Delft University of Technology
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft Institute of Applied Mathematics

Numerical Methods for Differential-Algebraic Equations

A thesis submitted to the
Delft Institute of Applied Mathematics
in partial fulfillment of the requirements

for the degree

MASTER OF SCIENCE
in
APPLIED MATHEMATICS

by

Kristin Altmann

Delft, The Netherlands

February 6, 2015

MSc THESIS APPLIED MATHEMATICS

“Numerical Methods for Differential-Algebraic Equations”

Kristin Altmann

Delft University of Technology

Daily supervisors

Dr. Matthias Möller

Prof. dr. ir. Cornelis Vuik

Committee members

Prof. dr. ir. Cornelis Vuik

Dr. Matthias Möller

Dr. Henk Schuttelaars

Dr. ir. Jan Schuurmans

Delft, The Netherlands

February 6, 2015

Abstract

In recent years, the use of differential equations in connection with algebraic constraints on the variables has become a widely accepted tool for modeling the dynamical behaviour of physical processes. Compared to ordinary differential equations (ODEs), it has been observed that a number of difficulties can arise when numerical methods are used to solve differential-algebraic equations (DAEs), for instance order reduction phenomena, drift-off effects or instabilities. DAEs arise naturally and have to be solved in a variety of applications such as the mathematical pendulum and a reheat furnace model, both of which are used in this thesis to demonstrate the application of numerical methods and the arising difficulties. Working towards the prospective development of an applicable NMPC algorithm with DAEs as system models, this thesis is mainly concerned with the analysis and numerical treatment of DAEs. A particular focus is put on the topics of the strangeness index, an iterative procedure for determining all hidden constraints, regularisation techniques and numerical methods for solving DAEs. This includes the implementation of the RadauIIa and BDF methods. Due to a multitude of examples, this thesis may serve as an accessible introduction to DAEs and as a foundation for future research into this field.

Key words: differential-algebraic equations, strangeness index, numerical methods, regularisation, RadauIIa methods, BDF methods, mathematical pendulum, reheat furnace model

Contents

List of Figures

List of Tables

Abbreviations

BDF	backward differentiation formula
DAE	differential-algebraic equation
DE-step	differentiation-elimination step
d-index	differentiation index
IVP	initial value problem
JCF	Jordan canonical form
KCF	Kronecker canonical form
NMPC	nonlinear model predictive control
ODE	ordinary differential equation
PDE	partial differential equation
s-index	strangeness index
SVD	singular value decomposition
WCF	Weierstraß canonical form

Notation

$\dot{x}$	total derivative of x with respect to t, i.e. $\dot{x}(t) = \frac{d}{dt}x(t)$
$\ddot{x}$	second total derivative of x with respect to t, i.e. $\ddot{x}(t) = \frac{d^2}{dt^2}x(t)$
$x^{(i)}$	i-th total derivative of $x(t)$ with respect to t, i.e. $x^{(i)}(t) = \frac{d^i}{dt^i}x(t)$
$g_x(x,y)$	partial derivative of $g(x,y)$ with respect to x, i.e. $g_x(x,y) = \frac{\partial}{\partial x}g(x,y)$
E^*	Hermitian conjugate of E, i.e. $E^* = \overline{E^T}$
T'	matrix that completes T to a nonsingular matrix $[T \;\; T']$
$\oplus$	direct sum
$\Vert\cdot\Vert$	norm
$\emptyset_{n,0}$	empty matrix $\in \mathbb{C}^{n,0}$
a_{ij}, b_i, c_i	coefficients of a Runge-Kutta method
b	weight vector of a Runge-Kutta method
c	node vector of a Runge-Kutta method
c_{air}	heat capacity of air
c_f	heat capacity of fuel
c_{fuel}	cost of fuel
c_g	heat capacity of waste gas
c_{prod}	production profit
c_s	heat capacity of steel products
c_w	heat capacity of furnace wall
e	Euler's number
$f(t)$	inhomogeneity of a linear DAE
$\tilde{f}(t)$	transformed inhomogeneity of a linear DAE
$f(x(t),t)$	right-hand side of a quasi-linear DAE
$f(t,x)$	right-hand side of an ODE
g	gravitational acceleration
$g(x,t)$	set of constraints
h	step size
$h_i(x,t)$	hidden constraint of level i
k_j	stage of a Runge-Kutta method for ODEs
l	length of the pendulum
m	number of equations

m	mass of the pendulum
m_c	number of constraints
m_i	size of the Jordan block J_i
m_1, m_2	number of equations of the differential part and of the constraint part of a semi-implicit DAE, respectively
n	number of unknowns
p	order of consistency or convergence
r, a, s, d, u, v	characteristic values of a linear DAE with constant coefficients
s	number of stages of a Runge-Kutta method
t	independent variable of a DAE
t_0	initial time
t_i	grid points
tol	prescribed tolerance
u_1, u_2, u_3	fuel flows to burners 1, 2, 3, respectively
v	speed of the steel products
v	index of nilpotency of a matrix
v_c	maximal constraint level of a DAE
v_d	differentiation index of a DAE
$v_{\max}$	maximal speed of the steel products
v_n	index of nilpotency of a DAE
v_s	strangeness index of a DAE
$x(t)$	state variable of a DAE depending on t
$\tilde{x}$	transformed state variable of a DAE depending on t
$\dot{x} = \Phi(x, t)$	underlying ODE
x^0	initial guess of the Newton method
x_0	initial value of an IVP
x_i	approximation of $x(t_i)$
x^k	Newton iterates
Δx	length of a section m of the furnace
Δx^k	correction term of the Newton iteration
A	coefficient matrix of a linear DAE with constant coefficients
A	Runge-Kutta matrix
$\tilde{A}$	transformed coefficient matrix of a linear DAE with constant coefficients
$A(t)$	coefficient matrix of a linear DAE with variable coefficients
$A_{s,m}$	surface area of the steel products in section m
$A_{w,m}$	surface area of the furnace wall in section m
B	width of the furnace
$\mathbb{C}$	set of complex numbers
$\mathcal{C}$	set of continuous functions
$\mathcal{C}^i$	set of i-times continuously differentiable functions
D_s	thickness of steel products

D_w	thickness of furnace wall
$\mathbb{D}_x$	domain of the state variable x
$\mathbb{D}_{\dot{x}}$	domain of the derivative $\dot{x}$
E	leading matrix of a linear DAE with constant coefficients
$\tilde{E}$	transformed leading matrix of a linear DAE with constant coefficients
E_k	kinetic energy
E_p	potential energy
$E(t)$	leading matrix of a linear DAE with variable coefficients
$E(x(t), t)$	leading matrix of a quasi-linear DAE
$(E_{\text{mod}}, A_{\text{mod}})$	modified matrix pair of (E, A)
$F(\dot{x}(t), x(t), t)$	right-hand side of a DAE
$\hat{F}(\dot{x}(t), x(t), t)$	right-hand side of a regularisation of a DAE
$\mathcal{F}_l$	derivative array of order l
H	height of the furnace
H_0	lower calorific value of fuel
I, I_d	identity matrix ($\in \mathbb{R}^{d,d}$)
$\mathbb{I}$	domain of the independent variable t
J	Jordan block of the WCF
J	objective function
J_f	Jacobian of the function f
J_i	Jordan block of the JCF
$\mathcal{J}_{\rho_j}$	Jordan block of the KCF
L	Lipschitz constant
L	length of the furnace
L_s	length of steel products
$\mathbb{L}_l$	set of solutions of the derivative array of order l
$\mathcal{L}$	Lagrange function
$\mathcal{L}_{\varepsilon_j}$	Jordan block of the KCF
$\mathcal{L}_*$	underdetermined block of a DAE in KCF
$\mathbb{M}$	solution manifold or set of consistency
$\mathcal{M}_{\eta_j}$	Jordan block of the KCF
$\mathcal{M}_*$	overdetermined block of a DAE in KCF
(M_l, N_l)	inflated pair of order l
N	nilpotent Jordan block of the WCF
N	number of discrete grid points
N	number of sections of the furnace
$\mathbb{N}$	set of natural numbers
$\mathbb{N}_0$	set of natural numbers (including 0)
$\mathcal{N}_{\sigma_j}$	Jordan block of the KCF
P	left transformation matrix
$P(t)$	left transformation matrix function

