations to the solution $x(t_i)$ at time t_i, for $i = 1, \ldots, N$. Here, we concentrate on a fixed stepsize $h = \frac{t_f - t_0}{N}$ in order to present the main results. A *discretization method* for the solution of (6.6) is given by an iteration

$$\mathcal{X}_{i+1} = \Phi(t_i, \mathcal{X}_i, h), \tag{6.7}$$

where $\mathcal{X}_i \in \mathbb{R}^n$ is an approximation to the solution at t_i, and $\mathcal{X}(t_i) \in \mathbb{R}^n$ represents the actual solution at t_i.

Definition 6.5 (**Consistency of a discretization method).** A discretization method (6.7) is said to be *consistent of order p* if

$$\|\mathcal{X}(t_{i+1}) - \Phi(t_i, \mathcal{X}(t_i), h)\| \leq Ch^{p+1},$$

for a constant C independent of h.

Definition 6.6 (**Stability of a discretization method).** A discretization method (6.7) is said to be *stable* if there exists a vector norm $\|.\|$ such that

$$\|\Phi(t_i, \mathcal{X}(t_i), h) - \Phi(t_i, \mathcal{X}_i, h)\| \leq (1 + hK)\|\mathcal{X}(t_i) - \mathcal{X}_i\|$$

in this vector norm, with a constant K independent of h.

Definition 6.7 (**Convergence of a discretization method).** A discretization method (6.7) is said to be *convergent of order p* if

$$\|\mathcal{X}(t_N) - \mathcal{X}_N\| \leq Ch^p,$$

with a constant C independent of h, provided that

$$\|\mathcal{X}(t_0) - \mathcal{X}_0\| \leq \tilde{C}h^p,$$

with a constant $\tilde{C}$ independent of h.

Theorem 6.8. *If the discretization method* (6.7) *is stable and consistent of order p, then it is convergent of order p.*

Proof. See [82, Theorem 5.4]. □

In the following we consider two types of discretization methods (6.7), namely Runge-Kutta methods and BDF methods.

6.2.1 Runge-Kutta Methods

In this section we consider the discretization of the strangeness-free DAE (6.6) via Runge-Kutta methods. An *s-stage Runge-Kutta method* for the computation of numerical approximations x_i to the values $x(t_i)$ of a solution x of (6.6) has the form

$$x_{i+1} = x_i + h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j}, \tag{6.8a}$$

where for $j = 1, \ldots, s$ the values $\dot{X}_{i,j}$ are given as solutions of the nonlinear system

$$\begin{aligned} \hat{F}_1(t_i + \gamma_j h, X_{i,j}, \dot{X}_{i,j}) &= 0, \\ \hat{F}_2(t_i + \gamma_j h, X_{i,j}) &= 0, \end{aligned} \tag{6.8b}$$

and the so-called *internal stages* $X_{i,j}$ are given by

$$X_{i,j} = x_i + h \sum_{l=1}^{s} \alpha_{jl} \dot{X}_{i,l}, \quad j = 1, \ldots, s. \tag{6.8c}$$

The coefficients α_{jl}, β_j and γ_j determine a particular Runge-Kutta method. In general, the coefficients are assumed to satisfy the condition

$$\gamma_j = \sum_{l=1}^{s} \alpha_{jl}, \quad i = 1, \ldots, s,$$

and the remaining freedom in the coefficients is used to obtain a certain order of consistency. Setting $\mathcal{X}_i = x_i$, $\mathcal{X}(t_i) = x(t_i)$, and $\Phi(t_i, \mathcal{X}_i, h) = x_i + h \sum_{j=1}^{s} \beta_j \dot{X}_{i,j}$, the Runge-Kutta method (6.8) can be seen as a general discretization method (6.7).

Theorem 6.9. *If the coefficients α_{jl}, β_j and γ_j of the Runge-Kutta method given by* (6.8) *satisfy the conditions*

$$\begin{aligned} B(p): &\quad \sum_{j=1}^{s} \beta_j \gamma_j^{k-1} = \frac{1}{k}, && k = 1, \ldots, p, \\ C(q): &\quad \sum_{l=1}^{s} \alpha_{jl} \gamma_l^{k-1} = \frac{1}{k} \gamma_j^k, && j = 1, \ldots, s,\ k = 1, \ldots, q, \\ D(r): &\quad \sum_{j=1}^{s} \beta_j \gamma_j^{k-1} \alpha_{jl} = \frac{1}{k} \beta_l (1 - \gamma_l^k), && l = 1, \ldots, s,\ k = 1, \ldots, r, \end{aligned} \tag{6.9}$$

with $p \leq q + r + 1$ and $p \leq 2q + 2$, then the method is consistent and convergent of order p.

Proof. See e.g. [58, p. 208]. □

An important class of Runge-Kutta methods for differential-algebraic equations are the so-called *stiffly accurate Runge-Kutta methods*, see e.g. [59]. These are defined to satisfy

$$\beta_j = \alpha_{sj} \text{ for all } j = 1, \dots, s. \tag{6.10}$$

From (6.10) it follows that the numerical solution x_{i+1} coincides with the last stage $X_{i,s}$. Therefore, it can be guaranteed that the numerical solution obtained by a stiffly accurate Runge-Kutta method is consistent with the strangeness-free DAE (6.6).
Another important class of Runge-Kutta methods are the so-called *collocation Runge-Kutta methods*. Starting with parameters γ_j, $j = 1, \dots, s$, that satisfy

$$0 < \gamma_1 < \cdots < \gamma_s = 1, \tag{6.11}$$

and setting $\gamma_0 = 0$, we can define the Lagrange interpolation polynomials corresponding to the parameters γ_j

$$L_l(\xi) = \prod_{\substack{j=0 \\ j \neq l}}^{k} \frac{\xi - \gamma_j}{\gamma_l - \gamma_j}, \quad \tilde{L}_l(\xi) = \prod_{\substack{m=l \\ m \neq l}}^{k} \frac{\xi - \gamma_m}{\gamma_l - \gamma_m}, \tag{6.12}$$

and the coefficients

$$\alpha_{jl} = \int_0^{\gamma_j} \tilde{L}_l(\xi) d\xi, \quad \beta_j = \int_0^1 \tilde{L}_j(\xi) d\xi, \quad j, l = 1, \dots, s. \tag{6.13}$$

This choice fixes a Runge-Kutta method with $\beta_j = \alpha_{sj}$ for $j = 1, \dots, s$, and thus the collocation methods are stiffly accurate. The stage values $X_{i,l}$, $l = 1, \dots, s$, together with $X_{i,0} = x_i$ define a polynomial $\pi_s \in \Pi_s$ via

$$\pi_s(t) = \sum_{l=0}^{s} X_{i,l} L_l \left(\frac{t - t_i}{h} \right), \tag{6.14}$$

and the derivative of π_s at the internal stages is given by

$$\dot{X}_{i,j} = \dot{\pi}_s(t_i + \gamma_j h) = \frac{1}{h} \sum_{l=0}^{s} X_{i,l} \dot{L}_l(\gamma_j), \quad j = 1, \dots, s.$$

Thus, in order to fix the new approximation $x_{i+1} = \pi_s(t_{i+1}) = X_{i,s}$ we require that the polynomial π_s given by (6.14) satisfies the strangeness-free system (6.6) at the so-called *collocation points* $t_{ij} = t_i + \gamma_j h$, $j = 1, \dots, s$.

Theorem 6.10. *The collocation Runge-Kutta methods defined by* (6.8) *and* $x_{i+1} = X_{i,s}$ *with coefficients* (6.13) *and collocation points as in* (6.11) *are convergent of order* $p = s$.

Proof. See [82, Theorem 5.17]. □

A special class of Runge-Kutta methods that are covered by Theorem 6.10 are the so-called *Radau IIa methods* defined by the conditions $B(2s-1)$, $C(s)$, and $D(s-1)$ of (6.9) together with $\gamma_s = 1$.

Theorem 6.11. *Choosing the nodes γ_j, $j = 1, \ldots, s$, in* (6.11) *and the coefficient α_{jl} and β_j such that $B(2s-1)$, $C(s)$, and $D(s-1)$ of* (6.9) *are satisfied, then the corresponding collocation Runge-Kutta method defined by* (6.8) *and $x_{i+1} = X_{i,s}$ is convergent of order $p = 2s - 1$.*

Proof. See [82, Theorem 5.18]. □

Due to the special choice of the nodes, the order of the Radau IIa methods is higher than suggested by Theorem 6.10. This effect is also called *superconvergence*. Therefore, Radau IIa methods have a large number of advantages as a high convergence rate and excellent stability properties, see also [59].
Further, Radau IIa methods are stiffly accurate by construction such that the consistency of the approximation x_i obtained from a Radau IIa method applied to strangeness-free DAEs of the form (6.6) follows from the consistency of the stages. These facts make the Radau IIa methods excellent candidates for the numerical integration of initial value problems for strangeness-free differential-algebraic equations of the form (6.6). All presented convergence results are based on the assumption that we use a constant stepsize. In the case of regular strangeness-free problems it is possible to use the same stepsize selection techniques as in the case of ordinary differential equations, see, e.g., [58]. Thus, the easy development of a suitable and efficient stepsize control, see e.g. [59], is a further advantage of Radau IIa methods. On the other hand, the discretization and the implementation of implicit Runge-Kutta methods is very technical and the solution of the nonlinear systems (6.8b) arising in every integration step is expensive.

6.2.2 BDF Methods

Besides one-step methods as the Runge-Kutta methods presented in the previous section also multi-step methods are frequently used for the numerical integration of DAEs. The general idea of a multi-step method is to use several previous approximations $x_i, \ldots, x_{i-k+1}$ for the computation of the approximation x_{i+1} to the solution value $x(t_{i+1})$. In the context of differential-algebraic equations, the most popular linear multi-step methods are the so-called *BDF methods*, where the abbreviation BDF stands for *backward differentiation formulae*. For the numerical solution of a regular strangeness-free differential-algebraic equation of the form (6.6) a BDF discretization is given by

$$\begin{aligned} \hat{F}_1(t_{i+1}, x_{i+1}, D_h x_{i+1}) &= 0, \\ \hat{F}_2(t_{i+1}, x_{i+1}) &= 0, \end{aligned} \tag{6.15}$$

α_{k-l}	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$	$l=6$
$k=1$	1	-1					
$k=2$	$\frac{3}{2}$	-2	$\frac{1}{2}$				
$k=3$	$\frac{11}{6}$	-3	$\frac{3}{2}$	$-\frac{1}{3}$			
$k=4$	$\frac{25}{12}$	-4	3	$-\frac{4}{3}$	$\frac{1}{4}$		
$k=5$	$\frac{137}{60}$	-5	5	$-\frac{10}{3}$	$\frac{5}{4}$	$-\frac{1}{5}$	
$k=6$	$\frac{147}{60}$	-6	$\frac{15}{2}$	$-\frac{20}{3}$	$\frac{15}{4}$	$-\frac{6}{5}$	$\frac{1}{6}$

Table 6.1: Coefficients for BDF methods

where

$$D_h x_{i+1} = \frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i+1-l}, \tag{6.16}$$

with coefficients α_{k-l} given in Table 6.1 for the simplest BDF methods. By the implicit function theorem $\hat{F}_1(t_{i+1}, x_{i+1}, D_h x_{i+1}) = 0$ can be locally solved for x_{i+1} by

$$x_{i+1} = \mathcal{S}(t_{i+1}, x_i, \ldots, x_{i-k+1}, h).$$

Then, by setting

$$\mathcal{X}_i = \begin{bmatrix} x_i \\ x_{i-1} \\ \vdots \\ x_{i-k+1} \end{bmatrix}, \ \mathcal{X}(t_i) = \begin{bmatrix} x(t_i) \\ x(t_{i-1}) \\ \vdots \\ x(t_{i-k+1}) \end{bmatrix}, \ \Phi(t_i, \mathcal{X}_i, h) = \begin{bmatrix} \mathcal{S}(t_{i+1}, x_i, \ldots, x_{i-k+1}, h) \\ x_i \\ \vdots \\ x_{i-k+2} \end{bmatrix},$$

the BDF method (6.16) can also be seen as a general discretization method (6.7). Note that we must provide $x_0, \ldots, x_{k-1}$ to initialize the iteration. These starting values are usually generated via appropriate one-step methods or within a combined order and stepsize control.

The basic idea for the derivation of BDF methods is to differentiate a polynomial $\pi_k \in \Pi_k$, which interpolates values $x_{i-k+1}, \ldots, x_{i+1}$ of x. We assume that the approximations $x_{i-k+1}, \ldots, x_i$ to the exact solution $x(t)$ at $t_{i-k+1}, \ldots, t_i$ are known and we consider the polynomial $\pi_k(t)$ of order k, which interpolates these values , i.e.,

$$\pi_k(t_{i-j+1}) = x_{i-j+1}, \quad j = 0, \ldots, k. \tag{6.17}$$

By Newton's interpolation formula (6.5) this polynomial can be expressed in terms of backward differences, see e.g. [59], as

$$\pi_k(t) = \pi_k(t_i + sh) = \sum_{j=0}^{k} (-1)^j \binom{-s+1}{j} \nabla^j x_{i+1}, \tag{6.18}$$

where the backward differences are defined as follows.

Definition 6.12 (Backward difference). Let $x_n, x_{n-1}, \ldots, x_{n-j-1} \in \mathbb{R}^n$ be given. Then the *j-th backward difference* $\nabla^j x_n$ is recursively defined by

$$\begin{aligned} \nabla^0 x_n &= x_n, \\ \nabla^{j+1} x_n &= \nabla^j x_n - \nabla^j x_{n-1} \text{ for } j \geq 0. \end{aligned}$$

The unknown value x_{i+1} will now be determined in such a way that the polynomial $\pi_k(t)$ satisfies the differential-algebraic equation at t_{i+1}. We have

$$\begin{aligned} \left.\frac{d\pi_k(t)}{dt}\right|_{t=t_{i+1}} &= \frac{1}{h}\sum_{j=0}^{k}(-1)^j \tfrac{d}{ds}\left.\binom{-s+1}{j}\right|_{s=1} \nabla^j x_{i+1} \\ &= \frac{1}{h}\sum_{j=0}^{k} \delta_j \nabla^j x_{i+1}, \end{aligned} \tag{6.19}$$

with coefficients

$$\delta_j := (-1)^j \tfrac{d}{ds}\left.\binom{-s+1}{j}\right|_{s=1}.$$

Using the definition of the binomial coefficients

$$(-1)^j\tbinom{-s+1}{j} = \frac{1}{j!}(s-1)s(s+1)\ldots(s+j-2) \text{ for } j > 0$$

and $\binom{-s+1}{0} = 1$, the coefficients δ_j are given by

$$\delta_0 = 0, \quad \delta_j = \frac{1}{j} \text{ for } j \geq 1.$$

Formula (6.19), therefore, becomes

$$\dot{x}_{i+1} = \dot{\pi}_k(t_{i+1}) = \frac{1}{h}\sum_{j=1}^{k}\frac{1}{j}\nabla^j x_{i+1},$$

and expressing this formula in terms of x_{i+1-l}, for $l = 0, \ldots, k$ gives (6.16).
In contrast to Runge-Kutta methods, the stability properties of BDF methods are somewhat restricted.

Theorem 6.13. *The k-step BDF method* (6.16) *is stable for* $1 \leq k \leq 6$*, and unstable for* $k \geq 7$*.*

Proof. See, e.g. [58, p. 381]. □

Theorem 6.14. *The BDF discretization* (6.15) *of* (6.6) *is convergent of order* $p = k$ *for* $1 \leq k \leq 6$ *provided that the initial values* $x_0, \ldots, x_{k-1}$ *are consistent.*

Proof. See [82, Theorem 5.27]. □

A nice property of BDF methods is that all numerical approximations x_{i+1} satisfy the algebraic constraints if the starting values $x_0, \ldots, x_{k-1}$ satisfy the algebraic constraints as well, which follows immediately from (6.15), such that consistency of the numerical solution is guaranteed.
The BDF-formulas (6.16) can be extended in a natural way to variable stepsizes $h_i = t_i - t_{i-1}$. In this case, the polynomial $\pi_k(t)$ of degree k that interpolates (t_j, x_j) for $j = i+1, i, \ldots, i-k+1$ can be expressed by Newton's interpolation formula (6.5) as

$$\pi_k(t) = \sum_{j=0}^{k} \prod_{l=0}^{j-1} (t - t_{i+1-l}) x[t_{i+1}, \ldots t_{i-j+1}]. \tag{6.20}$$

The variable stepsize BDF method is then given by

$$D_{h_i} x_{i+1} = \sum_{j=1}^{k} \prod_{l=1}^{j-1} (t_{i+1} - t_{i+1-l}) x[t_{i+1}, \ldots, t_{i-j+1}]. \tag{6.21}$$

Note, that the coefficients of the BDF method now depend on the past and current stepsizes. In the same way as for Runge-Kutta methods, there are no difficulties to supply BDF methods with a stepsize and order control if we restrict to regular strangeness-free problems of the form (6.6), see also [17] for more details.
Altogether, BDF methods provide an easy discretization and implementation for the numerical integration of DAEs and the solution of the nonlinear systems (6.15) arising in every integration step is relatively cheap. On the other hand, the development of a stepsize control is complicated and it is not very flexible in the choice of the stepsizes, see [48, 49]. A further disadvantage of BDF methods is that they are not stable for $k > 6$.

6.2.3 Interpolation

During the numerical solution of a hybrid system, the state of the system at the switch points is required. In addition, in order to evaluate the switching functions in the root-finding process, the numerical solution is required between gridpoints. Therefore, we need a continuous representation of the solution that enables an efficient interpolation of the approximate solution between gridpoints.
In standard applications a continuous representation of the solution is used only for output purpose, such that it is sufficient to require that the order q of the interpolation is such that the error in the interpolated values is of the size of the global error of the method. Thus, for a discretization method of order p it is sufficient to require that $q = p - 1$. But, in the case of hybrid system simulation the interpolated values are used for continuing the integration, such that we should require that $q = p$. Further, the interpolation error is propagated as it influences the determination of the switch points and the computation of the initial values used to restart the integration method. Thus, we need an error controlled continuous representation of the solution to ensure that the interpolation error is controlled.

For discretization methods based on a polynomial representation of the solution or its derivatives, like BDF methods or Runge-Kutta methods based on collocation, the construction of a continuous solution representation is straightforward as it is given by construction of the method. In the case of collocation Runge-Kutta methods the collocation polynomials $\pi_s \in \Pi_s$ as in (6.14) can be used for the continuous representation of the solution, and for BDF methods the polynomials $\pi_k \in \Pi_k$ as in (6.20) can be used.

Theorem 6.15. *The continuous representation based on a collocation polynomial π_s of an s-stage Runge-Kutta method is of order $q = s$, i.e., for all $t_i \le \hat{t} \le t_i + h$ we have*

$$\|x(\hat{t}) - \pi_s(\hat{t})\| \le Ch^{s+1}.$$

Moreover, for the derivatives of π_s we have

$$\|x^{(j)}(\hat{t}) - \pi_s^{(j)}(\hat{t})\| \le Ch^{s+1-j}, \quad j = 0, \ldots, s.$$

Proof. Setting $\dot{\pi}_s(t_0 + \gamma_j h) = k_j$, then from the Lagrange interpolation formula (6.4) we have

$$\dot{\pi}_s(t_0 + th) = \sum_{j=1}^{s} k_j L_j(t),$$

with L_j as in (6.12). Then, integration from 0 to γ_j yields

$$\pi_s(t_0 + \gamma_i h) = \pi_s(t_0) + h \int_0^{\gamma_i} \sum_{j=1}^{s} k_j L_j(t) dt. \tag{6.22}$$

The exact solution $x(t)$ satisfies the strangeness-free system (6.6) everywhere, hence also at the collocation points $t_0 + \gamma_i h$. We can apply the Lagrange interpolation formula (6.4) also to $\dot{x}(t)$ to get

$$\dot{x}(t_0 + th) = \sum_{j=1}^{s} \dot{x}(t_0 + \gamma_j h) L_j(t) + h^s R(t, h),$$

where the rest term $R(t, h)$ is a smooth function. Integration from 0 to γ_j yields

$$x(t_0 + \gamma_i h) = x(t_0) + h \int_0^{\gamma_i} \sum_{j=1}^{s} \dot{x}(t_0 + \gamma_j h) L_j(t) dt + h \int_0^{\gamma_i} h^s R(t, h) dt \tag{6.23}$$

and subtracting (6.22) from (6.23) with $\gamma_i = t$ yields

$$x(t_0 + th) - \pi_s(t_0 + th) = h \sum_{j=1}^{s} \Delta_j \int_0^t L_j(\tau) d\tau + h^{s+1} \int_0^t R(\tau, h) d\tau, \tag{6.24}$$

with $\Delta_j = \dot{x}(t_0 + \gamma_j h) - \dot{\pi}_s(t_0 + \gamma_j h)$. The k-th derivative of (6.24) with respect to t is given by

$$h^k \left(x^{(k)}(t_0 + th) - \pi_s^{(k)}(t_0 + th) \right) = h \sum_{j=1}^{s} \Delta_j L_j^{(k-1)}(t) + h^{s+1} \frac{\partial^{k-1}}{\partial t^{k-1}} R(t, h),$$

such that

$$\|x^{(k)}(t_0 + th) - \pi_s^{(k)}(t_0 + th)\| \leq h^{1-k} \| \sum_{j=1}^{s} \Delta_j L_j^{(k-1)}(t)\| + h^{s+1-k} \| \frac{\partial^{k-1}}{\partial t^{k-1}} R(t, h)\|.$$

Because of $C(s)$ the exact solution satisfies

$$x(t_0 + \gamma_j h) = x(t_0) + h \sum_{l=1}^{s} \alpha_{jl} \dot{x}(t_0 + \gamma_l h) + O(h^{s+1})$$

by Taylor expansion, such that for the internal stages $X_{i,j} = x_0 + h \sum_{l=1}^{s} \alpha_{jl} k_l$ we get

$$x(t_0 + \gamma_j h) - X_{i,j} = h \sum_{l=1}^{s} \alpha_{jl} (\dot{x}(t_0 + \gamma_l h) - k_l) + O(h^{s+1}), \quad \text{for } j = 1, \ldots, s.$$

Thus, $x(t_0 + \gamma_j h) - X_{i,j} = O(h^{s+1})$ and the result follows from the boundedness of the derivative of $R(t, h)$ and from $\Delta_j = O(h^{s+1})$. See also [58, Theorem 7.10, p. 213]. □

Thus, unfortunately for many collocation methods the order of the continuous representation is lower that the convergence order of the method, e.g., the continuous representation of the 5th order Radau IIa method has order $q = 3$. Alternatively, other interpolation schemes, as e.g. Hermite interpolation, might be applied if collocation Runge-Kutta methods are used as integration methods.