Q	right transformation matrix
$Q(t)$	right transformation matrix function
$Q_{air,m}$	heat brought in by air in section m
$Q_{c,m}$	heat brought in by combustion in section m
$Q_{f,m}$	heat brought in by fuel in section m
$Q_{g,m+1}$	heat brought in by waste gas of section $m+1$
$Q_{o,m}$	heat leaving the furnace wall to outside air in section m
$Q_{s,m}$	heat entering steel products in section m
$Q_{w,m}$	heat entering the furnace wall in section m
R_{af}	ratio of air to fuel
$\mathbb{R}$	set of real numbers
$S(x,t)$	selector matrix function
T	end point of the domain $\mathbb{I} = [t_0 T]$
T_o	temperature of air outside the furnace
T_{air}	temperature of air used in combustion
T_f	temperature of fuel
$T_{g,m}$	temperature of the waste gas in section m
$T_{s,m}$	temperature of the steel products in section m
$V_{s,m}$	volume of the steel products in section m
X_j, X_j'	stages of a Runge-Kutta method for DAEs
α_i, β_i	coefficients of a linear multi-step method
ε_j	size of the Jordan block $\mathcal{L}_{\varepsilon_j}$
ε_s	emissivity of steel products
ε_w	emissivity of furnace wall
η_j	size of the Jordan block $\mathcal{M}_{\eta_j}$
λ	Lagrange multiplier
ϱ	first characteristic polynomial of a linear multi-step method
ρ	density of steel products
ρ_j	size of the Jordan block $\mathcal{J}_{\rho_j}$
ρ_w	density of furnace wall
σ	Stefan-Boltzmann constant
σ_j	size of the Jordan block $\mathcal{N}_{\sigma_j}$
Φ	increment function of a one-step method
$\Phi_{f,m}$	fuel flow into section m
$\Phi_{g,m}$	waste gas flow from section m into $m-1$

Chapter 1

Introduction

In recent years, the use of differential equations in connection with algebraic constraints on the variables, for example due to laws of conservation or position constraints, has become a widely accepted tool for modeling the dynamical behaviour of physical processes. Such combinations of both differential and algebraic equations are called differential-algebraic equations (DAEs).

DAEs arise naturally and have to be solved in a variety of applications, such as mechanical multibody systems (vehicle dynamics, aeronautics, biomechanics, robotics, etc.), chemical process simulation, control theory, simulation of electrical networks, fluid dynamics and many other areas (see for instance Steinbrecher 2006, section 4.1.3; Kunkel/Mehrmann 2006, section 1.3; Brenan et al. 1996, section 1.3). Furthermore, solving partial differential equations (PDEs) by using the method of lines and discretizing the spatial derivatives first, for example by finite element or finite difference methods, can also lead to DAEs (see Kunkel/Mehrmann 2006, p. 10; Brenan et al. 1996, pp. 10-12).

The present master thesis project is conducted in collaboration with DotX Control Solutions BV, a Dutch company based in Alkmaar. DotX Control Solutions BV is specialised in designing and implementing complex control software for industrial processes and has worked on projects in water management, steel production, wind turbine design and paper drying (see DotX Control Solutions BV 2014a, 2014c).

DotX Control Solutions BV operates in a number of research fields, including nonlinear model predictive control (NMPC). NMPC is an advanced method of process control that typically computes a (sub-) optimal control based on predictions of the system dynamics. The predictions rely on a mathematical model of the complex dynamical process. As the name suggests, NMPC can handle nonlinearities in the underlying model. Nonlinear modeling realised by ordinary differential equations (ODEs) as system models has been an important and successful research field of DotX Control Solution BV. However, advanced industrial applications require models of even greater realism and complexity. Thus, it is intended to use DAEs as system models in NMPC in order to improve the control performance.

While the numerical solution techniques and the theoretical analysis of ODEs have reached maturity, many difficult and complex questions concerning both the analytical and the numerical behaviour of DAEs remain untreated or unsolved despite an enormous increase in the research of DAEs during the past 40 years (see Steinbrecher 2006, p. 21; Kunkel/Mehrmann 2006, pp. 4-5). Compared to ODEs, it has been observed that a number of difficulties can arise when numerical methods are used to solve DAEs, for instance order reduction phenomena, drift-off effects or instabilities (see Steinbrecher 2006, p. 106). Therefore, the main purpose of this thesis is to investigate the analytical and numerical behaviour of DAEs. In particular, efficient numerical methods for solving DAEs are required in order to prospectively develop an applicable NMPC algorithm with DAEs as system models.

In order to achieve the desired goals of this thesis, we are interested in answering the following questions:

- What types of DAEs exist and how can DAEs be characterized with respect to their analytical and numerical behaviour?
- Which numerical methods exist for DAEs? What are their differences regarding accuracy and computation time?
- Which classes of DAEs create no or only slight difficulties while solving them numerically?
- What strategies exist for classes of DAEs which cause significant problems in their numerical treatment?

Therefore, this thesis is structured as follows: After this introduction, chapter 2 covers the analysis of DAEs by giving some introductory examples, introducing necessary basic definitions and treating first linear DAEs and afterwards general nonlinear DAEs. Chapter 3 is concerned with the numerical treatment of DAEs. This includes a brief discussion of the difficulties which can arise when numerical methods for ODEs are used to solve DAEs. Subsequently, index reduction and regularisation techniques are surveyed, followed by different methods for the numerical solution of strangeness-free DAEs as well as numerical methods for solving systems of nonlinear equations. In chapter 4, two test problems, the mathematical pendulum and a reheat furnace model, are introduced. In chapter 5, the results of several implementations of numerical methods applied to the two test problems are presented and discussed. Finally, chapter 6 concludes this master thesis and gives an outlook on future research in this field.

Chapter 2

Analysis of differential-algebraic equations

This chapter reviews important facts on the analysis of DAEs. In the first section 2.1, some introductory examples illustrate that DAEs can be seen as combining the properties of ODEs and purely algebraic equations. Necessary basic definitions are introduced in section 2.2. The following sections cover the range of DAEs from simple types to more general ones, starting with linear DAEs with constant coefficients in section 2.3. Linear DAEs with variable coefficients are considered in section 2.4, while nonlinear DAEs are treated in section 2.5. In each of these sections, respectively, existence and uniqueness results are discussed. This includes the development of several canonical forms and the determination of characteristic quantities of a DAE, in particular different concepts of the so-called index.