Theorem 6.16. *The continuous representation based on the collocation polynomial π_k of a k-step BDF method is of order $q = k$, i.e., for all $t_i \leq \hat{t} \leq t_i + h$ we have*

$$\|x(\hat{t}) - \pi_s(\hat{t})\| \leq C h^{k+1}.$$

Moreover, for the derivatives of π_k we have

$$\|x^{(j)}(\hat{t}) - \pi_k^{(j)}(\hat{t})\| \leq C h^{k+1-j}, \quad j = 0, \ldots, k.$$

Proof. The results can be proved similar as in Theorem 6.15 using the interpolation polynomial π_k given by (6.18) instead of π_s. □

Thus, for BDF methods the interpolant between t_i and t_{i+1} is of the same order k as the method that was used to advance the solution from t_i to t_{i+1}. Note that for integration methods based on a polynomial representation as the presented BDF or Runge-Kutta

methods the interpolation error is automatically error controlled by the error control of the numerical solution. The interpolant is continuous, but it has a discontinuous derivative at t_n and t_{n+1}. It is also possible to define a continuously differentiable interpolant, see [15, 148], which leads to a more robust code for root finding. Further, note that in general, the algebraic constraints for DAEs are not automatically satisfied at interpolated points. To restart the integration at the interpolated points consistent values need to be computed since the interpolated values at the switch points are in general not consistent with the algebraic equations.

6.3 Detection and Location of Events

A hybrid system $\mathcal{H}$ changes between different modes whenever a transition condition $L_j^l(t, x^l, \dot{x}^l)$ is satisfied, i.e., at the occurrence of an event that is implicitly described in terms of roots of switching functions. During the simulation, the numerical solution must be advanced speculatively until a transition condition is satisfied and the integration is stopped. This means that after each integration step from t_i to t_{i+1} changes in the logical values of the transition conditions $L_j^l(t, x^l, \dot{x}^l)$ are detected. Then the exact event time is determined by a root finding procedure as the root of a switching function in the interval $[t_i, t_{i+1}]$ in order to permit a re-initialization at the switch point. In general, during the numerical integration of a hybrid system all events have to be detected and located in strict temporal sequence and particularly no event should be missed in order to determine the correct mode changes. If more than one root is found in an integration interval, then the earliest event time is required. Note, that due to the assumptions in the Definition 5.3 of a hybrid system only one transition should be satisfied at a time.

Remark 6.17. *In [13] it has been noted that events might be missed during the integration of a hybrid system, since the stepsize selection is only sensitive to the solution behavior of the DAE, but not to the behavior of the switching functions. Therefore, it is proposed to append the switching functions to the system of DAEs in mode l and introduce additional variables, the so-called* switching variables *z^l, such that an augmented system*

$$\begin{aligned} F^l(t, x^l, \dot{x}^l) &= 0, \\ g^l(t, x^l, \dot{x}^l) &= z^l \end{aligned} \tag{6.25}$$

is integrated. In this way, interpolation polynomials are constructed that interpolate z^l with the same accuracy as x^l. The drawback of this approach is that adding the switching functions to the system of DAEs might result in an increase in the index. For example, consider the linear DAE

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1^l \\ \dot{x}_2^l \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1^l \\ x_2^l \end{bmatrix} + \begin{bmatrix} f_1^l \\ f_2^l \end{bmatrix}$$

in some mode l of a hybrid system and a switching function $g^l(t, x^l, \dot{x}^l) = \dot{x}_2^l - f_3^l(t)$. Then,

the augmented system

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{x}_1^l \\ \dot{x}_2^l \\ \dot{z}^l \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1^l \\ x_2^l \\ z^l \end{bmatrix} + \begin{bmatrix} f_1^l \\ f_2^l \\ f_3^l \end{bmatrix}$$

has strangeness index $\mu = 1$, while the original system is strangeness-free.
Further, the integration of the augmented system may require more steps because properties of discontinuity functions may limit the stepsize. However, to locate zero crossings accurately it is unavoidable to adapt the stepsize to the behavior of the switching functions. Nevertheless, we prefer a direct adaption of the stepsize selection to the behavior of the switching functions instead of the numerical integration of an augmented system (6.25).

6.3.1 The Root Finding Procedure

In order to localize an event, i.e., the earliest time $t^\star$ in an integration interval $[t_i, t_{i+1}]$ at which a transition condition is satisfied, the roots of the corresponding switching functions must be found with sufficiently high accuracy. In particular, the root finding procedure must ensure that all events are located precisely and if multiple roots exist in an interval the earliest event time must be located. Since the exact solution is not known, the roots of the switching functions are computed for a numerical solution. As the switching functions are evaluated over the whole integration interval, the system state has to be interpolated between meshpoints. Thus, the root finding procedure depends on interpolation formulas which are provided by the BDF formulas or by the Runge-Kutta method, see Section 6.2.3. Note, that BDF and Runge-Kutta methods guarantee consistency of differential and algebraic variables at meshpoints, but not necessarily at interpolated points.

In the following, we describe the procedure used to find the roots of a set of switching functions

$$g_j^l(t, x^l, \dot{x}^l), \quad \text{with } j = 1, \ldots, n_T^l,\ l \in \mathbb{M}$$

in an interval $[t_i, t_{i+1}]$. The chosen method is an adapted version of the root finding procedure of the SUNDIALS code IDA [64] that checks for sign changes of any $g_j^l(t, x^l, \dot{x}^l)$ in $[t_i, t_{i+1}]$ and then computes the roots with a modified secant method. The basic idea of the secant method is derived from Newton's method by approximating the derivatives using a finite difference approximation

$$\frac{d}{dt} g_j^l(t_m, x^l, \dot{x}^l) \approx \frac{g_j^l(t_m, x^l, \dot{x}^l) - g_j^l(t_{m-1}, x^l, \dot{x}^l)}{t_m - t_{m-1}}.$$

Then, the secant method is defined by the iterative relation

$$t_{m+1} = t_m - g_j^l(t_m, x^l, \dot{x}^l) \frac{t_m - t_{m-1}}{g_j^l(t_m, x^l, \dot{x}^l) - g_j^l(t_{m-1}, x^l, \dot{x}^l)}. \tag{6.26}$$

The iterates $\{t_m\}$ of the secant method (6.26) converge to a simple root of g_j^l, if the starting values t_0 and t_1 are sufficiently close to the root $t^\star$, and if $\frac{d}{dt}g_j^l(t^\star, x^l, \dot{x}^l) \neq 0$, as well as $g_j^l \in C^2$. Then, the order of convergence of (6.26) is $p = \frac{1+\sqrt{5}}{2}$, see e.g. [134, p. 201].
At first, the algorithm checks if g_j^l has an exact root at t_i. If an exact root of any g_j^l is found at time t_i, then $g_j^l(t_i + \delta, x^l, \dot{x}^l)$ is computed for a small increment $\delta > 0$. If $g_j^l(t_i + \delta, x^l, \dot{x}^l) = 0$ also has an exact root, then the procedure stops with an error message. In this way, it is guaranteed that the values of all g_j^l are nonzero at some past value of t_i, beyond which a search for roots is done. In the next step, if no roots at t_i were found, the algorithm checks g_j^l at t_{i+1} for exact zeros and detect sign changes in (t_i, t_{i+1}). If no sign changes are found, then either a root is reported if some $g_j^l(t_{i+1}, x^l, \dot{x}^l) = 0$, or no root is found in (t_i, t_{i+1}). If one or more sign changes were found, then a loop is entered to locate the roots within a tolerance $TTOL$, given by

$$TTOL = 100\,U(|t_{i+1}| + |h|), \tag{6.27}$$

where U is the rounding unit of the machine. When sign changes are found in two or more switching functions g_j^l, $j = 1, \ldots, n_T^l$, then the one with the largest value of

$$\frac{|g_j^l(t_{i+1}, x^l, \dot{x}^l)|}{|g_j^l(t_{i+1}, x^l, \dot{x}^l) - g_j^l(t_i, x^l, \dot{x}^l)|},$$

corresponding to the secant method value closest to t_i, is the one where most likely the sign change occurs first. At each pass through the loop, a new value t_{mid} within the search interval is set and the values of $g_j^l(t_{mid}, x^l, \dot{x}^l)$ are checked. The point t_{mid} is computed via

$$t_{mid} = t_{i+1} - g_j^l(t_{i+1}, x^l, \dot{x}^l)\frac{t_{i+1} - t_i}{g_j^l(t_{i+1}, x^l, \dot{x}^l) - \alpha g_j^l(t_i, x^l, \dot{x}^l)}, \tag{6.28}$$

where α is a weight parameter. On the first two passes through the loop, α is set to 1, such that t_{mid} is the classical secant method value. Afterwards, α is reset according to the side of the subinterval in which the sign change was found in the previous two steps. The value of t_{mid} is closer to t_i when $\alpha < 1$ and closer to t_{i+1} when $\alpha > 1$. If the value of t_{mid} in (6.28) is within $TTOL/2$ of t_i or t_{i+1}, it is adjusted inward, such that its distance from the endpoint relatively to the interval size is between 0.1 and 0.5, with 0.5 being the midpoint, and the actual distance from the endpoint is at least $TTOL/2$. Then, either t_i or t_{i+1} is reset to t_{mid} according to which subinterval is found to have the sign change. If there is no sign change in (t_i, t_{mid}), then that root is reported. The loop continues until $|t_{i+1} - t_i| < TTOL$, and then the reported root location is t_{i+1}.
In general, the root finding procedure is only able to find roots of odd multiplicity that corresponds to a sign change in one of the switching functions $g_j^l(t, x^l, \dot{x}^l)$, or exact zeros at t_i or t_{i+1}. If more than one switching function g_j^l has a root in the given interval or if multiple roots are found for one switching function, then the one closest to t_i is returned. If a switching function has a root of even multiplicity, it will probably be missed. If such a root is desired, the switching function should be reformulated such that it changes sign at

the desired root. In general, the switching functions should be chosen as simple as possible, e.g. linear, and if possible, different switching functions should be used for different mode transitions.

Remark 6.18. *Numerical integration methods for DAEs require continuity of the solutions and sometimes also of the derivatives of the solution depending on the order of the method and the error control. If discontinuities are present in the solution, wrong error estimates and a failure of stepsize and error control can result. Missing events during the numerical integration of a hybrid system can lead to inefficient behavior or even to the failure of integration methods, since large error estimates can cause repeated step rejections and a drastic reduction of the stepsize until eventually a discontinuity resulting from the event is passed. Thus, a hybrid system integrator and the root finding procedure should provide a reliable event detection and location to ensure an efficient integration of hybrid systems.*

6.4 A Hybrid Mode Controller

In this section we describe the basic ideas for the construction of a hybrid mode controller for the numerical solution of hybrid differential-algebraic systems. For the numerical integration of the DAEs in each mode existing integration methods for continuous differential-algebraic systems can be embedded in the mode controller. These integration methods are used to proceed the solution in time, while the mode controller determines the switch points, organizes the mode switching, and provides consistent initial values after mode switching to restart the integration method at the switch point.
In general, the mode controller can be designed and implemented independently of the integration method, the index reduction procedure, or the switch point location method. Ideally, the index reduction technique and the integration method should be chosen according to the type of the DAE in the current mode. Otherwise, any integration method suited for strangeness-free differential-algebraic systems that provides a continuous representation of the solution, as e.g. the BDF or Runge-Kutta methods introduced in Section 6.2, can be used for the numerical integration of the DAEs in each mode.

In the following, we assume that the DAEs in each mode are mathematically well-behaved in a small interval following an event. This assumption is necessary to ensure that the solution of each DAE is a smooth function over the whole integration step and, in particular, that a solution of the DAE in mode l exists until the end of the integration step, such that the DAE in mode l is not allowed to have a singularity at the switch point. Using this assumption we can lock the function evaluation for the DAE solver during an integration step. This means that the equations evaluated cannot be changed while a time step is taken even if the transition condition is satisfied (this is sometimes called *discontinuity locking* [12]). Once a successful time step from t_i to t_{i+1} has been taken, the integration routine determines if events have occurred during the time step. If so, then the control is returned to the mode controller and the exact event time in the integration interval $[t_i, t_{i+1}]$ is located. The switch point is determined within a certain tolerance as root of the

switching functions using a root finding procedure. Then the mode controller backtracks to the earliest event and performs the required transition. This means that the successor mode k is determined by the mode allocation function $S^l(j) = k$ and the system state at the switch point, determined by interpolation, is transfered to the initial state of the next mode k with the help of the transition function T_l^k defined in (5.6) via

$$T_l^k(x^l(\tau_i), \dot{x}^l(\tau_i)) = [x^\star, \dot{x}^\star].$$

Before the integration can be restarted in the new mode k the initial value $x^\star$ has to be checked for consistency, since $x^\star$ is not necessarily consistent with the DAE in mode k. If $x^\star$ is not consistent, then a consistent initial value $x^k(\tau_i)$ at τ_i is computed on the basis of $x^\star$ in such a way that the solution x^k extends the past solution x^l in a physically reasonable way by fixing the differential variables, see Section 5.4. Note that we may have different characteristic values for the DAE in the new mode. Then the integration can be restarted at the switch point τ_i in the new mode k with initial value $x^k(\tau_i)$. Note that one-step methods like Runge-Kutta methods can be restarted at any point τ_i since no past values are used. This is different for multi-step method like BDF methods which use past values to approximate the solution. In this case, changing to a lower order method is necessary if discontinuities in the solution due to an event occur. If the discontinuity is of order m, i.e., the m-th derivative of the solution x exhibit a jump, then the order of the discretization method has to be reduced in order to meet the smoothness requirements of the method. For a k-step BDF method we need a $(k+1)$-times continuously differentiable solution such that the order has to be reduced to $m-2$. This implies the knowledge of the order of discontinuity which is usually not known. Therefore, and also for safety reasons, we always restart the BDF method with $k = 1$ and use the information available from last interval before the discontinuity, e.g., we can use the backward differences to estimate an appropriate stepsize to restart the integration.
Further, chattering behavior, i.e., repeated switching between modes, is to be detected during the numerical integration of hybrid systems. If chattering occurs, e.g. if a maximal allowed number of successive mode switching in a short time period is exceeded or if the sliding condition (5.42) approximated by (5.43) is fulfilled, then the mode controller should enable sliding mode simulation. This means that the system behavior is approximated by the corresponding DAE in sliding mode (5.38) which is alternatively integrated until the system leaves the sliding region.

Altogether, a hybrid mode controller can be realized as follows:

1. Check if the given initial values are consistent with the DAE in the current mode l. If necessary, determine new consistent initial values.

2. Integration from t_i to t_{i+1} using an appropriate discretization method:

   (a) Index reduction.

   (b) Integration of the reduced system.

   (c) Detection of changes in the index or characteristic values. If changes occur, then stop the integration.

   (d) Check whether a transition condition L_j^l is satisfied. If a transition condition is satisfied, then return to the mode controller (Goto 3.), otherwise continue the integration for the next time step (Goto 2.).

3. Localization of the switch point $t^\star \in [t_i, t_{i+1}]$ as root of a switching function g_j^l.

4. Determination of the system state at the switch point $t^\star$ by interpolation, i.e., $x^l(t^\star) = \pi(t^\star)$, where π is an interpolation polynomial.

5. Determination of the successor mode by the mode allocation function $S^l(j) = k$.

6. Check if chattering has occurred. If so, provision of the possibility to switch to sliding mode.

7. Transfer of the system state to the new initial values in mode k using the transition function.

8. Set current mode $l := k$. Restart of the integration method (Goto 1.).

Remark 6.19. *We have assumed that it is possible to extend the solution of the differential-algebraic equation in the current mode in a small interval beyond the event time. This is sometimes not possible or not in a unique way, e.g., at impasse points [123], or when characteristic values change. Sometimes events are employed to switch the DAE system at these critical points and thus continue the simulation. However, if the solution cannot be extended past the event, the proposed event location algorithm does not apply. In principle, the event can be moved slightly to the left of the critical point, but even this can effect the error and stepsize control if the integrator attempts to locate a point beyond the event.*

Remark 6.20. *The concept of the proposed hybrid mode controller is based on the assumption that only one transition condition can be satisfied at the same point in time. Further, we assume that the successor mode is uniquely determined by the mode allocation function. In practical applications this might be difficult to realize since the successor mode can depend on the system state and multiple transitions at a switch point occur. It also might happen that an immediate transition occurs if a transition condition is satisfied directly after the transfer of the state or the computation of consistent initial values. Another problem that arises is that the transition condition that causes the mode switching might not be fulfilled anymore after the consistent reinitialization. This problem of so-called* discontinuity sticking *as well as consistent event location is also treated in [13].*

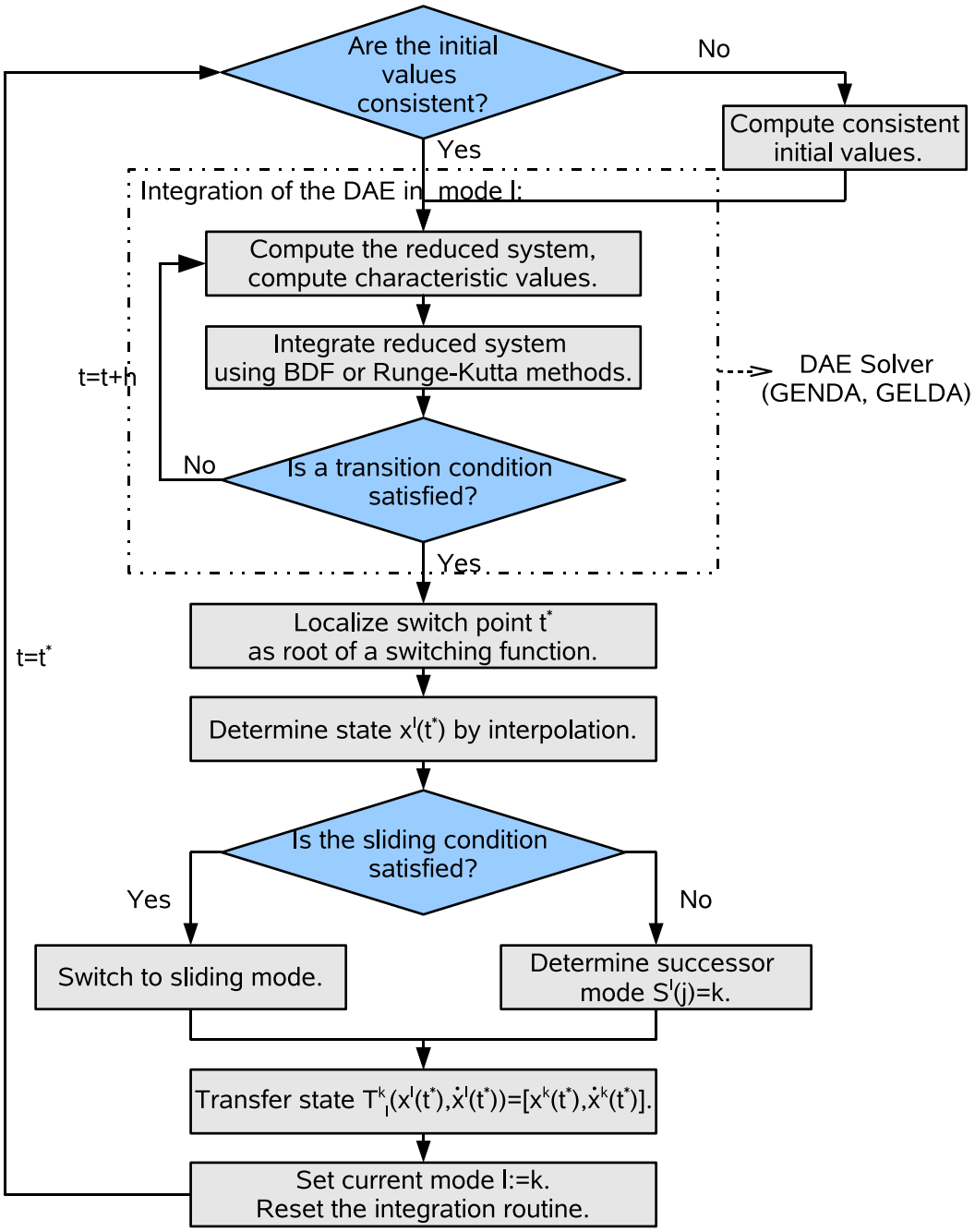

Figure 6.1: The hybrid mode controller

Chapter 7

A Mode Controller for Switched Differential-Algebraic Systems

In this chapter we describe the hybrid mode controller that has been implemented for the numerical solution of general nonlinear hybrid DAE systems. The mode controller is based on the numerical schemes presented in Chapter 6 and has been implemented in the FORTRAN code GESDA as a solver for **GE**neral **S**witched **D**ifferential-**A**lgebraic equations. For the numerical integration of the differential-algebraic systems inside each mode the differential-algebraic system solvers GELDA [85] and GENDA [87] have been embedded. In the following, we will describe the procedures used in the solver GESDA and discuss their features in detail. In Section 7.1.1, we describe the DAE solvers GELDA and GENDA that are used inside the mode controller for the numerical integration of the underlying DAEs. In Section 7.1.2, we describe the implementation of the sliding mode simulation. Finally, in Section 7.2 we present some numerical examples to demonstrate the applicability of the solver, and in Section 7.3 we give a short overview of further available DAE and hybrid system solvers.

7.1 The Hybrid System Solver GESDA

The solver GESDA is designed for the numerical solution of initial value problems for hybrid differential-algebraic systems as introduced in Definition 5.3 that consist of a number of either nonlinear DAEs of the form (5.3) or linear DAEs of the form (5.7) for each mode $l \in \mathbb{M}$. Mode switching occurs on the basis of transition conditions and the exact switch points are determined as the roots of switching functions. The user has to provide the descriptions of the DAEs in each mode together with a sufficient number of derivatives of the system depending on the index of the DAE in each mode. For linear DAEs of the form (5.7) the matrices E^l, A^l, and the right-hand sides b^l for all $l \in \mathbb{M}$ together with the corresponding derivatives up to a certain order k have to be provided in the subroutine MATSUB. For nonlinear DAE systems of the form (5.3) the functions F^l together with the corresponding derivatives up to order k have to be provided in the subroutine FUN, and the corresponding Jacobians in the subroutine DFUN. Further, the user has to specify the transition conditions for each mode in the subroutine UINTER and the corresponding switching functions in the subroutine GFUN. The roots of the switching functions defined in GFUN have to correspond to the points in time where a transition condition defined in UINTER changes its logical value. The successor modes are determined by the

mode allocation functions implemented in a user-provided subroutine MCHNG. Finally, the transition functions have to be provided in the subroutine TRANS that transfer the state of the system at a switch point to the initial state in the successor mode. For a detailed description of the user-supplied subroutines see the documentation of the code in the appendix.
The implemented solver follows the lines of the hybrid mode controller described in Section 6.4. After the possible computation of consistent initial conditions in the initial mode, the DAE in the current mode is numerically integrated according to its structure with an appropriate DAE solver, e.g. GELDA or GENDA, until an event occurs. After each successful time step of the numerical integration in the current mode $l \in \mathbb{M}$, the DAE solver checks whether a transition condition is satisfied by calls to the user-provided subroutine UINTER. Further, the DAE solver checks for changes in the index or in the characteristic values of the current DAE. If a transition condition is satisfied or changes in the characteristic values occur, then the integration is stopped. Changes in the characteristic values can only be handled by introducing further mode changes and additional modes, such that each DAE is well-defined in a small interval following a switch point. Otherwise, if a transition condition is satisfied, the switch point is localized as the root of a switching function. The root finding procedure implemented in the subroutine DRTFND uses the modified secant method described in Section 6.3.1. The default value for the tolerance TTOL used in the root finding procedure is computed by (6.27). This value can be adapted to a specific problem by the user. After the computation of the switch point, the state of the system at the switch point is determined by interpolation. The subroutines DCONTS and DBDTRP provide interpolation routines for Runge-Kutta and BDF methods, respectively, as described in Section 6.2.3. Next, the successor mode is determined using the subroutine MCHNG and the state is transfered to the new mode using the transition functions implemented in the subroutine TRANS. If the transfered state is consistent with the DAE in the new mode, then the DAE solver is restarted for the DAE in the new mode. Otherwise, new consistent values are determined in a least squares sense as described in Section 5.4 and then the DAE solver is restarted for the DAE in the successor mode. To restart the integration the order of the BDF method used in the next step is reset to $k = 1$ and the initial stepsize is set to the stepsize used in the last successful step in the predecessor mode. If no smooth transition at a mode change is possible, due to inconsistency of initial values or due to the user-defined transition functions, the code displays a warning. It might happen that after a mode switching event an immediate mode change is detected in the new mode. If only one transition is possible from the new mode, then this transition is carried out and the integration is resumed. If several transitions are possible, it is not clear which transition should be carried out, such that the code exits with an error message. If immediate mode changes occur repeatedly, i.e., if more than a maximal number MAXCGN of successive mode changes occur, the integration is stuck and the code exits with an error message. In this case, the transition conditions should be reformulated, e.g., if applicable, hysteresis can be used. The default value for the maximal number of immediate mode changes is given by MAXCGN=100, it can be adapted by the user. In addition, the maximal stepsize is restricted to HMAX in order to avoid stepping over regions. By default HMAX is set to

Subroutines within GESDA	
DGELDA	general linear differential-algebraic equation solver
DGENDA	general nonlinear differential-algebraic equation solver
DRTFND	root finding procedure
DBDTRP	computes a solution at a fixed time by backward differentiation interpolation
DCONTS	computes a solution at a fixed time by interpolation
DCKCON	checks consistency of initial values and if necessary computes consistent initial values

User-supplied subroutines	
USCAL	user-supplied scaling routine
UINTER	user-supplied subroutine defining transition conditions
TRANS	user-supplied subroutine defining transition functions
MCHNG	user-supplied subroutine defining mode allocation functions
GFUN	user-supplied subroutine defining switching functions
MATSUB	user-supplied subroutine for linear problems
FUN/DFUN	user-supplied subroutines for nonlinear problems

Table 7.1: The subroutines of GESDA and their purposes

HMAX=1.0. Also the maximal stepsize can be adapted by the user.

In Table 7.1 the subroutines of GESDA and their purposes as well as the user-supplied subroutines are summarized. For a more detailed description of the usage and the implementation of the code GESDA see the documentation of the code in the appendix.

7.1.1 The Embedded DAE Solvers

For the numerical integration of the DAEs inside each mode two DAE solvers have been embedded in the hybrid system solver GESDA. For general linear DAE systems the solver GELDA is used and for general nonlinear systems we use the solver GENDA. Note, that no restrictions on the index of the DAEs (5.3) or (5.7) are needed, since both solvers are constructed for the solution of DAE systems of arbitrary high index. There are several further routines for the numerical solution of differential-algebraic systems, which robustly solve continuous systems, see also Section 7.3, including solvers adapted to special structures arising e.g. in the equations of motion of multibody systems or in circuit equations. In general, further DAE solvers can be appended, in particular those suited for specially structured systems, such that the solution of the DAEs inside each mode can be adapted to the structure of the equations. Depending on the type of the differential-algebraic equations in each mode, an appropriate differential-algebraic system solver should be chosen.

The DAE solver GELDA

The first DAE solver that is embedded in GESDA is the solver GELDA [85] that was designed for the numerical integration of linear DAEs with variable coefficients of the form (2.5) with initial values $x(t_0) = x_0$ not necessarily consistent. While most of the standard integration methods are suitable only for regular strangeness-free DAEs, GELDA is suitable for the numerical integration of linear DAEs of arbitrary index and also allows the solution of over- and underdetermined systems. The implementation of GELDA is based on the combination of an index reduction procedure introduced in [72], that was shortly presented in Section 2.2.2, which first determines all the local invariants and then transforms the linear DAE (2.5) into an equivalent strangeness-free DAE (2.24) with the same solution set, followed by a discretization of the strangeness-free DAE either using a Runge-Kutta scheme adapted from the code RADAU5 [59] or a BDF method adapted from DASSL [17, 115]. The user has to provide the necessary number of derivatives of all system matrices. At each time step t_i the strangeness index μ and the characteristic quantities d_μ, a_μ, u_μ are computed iteratively using the derivative arrays $(\mathcal{M}_l, \mathcal{N}_l)$ as in (2.22), for $l = 0, 1, \ldots, \mu$. Then, unitary projections $Z_1(t_i)$, $Z_2(t_i)$ and $T_2(t_i)$, as defined in Theorem 2.41, are computed via singular value decompositions, that are used to extract the strangeness-free system. By computing the characteristic values at each time step t_i the code checks if changes in the characteristic values occur, and if so returns control to the calling program, setting an error flag. Consistent initial conditions are computed via correction of the given initial values by solving a minimization problem as described in Remark 5.27. For a detailed description of the code GELDA and of the order and stepsize control see [85]. Note that the code GELDA can also be used to solve descriptor systems of the form (5.49) using the behavior approach.

The DAE solver GENDA

The second embedded DAE solver is the solver GENDA [87], a nonlinear version of the code GELDA, that has been developed for nonlinear square problems of the form (2.3) of arbitrary index with initial values $x(t_0) = x_0$ not necessarily consistent. The code GENDA combines the index reduction technique described in Section 2.2.2 with the discretization of an equivalent strangeness-free formulation (2.19) of the DAE by use of BDF methods as described in Section 6.2. The user has to provide the necessary number of derivatives of the whole DAE, in particular, the whole derivative array $\mathcal{F}_l$ of level l given in (2.15), as well as the corresponding Jacobians. In addition, the characteristic quantities have to be provided by the user. Consistent initial values are computed as described in Section 5.4. In every integration step a nonlinear system of equations of the form

$$\begin{aligned} \mathcal{F}_\mu(t_i + h, x, \dot{x}, ..., x^{(\mu+1)}) &= 0, && \text{(7.1a)} \\ \tilde{Z}_1^T F(t_i + h, x, D_h x) &= 0, && \text{(7.1b)} \end{aligned}$$

is solved, where $\tilde{Z}_1$ denotes some approximation to Z_1 at the desired solution. Equation (7.1a) ensures that the algebraic constraints are satisfied and (7.1b) is a discretization of

the differential part of the reduced system (2.19). The solution of the underdetermined system of nonlinear equations (7.1a) is computed in a least squares sense using the subroutine NLSCON [111] which is an implementation of the Gauss-Newton method. In the computation of consistent initial values the solver enables the user to choose the differential variables to be kept fixed during the Gauss-Newton iterations. If any of the algebraic variables are chosen to be kept fixed the code returns an error message. Note, that the Jacobian of (7.1) is generally nonsquare but has full row rank at every solution $(x, \dot{x}, ..., x^{(\mu+1)})$ of the DAE (2.3) if $\tilde{Z}_1$ is a sufficiently good approximation to Z_1 and the stepsize is sufficiently small. This property extends to a neighborhood of the solution set, thus we get quadratic convergence to a solution and a simplified Gauss-Newton method [113] can be applied by fixing the Jacobian at any timestep. Before any rank decisions are made during the computation, the matrix $[-\mathcal{N}_\mu \;\; \mathcal{M}_\mu]$, with $\mathcal{M}_\mu$ and $\mathcal{N}_\mu$ as in (2.17), is equilibrated to lower its condition number by computing appropriate row and column scaling vectors. Optionally, the user can supply a scaling subroutine USCAL if an appropriate scaling method is known. On the other hand, it is also possible to completely deactivate scaling. Further, the user can require the code to verify the given characteristic values after consistent initial values have been computed or after the BDF solver successfully completed an iteration. In this case, GENDA returns an error message if any changes in the characteristic values are detected. For details of the integration routine see [87].

7.1.2 Sliding Mode Simulation

The code GESDA enables sliding mode simulation. During the simulation the code detects chattering, either by checking the sliding condition (5.42) using the approximation (5.43), or by comparing the distance between the last detected switch points. That means, if the distance between the last three detected switch points τ_{i+1}, τ_i, τ_{i-1} is lower than a chattering tolerance $TOLC$, i.e., if $\tau_{i+1} - \tau_i < TOLC$ and $\tau_i - \tau_{i-1} < TOLC$ with τ_{i+1} the last detected switch point, then the code assumes that chattering occurs. The default value for the chattering tolerance $TOLC$ is given by $TOLC = 10^{-3}$ and the default value for the parameter δ that is used in the approximation of the directional derivatives (5.43) is given by $\delta = 10^{-5}$. The default values can be adapted to the problem by the user.

Sliding mode simulation is initiated by the user using reverse communication. If chattering occurs, then the code stops with a warning message. The user can decide to switch to sliding mode, otherwise, the integration is resumed until a user-defined maximal number of switchings has occurred and the code stops with an error message. In the case of sliding mode simulation, the user has to provide the DAE describing the system during sliding motion that can be defined as in (5.46). Further, the user can completely disable the detection of chattering and the checking of the sliding condition.

Mode l	Transition j	Transition condition L_j^l	Successor mode $k = S^l(j)$
1	1	$i_D < -\eta$	3
	2	switch S is closed	4
2	1	$v_D < -\eta$	4
	2	switch S is opened	3
3	1	$v_D > \eta$	1
	2	switch S is closed	2
4	1	$i_D < -\eta$	2
	2	switch S is opened	1

Table 7.2: Transition conditions for the boost converter

7.2 Numerical Examples

In this section, we give a number of examples for hybrid differential-algebraic systems that were solved using the code GESDA. The first example is the boost converter introduced in Example 5.1. Further canonical examples are stick-slip vibrations of mechanical systems with dry friction that are considered in Sections 7.2.2 and 7.2.3. Finally, we consider the model of a bowed string exhibiting slip-stick transition behavior in Section 7.2.4. Note that throughout the whole section we omit physical units like meters or seconds.

7.2.1 The Boost Converter

As a first example we consider the boost converter that has been introduced in Example 5.1, consisting of an inductor L, a diode D, a capacitor C, a resistor R, a switch S, and a voltage source V_i. The system is modeled as a hybrid system with four operation modes according to Example 5.1. In each mode we have a linear DAE of size $m = n = 7$, consisting of the equations (5.1) together with the algebraic constraints (5.2), with characteristic values given in the following tabular.

Mode l	Characteristic values $(\mu^l, d_\mu^l, a_\mu^l, v_\mu^l, u_\mu^l)$
1	$(0, 2, 5, 0, 0)$
2	$(0, 2, 5, 0, 0)$
3	$(1, 1, 6, 0, 0)$
4	$(1, 1, 6, 0, 0)$

The transition conditions for the four modes depend on the state of the switch, on the current through the diode, and on the voltage across the diode. They are given in Table 7.2, where i_D and v_D denote the current through the diode and the voltage across the diode, respectively, and η is a small tolerance. The hybrid system is solved with the solver GESDA in the interval $[0, 1]$, using the BDF method of the DAE solver GELDA to integrate the DAEs in each mode, with the parameters $V_i = 48$, $R = 2$, $C = 10^{-2}$, $L = 1$,

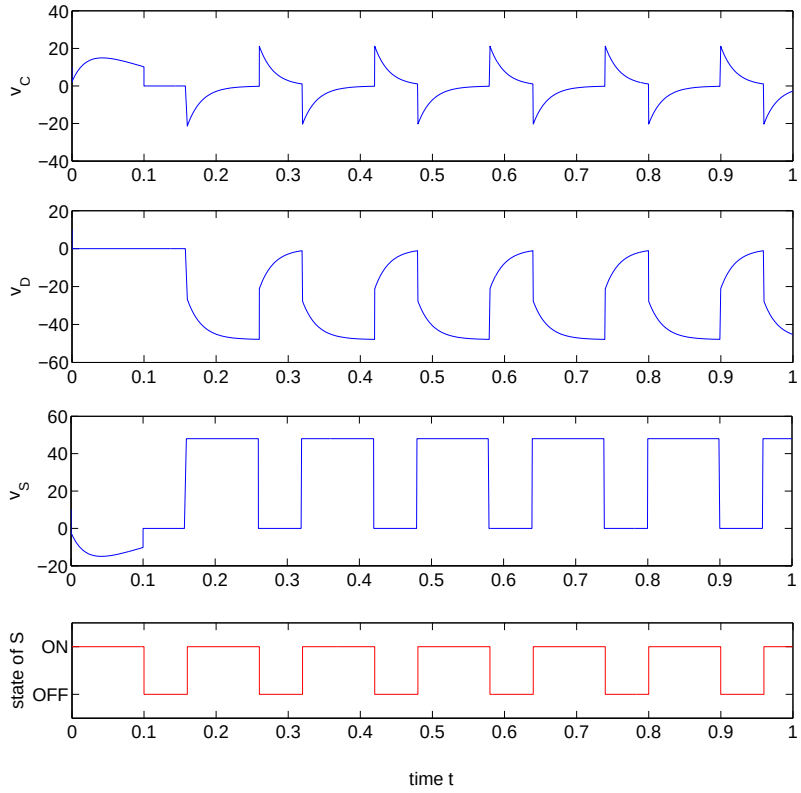

Figure 7.1: Simulation results for the boost converter

and $\eta = 10^{-6}$, the error tolerances $RTOL = ATOL = 10^{-8}$ and the initial value vector $[10, 10, 10, 10, 10, 10, 10]^T$ in the initial mode 1. Note that the code checks the initial values for consistency and corrects possible inconsistencies in a least squares sense. The results of the numerical simulation for the voltages v_C, v_D and v_S are given in Figure 7.1 together with the prescribed conducting and blocking time periods for the switch S. The switch is off for a period of length 0.1 and on for a period of length 0.06. After an initial phase the voltages adapt a periodic solution depending on the state of the switch. Further, the numerical error in the computed voltage v_D compared with the exact solution is plotted in Figure 7.2. We can observe higher errors at the switch points, i.e., after each restart of the numerical integration, due to the computation of the initial values. Altogether, the code detects 13 switch points during the numerical integration.