Generally, the idea of all these index concepts is to classify DAEs with respect to their difficulty in the analytical as well as the numerical solution. In particular, they measure the degree of smoothness of the problem that is needed to obtain existence and uniqueness results. In this thesis, the focus is set on the index of nilpotency for linear DAEs with constant coefficients, the most widely used differentiation index (d-index) and the recently developed strangeness index (s-index). Other index concepts, such as the perturbation index (see Hairer et al. 1989), the geometric index (see Rheinboldt 1984), the tractability index (see Griepentrog/März 1986) or the structural index (see Pantelides 1988), will not be discussed.

The contents of the present chapter are covered in a more detailed way in Kunkel/Mehrmann (2006) and partially in Steinbrecher (2006).

2.1 Introductory examples

Example 1: Differential equations

Consider the linear (ordinary) differential equations

$$\dot{x}_1 = 3x_1 - x_2,$$
$$\dot{x}_2 = 4x_1 - 2x_2.$$

The coefficient matrix $A = \begin{pmatrix} 3 & -1 \\ 4 & -2 \end{pmatrix}$ has eigenvalues 2 and -1, and corresponding eigenvectors $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ and $\begin{pmatrix} 1 \\ 4 \end{pmatrix}$, respectively. Thus, the general solution of the system is given by

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = c_1 \begin{pmatrix} 1 \\ 1 \end{pmatrix} \cdot e^{2t} + c_2 \begin{pmatrix} 1 \\ 4 \end{pmatrix} \cdot e^{-t} \qquad \text{with} \quad c_1, c_2 \in \mathbb{R}.$$

By prescribing initial values, the parameters c_1, c_2 can be determined, e.g. $x(0) = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$ yields $c_1 = 2$, $c_2 = 1$. Hence, the solution of the initial value problem is

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix} \cdot e^{2t} + \begin{pmatrix} 1 \\ 4 \end{pmatrix} \cdot e^{-t}.$$

In summary, the initial values can *freely* be chosen in $\mathbb{R}^2$, i.e. there are 2 degrees of freedom and the solution manifold, i.e. the space every solution trajectory $(x(t), t)$ lies in, is the whole $\mathbb{R}^3$ (3-dimensional).

Example 2: Algebraic equations

Consider the algebraic equations

$$0 = 2x_1 + x_2 + 3\sin(t),$$
$$0 = x_1 - x_2 + 6\cos(t).$$

Adding both equations yields

$$0 = 3x_1 + 3\sin(t) + 6\cos(t) \qquad \Leftrightarrow \qquad x_1 = -\sin(t) - 2\cos(t).$$

By inserting this into the second equation we obtain

$$0 = -\sin(t) - 2\cos(t) - x_2 + 6\cos(t) \qquad \Leftrightarrow \qquad x_2 = -\sin(t) + 4\cos(t).$$

Thus, the unique solution of the system is given by

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -1 \\ -1 \end{pmatrix} \sin(t) + \begin{pmatrix} -2 \\ 4 \end{pmatrix} \cos(t).$$

In summary, initial values *cannot* be chosen, i.e. there are 0 degrees of freedom and the solution manifold is one trajectory (1-dimensional).

Example 3: Differential-algebraic equations I

Consider the differential-algebraic equations

$$\begin{aligned} \dot{x}_1 &= 2x_1 + 2x_2, \\ 0 &= x_1 + x_2 - \cos(t). \end{aligned}$$

From the second equation we get $x_2 = -x_1 + \cos(t)$. Inserting this into the first equation yields

$$\dot{x}_1 = 2x_1 - 2x_1 + 2\cos(t) = 2\cos(t) \quad \Rightarrow \quad x_1 = 2\sin(t) + c \quad \text{with} \quad c \in \mathbb{R}.$$

By inserting this into the second equation we obtain $x_2 = -2\sin(t) + \cos(t) - c$.

Thus, the general solution of the system is given by

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \sin(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cos(t) + \begin{pmatrix} 1 \\ -1 \end{pmatrix} c \qquad \text{with} \quad c \in \mathbb{R}.$$

The parameter c can be determined from prescribed initial values, but one has to be careful since the initial values have to satisfy the algebraic equations. For example $x(0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is not possible since it yields $\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} c \\ 1-c \end{pmatrix}$ which is contradictory.

The initial values $x(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ are a possible choice and yield $c = 1$. Hence, the solution of the initial value problem is

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \end{pmatrix} \sin(t) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} \cos(t) + \begin{pmatrix} 1 \\ -1 \end{pmatrix}.$$

In summary, the initial values *cannot* be chosen *freely*, i.e. there is 1 degree of freedom and the solution manifold is a surface (2-dimensional).

The previous examples show that DAEs reflect the behaviour of differential equations as well as the behaviour of algebraic equations. In particular, the degrees of freedom and the dimension of the solution manifold of DAEs lie in between those of differential equations and algebraic equations.

The conjecture that the degrees of freedom of a DAE are equal to the number of unknowns minus the number of algebraic equations is not true in general. It is possible that besides the explicitly given algebraic equations, there are further algebraic constraints hidden in the DAE. These are so-called hidden constraints. This situation is illustrated in the following example.

Example 4: Differential-algebraic equations II

Consider the differential-algebraic equations

$$\begin{aligned} \dot{x}_1 &= x_1 + 2x_2 + 3x_3 + \sin(t) + \cos(t), \\ \dot{x}_2 &= x_1 + x_2 + 2x_3 + \cos(t), \\ 0 &= x_1 - x_2 + \cos(t). \end{aligned}$$

Subtracting the second from the first equation yields $\dot{x}_1 - \dot{x}_2 = x_2 + x_3 + \sin(t)$. By differentiating the third equation we get $\dot{x}_1 - \dot{x}_2 = \sin(t)$. Thus, it follows that $0 = x_2 + x_3$, which is the hidden constraint.

Inserting the algebraic constraints into the second equation yields

$$\dot{x}_2 = x_2 - \cos(t) + x_2 - 2x_2 + \cos(t) = 0 \quad \Rightarrow \quad x_2 = c \quad \text{with} \quad c \in \mathbb{R}.$$

Hence, it follows that $x_1 = x_2 - \cos(t) = c - \cos(t)$ and $x_3 = -x_2 = -c$. Therefore, the general solution of the system is given by

$$x(t) = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix} \cos(t) + \begin{pmatrix} 1 \\ 1 \\ -1 \end{pmatrix} c \qquad \text{with} \quad c \in \mathbb{R}.$$

Thus, there is only 1 degree of freedom and possible initial values have to satisfy both algebraic equations, the explicitly given constraint $0 = x_1 - x_2 + \cos(t)$ and the hidden constraint $0 = x_2 + x_3$.

2.2 Basic definitions

First of all, it is necessary to formulate a general definition of DAEs:

Definition 1: Differential-algebraic equations

A set of equations of the form

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

with $F : \mathbb{D}_{\dot{x}} \times \mathbb{D}_x \times \mathbb{I} \to \mathbb{C}^m$ where $\mathbb{I} \subseteq \mathbb{R}$ is a compact interval and $\mathbb{D}_{\dot{x}}, \mathbb{D}_x \subseteq \mathbb{C}^n$ are open, $n, m \in \mathbb{N}$, is called a set of *differential-algebraic equations* (DAEs).