7.2.2 Stick-slip Friction Between Rigid Bodies

The second example considers stick-slip friction between two rigid bodies as depicted in Figure 7.3, see also Example 5.2 and [34]. The equations of motion of the multibody system

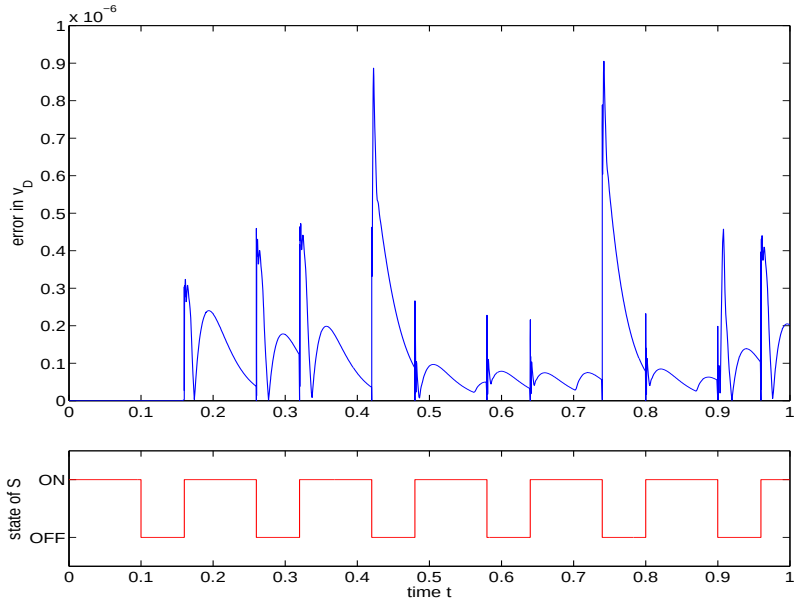

Figure 7.2: Numerical error in the computed voltage v_D

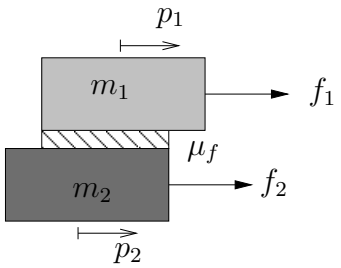

Figure 7.3: Stick-slip friction between rigid bodies

consisting of two masses with dry friction between them are given by

$$\begin{aligned} \dot{p}_1 &= v_1, \\ \dot{p}_2 &= v_2, \\ m_1\dot{v}_1 &= f_1 - \mu_f \|F_N\| \operatorname{sign}(v_1 - v_2), \\ m_2\dot{v}_2 &= f_2 + \mu_f \|F_N\| \operatorname{sign}(v_1 - v_2), \end{aligned} \tag{7.2}$$

where p_1, p_2 describe the positions of the bodies and v_1, v_2 are the corresponding velocities. Further, f_1 and f_2 are the applied forces, μ_f is the coefficient of friction, and F_N is the normal force on the surface between the two bodies. Thus, the system (7.2) represents a hybrid system consisting of two modes depending on the direction of the relative velocity between the bodies, i.e., the system is in mode 1 if $\operatorname{sign}(v_1 - v_2) = 1$ and the system is in

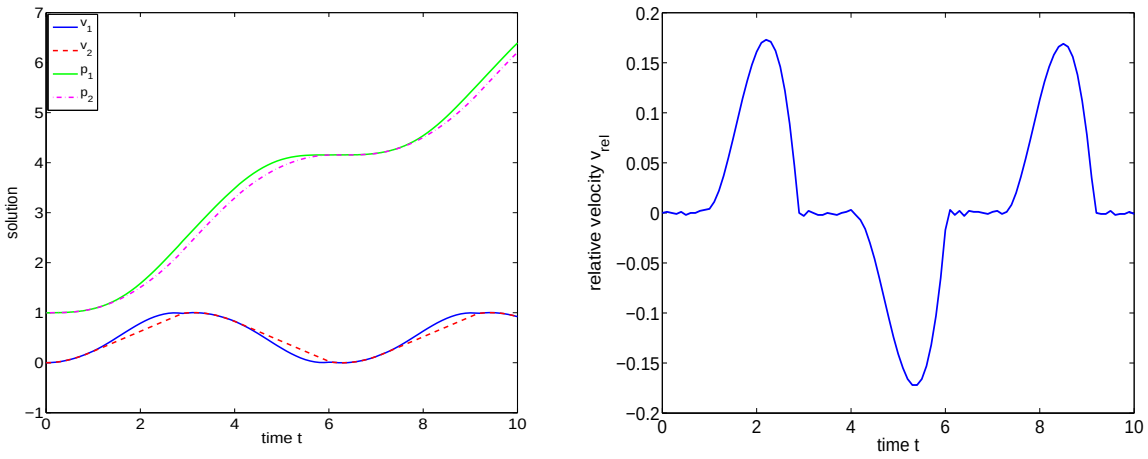

Figure 7.4: Solution of (7.2) and relative velocities between the two bodies

mode 2 if $\text{sign}(v_1 - v_2) = -1$. Note that the model equations (7.2) cannot be applied if the relative velocity is $v_{rel} = v_1 - v_2 = 0$, since sign(0) is not defined. We apply a small hysteresis band $[-\varepsilon, \varepsilon]$ around zero relative velocity to define the transition conditions

$$L_1^1(v_1, v_2) = v_1 - v_2 < -\varepsilon,$$
$$L_1^2(v_1, v_2) = v_1 - v_2 > \varepsilon.$$

The hybrid system (7.2) is solved with the solver GESDA using the BDF method of the DAE solver GELDA. Since (7.2) is an ordinary differential equation, the characteristic values in both modes are given by $\mu = 0$, $d_\mu = 4$, $a_\mu = v_\mu = u_\mu = 0$. The solution is computed in the interval $[0, 10]$ using $m_1 = m_2 = 1$, $f_1 = \sin(t)$, $f_2 = 0$, $F_N = 1$, $\mu_f = 0.4$, and $\varepsilon = 0.003$ with relative and absolute error tolerance $ATOL = RTOL = 10^{-8}$ and initial values $p_1(0) = p_2(0) = 1$, $v_1(0) = v_2(0) = 0$ in initial mode 1. The computed solution of the system and the relative velocity v_{rel} between the two bodies are given in Figure 7.4. We can see that in the beginning both bodies move together, since the applied force is less than the friction force. If the applied force exceeds the friction force, the velocity of the first body becomes greater than the velocity of the second body and the bodies slid over each other until the applied force is again lower than the friction force and the two bodies stick together once more. The relative velocity between the bodies can be seen to oscillate around zero during stiction, resulting in a large number of integration steps and a high computational effort. Altogether, the code requires 8209 integration steps, and 396 mode switches are detected. Physically, these oscillations do not occur, since in the case of stiction both bodies move together and the relative velocity is zero. Using sliding mode

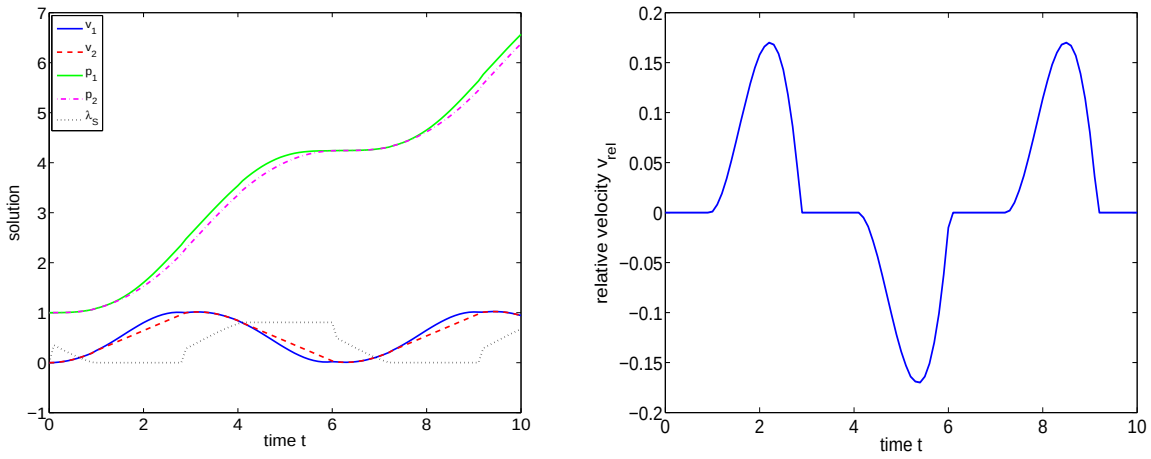

Figure 7.5: Solution and relative velocity using sliding mode simulation

simulation we can define the system behavior during stiction for $v_{rel} = 0$ by

$$\begin{aligned} \dot{p}_1 &= v_1, \\ \dot{p}_2 &= v_2, \\ m_1\dot{v}_1 &= f_1 - \mu_f\|F_N\| + \lambda_S, \\ m_2\dot{v}_2 &= f_2 + \mu_f\|F_N\| - \lambda_S, \\ 0 &= v_1 - v_2, \end{aligned} \tag{7.3}$$

defining the DAE in sliding mode (mode 3). Now, system (7.3) represents a DAE with characteristic quantities $\mu = 1, d_\mu = 4$, and $a_\mu = 1$. The system stays in sliding mode as long as the applied force is less than the friction force. If the applied force exceeds the friction force, then the system leaves the sliding mode, i.e., if $\sin(t) - 2\mu_s > 0$, then the system switches back to mode 1, and if $\sin(t) + 2\mu_s < 0$, then the system switches back to mode 2. The system is solved with GESDA using sliding mode simulation with chattering tolerance $TOLC = 0.01$ and $\delta = 10^{-5}$ used in the approximation of the sliding condition (5.43). The solution of the system together with the relative velocity v_{rel} using sliding mode simulation whenever chattering is detected by the code is given in Figure 7.5. We can see that the oscillations in the relative velocity during stiction disappear. Altogether, 12 switch points are detected during the numerical integration using sliding mode simulation and the number of integration step is reduced to 4737. The switch points together with the corresponding mode changes are given in Table 7.3. Note that in Figure 7.4 and in Figure 7.5 the solution is plotted only at predefined output points.

7.2.3 Stick-Slip Vibrations

In this example we consider stick-slip vibrations in a single-degree-of-freedom model given in Figure 7.6, see also [91]. We consider a mass m attached to the inertial space by a spring

Switch points $t^\star$	Mode changes
0.0038	$1 \to 2$
0.0112	$2 \to 1$
0.0188	$1 \to 3$
0.9273	$3 \to 1$
2.8719	$1 \to 2$
2.8776	$2 \to 1$
2.8886	$1 \to 3$
4.0689	$3 \to 2$
6.0340	$2 \to 3$
7.2105	$3 \to 1$
9.1465	$1 \to 2$
9.1521	$2 \to 3$

Table 7.3: Detected switch points in the solution of (7.2) using sliding mode simulation

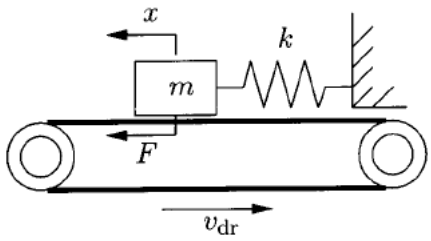

Figure 7.6: Model of a mass riding on a belt

of stiffness k that is riding on a driving belt moving at a constant velocity v_{dr}. Between mass and belt dry friction occurs, with a friction force F depending on the relative velocity between mass and belt. The equations of motion describing this system are given by

$$\begin{aligned} \dot{x} &= v, \\ m\dot{v} &= -kx - F(v_{rel}), \end{aligned} \tag{7.4}$$

where $v_{rel} = v - v_{dr}$ denotes the relative velocity of the mass with respect to the belt and the friction force F is given by

$$F(v_{rel}) = \begin{cases} -\mu_d \|F_N\| = \frac{-F_s}{1+\gamma|v_{rel}|} & \text{if } v_{rel} > 0, \\ \mu_d \|F_N\| = \frac{F_s}{1+\gamma|v_{rel}|} & \text{if } v_{rel} < 0. \end{cases}$$

Here, the dynamic friction coefficient μ_d is given by

$$\mu_d = \frac{\mu_s}{1 + \gamma|v_{rel}|},$$

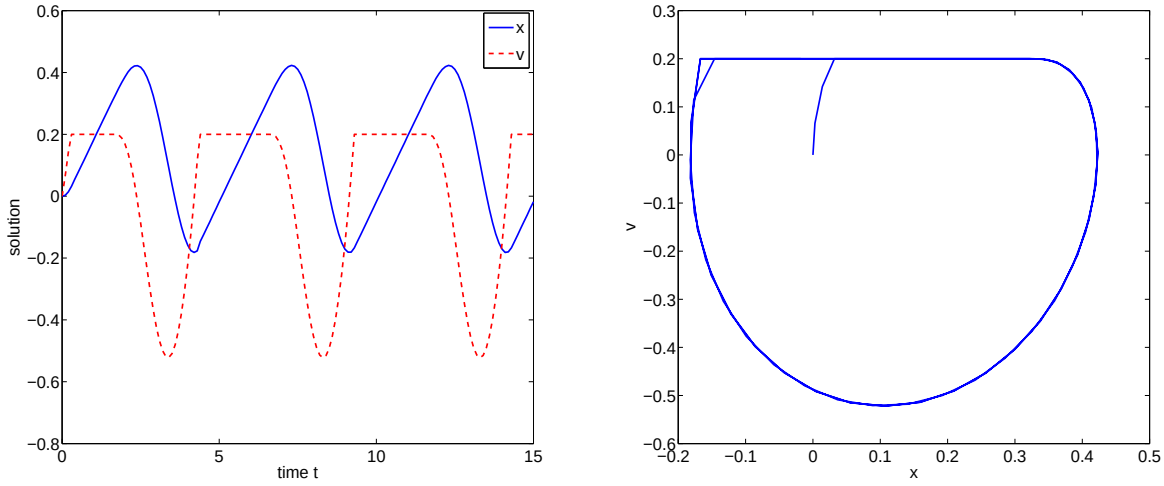

Figure 7.7: Solution of system (7.4) and phase portrait

where the positive parameter γ measures the rate at which μ_d decreases with an increase in $|v_{rel}|$, μ_s denotes the constant static friction coefficient, F_N is the normal force, and F_s denotes the maximum static friction force given by $F_s = \mu_s \|F_N\|$. Thus, the system (7.4) is a hybrid system consisting of two modes depending on the sign of the relative velocity, i.e., the system is in mode 1 if $\text{sign}(v_{rel}) = 1$ and the system is in mode 2 if $\text{sign}(v_{rel}) = -1$. The transition conditions for the two modes are given by

$$L_1^1(v) = v - v_{dr} \leq -\varepsilon,$$
$$L_1^2(v) = v - v_{dr} \geq \varepsilon,$$

applying a small hysteresis band $[-\varepsilon, \varepsilon]$ around zero relative velocity, where $\varepsilon \ll v_{dr}$. Again, system (7.4) is an ordinary differential equation with characteristic values $\mu = 0$, $d_\mu = 2$, $a_\mu = v_\mu = u_\mu = 0$. The hybrid system is solved with GESDA using the BDF method of the DAE solver GENDA in the interval $[0, 15]$ with parameters $m = 1, k = 3, F_s = 1, v_{dr} = 0.2, \gamma = 3.0, \varepsilon = 10^{-5}$, with relative and absolute error tolerance $RTOL = ATOL = 10^{-8}$ and initial values $x(0) = 0$, $v(0) = 1$ in initial mode 1. The computed solution of the system and the phase portrait are given in Figure 7.7. We can see a regular cycle of stick-slip phases and a stable stick-slip periodic solution of the system. During the numerical integration chattering occurs in regions of near-zero relative velocity. Altogether, the code detects 247709 switch points and requires 272864 integration steps to solve the system.
Again, we can use sliding mode simulation to handle the chattering behavior. We define the system behavior during stiction by

$$\begin{aligned} \dot{x} &= v, \\ m\dot{v} &= -kx - (2\lambda - 1)F_s, \\ 0 &= v - v_{dr}, \end{aligned} \tag{7.5}$$

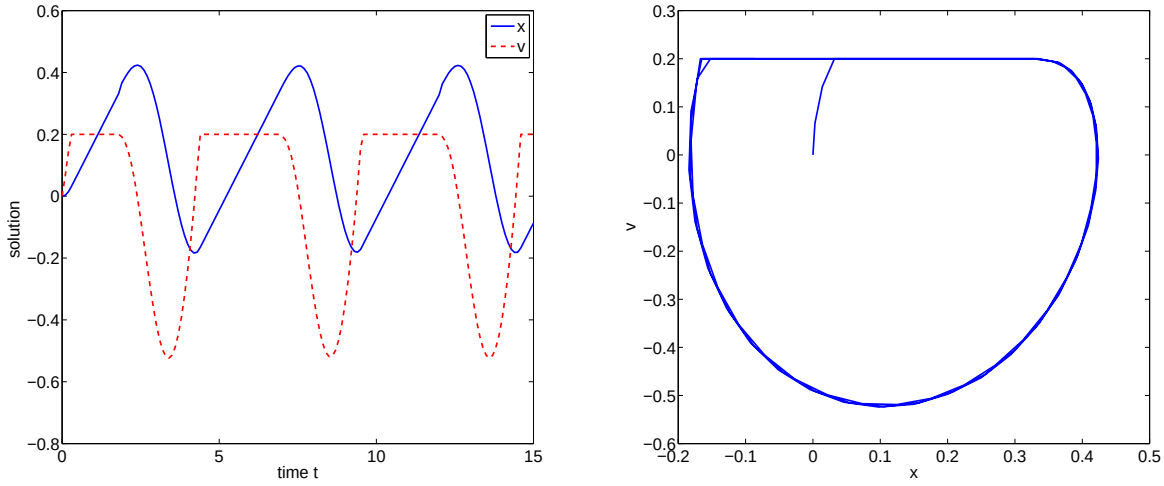

Figure 7.8: Solution of system (7.4) and phase portrait using sliding mode simulation

see also Example 5.30. The system (7.5) represents a DAE with characteristic quantities $\mu = 1, d_\mu = 1, a_\mu = 2, v_\mu = u_\mu = 0$. The system is solved with GESDA using sliding mode simulation with chattering tolerance $TOLC = 10^{-3}$ and $\delta = 10^{-8}$. The system stays in sliding mode as long as the applied force is less than the friction force. If the applied force $f_a = kx$ exceeds the friction force, then the system leaves the sliding mode, i.e., if $F_s < -f_a$ the system switches back to mode 1, and if $F_s < f_a$, then the system switches back to mode 2. The computed solution of the system (7.4) with DAE in sliding mode (7.5) together with the phase portrait using sliding mode simulation whenever chattering is detected by the code is given in Figure 7.8. Also in this case, the number of integration steps is reduced drastically to 9276 and only 9 switch points occur. The detected switch points and the corresponding mode changes are given in Table 7.4. We can see that an immediate mode transition occurs at the switch point $t^\star = 9.4820$. Further, at the beginning of the integration, an immediate mode change from the initial mode 1 to mode 2 occurs. In both simulations, i.e., without sliding mode and using sliding mode simulation, the maximal stepsize is restricted to $HMAX = 0.1$ in order to avoid stepping over regions.

7.2.4 The Bowed String

In this example we consider a one-dimensional violin string that is scraped by a bow as depicted in Figure 7.9. We assume an idealized string of length $L > 0$ and mass m_S, which is clamped at position $x = 0$ and $x = L$ under tension T. The string is bowed with a bow of mass m_b of negligible width at a bowing point x_c. This model of a bowed string is also called the *Raman-Model*, see also [43, 68, 122]. The motion of the string is governed by

Switch points $t^\star$	Mode changes
0.2676	$2 \to 3$
1.8000	$3 \to 2$
4.4263	$2 \to 3$
6.9000	$3 \to 2$
9.4819	$2 \to 1$
9.4820	$1 \to 2$
9.4820	$2 \to 3$
11.998	$3 \to 2$
14.638	$2 \to 3$

Table 7.4: Detected switch points in the solution of (7.4) using sliding mode simulation

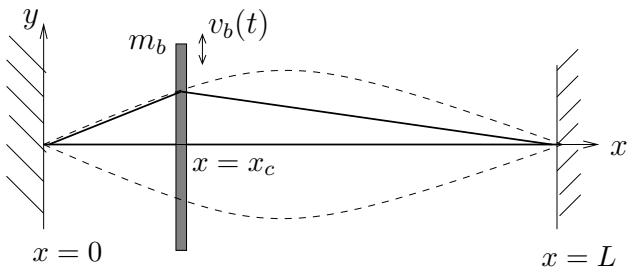

Figure 7.9: The bowed string

the one-dimensional wave equation

$$\rho \frac{\partial^2 u(x,t)}{\partial t^2} = T \frac{\partial^2 u(x,t)}{\partial x^2} + P(x,t), \tag{7.6a}$$

where $u(x,t)$ denotes the transversal displacement of the string at position $x \in [0,L]$ and time $t \in \mathbb{I} = [0,t_f]$ and $\rho > 0$ is the mass density, which is assumed to be constant $\rho = \frac{m_S}{L}$. Further, $P(x,t)$ denotes the external frictional force exerted by the bow on the string. We assume that the string is initially at rest in the undeformed configuration, i.e., we have the boundary conditions

$$u(0,t) = u(L,t) = 0 \quad \text{for all } t > 0, \tag{7.6b}$$

and the initial conditions

$$\begin{aligned} u(x,0) &= 0 \quad \text{for all } x \in [0,L], \\ \frac{\partial u}{\partial t}(x,0) &= 0 \quad \text{for all } x \in [0,L]. \end{aligned} \tag{7.6c}$$

The partial differential equation (7.6a) is discretized by the method of lines, see e.g. [131], i.e., we semidiscretize the spatial derivatives by second order central finite differences

$$\frac{\partial^2 u(x,t)}{\partial x^2} \approx \frac{u(x+\Delta x,t) - 2u(x,t) + u(x-\Delta x,t)}{\Delta x^2} + O(\Delta x^2),$$

to reduce the partial differential equation (7.6a) to a DAE system. Considering an equidistant grid $x_i = \frac{iL}{N}$, $i = 0, \ldots, N$ for the spatial variable x with spacing $\Delta x = \frac{L}{N}$, we get approximations $u_i(t)$ to $u(x_i, t)$. Incorporating the zero boundary conditions the semidiscretized equations are give by

$$\begin{aligned}
\rho \ddot{u}_0(t) &= T \frac{u_1(t) - 2u_0(t)}{\Delta x^2} + P(x_0, t), \\
\rho \ddot{u}_i(t) &= T \frac{u_{i+1}(t) - 2u_i(t) + u_{i-1}(t)}{\Delta x^2} + P(x_i, t), \quad \text{for } 1 \leq i \leq N-1, \\
\rho \ddot{u}_N(t) &= T \frac{-2u_N(t) + u_{N-1}(t)}{\Delta x^2} + P(x_N, t).
\end{aligned} \tag{7.7a}$$

With the boundary conditions (7.6b) and the initial conditions (7.6c) we get in addition the algebraic equations

$$\begin{aligned}
u_0(t) &= 0, \\
u_N(t) &= 0,
\end{aligned} \tag{7.7b}$$

as well as the initial conditions

$$u_i(0) = 0, \quad \dot{u}_i(0) = 0 \quad \text{for } i = 0, \ldots, N. \tag{7.7c}$$