Furthermore, $x : \mathbb{I} \to \mathbb{C}^n$ are called *state variables* or *unknown variables* and $t \in \mathbb{I}$ is called the *independent variable.*

Note that this most general form of DAEs includes ODEs ($\dot{x} = f(t, x)$) as well as purely algebraic equations ($0 = f(t, x)$) as special cases. Other special forms of DAEs are:

- *Linear DAEs with constant coefficients*:

  $$E\dot{x}(t) = Ax(t) + f(t), \tag{2.2}$$

  with $E, A \in \mathbb{C}^{m,n}$ and $f : \mathbb{I} \to \mathbb{C}^m$.

- *Linear DAEs with variable coefficients*:

  $$E(t)\dot{x}(t) = A(t)x(t) + f(t), \tag{2.3}$$

  with $E, A : \mathbb{I} \to \mathbb{C}^{m,n}$ and $f : \mathbb{I} \to \mathbb{C}^m$.

- *Quasi-linear DAEs*:

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

  with $E : \mathbb{C}^n \times \mathbb{I} \to \mathbb{C}^{m,n}$ and $f : \mathbb{C}^n \times \mathbb{I} \to \mathbb{C}^m$.

Uniqueness of solutions is usually considered in the context of initial value problems:

Definition 2: Initial value problem

If in addition to the DAE (2.1) an *initial condition*

$$x(t_0) = x_0 \tag{2.5}$$

with given $t_0 \in \mathbb{I}$ and $x_0 \in \mathbb{C}^n$ is prescribed, we have an *initial value problem* (IVP) where x_0 is called the *initial value.*

In the following, only the classical solvability concept is considered:

Definition 3: Solution of DAEs

A function $x : \mathbb{I} \to \mathbb{C}^n$ is called a *solution of the DAE* (2.1) if x is continuously differentiable and satisfies (2.1) pointwise for all $t \in \mathbb{I}$.

A solution x of the DAE (2.1) is called a *solution of the IVP* (2.1), (2.5) if x furthermore satisfies the initial condition (2.5).

Definition 4: Consistency

The DAE (2.1) is called *solvable* if there exists a solution of the DAE (2.1).

Values $y \in \mathbb{C}^n$ are called *consistent* to the DAE (2.1) if there exists a solution x of the DAE (2.1) and a $t \in \mathbb{I}$ with $x(t) = y$.

The initial values x_0 in (2.5) are called *consistent* if there exists a solution of the IVP (2.1), (2.5).

Example 5: Consistent initial values

In Example 1, we have seen that any initial value $x_0 \in \mathbb{R}^2$ is consistent.

In Example 2, there is only one consistent initial value $x_0 = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$.

In Example 3, we have seen that $x_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ is not consistent, and that $x_0 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ is consistent.

In general, consistent initial values for this example are of the form $x_0 = \begin{pmatrix} c \\ 1-c \end{pmatrix}$ with $c \in \mathbb{R}$.

2.3 Linear differential-algebraic equations with constant coefficients

In this section, we consider linear DAEs with constant coefficients of the form

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

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

$$x(t_0) = x_0 \qquad \text{with} \qquad t_0 \in \mathbb{I},\ x_0 \in \mathbb{C}^n. \tag{2.7}$$

2.3.1 Equivalence and regularity

Let $P \in \mathbb{C}^{m,m}$ and $Q \in \mathbb{C}^{n,n}$ be nonsingular. Then, multiplication of the DAE (2.6) with P from the left and a change of variables $\tilde{x} = Q^{-1}x$ yields:

$$E\dot{x} = Ax + f(t) \quad \Leftrightarrow \quad PE\dot{x} = PAx + Pf(t) \quad \Leftrightarrow \quad PEQ\dot{\tilde{x}} = PAQ\tilde{x} + Pf(t). \tag{2.8}$$

Thus, the DAE (2.6) can be transformed into another linear DAE with constant coefficients of the form

$$\tilde{E}\dot{x} = \tilde{A}x + \tilde{f}(t) \quad \text{with} \quad \tilde{E} = PEQ,\ \tilde{A} = PAQ,\ \tilde{f} = Pf. \tag{2.9}$$

Note that the relation $x = Q\tilde{x}$ gives a 1-to-1 correspondence between the corresponding solution spaces. Hence, we can also investigate the transformed DAE (2.9) instead of the original DAE (2.6) with respect to existence and uniqueness of solutions.

This motivates the following definition:

Definition 5: Strong equivalence

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

$$E_2 = PE_1Q \qquad \text{and} \qquad A_2 = PA_1Q. \tag{2.10}$$

Remark:
As the name suggests, it can be shown that the relation introduced in Definition 5 is indeed an equivalence relation (see Kunkel/Mehrmann 2006, Lemma 2.2).

One important property of matrix pairs concerning the solution behaviour of the corresponding linear DAE is the following:

Definition 6: Regular and singular matrix pairs

Let $E, A \in \mathbb{C}^{m,n}$. The matrix pair (E, A) is called *regular* if $m = n$ and if the polynomial $\det(\lambda E - A)$ is not the zero polynomial.
Otherwise, (E, A) is called *singular*.

Lemma 1

If (E, A) with $E, A \in \mathbb{C}^{m,n}$ is strongly equivalent to a regular matrix pair, then (E, A) is regular.

Proof: See Kunkel/Mehrmann (2006), Lemma 2.6. ■

With the help of the equivalence relation for matrix pairs defined above, the next step consists in finding an appropriate canonical form in such a way that the properties and invariants of the corresponding DAE can be easily determined. Thus, in the following two subsections we discuss canonical forms of matrix pairs which can be traced back to the fundamental works of Weierstraß (1858) and Kronecker (1890), and the resulting solvability statements for the corresponding DAEs, first for regular and then for singular matrix pairs.

2.3.2 Solvability of regular linear differential-algebraic equations with constant coefficients

Definition 7: Nilpotent matrix, index of nilpotency

A matrix $A \in \mathbb{C}^{n,n}$ is called *nilpotent* if there exists a $v \in \mathbb{N}$ with $A^v = 0$ and $A^{v-1} \neq 0$. The number v is called the *index of nilpotency* of A.

Theorem 1: Jordan canonical form (JCF)

Let $A \in \mathbb{C}^{n,n}$. Then there exists a nonsingular matrix $P \in \mathbb{C}^{n,n}$ such that

$$P^{-1}AP = \operatorname{diag}(J_1, \dots, J_k) \quad \text{where} \quad J_i = \begin{pmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{pmatrix} \in \mathbb{C}^{m_i,m_i} \tag{2.11}$$

are so-called Jordan blocks with $\sum\limits_{i=1}^{k} m_i = n$.

The JCF is unique except for permutation of the Jordan blocks.

Proof: See Weintraub (2009). ■

Theorem 2: Weierstraß canonical form (WCF)

Let $E, A \in \mathbb{C}^{n,n}$ such that (E, A) is regular.
Then there exist a matrix $J \in \mathbb{C}^{d,d}$ and a nilpotent matrix $N \in \mathbb{C}^{n-d,n-d}$ with $d \in \mathbb{N}$, $0 \leq d \leq n$, both in Jordan canonical form such that (E, A) is strongly equivalent to

$$(\begin{pmatrix} I_d & 0 \\ 0 & N \end{pmatrix}, \begin{pmatrix} J & 0 \\ 0 & I_{n-d} \end{pmatrix}). \tag{2.12}$$

The WCF is unique except for permutation of the Jordan blocks in J and N.