The external frictional force $P(x, t)$ is of the form

$$P(x,t) = \begin{cases} F(v_{rel}(t)) & \text{for } x = x_c \\ 0 & \text{else} \end{cases},$$

since the bow exerts frictional force on the string only at the contact point x_c, where $F(v_{rel}(t))$ is a function depending on the relative velocity between bow and string at the bowing point x_c given by

$$v_{rel}(t) = \frac{\partial u(x_c, t)}{\partial t} - v_b(t) = \dot{u}_c(t) - v_b(t),$$

where $x_c = \frac{cL}{N}$, $1 \leq c \leq N-1$ is defined to be a grid point of the spatial discretization, and v_b denotes the bow velocity. The friction force $F(v_{rel}(t))$ is of the form

$$F(v_{rel}(t)) = \begin{cases} \mu_f \|F_N\| & \text{if } v_{rel} > 0, \\ -\mu_f \|F_N\| & \text{if } v_{rel} < 0, \end{cases}$$

where F_N is the normal bow force acting on the string and μ_f is the coefficient of friction. The equation of motion of the bow is given by

$$m_b \ddot{y}_b(t) = f_k(t) - F(v_{rel}(t)),$$

where $f_k(t)$ is the excitation of the bow. Altogether, after transformation to first order by introducing new variables $v_i = \dot{u}_i$ and $v_b = \dot{y}_b$ for the velocities, the bow-string system can be modeled as a multibody system of the form

$$\begin{aligned} \dot{y}_b &= v_b, \\ \dot{y}_s &= v_s, \\ m_b \dot{v}_b(t) &= f_k(t) - F(v_{rel}(t)), \\ \rho \dot{v}_s &= A y_s - G^T \lambda + e_c F(v_{rel}(t)), \\ 0 &= G y_s, \end{aligned} \tag{7.8}$$

with $y_s = [u_0, \ldots, u_N]^T \in \mathbb{R}^{N+1}$, $v_s = [v_0, \ldots, v_N]^T \in \mathbb{R}^{N+1}$, and $\lambda \in \mathbb{R}^2$, as well as

$$A = \frac{T}{\Delta x^2} \begin{bmatrix} -2 & 1 & 0 & \ldots & 0 \\ 1 & -2 & 1 & & \vdots \\ 0 & 1 & \ddots & \ddots & 0 \\ \vdots & & \ddots & -2 & 1 \\ 0 & \ldots & 0 & 1 & -2 \end{bmatrix} \in \mathbb{R}^{N+1,N+1}, \quad G = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 1 \end{bmatrix} \in \mathbb{R}^{2,N+1},$$

$$e_c = [0 \ \ldots \ 0 \ 1 \ 0 \ \ldots \ 0]^T \in \mathbb{R}^{N+1},$$

where e_c is the unit vector with 1 at position c and zeros otherwise, together with the initial conditions

$$y_s(0) = 0, \quad v_s(0) = 0, \quad y_b(0) = 0, \quad v_b(0) = v_0, \quad \lambda(0) = 0. \tag{7.9}$$

The characteristic values of the differential-algebraic system (7.8) are given by $\mu = 2, d_\mu = 2N, a_\mu = 6$, and $v_\mu = u_\mu = 0$. Thus, we have a hybrid DAE system with two modes $\mathbb{M} = \{1, 2\}$ corresponding to the states of $F(v_{rel}(t))$. The transition conditions are given by

$$\begin{aligned} L_1^1 &= v_c - v_b \leq -\varepsilon, \\ L_1^2 &= v_c - v_b > \varepsilon, \end{aligned}$$

where ε is a small constant defining a hysteresis band around zero relative velocity. The bowed string system (7.8) is solved with GESDA in the interval $[0, 6]$ using the BDF method of the DAE solver GELDA with error tolerance $ATOL = RTOL = 10^{-10}$, initial values (7.9) in initial mode 1, and using the parameters given in Table 7.5. The displacement and the velocity of the bow together with the current mode as well as the phase portrait of the bow motion are given in Figure 7.10. The displacement of the string at the bowing point and at an intermediate point on the string together with the current mode of the hybrid system are given in Figure 7.11. We can see that the motion of the bow results in a stable periodic solution. Further, fast changing between the two modes in regions of near zero

$L = 0.7$	length of the string
$T = 5.5$	tension on the string
$\rho = 0.5$	mass density of the string
$v_0 = 0.5$	initial bow velocity
$\Omega = 2$	frequency of excitation
$f_k = -v_0\Omega\sin(\Omega t)$	sinusoidal excitation
$\mu_f = 0.4$	coefficients of friction
$F_N = 1.8$	bow pressure
$m_b = 1$	bow mass
$N = 100$	grid size for spatial discretization
$x_c = \frac{10}{N}L = 0.07$	bowing point
$\varepsilon = 0.005$	capture and break-away velocity

Table 7.5: Parameters for the bowed string

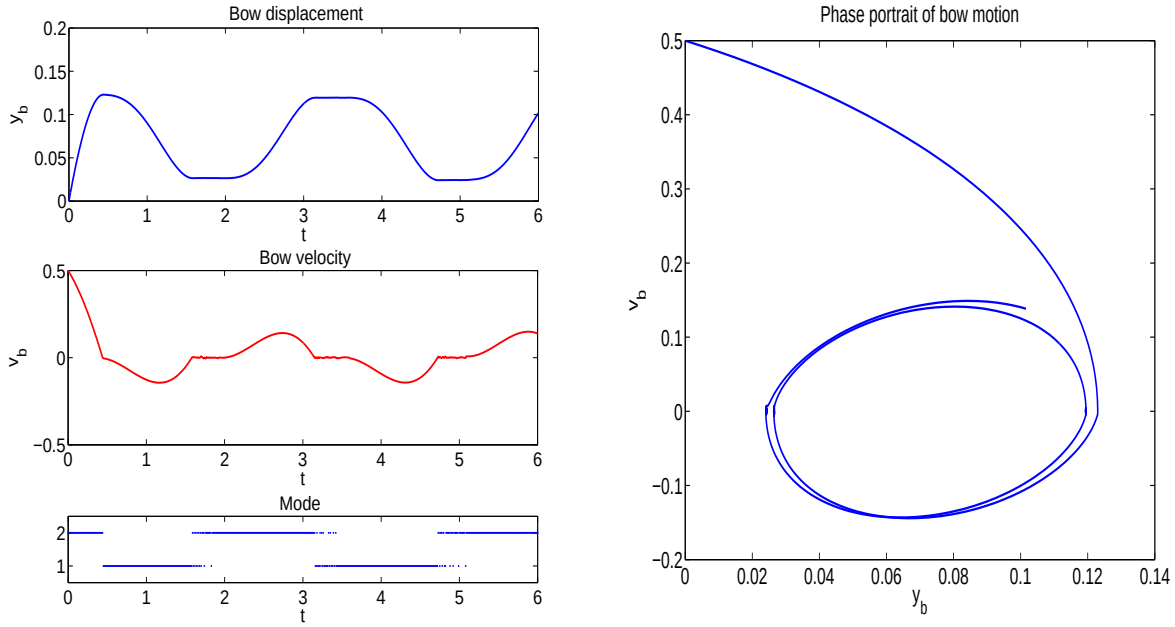

Figure 7.10: Motion of the bow and phase portrait

relative velocity can be observed, i.e., chattering occurs. In the plots of the current mode overlapping regions point to fast alternating changes between the two modes. Further, the computed relative velocity between string and bow is given in Figure 7.12. Again, we can see the oscillations around zero relative velocity. Altogether, the code detects 88 switch points. In the second case the bow-string system (7.8) is solved with GESDA using sliding

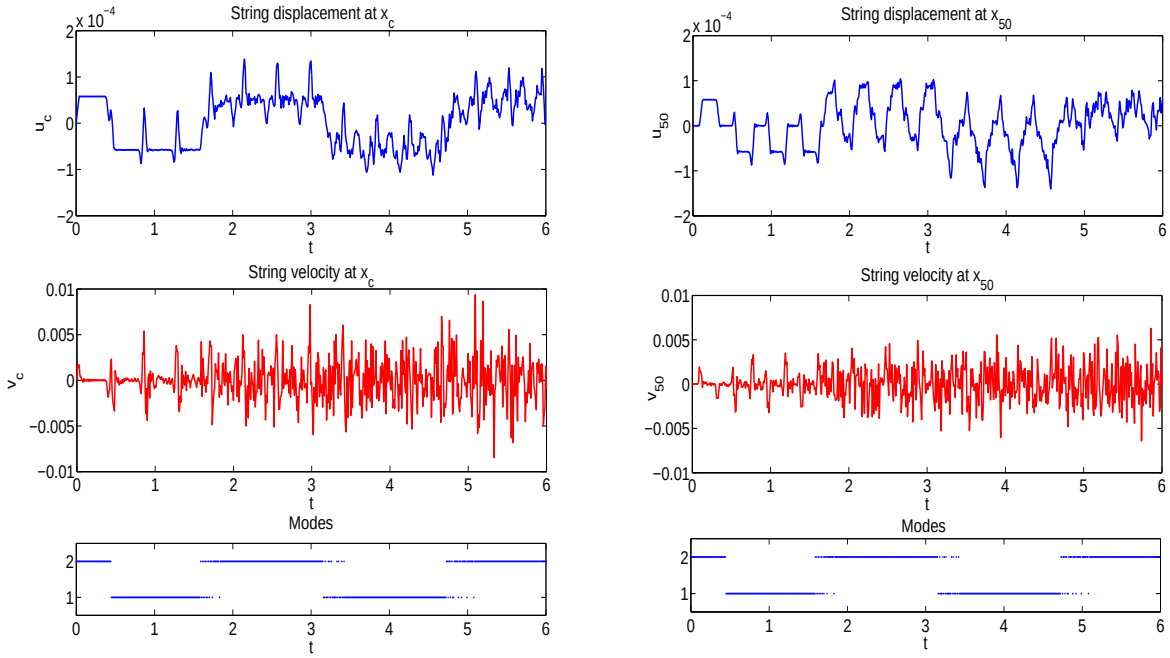

Figure 7.11: Displacement and velocity of the string

mode simulation. The system during sliding motion is defined by

$$\begin{aligned} \dot{y}_b &= v_b, \\ \dot{y}_s &= v_s, \\ m_b \dot{v}_b(t) &= f_k(t) + \lambda_S, \\ \rho \dot{v}_s &= A y_s - G^T \lambda - e_C \lambda_S, \\ 0 &= G y_s, \\ 0 &= v_b - v_c, \end{aligned} \tag{7.10}$$

with characteristic values $\mu = 2, d_\mu = 2N, a_\mu = 7$, and $v_\mu = u_\mu = 0$. Once chattering is detected or the sliding condition is fulfilled during the simulation, the system switches to sliding mode and stays within sliding mode as long as the applied force is less than the friction force, i.e., as long as

$$|f_k(t)| \leq \mu_f \|F_N\|.$$

The relative velocity between string and bow using sliding mode simulation is given in Figure 7.13. We can see that the oscillations around zero relative velocity observed in the previous simulation do not occur. Further, the displacement and the velocity of the bow and the phase portrait of the bow motion are given in Figure 7.14, and the displacement of the string at the bowing point and at an intermediate point on the string are given in Figure 7.15. In the regions of zero relative velocity the system is integrated in the sliding mode and alternating changes between modes are prevented. During the simulation only

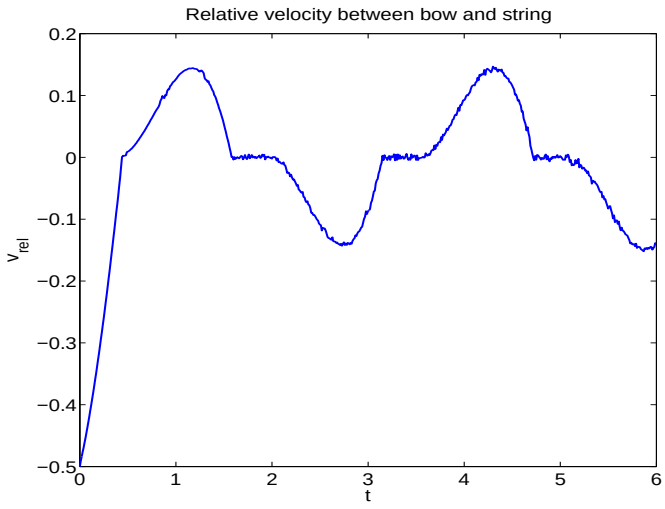

Figure 7.12: Relative velocity between bow and string

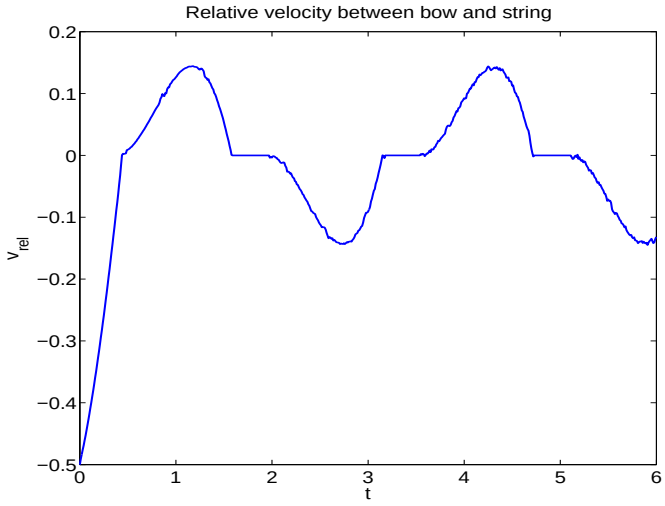

Figure 7.13: Relative velocity between bow and string using sliding mode simulation

8 switch points are detected and the solution of the bow motion exhibit a more regular periodic cycle as in the previous simulation.

Remark 7.1. *The idealized motion of a bowed string was examined experimentally by Helmholtz. These investigations were extended by Raman [122]. Under ideal bowing conditions, the bow and string interaction results in a motion of the string of a regular cycle of stick-slip phases. In the beginning, the shear force causes the string to move under the bow until the part of the string that is connected with the bow has a velocity equal to that of the*

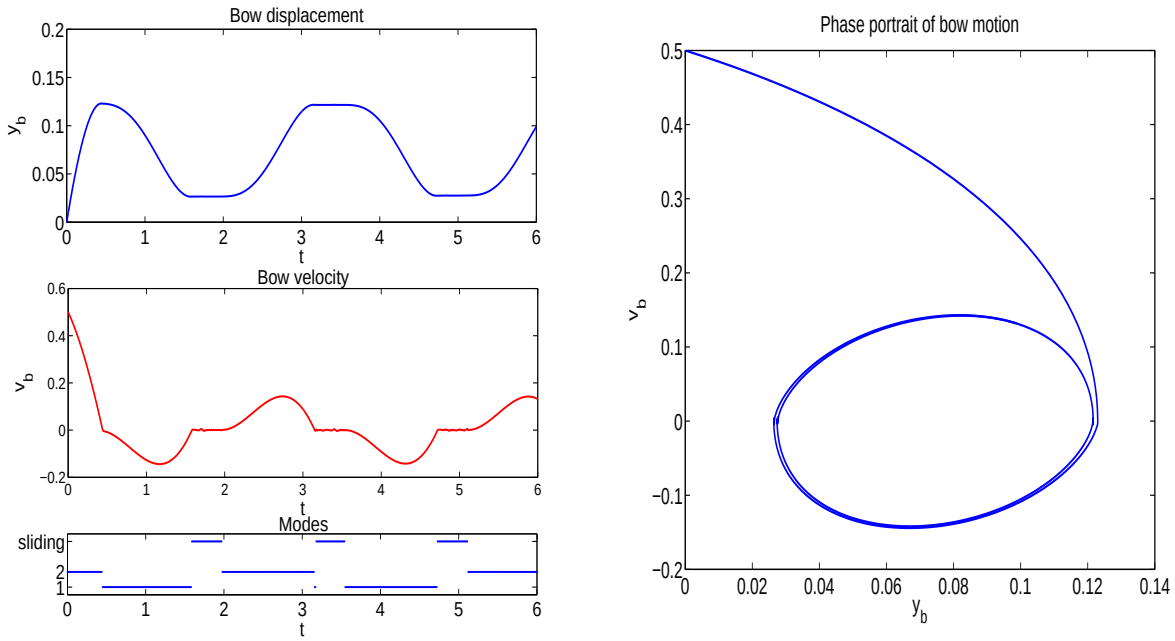

Figure 7.14: Motion of the bow and phase portrait using sliding mode simulation

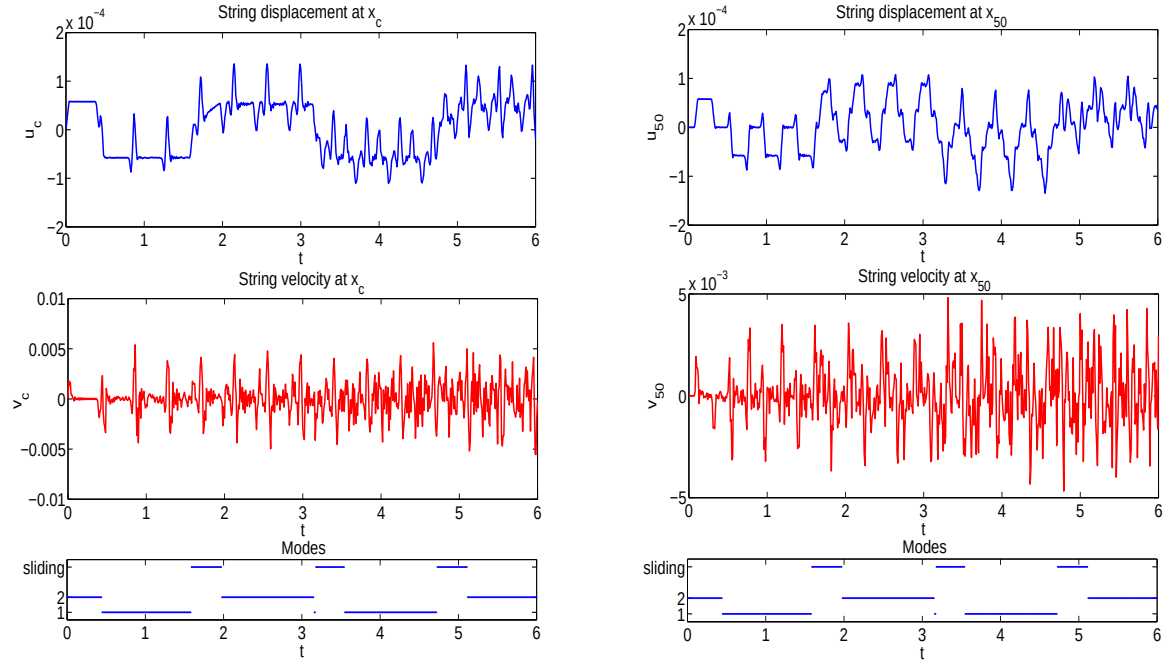

Figure 7.15: Displacement and velocity of the string using sliding mode simulation

bow. Then the string will stick for a while to the bow until the tension force on the string becomes too strong and the string moves back slipping again under the bow, and so on. The beginning and end of the slip phase are triggered by the arrival of a propagating bend or so-called Helmholtz corner*. This idealized motion of an one-dimensional bowed string is also called the* Helmholtz motion*, where the Helmholtz corner is traveling back and forth on the string under an approximately parabolic envelope.*

7.3 Further DAE Solvers

There is a wide number of software packages available that have been designed for the numerical solution of differential-algebraic equations, many of which are specially designed for DAE systems of a certain application class. In the following, we give an overview over available and commonly used numerical solvers for the numerical integration of DAEs. In general, any of the following DAE solvers that enables an interruption of the integration via a user-supplied subroutine can be embedded in the designed mode controller.
One of the first software packages developed for the numerical solution of DAEs is the software package DASSL [17, 115] based on backward differentiation formulas with stepsize and order control that was designed to integrate nonlinear differential-algebraic equations of the form (2.3) of d-index at most one. This code is widely used in numerous applications and works efficiently for nonstiff systems. In addition, the solvers DASPK and DASRT [18], extensions of DASSL, have been designed for large scale DAEs. They are also based on backward differentiation formulas with stepsize and order control, but in contrast to DASSL the numerical solution of the arising linear systems is done via iterative methods. The latest versions of DASPK and DASRT also include a sensitivity analysis and a root finding procedure.
Another solver for DAEs is the code RADAU5 [59] that has been designed for the numerical integration of differential-algebraic initial value problems of the form

$$E\dot{x} = f(t, x), \quad x(t_0) = x_0, \tag{7.11}$$

where E is a constant square and possibly singular matrix. The code RADAU5 is based on the 2-stage implicit Runge-Kutta method of Radau IIa type of order 5 and allows the numerical integration of DAEs of the form (7.11) up to d-index 3. Further solvers available for the solution of (7.11) are the solvers SDIRK4, a diagonally-implicit Runge-Kutta method of order 4, and the solver RODAS, a Rosenbrock method of order 4(3), see [59]. In addition, the package DAESOLVE [121, 125] has been implemented for different classes of quasilinear problems. Note that in contrast to the solvers GELDA and GENDA, all the above listed codes have limitations to the index that they can handle, which is typically a differentiation index of at most three.
There are also many algorithms that have been implemented for special application classes. For the numerical integration of the equations of motion (1.1) of mechanical multibody systems a large number of numerical methods have been developed. First, we want to mention the code ODASSL [34, 44] that uses a backward differences discretization, similar as the

code DASSL, to solve systems of overdetermined DAEs that stem from the equations of motions augmented with the time derivatives of the algebraic constraints. Further, the numerical integrator MEXAX [93] (originally called MEXX) has been designed for the numerical solution of constrained mechanical systems, including dry friction and external dynamics. The code is based on coordinate projection and uses relatively expensive but very accurate extrapolation methods for the integration. Further, the subroutine library MBSpack [136] provides a collection of numerical integration methods based on explicit Runge-Kutta methods for the equations of motion (1.1). Recently, the code GEOMS [137, 138] has been developed for the numerical integration of general equations of motion of multibody systems that also allows redundant constraints and takes into account possibly existing information concerning solution invariants. The code GEOMS combines a stabilization technique with an implicit Runge-Kutta scheme (a Radau IIa method of order 5) as discretization of a strangeness-free formulation of the equations of motion. Further, commercial software packages for the dynamic analysis of mechanical systems with the multibody system method are available in packages like SIMPACK [127, 130], ADAMS [128, 130] or DYMOLA [114]. They comprise the computation of the symbolic equations of motion or the evaluation of the residuals of the model equations and the simulation of the dynamical behavior.
Also, for the numerical simulation of electrical circuits a number of software packages has been developed. The most prominent code is the general purpose electronic circuit simulator SPICE [109]. Further, the circuit simulator TITAN [38] has been developed. Usually, circuit simulation programs take a text netlist describing the circuit elements (transistors, resistors, capacitors, etc.) and their connections, and translate this description into equations to be solved. The general equations produced are nonlinear differential-algebraic equations which are solved using implicit integration methods, Newton's method and sparse matrix techniques.