Proof: See Kunkel/Mehrmann (2006), Theorem 2.7. ■

Example 6: Weierstraß canonical form

In Example 4, we have a linear DAE with constant coefficients with corresponding matrix pair

$$(E, A) = (\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 1 & -1 & 0 \end{pmatrix}).$$

(E, A) is regular since

$$\det(\lambda E - A) = \det \begin{pmatrix} \lambda - 1 & -2 & -3 \\ -1 & \lambda - 1 & -2 \\ -1 & 1 & 0 \end{pmatrix} = -\lambda$$

is not the zero polynomial.

With

$$P = \begin{pmatrix} 1 & -\frac{3}{2} & \frac{1}{2} \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} -2 & 0 & 3 \\ -2 & 0 & 2 \\ 2 & 1 & -2 \end{pmatrix}$$

we get the following WCF of (E, A)

$$PEQ = \left(\begin{array}{c|cc} 1 & 0 & 0 \\ \hline 0 & 0 & 1 \\ 0 & 0 & 0 \end{array}\right) = \left(\begin{array}{c|c} I & \\ \hline & N \end{array}\right), \quad PAQ = \left(\begin{array}{c|cc} 0 & 0 & 0 \\ \hline 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right) = \left(\begin{array}{c|c} J & \\ \hline & I \end{array}\right).$$

By transforming the regular matrix pair (E, A) into WCF the corresponding DAE $E\dot{x} = Ax + f$ can be cast into the form

$$\begin{aligned} \dot{x_1} &= Jx_1 + f_1 \\ N\dot{x_2} &= \quad x_2 + f_2 \end{aligned} \tag{2.13}$$

with x_1 and x_2 separated.

The first subproblem $\dot{x_1} = Jx_1 + f_1$ is a linear ODE whose corresponding IVP is uniquely solvable for $f \in \mathcal{C}(\mathbb{I}, \mathbb{C}^n)$. Therefore, we only have to investigate DAEs of the form

$$N\dot{x} = x + f \tag{2.14}$$

where $N \in \mathbb{C}^{n,n}$ is nilpotent.

Lemma 2

Consider the DAE (2.14). Let $v \in \mathbb{N}$ be the index of nilpotency of N. If $f \in \mathcal{C}^v(\mathbb{I}, \mathbb{C}^n)$, then (2.14) has the unique solution

$$x = -\sum_{i=0}^{v-1} N^i f^{(i)}. \tag{2.15}$$

Proof: See Kunkel/Mehrmann (2006), Lemma 2.8. ■

Remarks:

- The solution of the DAE (2.14) is unique without specifying initial values. Thus, the only consistent initial condition at t_0 is given by the value of x from (2.15) at t_0.
- Formally, the inhomogeneity f has to be v times continuously differentiable to obtain a continuously differentiable solution x, but in general, this is an overly restrictive assumption.

The following theorem summarises the obtained results.

Theorem 3

Consider the DAE (2.6) with the initial condition (2.7). Let $E, A \in \mathbb{C}^{n,n}$ with (E, A) being regular and let $P, Q \in \mathbb{C}^{n,n}$ be nonsingular matrices which transform (E, A) into WCF, i.e.

$$PEQ = \begin{pmatrix} I & 0 \\ 0 & N \end{pmatrix}, \qquad PAQ = \begin{pmatrix} J & 0 \\ 0 & I \end{pmatrix}, \qquad Pf = \begin{pmatrix} \tilde{f}_1 \\ \tilde{f}_2 \end{pmatrix}. \tag{2.16}$$

Set

$$Q^{-1}x = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix}, \qquad Q^{-1}x_0 = \begin{pmatrix} \tilde{x}_{1,0} \\ \tilde{x}_{2,0} \end{pmatrix}. \tag{2.17}$$

Furthermore, let v be the index of nilpotency of N and let $f \in \mathcal{C}^v(\mathbb{I}, \mathbb{C}^n)$.

Then we have the following:

1) The DAE (2.6) is solvable with the general solution

$$x = Q \begin{pmatrix} \tilde{x}_1 \\ -\sum_{i=0}^{v-1} N^i \tilde{f}_2^{(i)} \end{pmatrix} \tag{2.18}$$

where $\tilde{x}_1$ is a solution of $\dot{\tilde{x}}_1 = J\tilde{x}_1 + \tilde{f}_1$.

2) An initial condition (2.7) is consistent if and only if

$$\tilde{x}_{2,0} = -\sum_{i=0}^{v-1} N^i \tilde{f}_2^{(i)}(t_0). \tag{2.19}$$

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

3) Any IVP (2.6)-(2.7) with consistent initial values has the unique solution (2.18) where $\tilde{x}_1$ satisfies the initial condition $\tilde{x}_1(t_0) = \tilde{x}_{1,0}$.

Example 7

Consider the linear DAE from Example 4:

$$\begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix}\dot{x} = \begin{pmatrix}1&2&3\\1&1&2\\1&-1&0\end{pmatrix}x + \begin{pmatrix}\sin(t)+\cos(t)\\ \cos(t)\\ \cos(t)\end{pmatrix} \quad\Leftrightarrow\quad E\dot{x} = Ax + f.$$

In the previous example, we had the WCF with

$$P = \begin{pmatrix}1&-\frac{3}{2}&\frac{1}{2}\\1&-1&0\\0&0&1\end{pmatrix}, \quad Q = \begin{pmatrix}-2&0&3\\-2&0&2\\2&1&-2\end{pmatrix}, \quad Q^{-1} = \begin{pmatrix}1&-\frac{3}{2}&0\\0&1&1\\1&-1&0\end{pmatrix}$$

such that we have the equivalent DAE in WCF

$$\tilde{E}\dot{\tilde{x}} = \tilde{A}\tilde{x} + \tilde{f} \quad \text{with} \quad \tilde{E} = PEQ = \left(\begin{array}{c|cc}1&0&0\\\hline 0&0&1\\0&0&0\end{array}\right), \quad \tilde{A} = PAQ = \left(\begin{array}{c|cc}0&0&0\\\hline 0&1&0\\0&0&1\end{array}\right),$$

$$\tilde{f} = Pf = \begin{pmatrix}\sin(t)\\ \sin(t)\\ \cos(t)\end{pmatrix}, \quad \tilde{x} = \begin{pmatrix}\tilde{x}_1\\ \tilde{x}_2\\ \tilde{x}_3\end{pmatrix} = Q^{-1}x = \begin{pmatrix}x_1 - \frac{3}{2}x_2\\ x_2 + x_3\\ x_1 - x_2\end{pmatrix}, \quad v = 2.$$

Solving the two subproblems yields:

$$\dot{\tilde{x}}_1 = \sin(t) \qquad \Rightarrow \qquad \tilde{x}_1 = k - \cos(t) \quad \text{with} \quad k \in \mathbb{R},$$

$$\begin{pmatrix}\tilde{x}_2\\ \tilde{x}_3\end{pmatrix} = -\sum_{i=0}^{v-1} N^i \tilde{f}_2^{(i)} = -\begin{pmatrix}1&0\\0&1\end{pmatrix}\begin{pmatrix}\sin(t)\\ \cos(t)\end{pmatrix} - \begin{pmatrix}0&1\\0&0\end{pmatrix}\begin{pmatrix}\cos(t)\\ -\sin(t)\end{pmatrix} = \begin{pmatrix}0\\ -\cos(t)\end{pmatrix}.$$

Back-transformation yields the general solution

$$x = Q\tilde{x} = \begin{pmatrix}-2&0&3\\-2&0&2\\2&1&-2\end{pmatrix}\begin{pmatrix}k-\cos(t)\\ 0\\ -\cos(t)\end{pmatrix} = \begin{pmatrix}-2k-\cos(t)\\ -2k\\ 2k\end{pmatrix} = \begin{pmatrix}c-\cos(t)\\ c\\ -c\end{pmatrix}$$

with $c = -2k \in \mathbb{R}$ which is the same result we obtained in Example 4.