For the numerical simulation of switched systems not only the numerical integration but also the modeling has to be taken into account. There are some, mostly commercial, software packages which combine the modeling with the numerical simulation and also allow the treatment of switched systems. The modeling environment and simulator ABACUSS II [29], that was developed for chemical engineering applications, supports hybrid DAE models. It uses the software library DAEPACK [143] that uses automatic differentiation to determine the analytical derivatives of the original model to reformulate DAE systems as ODEs. The simulator ABACUSS II incorporates hybrid discrete/continuous dynamic simulation with state event location, but it cannot solve DAEs of d-index higher that one and does not treat chattering behavior. Further, there is the software package DYMOLA [114], a modeling and simulation environment for integrated and complex systems within a wide field of applications, based on the object-oriented modeling language *Modelica* [100], that allows the convenient modeling of complex physical systems, e.g., systems containing mechanical, electrical, hydraulic, thermal or control components. Within DYMOLA the differential-algebraic systems are converted symbolically to state-space form. Graph-theoretical algorithms are used to determine which variable to solve for in each equation

and to find minimal systems of equations that have to be solved simultaneously. The equations are then solved symbolically. DYMOLA also supports instantaneous and discontinuous equations. Next, we want to mention HYBRSIM [108], a modeling and simulation environment for hybrid systems using hybrid bond graph representations and allowing DAE models. HYBRSIM performs event detection and location based on a bisectional search, handles runtime causality changes, including derivative causality, and performs consistent reinitialization, but it does not support the handling of chattering behavior. The continuous behavior is handled by a simple Forward Euler integration scheme. Further, OmSim [3], based on the object-oriented modeling language *Omola* [3], is an environment for the modeling and simulation of continuous time and discrete event dynamics. Omola supports behavioral descriptions in terms of differential-algebraic equations. OmSim analyzes and manipulates the model including elimination of algebraic variables from the dynamic problem when feasible, to reduce the index of the DAE, that is then solved using the solvers DASRT and RADAU5. OmSim does not support reinitialization during the integration or the treatment of chattering. Finally, we want to mention MOSILAB [112], a simulation tool for the modeling and simulation of complex technical systems also based on the modeling language *Modelica*, that also supports hybrid models, and SIMULINK the block diagram based modeling and simulation environment of MathWorks. The treatment of sliding modes is currently not supported in available simulators, see also [104], but a sliding mode simulation algorithm has been proposed e.g. in [106]. For an overview of simulation packages for hybrid systems see also [104].

Remark 7.2. *In the numerical simulation of DAEs essentially two approaches are taken. The first approach [3, 114, 143] uses a reformulation of the system that solves all the constraint equations (e.g., via computer algebra packages) and generates an ordinary differential equation for the dynamics, for which standard simulation methods are available. There are several disadvantages of this approach. First of all, the methods are expensive and the output of the manipulations are often huge formulas that have doubtful numerical properties. Second, all the constraints are solved in finite precision arithmetic such that the approximate solution deviates from the constraints leading to physically meaningless results. Furthermore, the resulting variables are usually non-physical and thus the results are difficult to interprete. The second approach [34, 82, 137] that we have used in this thesis uses a reformulation as strangeness-free or low index DAE, while preserving the constraints and keeping the original variables and their physical meaning. This makes initialization easy and avoids that the numerical solution drifts off from the solution manifold. On the other hand, this reformulation is often costly as well and it is frequently necessary to introduce further variables in this process, which increases the system size. In circuit simulation and multibody dynamics [7, 8, 36, 137] network based methods have been derived that yields efficient ways to do this reformulation.*

7.4 Future Work

The solver GESDA enables the solution of general hybrid differential-algebraic systems consisting of linear or nonlinear DAEs of arbitrary high index in each mode. Furthermore, even over- and underdetermined linear differential-algebraic systems can be solved within GESDA using the DAE solver GELDA. The solver GESDA provides wide freedom in the definition of the transition conditions as well as in the transition functions. Further, the consistent reinitialization provides the possibility to fix initial values for differential components while consistent values for algebraic components are computed, thus providing a solution as smooth as possible. A particular feature of the solver GESDA is the use of sliding mode simulation that allows an efficient treatment of chattering behavior during the numerical simulation and thus provides the possibility to reduce the computational effort drastically. Nevertheless, not all phenomena in hybrid system behavior can be treated with GESDA. In particular, the possibility that multiple transitions can occur at the same time should be included in the solver. In this case, we need to find ways to decide which mode change should be performed. Further, we have to investigate a stepsize selection strategy to make sure that the numerical solution is adapted to the behavior of the switching functions such that in particular no event is missed during the numerical integration. So far, we only include the possibility to restrict the maximal allowed stepsize in order to avoid stepping over regions. In addition, further DAE solvers can be embedded in the mode controller, e.g. solvers for the equations of motion of multibody systems.

Chapter 8

Conclusion

The numerical simulation of complex dynamical systems described by differential-algebraic equations plays an important role in technical applications. In this thesis, we have considered the analysis as well as the numerical solution of structured and switched differential-algebraic systems. Basically, we have focused on three topics. The achievements of each topic are summarized in the following.

First, we have considered higher order differential-algebraic systems, in particular second order DAEs that arise frequently in technical applications. It is known that the classical order reduction, that is usually used to transform ordinary differential equations of higher order into first order systems, leads to a number of difficulties when applied to DAEs, as e.g. an increase in the index of the DAE, see [102]. In [102, 135] condensed forms for linear higher order DAEs and a stepwise index reduction procedure based on global equivalence transformations have been derived that allow to transform a linear higher order DAE into a so-called strangeness-free higher order system. From this strangeness-free system we can read off which higher order derivatives of variables can be replaced by new variables to transform the system into a first order system without increasing the index. Unfortunately, this algebraic approach is not numerically computable as it involves the derivatives of computed transformation matrices. In this thesis, we have derived a new index reduction method for second order DAEs based on a derivative array approach using the condensed forms given in [102]. At first, a characterization of the relationship between the characteristic values in each step of the index reduction procedure in terms of ranks of submatrices of the corresponding matrix triple is given in Lemma 3.15. Here, we use a slightly different index reduction procedure compared to [102, 135] concerning the sequence of differentiations of equations. This difference also effects the definition of the strangeness index, see Remark 3.10. Next, we use a derivative array approach to derive an index reduction method that allows to transform the second order system locally at each point into a strangeness-free system in a numerical feasible way. In Theorem 3.18 it is shown that the local characteristic quantities of the inflated system, given by the derivatives of the original second order system, are invariant under global equivalence transformations of the original system. In Theorem 3.19 we have shown how the local quantities of the inflated system are related to the global characteristic quantities of the original system for differential-algebraic systems of strangeness index $\mu \leq 2$. Using these results we can determine a number of projections (Theorem 3.21) that allow to extract a reduced second order system from the enlarged derivative array system locally at each point. This reduced second order system can be shown to be strangeness-free with the same characteristic

values as the strangeness-free system obtained by the stepwise index reduction procedure and locally with the same solution as the original system, see Theorem 3.22. In Section 3.2, the derived index reduction method is extended to nonlinear second order system using linearization along solution trajectories. In particular, linearization of the nonlinear second order system and differentiation of the system commute, see Theorem 3.25, such that the results obtained for linear second order systems can also be applied in the nonlinear case. Hypothesis 3.26 is formulated that is invariant under equivalence transformations of the original system and that allows to extract locally a reduced nonlinear second order system that is strangeness-free and has locally the same solution as the original second order system, see Theorems 3.29 and 3.30. Further, the approach allows the formulation of a numerically computable trimmed first order system given in Lemma 3.35 that is also strangeness-free. For linear time-invariant second order systems this trimmed first order formulation also allows an explicit solution representation in terms of the original matrix-valued functions. Theorems 3.19 and 3.21 have been proven only for the case that the strangeness index is $\mu \leq 2$. Nevertheless, the assertions are expected to be also valid for higher index problems. If a condensed form similar to the form given in [82, Theorem 3.21] could be proved, then we might be able to prove Theorems 3.19 and 3.21 also for the case that $\mu > 2$. With respect to this, Hypothesis 3.26 in the nonlinear case has been formulated for arbitrary index μ. In general, all results obtained in Chapter 3 can also be extended to arbitrary high order systems in an analogous way, see also Remark 3.24.

The second part of this thesis involves structure preservation for symmetric and self-adjoint linear differential-algebraic systems that arise e.g. in mechanical systems or in optimal control problems. Since the structure of a system describes its physical properties, the structure should be preserved during the numerical solution. Otherwise, physical meaningless solutions may result as rounding errors obscure the physical properties. In Theorem 4.10 a structure preserving condensed form for symmetric pairs of matrix-valued functions has been derived. It has emerged that for symmetric time-variant systems strong assumptions on the coefficient matrices, in particular on the kernel of $E(t)$ given in Assumption 4.8, are needed in order to be able to obtain a structure preserving condensed form using global congruence transformations. It does not seem to be possible to lessen these assumptions while preserving the symmetric structure during the transformation into condensed form using global congruence transformations. Nevertheless, it has emerged that the self-adjoint structure is better suited for studying structure preserving condensed forms, since global congruence transformations preserve the self-adjoint structure of pairs of matrix-valued functions without further assumptions. A structure preserving condensed form for self-adjoint pairs of matrix-valued functions is given in Theorem 4.25. We have seen that in both cases, i.e., for symmetric and self-adjoint linear systems, even under Assumption 4.8, a structure preserving strangeness-free formulation only exists if the strangeness-index of the system is lower or equal to 1. Furthermore, an index reduction methods based on minimal extension of the original system has been derived for self-adjoint systems of strangeness index $\mu = 1$ that allows the formulation of a structure preserving strangeness-free system and therefore enables a structure preserving numerical treatment of linear self-adjoint

DAEs. This approach is also applicable in the case of linear symmetric differential-algebraic systems of strangeness index $\mu = 1$ without the need of Assumption 4.8.

The third part of this thesis involves differential-algebraic systems that switch between different modes of operation. As a framework for the analysis and numerical solution we choose the formulation of switched DAE systems as so-called hybrid systems as given in Definition 5.3, i.e., a combination of DAEs describing the continuous dynamics of the system and transition conditions describing the discrete mode changes. This formulation allows to describe a huge number of possible system configurations and a wide scope of definitions of switching conditions and therefore enables the description of different phenomena in hybrid system behavior. An index reduction for hybrid systems can be realized in the same way as for general nonlinear DAEs separately in each mode. Since the index reduction is done locally at each time point, a reduced system can be extracted in each mode independently of the previous or future system behavior. Further, a strangeness-free hybrid system with the same solution as the original hybrid system can be obtained, see Theorem 5.9. Next, the existence and uniqueness of solutions of hybrid systems after mode switching are considered. Under the assumption that the differential-algebraic system in each mode is solvable and has a well-defined strangeness index, conditions for the existence and uniqueness of continuous solutions of a hybrid system are given in Theorem 5.18. Further, conditions for the existence and uniqueness of generalized solutions that also allow jumps or discontinuities in the state vector at the mode switches are given in Theorem 5.21. In general, in order to ensure the existence of a solution, the transition functions have to guarantee that the initial values after mode switching are consistent with the DAE in the new mode. The chosen concept of impulsive smooth distributions allows the treatment of discontinuities, but it requires that the coefficient matrices are infinitely often continuously differentiable which is not always guaranteed. Another distributional approach is presented in [144] that also allows discontinuities in the coefficients. For nonlinear hybrid systems conditions for the solvability are given in Theorem 5.24.
One of the difficulties in the numerical solution of DAEs is to compute consistent initial values before starting the integration. In contrast to the numerical solution of standard DAEs, for switched DAE systems consistent initial values are needed not only at the initial time t_0 but also at every switch point τ_i. Consistent initial values for nonlinear DAEs can be obtained by solving a nonlinear system of equations arising from the derivative arrays. In addition, we fix certain components of the initial value vector during the solution of the nonlinear system in order to find a reasonable continuation of the previous solution, i.e., if possible, differential variables are continued smoothly over a switch point, whereas algebraic variables are chosen consistently with the DAE in the current mode.
Another problem in the numerical simulation of switched systems is the occurrence of numerical chattering, i.e., fast changing between different modes of operation. The numerical solution of a switched system exhibiting chattering behavior requires high computational costs due to repeated mode switchings and reinitializations and since small stepsizes are required to restart the integration after each mode change. In the worst case, the numerical integration breaks down as it does not proceed in time, but chatters between modes. One

possibility to prevent numerical chattering is the introduction of hysteresis such that the integration in each mode is done in an interval of a length bounded from below. Another way to avoid oscillations around switching surfaces is to approximate the system dynamics along the switching surface in the sliding region. An additional mode, the so-called *sliding mode*, can be inserted into the system that represents the dynamics during sliding. For discontinuous and switched differential-algebraic systems the behavior of the system during sliding is approximated via a convex combination such that the solution trajectory stays on the switching surface along which chattering occurs. The corresponding DAE in sliding motion, that is based on the reduced systems in each mode, is of strangeness index $\mu = 1$, see Theorem 5.29. In this way, high computational costs due to fast oscillations along the switching surface can be prevented.

Further, we have considered hybrid control problems for linear time-invariant descriptor systems. In general, classical control concepts for DAEs can be applied to hybrid systems locally in every mode, but some attention has to be paid to the transitions of the system state between modes. Choosing a control u^l in some mode l influences the transition conditions and mode changes of the hybrid system as well as the points in time at with switching occur. Thus, changes in the controls lead to a huge number of possible hybrid mode trajectories and hybrid time trajectories and mode transitions often cause nonsmoothness of the solution which complicates the minimization problem used in the optimal control theory. Under the assumption that the descriptor system in each mode is controllable and observable within the reachable set, necessary conditions for the controllability and observability of a linear hybrid system are given in Theorem 5.42. Further, necessary conditions for optimality of a hybrid optimal control problem under some transversality conditions are given in Theorem 5.46.

Finally, a hybrid mode controller has been implemented that enables the numerical solution of general nonlinear switched DAE systems. For the numerical integration of the DAEs in each mode the two DAE solvers GELDA and GENDA have been embedded. During the numerical integration, the solver checks if a transition condition is satisfied, and if so, the control is returned to the mode controller which organizes the mode switching and restarts of the integration method as well as the consistent reinitialization. The switch points are determined as the roots of the switching functions using a modified secant method. Further, the state of the system at a switch point is determined by interpolation using the interpolation polynomials given by a Runge-Kutta or BDF method, respectively. In particular, the code detects if chattering occurs during the simulation and enables the use of sliding mode simulation. Concluding, we have demonstrated the applicability of the implemented solver at several examples given in Section 7.2.