For initial values x_0 with $t_0 = 0$ we get the consistency condition

$$\begin{pmatrix}\tilde{x}_{2,0}\\ \tilde{x}_{3,0}\end{pmatrix} \overset{!}{=} -\sum_{i=0}^{v-1} N^i \tilde{f}_2^{(i)}(0) = \begin{pmatrix}0\\ -1\end{pmatrix} \quad \Rightarrow \quad x_0 = Q\tilde{x}_0 = \begin{pmatrix}-2\tilde{x}_{1,0}-3\\ -2\tilde{x}_{1,0}-2\\ 2\tilde{x}_{1,0}+2\end{pmatrix}.$$

2.3.3 Solvability of singular linear differential-algebraic equations with constant coefficients

Consider linear DAEs with constant coefficients $E\dot{x} = Ax + f$ with (E, A) being singular, i.e. $m \neq n$ or $\det(\lambda E - A) = 0$.

Then, we have the following existence and uniqueness results for the corresponding IVP:

Theorem 4

Let $E, A \in \mathbb{C}^{m,n}$ such that (E, A) is singular. Then:

1) If $\operatorname{rank}(\lambda E - A) < n$ for all $\lambda \in \mathbb{C}$, the the homogeneous IVP

$$E\dot{x} = Ax, \qquad x(t_0) = 0 \tag{2.20}$$

has a nontrivial solution. Thus, the solution of the corresponding inhomogeneous IVP (2.6)-(2.7) is not unique.

2) If $\operatorname{rank}(\lambda E - A) = n$ for some $\lambda \in \mathbb{C}$ (and hence $m > n$), then there exist arbitrary smooth inhomogeneities f for which the corresponding DAE (2.6) is not solvable.
If the DAE (2.6) is solvable for the stated f, then the corresponding IVP (2.6)-(2.7) is uniquely solvable for every consistent initial value.

Proof: See Kunkel/Mehrmann (2006), Theorem 2.14. ■

Similar to the regular case there exists a canonical form for general matrix pairs (E, A) which is an extended version of the WCF:

Theorem 5: Kronecker canonical form (KCF)

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

$$P(\lambda E - A)Q = \operatorname{diag}(\mathcal{J}_{\rho_1}, \ldots, \mathcal{J}_{\rho_r}, \mathcal{N}_{\sigma_1}, \ldots, \mathcal{N}_{\sigma_s}, \mathcal{L}_{\varepsilon_1}, \ldots, \mathcal{L}_{\varepsilon_p}, \mathcal{M}_{\eta_1}, \ldots, \mathcal{M}_{\eta_q}) \tag{2.21}$$

with $r, s, p, q \in \mathbb{N}_0$ and where the block entries on the diagonal have the following properties:

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

$$\lambda \begin{pmatrix} 1 & & \\ & \ddots & \\ & & \ddots & \\ & & & 1 \end{pmatrix} - \begin{pmatrix} \lambda_j & 1 & & \\ & \lambda_j & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_j \end{pmatrix}. \tag{2.22}$$

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

$$\lambda \begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix} - \begin{pmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{pmatrix}. \tag{2.23}$$

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

$$\lambda \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}. \tag{2.24}$$

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

$$\lambda \begin{pmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{pmatrix} - \begin{pmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{pmatrix}. \tag{2.25}$$

The KCF is unique except for permutation of the blocks, i.e. the kind, size and number of the blocks are characteristic for the matrix pair (E, A).

Proof: See Gantmacher (1959), Chapter XII. ■

Remarks:

- For regular matrix pairs (E, A) the blocks $\mathcal{L}_{\varepsilon_j}$ and $\mathcal{M}_{\eta_j}$ do not exist. Then the KCF corresponds to the WCF with

$$\lambda I - J = \begin{pmatrix} \mathcal{J}_{\rho_1} & & \\ & \ddots & \\ & & \mathcal{J}_{\rho_r} \end{pmatrix}, \qquad \lambda N - I = \begin{pmatrix} \mathcal{N}_{\sigma_1} & & \\ & \ddots & \\ & & \mathcal{N}_{\sigma_s} \end{pmatrix}. \tag{2.26}$$

- Blocks $\mathcal{L}_{\varepsilon_j}$ of size 0×1 and blocks $\mathcal{M}_{\eta_j}$ of size 1×0 are possible.
- The block notation in KCF implies that a pair $(0, 0)$ of size 1×1 actually consists of two blocks, $\mathcal{L}_0$ of size 0×1 and $\mathcal{M}_0$ of size 1×0.

By transforming the matrix pair (E, A) into KCF the corresponding DAE $E\dot{x} = Ax + f$ becomes

$$\left.\begin{aligned} \dot{x_1} &= Jx_1 + f_1 \\ N\dot{x_2} &= \;\; x_2 + f_2 \end{aligned}\right\} \begin{array}{c}\text{regular part}\\ \hat{=}\ \text{WCF}\end{array} \tag{2.27}$$

$$\left.\begin{array}{l}
\hline
\left.\begin{array}{rcl}
\begin{pmatrix} | & \\ 0 & I_{\varepsilon_1} \\ | & \end{pmatrix}\dot{x}_3 &=& \begin{pmatrix} & | \\ I_{\varepsilon_1} & 0 \\ & | \end{pmatrix} x_3 + f_3 \\
&\vdots& \\
\begin{pmatrix} | & \\ 0 & I_{\varepsilon_p} \\ | & \end{pmatrix}\dot{x}_{p+2} &=& \begin{pmatrix} & | \\ I_{\varepsilon_p} & 0 \\ & | \end{pmatrix} x_{p+2} + f_{p+2}
\end{array}\right\} \mathcal{L}_* \\
\hline
\left.\begin{array}{rcl}
\begin{pmatrix} I_{\eta_1} \\ -0- \end{pmatrix}\dot{x}_{p+3} &=& \begin{pmatrix} -0- \\ I_{\eta_1} \end{pmatrix} x_{p+3} + f_{p+3} \\
&\vdots& \\
\begin{pmatrix} I_{\eta_q} \\ -0- \end{pmatrix}\dot{x}_{q+p+2} &=& \begin{pmatrix} -0- \\ I_{\eta_q} \end{pmatrix} x_{q+p+2} + f_{q+p+2}
\end{array}\right\} \mathcal{M}_*
\end{array}\right\} \text{singular part.} \tag{2.28}$$

The regular part has already been investigated in the previous subsection. Hence, it suffices to investigate DAEs of the form

$$\begin{pmatrix} | & \\ 0 & I \\ | & \end{pmatrix}\dot{x} = \begin{pmatrix} & | \\ I & 0 \\ & | \end{pmatrix} x + f \qquad \text{of size } m \times (m+1) \tag{2.29}$$

$$\text{and} \qquad \begin{pmatrix} I \\ -\ 0\ - \end{pmatrix}\dot{x} = \begin{pmatrix} -\ 0\ - \\ I \end{pmatrix} x + f \qquad \text{of size } (n+1) \times n. \tag{2.30}$$

Consider a DAE according to an $\mathcal{L}_*$ block, i.e. a DAE of the form

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \end{pmatrix}\begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_{m+1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & & \\ & \ddots & \ddots & \\ & & 1 & 0 \end{pmatrix}\begin{pmatrix} x_1 \\ \vdots \\ x_{m+1} \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_m \end{pmatrix}. \tag{2.31}$$

This is an underdetermined system. Therefore, x_j can be chosen arbitrarily for a $j \in \{1, \ldots, m+1\}$. Then x_k, $k \neq j$, can be determined by $\dot{x}_{k+1} = x_k + f_k$ by solving differential equations for $k = j, \ldots, m$ and by solving algebraic equations for $k = j-1, \ldots, 1$.