Appendix

Documentation of the code GESDA

```
      SUBROUTINE DGESDA(INFO, MATSUB, FUN, DFUN, USCAL, UINTER, TRANSF,
     $                  MDCHNG, GFUN, MXTRAN, TRAN, SOLVER, METHOD,
     $                  M, N, NMAX, T, TOUT, TS, LTS, NSP, X, XPRIME,
     $                  CVAL, IPAR, LIPAR, RPAR, IFIX, IDIFCO, SCALC,
     $                  SCALR, RTOL, ATOL, MDHIST, IWORK, LIW, RWORK,
     $                  LRW, STATS, DWORK, ISMTH, IWARN, IERR)
C
C     PURPOSE
C
C     DGESDA solves nonlinear hybrid DAEs of the form
C
C         F_1(t,x_1(t),x_1'(t)) = 0,
C         ...
C         F_n(t,x_2(t),x_2'(t)) = 0,   x(t0) = x0,
C
C     consisting of n systems of nonlinear DAEs with transition conditions
C     between the systems, or linear hybrid DAEs of the form
C
C        E_1(t)x_1'=A_1(t)x_1 = f_1(t),
C        ...
C        E_n(t)x_n'=A_n(t)x_n = f_n(t),   x(t0) = x0,
C
C     consisting of n systems of linear DAEs, for x in a specified range of
C     the independent variable t.
C
C     ARGUMENT LIST
C
C       USER-SUPPLIED SUBROUTINES
C
C       MATSUB - User supplied SUBROUTINE.
C                This is a subroutine which the user provides to define the
C                matrices E_l(t) and A_l(t) and the right hand sides f_l(t) for
C                each mode l as well as their derivatives. It has the form
C
C                SUBROUTINE MATSUB(IMAT, M, N, T, IDIF, W, LDW, IPAR, RPAR, IERR).
C
C                The subroutine takes as input the number of equations M in the
C                current mode, the number of unknowns N in the current mode,
C                the time T and the integer parameters IMAT and IDIF. Further, the
C                integer and double precision arrays IPAR and RPAR that can be used
C                for communication between the calling program and the MATSUB
C                subroutine. Note that IPAR(1) gives the current mode of the hybrid
```

```
C               system. As output W, the subroutine produces the IDIF-th  derivative
C               of E_l(t) at time T if IMAT=1, the IDIF-th derivative of A_l(t) at
C               time T if IMAT=2, or the IDIF-th derivative of f_l(t) at time T if
C               IMAT=3 for the current mode l defined in IPAR(1).
C               In the calling program, MATSUB must be declared as external.
C
C       FUN     - User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the
C               nonlinear differential-algebraic equations F_l and its derivatives for
C               each mode l. It has the form
C
C               SUBROUTINE FUN (T, IDIF, F, X, IPAR, RPAR, IERR).
C
C               The subroutine takes as input the time T, the vector X containing an
C               approximation to the solution x and its first (MXINDX+2) derivatives,
C               and the integer parameter IDIF. Further, the integer and double
C               precision arrays IPAR and RPAR that can be used for communication
C               between the calling program and the FUN subroutine. Note that IPAR(1)
C               gives the current mode of the hybrid system. As output, the subroutine
C               produces the IDIF-th derivative of the DAE in the current mode at time
C               T and state X in the first M elements of the  1-dimensional array F.
C               In the calling program, FUN must be declared as external.
C
C       DFUN    - User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the
C               Jacobian of F_l and its derivatives for each mode l. It is of the form
C
C               SUBROUTINE DFUN (T, IDIF, JAC, LDJAC, X, IPAR, RPAR, IERR).
C
C               The subroutine takes as input the time T, the vector X containing an
C               approximation to the solution x and its first (MXINDX+2) derivatives,
C               and the integer parameter IDIF. Further, the integer and double
C               precision arrays IPAR and RPAR that can be used for communication
C               between the calling program and the DFUN subroutine. Note that IPAR(1)
C               gives the current mode of the system. As output, the subroutine
C               produces all partial derivatives of the IDIF-th derivative of the DAE
C               in the current mode IPAR(1) with respect to all entries of X in the
C               first M rows and (MXINDX+2)*N columns of the 2-dimensional array JAC.
C
C       USCAL   - User supplied scaling SUBROUTINE of the form
C
C               SUBROUTINE USCAL( MQ, NQ, A, LDA, SCALC, SCALR, IERR).
C
C               (See Documentation of GENDA.)
C
C       UINTER - User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the transition
C               conditions of the hybrid system. It is of the form
C
C               SUBROUTINE UINTER(T,X,XPRIME,RPAR,IPAR,ATOL,IRTRN).
C
```

```
C               The subroutine takes as input the time T, the vector X
C               containing an approximation to the solution x, the vector XPRIME
C               containing an approximation to the derivative x', the integer and
C               double precision arrays IPAR and RPAR used for communication between
C               the calling program and the UINTER subroutine, and the absolute error
C               tolerance ATOL. IPAR(1) defines the current mode. The subroutine is
C               called after each intermediate step in the integration by the solver
C               GELDA or GENDA to check if a transition conditions is satisfied.
C               As output the subroutine has to provide an integer value IRTRN that
C               indicates if a transition condition is satisfied. The integration
C               process is stopped if IRTRN =-1 and the control is return to GESDA.
C               If IRTRN=0 the integration is continued.
C
C       TRANSF - User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the transition
C               functions of the form
C
C               SUBROUTINE TRANSF(XNEW,XOLD,N,MODNEW,MODOLD,IERR).
C
C               The subroutine takes as input the vector XOLD containing an
C               approximation to the solution in the old mode, the integer N
C               containing the size of X, and the integers MODNEW and MODOLD defining
C               the successor and predecessor modes. As output the subroutine provides
C               the vector XNEW containing the solution after transfer to the new
C               mode. If an error occurs IERR should be < 0.
C
C       MDCHNG - User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the mode
C               allocation functions of the form
C
C               SUBROUTINE MDCHNG(MODE, J, NMODE, MXTRAN, NEWMOD, IERR).
C
C               As input the subroutine takes the integer MODE defining the current
C               mode, the integer J providing the active transition, the integer NMODE
C               providing the number of modes, the integer MXTRAN providing the
C               maximal number oftransitions. As output the subroutine provides the
C               integer NEWMOD defining the new mode. If an error occurs
C               IERR should be < 0.
C
C       GFUN  -  User supplied SUBROUTINE.
C               This is a subroutine which the user provides to define the switching
C               functions. It is of the form
C
C               SUBROUTINE GFUN(T,IDIF,  X,  XPRIME, GT, NRTFN, IPAR, RPAR, IERR).
C
C               The subroutine takes as input the time T, and integer IDIF, the
C               vector X containing an approximation to the solution x, the vector
C               XPRIME containing an approximation to the derivative x', the integer
C               and double precision arrays IPAR and RPAR used for communication
C               between the calling program and the GFUN subroutine, and the integer
C               NRTFN providing the number of switching functions in the current mode
```

```
C               defined by IPAR(1). As output the subroutine provides the array GT
C               containing the values of the switching functions in the current mode
C               at time T and state X. If an error occurs IERR should be < 0.
C
C
C     ARGUMENTS IN
C
C       MAXTRAN- INTEGER.
C               The maximal number of transition.
C
C         TRAN - INTEGER array of DIMENSION NMODE.
C               The number of possible transitions for each mode.
C
C       SOLVER - INTEGER array of DIMENSION NMODE.
C               Indicates which solver should be used in each mode as follows:
C               SOLVER(i)=1 the code uses the solver GENDA.
C               SOLVER(i)=2 the code uses the solver GELDA.
C
C       METHOD - INTEGER array of DIMENSION NMODE.
C               Indicates which integration method should be used in each mode as
C               follows:
C               METHOD(i)=1 the code uses the BDF solver.
C               METHOD(i)=2 the code uses the Runge-Kutta solver.
C
C         M -    INTEGER array of DIMENSION NMODE.
C               The number of equations in the DAE system for each mode.
C               M(i) .GE. 1.
C
C         N -    INTEGER array of DIMENSION NMODE.
C               The number of components of x for each mode.
C               N(i) = M(i) in this version of DGENDA.
C
C         NMAX-  INTEGER.
C               The maximal size of the vector X given by MAX[(MXINDX+2)*N(i)].
C
C         T -    DOUBLE PRECISION.
C               The initial point of the integration.
C               NOTE that this scalar is overwritten.
C
C         TOUT - DOUBLE PRECISION.
C               The point at which a solution is desired. Integration
C               either forward in T (TOUT > T) or backward in T (TOUT < T)
C               is allowed.
C
C         X -    DOUBLE PRECISION array of DIMENSION (NMAX).
C               If INFO(11)=0, this array can contain a guess for the
C               initial value. A consistent initial value close (in the
C               least square sense) to this guess is then computed. If no
C               guess is available, set all elements of X to zero.
C               If INFO(11)=1, this array must contain consistent initial
C               values of the NMAX solution components at the initial point
```

```
C               T. NOTE that this array is overwritten.
C
C         CVAL - INTEGER array of DIMENSION (5,NMODE).
C               Contains the characteristic values of the DAE in each mode:
C                    CVAL(1,i) contains the strangeness index MU.
C                    CVAL(2,i) contains the number DMU of differential
C                              components.
C                    CVAL(3,i) contains the number AMU of algebraic components.
C                    CVAL(4,i) contains the number UMU of undetermined
C                              components.
C                    CVAL(5,i) contains the number VMU of redundancies.
C
C         IPAR - INTEGER array of DIMENSION (LIPAR).
C               This integer array can be used for communication between
C               the calling program and the user-provided subroutines.
C               IPAR needs to contain the following parameters:
C                    IPAR(1) - initial mode,
C                    IPAR(2) - number of modes,
C                    IPAR(3) - sliding mode (if dot defined set =0).
C
C        LIPAR - The length of IPAR.
C
C         RPAR - DOUBLE PRECISION array of DIMENSION (*).
C               This real array can be used for communication between the
C               calling program and the user-provided subroutines.
C
C         IFIX - INTEGER array of DIMENSION (NMAX).
C
C       IDIFCO - INTEGER array of DIMENSION (NMODE*NMAX).
C               Contains information of differential components.
C               By setting IDIFCO(i)=1 the corresponding variables is kept
C               fixed during the computations of consistent initial values.
C
C         RTOL - DOUBLE PRECISION array of DIMENSION (*)
C               The relative error tolerances which the user provides to
C               indicate how accurately he wishes the solution to be
C               computed. (See documentation of GELD/GENDA.)
C
C         ATOL - DOUBLE PRECISION array of DIMENSION (*)
C               The absolute error tolerances which the user provides.
C               (See documentation of GELD/GENDA.)
C
C       ARGUMENTS OUT
C
C          T -   DOUBLE PRECISION.
C               The solution was successfully advanced to the output value
C               of T.
C
C          TS-   DOUBLE PRECISION array of DIMENSION (LTS).
C               Contains the detected switch points.
C
```

```
C         LTS-   The length of TS.
C
C         NSP-   INTEGER.
C                The number of detected switch points.
C
C          X -   DOUBLE PRECISION array of DIMENSION (NMAX).
C                Contains the computed solution approximation at T.
C
C     XPRIME -   DOUBLE PRECISION array of DIMENSION (NMAX).
C                Contains the computed first derivative of the solution
C                approximation at T.
C
C     MDHIST -   INTEGER array of DIMENSION (LTS+1).
C                Contains the mode history.
C
C      IWORK -   INTEGER array of DIMENSION at least (LIW).
C                (See documentation of GELD/GENDA.)
C
C        LIW -   The length of IWORK. (See documentation of GELD/GENDA.)
C
C      RWORK -   DOUBLE PRECISION array of DIMENSION at least (LRW).
C
C        LRW -   The length of RWORK. (See documentation of GELD/GENDA.)
C
C      STATS -   INTEGER array of DIMENSION(5).
C                The array STATS contains some statistic of the integration:
C                    STATS(1) = number of integration steps so far
C                    STATS(2) = number of calls to MATSUB or FUN so far
C                    STATS(3) = number of factorizations so far
C                    STATS(4) = number of error test failures
C                    STATS(5) = number of convergence test failures
C
C      DWORK -   DOUBLE PRECISION array of DIMENSION (4+NNMAX)
C                providing the workspace used in GESDA.
C                (See documentation of the INFO array).
C
C                 DWORK(1) contains the tolerance TTOL used in the root
C                          finding procedure DRTFND.
C                 DWORK(2) contains chattering tolerance.
C                 DWORK(3) contains tolerance delta in the approximation of
C                          the sliding condition.
C                 DWORK(4) contains maximal allowed number of immediate transitions.
C                 DWORK(5) contains the interpolated and transfered solution at
C                          the last detected switch point.
C
C       ISMTH -   INTEGER.
C                Integer indicating if nonsmooth transitions occur.
C                ISMTH =  0: only smooth transitions occur
C                ISMTH = -1: no smooth transition due to transition function
C                ISMTH = -2: no smooth transition due to consistent initialization
C
```

```
C       IWARN -  INTEGER.
C                Integer containing warnings:
C                IWARN >=0: the strangeness index has changed and IWARN contains
C                           the new value of the strangeness index.
C                IWARN =-1: a high number of switch points occur, possible chattering!
C                IWARN =-2: an immediate mode change occur.
C
C        IERR -  INTEGER.
C                Error indicator. Unless the code detects an error, IERR
C                contains a positive value on exit.
C                (See also documentation of GELDA/GENDA.)
C
C        INFO -  INTEGER array of DIMENSION (27)
C                The basic task of the code is to solve the system from T
C                to TOUT and return an answer at TOUT. INFO is an integer
C                array which is used to communicate exactly how the user
C                wants this task to be carried out. The simplest use of the
C                code corresponds to setting all entries of INFO to 0.
C
C         INFO(1)-INFO(22) are used in the solvers GELDA, GENDA,
C                          see documentation of GELDA and GENDA.
C            INFO(7) The user can specify a maximum (absolute value of)
C                    stepsize, so that the code will avoid passing over regions.
C                    DO YOU WANT THE CODE TO USE THE DEFAULT MAXIMUM
C                    STEPSIZE HMAX = 1.0D0 ...
C                    Yes - Set INFO(7)=0
C                    No  - Set INFO(7)=1
C                    and define HMAX by setting RWORK(2)=HMAX
C
C           INFO(23) This parameter determines if a routine UINTER is to be
C                    called after each successful integration step, i.e. if
C                    a switched system is to be solved. UINTER can be used
C                    for checking thresholds or intermediate outputs, etc.
C                    DO YOU WANT CALLS TO UINTER AFTER EACH SUCCESSFUL STEP
C                    Yes - SET INFO(23) = 0
C                    No  - SET INFO(23) = 1
C
C           INFO(24) The root finding procedure DRTFND determines the switch
C                    point within a tolerance of TTOL. By default the TTOL
C                    is computed by  TTOL = (|THI|+|H|)*UROUND*100
C                    where UROUND is the unit roundof.
C                    DO YOU WANT TO USE THE DEFAULT TOLERANCE ...
C                    Yes - Set INFO(24)=0
C                    No  - Set INFO(24)=1
C                          and define TTOL by setting DWORK(1)=TTOL.
C
C          INFO(25) The codes detect chattering behavior. by default the
C                   chattering tolerance TOLC is set to TOLC=10^{-5} and
C                   the parameter DELTA used for the approximation of the
C                   sliding condition is DELTA=10^{-5}.
C                   DO YOU WANT TO USE THE DEFAULT TOLERANCES ...
```

```
C                  Yes - Set INFO(25)=0
C                  No  - Set INFO(25)=1
C                        and define TOLC and DELTA by setting DWORK(2)=TOLC
C                        and DWORK(3)=DELTA
C
C          INFO(26) The codes enables sliding mode simulation. The detection
C                   of chattering can de disabled.
C                   DO YOU WANT THE CODE TO DETECT CHATTERING ...
C                   Yes - Set INFO(26)=0
C                   No  - Set INFO(26)=1
C
C          INFO(27) The codes allows a maximal number of immediate mode
C                   changes. The default value for the maximal allowed number
C                   of immediate transitions is given by MAXCGN=100.
C                   DO YOU WANT TO USE THE DEFAULT VALUE ...
C                   Yes - Set INFO(27)=0
C                   No  - Set INFO(27)=1
C                         and define MAXCGN by setting DWORK(4)=MAXCGN
C
C     ERRORS DETECTED BY THE ROUTINE (See also documentation of GELDA/GENDA.)
C
C     IERR = -201 : Some element of INFO vector is not zero or one.
C     IERR = -202 : NMODE .LE. 0.
C     IERR = -203 : Current mode exceeds NMODE.
C     IERR = -204 : N(I).LE. 0  OR  M(I).LE. 0.
C     IERR = -205 : Mode chattering occur.
C     IERR = -206 : METHOD(I)>1 not possible for the nonlinear case.
C     IERR = -207 : Wrong maximal transition.
C     IERR = -208 : TRAN(I) has an invalid value.
C     IERR = -209 : TTOL has an invalid value.
C     IERR = -211 : An error occurred in the subroutine DRTFND.
C     IERR = -212 : IERR in DMCHNG has a negative value.
C     IERR = -213 : IERR in DTRANS has a negative value.
C     IERR = -215 : Chosen solver not implemented yet.
C     IERR = -216 : Wrong transition function detected.
C     IERR = -217 : Number of switch points exceeds LTS.
C     IERR = -218 : Wrong arguments in DWORK array.
C     IERR = -219 : No root found by DRTFND, but an event was detected by UINTER.
C     IERR = -220 : Too many immediate mode changes occur.
C     IERR = -221 : An error occurred in the subroutine DCKCON
C     IERR = -222 : Invalid value in IDIFCO.
C     IERR = -223 : An immediate mode change occur and TRAN(i).GT.1.
C
C     VERSION      : July 1, 2008
C
C     AUTHOR       : L. Wunderlich (Technische Universitaet Berlin, Germany)
C                    wunder@math.tu-berlin.de
C
C     REFERENCES   : L. Wunderlich. Analysis and Numerical Solution of Structured
C                    and Switched Differential-Algebraic Systems. PhD Thesis,
C                    TU Berlin, Institut fuer Mathematik, 2008.
```

List of Figures

List of Tables

Bibliography

[1] J. Agrawal, K. M. Moudgalya, and A. K. Pani, *Sliding motion of discontinuous dynamical systems described by semi-implicit index one differential algebraic equations*, Chemical Engineering Science **61**, no. 14 (2006), pp. 4722–4731.

[2] R. Alur, C. Courcoubetis, T. A. Henzinger, and P. Ho, *Hybrid Systems I*, Lecture Notes in Computer Science 736, Springer, 1993, ch. Hybrid Automata: An Algorithmic Approach to the Specification and Verification of Hybrid Systems, pp. 209–229.

[3] M. Andersson, *Object-Oriented Modeling and Simulation of Hybrid Systems*, PhD thesis, Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, 1994.

[4] C. Arévalo and P. Lötstedt, *Improving the accuracy of BDF methods for index 3 differential-algebraic equations*, BIT **35** (1994), pp. 297–308.

[5] M. Avraam, N. Shan, and C. Pantelides, *Modelling and optimization of general hybrid systems in the continuous time domain*, Computational Chemical Engineering **22** (1998), pp. 221–228.

[6] S. Bächle, *Index reduction for differential-algebraic equations in circuit simulation*, Technical Report 141, Matheon - DFG Research Center "Mathematics for key techonolgies", 2004.

[7] S. Bächle and F. Ebert, *Element-based topological index reduction for differential-algebraic equations in circuit simulation*, Technical Report 246, Matheon - DFG Research Center "Mathematics for key techonolgies", 2005.

[8] S. Bächle and F. Ebert, *Graph theoretical algorithms for index reduction in circuit simulation*, Technical Report 245, Matheon - DFG Research Center "Mathematics for key techonolgies", 2005.

[9] A. Backes, *Extremalbedingungen für Optimierungs-Probleme mit Algebro-Differentialgleichungen*, PhD thesis, Humboldt-Universität zu Berlin, 2005.

[10] K. Balla, G. A. Kurina, and R. März, *Index criteria for differential algebraic equations arising from linear-quadratic optimal control problems*, Journal of Dynamics and Control Systems **12**, no. 3 (2006), pp. 289–311.

[11] K. Balla and R. März, *Linear differential algebraic equations of index 1 and their adjoint equations*, Results in Mathematics **37** (2000), pp. 13–35.

[12] P. Barton and C. Lee, *Modeling, simulation, sensitivity analysis, and optimization of hybrid systems*, ACM Transactions on Modeling and Computer Simulation **12**, no. 4 (2002), pp. 256–289.

[13] P. BARTON AND T. PARK, *State event location in differential-algebraic models*, ACM Transactions on Modeling and Computer Simulation **6**, no. 2 (1996), pp. 137–165.

[14] J. BAUMGARTE, *Stabilization of constraints and integrals of motion in dynamical systems*, Computational Methods in Applied Mechanical Engineering **1** (1972), pp. 1–16.

[15] M. BERZINS, *A C^1 interpolant for codes based on backward difference formulae*, Applied Numerical Mathematics **2** (1986), pp. 109–118.

[16] M. BRANICKY, V. BORKAR, AND S. MITTER, *A unified framework for hybrid control: Model ond optimal control theory*, IEEE Treansactions on Automatic Control **43**, no. 1 (1998), pp. 31–45.

[17] K. BRENAN, S. CAMPBELL, AND L. PETZOLD, *Numerical Solution of Initial-Value Problems in Differential Algebraic Equations*, Classics in Applied Mathematics 14, SIAM, Philadelphia, PA, 1996.

[18] P. BROWN, A. HINDMARSH, AND L. PETZOLD, *Using Krylov methods in the solution of large-scale differential-algebraic systems*, SIAM Journal on Scientific Computing **15**, no. 6 (1994), pp. 1467–1488.

[19] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. NICHOLS, *Feedback design for regularizing descriptor systems*, Linear Algebra and its Applications **299** (1999), pp. 119–151.

[20] R. BYERS, V. MEHRMANN, AND H. XU, *A structured staircase algorithm for skew-symmetric/symmetric pencils*, Electronic Transactions on Numerical Analysis **26** (2007), pp. 1–33.

[21] R. BYERS, V. MEHRMANN, AND H. XU, *Trimmed linearizations for structured matrix polynomials*, Linear Algebra and Its Applications (2008). To appear.

[22] S. CAMPBELL, *A general form for solvable linear time varying singular systems of differential equations*, SIAM Journal on Mathematical Analysis **18** (1987), pp. 1101–1115.

[23] S. CAMPBELL, *Linearization of DAE's along trajectories*, Zeitschrift für Angewandte Mathematik und Physik **46** (1995), pp. 70–84.

[24] S. CAMPBELL AND C. GEAR, *The index of general nonlinear DAEs*, Numerische Mathematik **72**, no. 2 (1995), pp. 173–196.

[25] S. CAMPBELL AND E. GRIEPENTROG, *Solvability of general differential algebraic equations*, SIAM Journal on Scientific Computing **16** (1995), pp. 257–270.

[26] S. L. CAMPBELL, *Singular Systems of Differential Equations*, Pitman, San Francisco, CA, 1980.

[27] S. L. CAMPBELL AND C. D. MEYER, *Generalized Inverses of Linear Transformations*, Pitman, San Fransisco, 1979.

[28] D. CHEN, *Stabilization of planar switched systems*, Systems Control Lett. **51** (2004), pp. 79–88.

[29] J. CLABAUGH, *The ABACUSS II syntax manual*, Massachussets Institute of Technology, 2001.

[30] L. DAI, *Singular Control Systems*, no. 118 in Lecture Notes in Control and Information Sciences, Springer-Verlag, Berlin, Heidelberg, 1989.

[31] R. DAVID AND H. ALLA, *On hybrid Petri nets*, Discrete Event Dynamics **11** (2001), pp. 9–40.

[32] C. DE BOOR AND H.-O. KREISS, *On the condition of the linear systems associated with discretized BVPs of ODEs*, SIAM Journal on Numerical Analysis **23**, no. 5 (1986), pp. 936–939.

[33] P. DEUFLHARD, *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, Springer-Verlag, Berlin, 2004.

[34] E. EICH-SOELLNER AND C. FÜHRER, *Numerical Methods in Multibody Dynamics*, B.G. Teubner, Stuttgart, 1998.

[35] D. ESTÉVEZ SCHWARZ, *Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation*, PhD thesis, Humboldt-Universität zu Berlin, 2000.

[36] D. ESTÉVEZ SCHWARZ AND C. TISCHENDORF, *Structural analysis of electric circuits and consequences for MNA*, International Journal of Circuit Theory and Applications **28**, no. 2 (2000), pp. 131–162.

[37] H. FASSBENDER AND D. KRESSNER, *Structured eigenvalue problems*, in GAMM Mitteilungen 29, Themenheft Applied and Numerical Linear Algebra, Part II, 2006, pp. 297–318.

[38] U. FELDMANN AND R. SCHULTZ, *TITAN: A universal circuit simulator with event control for latency exploitation*, ESSCIRC'88 (1988), p. 183.

[39] A. FILIPPOV, *Differential equations with discontinuous right-hand side*, Matematicheskii Sbornik **51** (1960), pp. 99–128.

[40] A. FILIPPOV, *Differential Equations with Discontinuous Right-hand Sides*, Kluwer, 1998.

[41] O. FORSTER, *Analysis 1 – Differential- und Integralrechnung einer Veränderlichen*, Vieweg-Studium: Grundkurs Mathematik, Vieweg, Braunschweig, 5 ed., 1999.

[42] O. FORSTER, *Analysis 2 – Differentialrechnung im $\mathbb{R}^n$, gewöhnliche Differentialgleichungen*, Vieweg-Studium: Grundkurs Mathematik, Vieweg, Braunschweig, 5 ed., 1999.

[43] F. G. FRIEDLANDER, *On the oscillations of the bowed string*, Proceedings of Cambridge Philosophical Society **49** (1953), pp. 516–530.

[44] C. FÜHRER, *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen - Theorie, numerische Ansätze und Anwendungen*, PhD thesis, TU München, 1988.

[45] C. FÜHRER, P. RENTROP, AND B. SIMEON, *The Drazin inverse in multibody system dynamics*, Numerische Mathematik **64** (1993), pp. 521–539.

[46] S. GALÁN AND P. BARTON, *Dynamic optimization of hybrid systems*, Computational Chemical Engineering **22** (1998), pp. 183–190.

[47] F. GANTMACHER, *The Theory of Matrices, volume II*, Chelsea Publishing Company, New York, 1959.

[48] C. GEAR AND K. TU, *The effect of variable mesh size on the stability of multistep methods*, SIAM Journal on Numerical Analysis **11**, no. 5 (1974), pp. 1025–1043.

[49] C. GEAR AND D. WATANABE, *Stability and convergence of variable order multistep methods*, SIAM Journal on Numerical Analysis **11** (1974), pp. 1044–1058.

[50] C. W. GEAR, B. LEIMKUHLER, AND G. K. GUPTA, *Automatic integration of Euler-Lagrange equations with constraints*, Journal of Computational and Applied Mathematics **12/13** (1985), pp. 77–90.

[51] T. GEERTS, *Solvability conditions, consistency, and weak consistency for linear differential-algebraic equations and time-invariant linear systems: The general case*, Linear Algebra and its Applications **181** (1993), pp. 111–130.

[52] T. GEERTS AND V. MEHRMANN, *Linear differential equations with constant coefficients: A distributional approach*, Technical Report SFB 343/90–073, Universität Bielefeld, 1990.

[53] I. GOHBERG, P. LANCASTER, AND L. RODMAN, *Matrix Polynomials*, Computer Science and Applied Mathematics, Academic Press, New York, 1982.

[54] G. GOLUB AND C. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press Baltimore and London, 3 ed., 1996.

[55] E. GRIEPENTROG AND R. MÄRZ, *Differential-Algebraic Equations and Their Numerical Treatment*, Teubner-Texte zur Mathematik 88, BSB B.G.Teubner Verlagsgesellschaft, Leipzig, 1986.

[56] M. GÜNTHER AND U. FELDMANN, *CAD-based electric-circuit modeling in industry, I. Mathematical structure and index of network equations*, Surveys on Mathematics for Industry **8** (1999), pp. 97–129.

[57] M. GÜNTHER AND U. FELDMANN, *CAD-based electric-circuit modeling in industry, II. Impact of circuit configuration and parameters*, Surveys on Mathematics for Industry **8** (1999), pp. 131–157.

[58] E. HAIRER, S. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations I - Nonstiff Problems*, Springer, Berlin Heidelberg, 2. ed., 1993.

[59] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer, Berlin, 2. ed., 1996.

[60] P. HAMANN, *Modellierung und Simulation von realen Planetengetrieben*, Diplomarbeit, TU Berlin, in cooperation with DaimlerChrysler AG, 2003.

[61] P. HAMANN AND V. MEHRMANN, *Numerical solution of hybrid systems of differential-algebraic equations*, Computational Methods in Applied Mechanical Engineering **197**, no. 6-8 (2008), pp. 693–705.

[62] M. HAUTUS AND L. SILVERMAN, *System structure and singular control*, Linear Algebra and its Applications **50** (1983), pp. 369–402.

[63] N. J. HIGHAM, D. S. MACKEY, AND F. TISSEUR, *The conditioning of linearizations of matrix polynomials*, SIAM Journal on Matrix Analysis and Applications **28**, no. 4 (2006), pp. 1005–1028.

[64] A. C. HINDMARSH, R. SERBAN, AND A. COLLIER, *User Documentation for ida v2.5.0*, Center for Applied Scientific Computing Lawrence Livermore National Laboratory, November 6 2006.

[65] T.-M. HWANG, W.-W. LIN, AND V. MEHRMANN, *Numerical solution of quadratic eigenvalue problems with structure-preserving methods*, SIAM Journal on Scientific Computing **24** (2003), pp. 1283–1302.

[66] A. ILCHMANN AND V. MEHRMANN, *A behavioural approach to linear time-varying descriptor system. Part 1. general theory*, SIAM Journal on Control **44**, no. 5 (2005), pp. 1725–1747.

[67] A. ILCHMANN AND V. MEHRMANN, *A behavioural approach to linear time-varying descriptor system. Part 2. descriptor systems*, SIAM Journal on Control **44**, no. 5 (2005), pp. 1748–1765.

[68] J. B. KELLER, *Bowing of violin strings*, Comments on Pure Applied Mathematics **6**, no. 4 (1953), pp. 483–495.

[69] D. E. KIRK, *Optimal Control Theory : an introduction*, Englewood Cliffs, NJ : Prentice-Hall, 1970.

[70] D. KRESSNER, *Numerical Methods for General and Structured Eigenvalue Problems*, Lecture Notes in Computational Science and Engineering 46, Springer, Heidelberg, 2005.

[71] L. KRONECKER, *Algebraische Reduktion der Schaaren bilinearer Formen*, in Sitzungsberichte Akademie der Wissenschaften, Berlin, 1890, pp. 763–776.

[72] P. KUNKEL AND V. MEHRMANN, *Canonical forms for linear differential-algebraic equations with variable coefficients*, Journal of Computational and Applied Mathematics **56** (1994), pp. 225–251.

[73] P. KUNKEL AND V. MEHRMANN, *A new look at pencils of matrix valued functions*, Linear Algebra and its Applications **212/213** (1994), pp. 215–248.

[74] P. KUNKEL AND V. MEHRMANN, *Generalized inverses of differential-algebraic operators*, SIAM Journal on Matrix Analysis and Applications **17** (1996), pp. 426–442.

[75] P. KUNKEL AND V. MEHRMANN, *Local and global invariants of linear differential-algebraic equations and their relation*, Electronic Transactions on Numerical Analysis **4** (1996), pp. 138–157.

[76] P. KUNKEL AND V. MEHRMANN, *A new class of discretization methods for the solution of linear differential-algebraic equations with variable coefficients*, SIAM Journal on Numerical Analysis **33**, no. 5 (1996), pp. 1941–1961.

[77] P. KUNKEL AND V. MEHRMANN, *The linear quadratic control problem for linear descriptor systems with variable coefficients*, Mathematics of Control, Signals, and Systems **10** (1997), pp. 247–264.

[78] P. KUNKEL AND V. MEHRMANN, *Regular solutions of nonlinear differential-algebraic equations and their numerical determination*, Numerische Mathematik **79** (1998), pp. 581–600.

[79] P. KUNKEL AND V. MEHRMANN, *Analysis of over- and underdetermined nonlinear differential-algebraic systems with application to nonlinear control problems*, Mathematics of Control, Signals, and Systems **14** (2001), pp. 233–256.

[80] P. KUNKEL AND V. MEHRMANN, *Index reduction for differential-algebraic equations by minimal extension*, Zeitschrift für Angewandte Mathematik und Mechanik **84** (2004), pp. 579–597.

[81] P. KUNKEL AND V. MEHRMANN, *Characterization of classes of singular linear differential-algebraic equations*, Electronic Journal on Linear Algebra **13** (2005), pp. 359–386.

[82] P. KUNKEL AND V. MEHRMANN, *Differential-Algebraic Equations — Analysis and Numerical Solution*, EMS Publishing House, Zürich, Switzerland, 2006.

[83] P. KUNKEL AND V. MEHRMANN, *Necessary and sufficient conditions in the optimal control for general nonlinear differential-algebraic equations*, Mathematics of Control, Signals, and Systems (2008). To appear.

[84] P. KUNKEL, V. MEHRMANN, AND W. RATH, *Analysis and numerical solution of control problems in descriptor form*, Mathematics of Control, Signals, and Systems **14** (2001), pp. 29–61.

[85] P. KUNKEL, V. MEHRMANN, W. RATH, AND J. WEICKERT, *GELDA: A software package for the solution of general linear differential algebraic equations*, SIAM Journal on Scientific Computing **18** (1997), pp. 115 – 138.

[86] P. KUNKEL, V. MEHRMANN, M. SCHMIDT, I. SEUFER, AND A. STEINBRECHER, *Weak formulations of linear differential-algebraic systems*, Preprint 16, Institut für Mathematik, Technische Universität Berlin, 2006.

[87] P. KUNKEL, V. MEHRMANN, AND I. SEUFER, *GENDA: A software package for the numerical solution of general nonlinear differential-algebraic equations*, Preprint 730, Institut für Mathematik, Technische Universität Berlin, 2002.

[88] C. LACOURSIÈRE, *Ghosts and Machines: Regularized Variational Methods for Interactive Simulation of Multibodies with Dry Frictional Contacts*, PhD thesis, Umeå University, Sweden, 2007.

[89] P. LANCASTER, *Lambda-Matrices and vibrating systems*, International series of monographs in pure and applied mathematics. 94, Pergamon Press, Oxford, 1966.

[90] P. LANCASTER, *Theory of matrices*, Academic Press, New York, 1969.

[91] R. LEINE, D. VAN CAMPEN, A. DE KRAKER, AND L. VAN DEN STEEN, *Stick-slip vibrations induced by alternate friction models*, Nonlinear Dynamics **16**, no. 1 (1998), pp. 41–54.

[92] P. LOSSE AND V. MEHRMANN, *Controllability and observability of second order descriptor systems*, SIAM Journal on Control **47** (2008), pp. 1351–1379.

[93] C. LUBICH, U. NOWAK, U. POEHLE, AND C. ENGSTLER, *MEXX - Numerical software for the integration of constrained mechanical systems*, Technical Report SC 92-12, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1992.

[94] J. LYGEROS, G. PAPPAS, AND S. SASTRY, *An introduction to hybrid system modeling, analysis, and control*, technical report, Preprints of the First Nonlinear Control Network Pedagogical School, Athens, Greece, 1999.

[95] D. S. MACKEY, N. MACKEY, C. MEHL, AND V. MEHRMANN, *Vector spaces of linearizations for matrix polynomials*, SIAM Journal on Matrix Analysis and Applications **28**, no. 4 (2006), pp. 971–1004.

[96] C. MAJER, W. MARQUARDT, AND E. GILLES, *Reinitialization of DAEs after discontinuities*, Tech. Report LPT-1995-02, RWTH Aachen, Lehrstuhl für Prozeßtechnik, 1995.

[97] R. MÄRZ, *The index of linear differential algebraic equations with properly stated leading terms*, Results in Mathematics **42**, no. 3-4 (2002), pp. 308–338.

[98] R. MÄRZ, *Characterizing differential-algebraic equations without the use of derivative arrays*, Computational Mathematics and Applications **50**, no. 7 (2005), pp. 1141–1156.

[99] R. MÄRZ AND R. RIAZA, *Linear differential-algebraic equations with properly stated leading term: A-critical points*, Journal on Mathemtical Computations in Modern Dynamical Systems **13**, no. 3 (2007), pp. 291 – 314.

[100] S. MATTSSON, H. ELMQVIST, AND J. BROENINK, *Modelica: An international effort to design the next generation modelling language*, Journal A, Benelux Quarterly Journal on Automatic Control **38**, no. 3 (1997), pp. 16–19.

[101] S. MCILRAITH, G. BISWAS, D. CLANCY, AND V. GUPTA, *Hybrid system diagnosis*, in Third Intl. Workshop on hybrid systems, 2000, pp. 282–295.

[102] V. MEHRMANN AND C. SHI, *Transformation of high order linear differential-algebraic systems to first order*, Numerical Algorithms **42**, no. 3-4 (2006), pp. 281–307.

[103] A. S. MORSE, *Trends in Control*, Springer, London, 1995, ch. Control using logic-based switching, pp. 69–114.

[104] P. MOSTERMAN, *An overview of hybrid simulation phenomena and their support by simulation packages*, in Hybrid Systems: Computation and Control, no. 1569 in Lecture Notes in Computer Science, Springer, 1999, pp. 165–177.

[105] P. MOSTERMAN, F. ZHAO, AND G. BISWAS, *A study of transitions in dynamic behavior of physical systems*, in 12th International Workshop on Qualitative Reasoning, Cape Cod, 1998, pp. 96 – 105.

[106] P. MOSTERMAN, F. ZHAO, AND G. BISWAS, *Sliding mode model semantics and simulation for hybrid systems*, in Hybrid Systems V, LNCS 1567, Springer-Verlag Berlin Heidelberg, 1999, pp. 218–237.

[107] P. J. MOSTERMAN, *Hybrid Dynamic Systems: A hybrid bond graph modeling paradigm and its application in diagnosis*, PhD thesis, Vanderbilt University, 1997.

[108] P. J. MOSTERMAN, *HYBRSIM - a modeling and simulation environment for hybrid bond graphs*, Journal of Systems and Control Engineering **216** (2002), pp. 35–46.

[109] L. W. NAGEL AND D. O. PEDERSON, *SPICE (simulation program with integrated circuit emphasis)*, memorandum no. erl-m382, University of California, Berkeley, 1973.

[110] K. NOMIZU, *Characteristic roots and vectors of a differentiable family of symmetric matrices*, Linear and Mulitilinear Algebra **1** (1973), pp. 159–162.

[111] U. NOWAK AND L. WEIMANN, *A family of Newton codes for systems of highly nonlinear equations - Algorithm, Implementation, Application*, technical Report 91-10, Konrad-Zuse-Zentrum für Informationstechnik Berlin, 1991.

[112] C. NYTSCH-GEUSEN, T. ERNST, A. NORDWIG, P. SCHNEIDER, P. SCHWARZ, M. VETTER, C. WITTWER, A. HOLM, T. NOUIDUI, J. LEOPOLD, G. SCHMIDT, U. DOLL, AND A. MATTES, *MOSILAB: Development of a modelica based generic simulation tool supporting model structural dynamics*, in Proceedings, 4th International Modelica Conference TU Hamburg-Harburg, 2005.

[113] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solutions of Nonlinear Equations in Several Variables*, Classics in Applied Mathematics 30, SIAM, Philadelphia, PA, 2000.

[114] M. OTTER, *Objektorientierte Modellierung mechatronischer Systeme am Beispiel geregelter Roboter*, Fortschritt-Berichte VDI Reihe 20, Nr. 147, VDI-Verlag, Düsseldorf, 1995.

[115] L. R. PETZOLD, *A description of DASSL: a differential/algebraic system solver*, IMACS Trans. Scientific Computing **1** (1983), pp. 65–68.

[116] V. N. PHAT, *Robust stability and stabilizability of uncertain linear hybrid systems with state delays*, IEEE Transactions on CAS II **52** (2005), pp. 94–98.

[117] V. N. PHAT AND S. PAIROTE, *Global stabilization of linear periodically time-varying switched systems via matrix inequalities*, Journal on Control Theory and Applications **1** (2006), pp. 24–29.

[118] P. RABIER AND W. C. RHEINBOLDT, *Classical and generalized solutions of time-dependent linear differential-algebraic equations*, Linear Algebra and its Applications **245** (1996), pp. 259–293.

[119] P. RABIER AND W. C. RHEINBOLDT, *Time-dependent linear DAEs with discontinuous inputs*, Linear Algebra and its Applications **247** (1996), pp. 1–29.

[120] P. J. RABIER AND W. C. RHEINBOLD, *On impasse points of quasilinear differential algebraic equations*, Journal of Mathematical Analysis and Applications **181** (1994), pp. 429–454.

[121] P. J. RABIER AND W. C. RHEINBOLDT, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, USA, 2000.

[122] C. V. RAMAN, *On the mechanical theory of the vibrations of bowed strings*, Indian Association for the Cultivation of Science Bulletin **15** (1918), pp. 1–158.

[123] G. REISSIG, *Differential-algebraic equations and impasse points*, IEEE Transactions on Circuits and Systems **43**, no. 2 (1996), pp. 122–133.

[124] G. REISSIG, H. BOCHE, AND P. BARTON, *On inconsistent initial conditions for linear-time-invariant differential-algebraic equations*, IEEE Transactions on Circuits and Systems **49** (2002), pp. 1646–1648.

[125] W. C. RHEINBOLDT, *MANPACK: A set of algorithms for computations on implicitly defined manifolds*, Applied Mathematical Computations **27** (1996), pp. 15–28.

[126] P. RIEDINGER, F. KRATZ, C. IUNG, AND C. ZANNE, *Linear quadratic optimization for hybrid systems*, in IEEE Conference on Decision and Control (CDC'99), Phoenix, 1999, pp. 3059–3064.

[127] W. RULKA, *SIMPACK-a computer program for simulation of large-motion multibody systems*, in Multibody system handbook, W. Schiehlen, ed., Springer-Verlag, Berlin, Germany, 1990, pp. 265–284.

[128] R. RYAN, *Multibody system handbook*, Springer-Verlag, Berlin, 1990, ch. Adams-multibody system analysis software, pp. 361–402.

[129] J. SAND, *On implicit Euler and related methods for high-order high-index DAEs*, Applied Numerical Mathematics **42** (2002), pp. 411–424.

[130] W. SCHIEHLEN, *Multibody System Handbook*, Springer, Berlin, 1990.

[131] W. E. SCHIESSER, *The Numerical Method of Lines*, San Diego: Academic Press, 1991.

[132] C. SCHRÖDER, *Palindromic and Even Eigenvalue Problems – Analysis and Numerical Methods.*, PhD thesis, Institut für Mathematik, Technische Universität Berlin, 2008.

[133] L. SCHWARTZ, *Theorie des Distributions*, Hermann, Paris, 1978.

[134] H. R. SCHWARZ, *Numerische Mathematik*, Teubner, 2 ed., 1988.

[135] C. SHI, *Linear differential-algebraic equations of higher-order and the regularity or singularity of matrix polynomials*, PhD thesis, Institut für Mathematik, Technische Universität Berlin, 2004.

[136] B. SIMEON, *MBSPACK – numerical integration software for constrained mechanical motion*, Surveys on Mathematics for Industry **5**, no. 3 (1995), pp. 169–202.

[137] A. STEINBRECHER, *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*, PhD thesis, Institut für Mathematik, Technische Universität Berlin, 2006.

[138] A. STEINBRECHER, *GEOMS: A software package for the numerical integration of general model equations of multibody systems*, Technical Report 400, MATHEON - DFG Research Center "Mathematics for key technologies", 2007.

[139] G. W. STEWART AND J.-G. SUN, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[140] R. THOMPSON, *The characteristic polynomial of a principal subpencil of a Hermitian matrix pencil*, Linear Algebra and its Applications **14** (1976), pp. 135–177.

[141] C. TISCHENDORF, *Topological index calculation of differential-algebraic equations in circuit simulation*, Surveys on Mathematics for Industry **8**, no. 3-4 (1999), pp. 187–199.

[142] F. TISSEUR, *Backward error and condition of polynomial eigenvalue problems*, Linear Algebra and its Applications **309** (2000), pp. 339–361.

[143] J. TOLSMA AND P. BARTON, *DAEPACK: An open modeling environment for legacy models*, Industrial and Engineering Chemistry Research **39**, no. 6 (2000), pp. 1826 –1839.

[144] S. TRENN, *Distributional solution theory for linear DAEs*, preprint, Technische Universität Ilmenau, 2008. Submitted to PAMM.

[145] V. UTKIN, *Sliding Modes in Control Optimization*, Springer, 1992.

[146] V. UTKIN, J. GULDNER, AND J. SHI, *Sliding mode control in electromechanical systems*, Taylor & Francis, 1999.

[147] V. UTKIN AND F. ZHAO, *Adaptive simulation and control of variable-structure control systems in sliding regimes*, Automatica: IFAC Journal **32**, no. 7 (1996), pp. 1037–1042.

[148] H. A. WATTS, *A smooth output interpolation process for BDF codes*, Journal of Applied Mathematics and Computing **31** (1989), pp. 397–418.

[149] J. WEICKERT, *Navier-Stokes equations as a differential-algebraic system*, Technical report SFB393/96-0, Fakultät für Mathematik, Technische Universität Chemnitz-Zwickau, Chemnitz, 1996.

[150] K. WEIERSTRASS, *Zur Theorie der bilinearen quadratischen Formen*, in Monatshefte Akademie der Wissenschaften, Berlin, 1867, pp. 310–338.

[151] L. WUNDERLICH, *Numerical solution of second order differential-algebraic equations*, Diplomarbeit, Technische Universität Berlin, 2004.

[152] L. WUNDERLICH, *Numerical solution of semi-explicit systems of second order differential-algebraic equations*, Preprint 24, Institut für Mathematik, Technische Universität Berlin, 2005.

[153] L. WUNDERLICH, *Structure preserving condensed forms for pairs of hermitian matrices and matrix valued functions*, Preprint 4, Institut für Mathematik, Technische Universität Berlin, 2006.

[154] X. XU AND P. J. ANTSAKLIS, *Hybrid Systems: Computation and Control*, Lecture Notes in Computer Science 2623, Springer Berlin / Heidelberg, 2003, ch. Results and Perspectives on Computational Methods for Optimal Control of Switched Systems, pp. 540–555.

[155] R. ZURMÜHL AND S. FALK, *Matrizen und ihre Anwendungen I, 5. Auflage*, Springer-Verlag, Berlin, 1986.

[156] R. ZURMÜHL AND S. FALK, *Matrizen und ihre Anwendungen II, 6. Auflage*, Springer-Verlag, Berlin, 1992.