Consider a DAE according to an $\mathcal{M}_*$ block, i.e. a DAE of the form

$$\begin{pmatrix} 1 & & \\ 0 & \ddots & \\ & \ddots & 1 \\ & & 0 \end{pmatrix} \begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} 0 & & \\ 1 & \ddots & \\ & \ddots & 0 \\ & & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_{n+1} \end{pmatrix}. \tag{2.32}$$

This is an overdetermined system. We get $x_n = -f_{n+1}$, $x_j = \dot{x}_{j+1} - f_{j+1}$ for $j = n-1, \dots, 1$ and in addition the consistency condition $0 = \dot{x}_1 - f_1$.

In summary, a DAE in KCF is solvable if and only if the consistency condition for every $\mathcal{M}_*$ block is satisfied. Furthermore, free parameters result from the J block (initial values) and from $\mathcal{L}_*$ blocks (whole function x_j).

Example 8: Kronecker canonical form

Consider the linear DAE

$$\left.\begin{aligned} \dot{x}_1 - \dot{x}_2 &= 2x_1 - x_2, \\ \dot{x}_1 &= 2x_1 + x_2, \\ \dot{x}_2 &= -x_1 + 3x_2 + x_3 - x_4 + f_3, \\ 0 &= x_1 - x_2 - x_3 + x_4 + f_4. \end{aligned}\right\} \quad \Leftrightarrow \quad E\dot{x} = Ax + f,$$

with

$$E = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 3 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ 0 \\ f_3 \\ f_4 \end{pmatrix}.$$

Checking the regularity:

$$\begin{aligned} \det(\lambda E - A) &= \det \begin{pmatrix} \lambda - 2 & -\lambda + 1 & 0 & 0 \\ \lambda - 2 & -1 & 0 & 0 \\ 1 & \lambda - 3 & -1 & 1 \\ -1 & 1 & 1 & -1 \end{pmatrix} \\ &= \det \begin{pmatrix} \lambda - 2 & -\lambda + 1 \\ \lambda - 2 & -1 \end{pmatrix} \cdot \underbrace{\det \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}}_{=0} = 0 \end{aligned}$$

Thus, the DAE is singular with $\operatorname{rank}(\lambda E - A) < n$. It follows from Theorem 4 that the solution of a corresponding IVP will not be unique.

With

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 1 & -1 & 0 \\ 1 & -1 & 1 & 1 \end{pmatrix}, \qquad Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} = Q^{-1}$$

we get the transformed DAE in KCF

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \\ \dot{\tilde{x}}_3 \\ \dot{\tilde{x}}_4 \end{pmatrix} = \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ -f_3 \\ f_3 + f_4 \end{pmatrix}$$

$$(\quad \Leftrightarrow \quad PEQ\dot{\tilde{x}} = PAQ\tilde{x} + Pf, \quad \tilde{x} = Q^{-1}x)$$

with $P(\lambda E - A)Q = \operatorname{diag}(\mathcal{J}_2, \mathcal{N}_1, \mathcal{L}_0, \mathcal{M}_0)$.

Block $\mathcal{J}_2$:
Solving the linear ODE yields:

$$\begin{pmatrix} \dot{\tilde{x}}_1 \\ \dot{\tilde{x}}_2 \end{pmatrix} = \begin{pmatrix} 2 & 1 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} \quad \Rightarrow \quad \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} c_1 te^{2t} + c_2 e^{2t} \\ c_1 e^{2t} \end{pmatrix} \quad \text{with} \quad c_1, c_2 \in \mathbb{R}.$$

Block $\mathcal{N}_1$:
$\tilde{x}_3 = f_3$ (consistency condition for initial values)

Block $\mathcal{L}_0$:
underdetermined system: one variable for zero equations ⇒ $\tilde{x}_4$ arbitrary

Block $\mathcal{M}_0$:
overdetermined system: zero variables for one equation ⇒ consistency condition $f_4 = -f_3$

Back-transformation yields the general solution

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = Q\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_1 - \tilde{x}_2 - \tilde{x}_3 + \tilde{x}_4 \\ \tilde{x}_4 \end{pmatrix} = \begin{pmatrix} c_1 te^{2t} + c_2 e^{2t} \\ c_1 e^{2t} \\ c_1 te^{2t} + c_2 e^{2t} - c_1 e^{2t} - f_3 + \tilde{x}_4 \\ \tilde{x}_4 \end{pmatrix},$$

with $c_1, c_2 \in \mathbb{R}$.

An initial condition $x(t_0) = x_0$ is consistent if x_0 satisfies the consistency condition $\tilde{x}_3(t_0) = f_3(t_0)$. With $\tilde{x} = Q^{-1}x$ we get

$$x_1(t_0) - x_2(t_0) - x_3(t_0) + x_4(t_0) = f_3(t_0).$$

2.3.4 Characteristic quantities

Index of nilpotency

We have seen the KCF of a matrix pair (E, A) which is unique except for permutation of blocks. In that we have the block $\tilde{N} = \operatorname{diag}(\mathcal{N}_{\sigma_1}, \dots, \mathcal{N}_{\sigma_s}) = \lambda N - I$ with the nilpotent matrix N with index of nilpotency v. This index is independent of block permutations and therefore also a characteristic quantity of the associated linear DAE (2.6).

Definition 8: Index of nilpotency for linear DAEs

Let $E, A \in \mathbb{C}^{m,n}$ be given and let $\operatorname{diag}(\tilde{J}, \tilde{N}, \tilde{L}, \tilde{M})$ be the KCF of $(\lambda E - A)$ with $\tilde{N} = \lambda N - I$ where N is nilpotent with index of nilpotency v_n.

Then v_n is also called the *index of nilpotency* of the DAE $E\dot{x} = Ax + f$. If $\tilde{N}$ is not present in the KCF, then $v_n = 0$.

In Lemma 2 we have seen that the index of nilpotency determines the smoothness requirements on the inhomogenity f in order to obtain a continuously differentiable solution x. The index of nilpotency v_n also has another meaning. Consider one block of $N\dot{x} = x + f$:

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix} \begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_n \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \\ f_n \end{pmatrix} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \text{differential equations} \\ \} = h_0(x,t) \text{ algebraic constraint} \end{array} \tag{2.33}$$

First differentiation of the algebraic constraints yields

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & -1 \end{pmatrix} \begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ 0 \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \\ \dot{f}_n \end{pmatrix}. \tag{2.34}$$

Transformation from left yields

$$\begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & \\ & & 0 & 1 \\ & & & 0 \end{pmatrix} \begin{pmatrix} \dot{x}_1 \\ \vdots \\ \dot{x}_{n-1} \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{n-1} \\ x_{n-1} \end{pmatrix} + \begin{pmatrix} f_1 \\ \vdots \\ f_{n-1} \\ f_{n-1} + \dot{f}_n \end{pmatrix} \begin{array}{l} \\ \\ \\ \} = h_1(x,t) \end{array} \tag{2.35}$$

Thus, a new algebraic constraint $h_1(x,t)$ is obtained which is contained in the DAE $N\dot{x} = x + f$, but hidden. $h_1(x,t)$ is called *hidden constraint of level 1*.

Continuing the differentiation and transformation procedure yields after i steps

$$\begin{pmatrix}0&1&&\\&\ddots&\ddots&\\&&0&1\\&&&0\end{pmatrix}\begin{pmatrix}\dot{x}_1\\\vdots\\\dot{x}_{n-1}\\\dot{x}_n\end{pmatrix}=\begin{pmatrix}x_1\\\vdots\\x_{n-1}\\x_{n-i}\end{pmatrix}+\begin{pmatrix}f_1\\\vdots\\f_{n-1}\\\sum\limits_{j=0}^{i} f_{n-j}^{(i-j)}\end{pmatrix}\quad\Big\}=h_i(x,t). \tag{2.36}$$

Thus, a new algebraic constraint $h_i(x,t)$ is obtained after i differentiations of (parts of) the DAE $N\dot{x}=x+f$ which is called *hidden constraint of level i.*

Finally, after n $(=v_n)$ differentiations and transformations we get

$$\begin{pmatrix}0&1&&\\&\ddots&\ddots&\\&&0&1\\-1&&&0\end{pmatrix}\begin{pmatrix}\dot{x}_1\\\vdots\\\dot{x}_{n-1}\\\dot{x}_n\end{pmatrix}=\begin{pmatrix}x_1\\\vdots\\x_{n-1}\\0\end{pmatrix}+\begin{pmatrix}f_1\\\vdots\\f_{n-1}\\\sum\limits_{j=0}^{n-1} f_{n-j}^{(n-j)}\end{pmatrix}. \tag{2.37}$$

By multiplying this DAE with $P=\begin{pmatrix}0&&&-1\\1&\ddots&&\\&\ddots&\ddots&\\&&1&0\end{pmatrix}$ from the left we get an ODE

$$\begin{pmatrix}\dot{x}_1\\\vdots\\\dot{x}_{n-1}\\\dot{x}_n\end{pmatrix}=\begin{pmatrix}0\\x_1\\\vdots\\x_{n-1}\end{pmatrix}+\begin{pmatrix}-\sum\limits_{j=0}^{n-1} f_{n-j}^{(n-j)}\\f_1\\\vdots\\f_{n-1}\end{pmatrix}, \tag{2.38}$$

which is called *underlying ODE.*

Thus, we obtained hidden constraints $0=h_i(x,t)$ of level $i=1,\dots,v_n-1$. Since the solution x has to satisfy all (hidden) constraints $0=h_i(x,t)$ for all $t\in\mathbb{I}$ we get the so-called *solution manifold* $\{\mathbb{C}^n\times\mathbb{I}:0=h_0(x,t),\dots,0=h_{v_n-1}(x,t)\}\ni(x(t),t)$. Note that here, x is not the state of the DAE (2.6), but the state of the corresponding KCF.

In summary, the larger the index of nilpotency v_n, the stronger the smoothness requirements on f and the deeper some constraints are hidden.

The hidden constraints are not explicitly stated in the DAE, and thus they impose additional consistency conditions on the initial values and cause difficulties in the numerical solution of the DAE. Therefore, the index of nilpotency v_n is a measurement of the numerical difficulties in solving the DAE.

Differentiation index

Another index approach is most common in the extant literature which is a generalisation to nonlinear DAEs of the form $0 = F(\dot{x}, x, t)$. It is due to an idea by Campbell (1987; also see Campbell/Griepentrog 1995) to differentiate the original DAE. Thus, summarising the nonlinear DAE $0 = F(\dot{x}, x, t)$ and all its derivatives up to a certain order $l \in \mathbb{N}_0$ in one large system of equations yields the following definition:

Definition 9: Derivative array, inflated DAEs

Let $F : (\dot{x}, x, t) \mapsto F(\dot{x}, x, t)$ be a function in $\mathcal{C}^s(\mathbb{C}^n \times \mathbb{C}^n \times \mathbb{I}, \mathbb{C}^m)$, $s \in \mathbb{N}$.
Then, the associated *derivative array of order* l with $l \in \mathbb{N}_0$, $l \leq s$, has the form

$$\mathcal{F}_l(x, \dot{x}, \dots, x^{(l+1)}, t) = \begin{pmatrix} F(\dot{x}, x, t) \\ \frac{d}{dt} F(\dot{x}, x, t) \\ \vdots \\ \frac{d^l}{dt^l} F(\dot{x}, x, t) \end{pmatrix}. \tag{2.39}$$

Then, the corresponding *derivative array equations* or *inflated DAEs* with respect to the DAE $0 = F(\dot{x}, x, t)$ are given by $0 = \mathcal{F}_l(x, \dot{x}, \dots, x^{(l+1)}, t)$.

Example 9: Derivative array for linear DAEs with constant coefficients

Differentiating a linear DAE with constant coefficients ($E\dot{x} = Ax + f$) yields

$$\begin{pmatrix} E & 0 & 0 & \cdots \\ -A & E & 0 & \cdots \\ 0 & -A & E & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \dot{x} \\ \ddot{x} \\ \vdots \end{pmatrix} = \begin{pmatrix} A & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & \dots \\ \vdots & \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} x \\ \dot{x} \\ \vdots \end{pmatrix} + \begin{pmatrix} f \\ \dot{f} \\ \vdots \end{pmatrix},$$

i.e. the derivative array equations of order l for linear DAEs with constant coefficients are given by

$$M_l \cdot \dot{z}_l = N_l \cdot z_l + g_l$$

where

$$(M_l)_{ij} = \begin{cases} E & \text{for } i = j \\ -A & \text{for } i = j + 1 \\ 0 & \text{else} \end{cases}, \qquad (N_l)_{ij} = \begin{cases} A & \text{for } i = j = 0 \\ 0 & \text{else} \end{cases}, \qquad i, j = 0, \dots, l,$$

$$(z_l)_j = x^{(j)}, \qquad (g_l)_j = f^{(j)}, \qquad j = 0, \dots, l.$$

The matrix pair (M_l, N_l) with $M_l, N_l \in \mathbb{C}^{m(l+1), n(l+1)}$ is called *inflated pair*.

Based on the derivative array, the differentiation index is defined as follows (see Gear 1988; Campbell/Gear 1995; Steinbrecher 2006):

Definition 10: Differentiation index (d-index), underlying ODE

Suppose that the DAE (2.1) is a solvable DAE. Let $\mathcal{F}_l(x, \dot{x}, w, t)$ be the corresponding derivative array of order l where $w = (\ddot{x}, \dots, x^{(l+1)})$. Let $\dot{x}$ be considered locally as an algebraic variable y.

The smallest number $v_d \in \mathbb{N}_0$ (if it exists) for which y is uniquely determined by (x, t) and $0 = \mathcal{F}_{v_d}(x, y, w, t)$ for all consistent values as $y = \Phi(x, t)$ is called the *differentiation index (d-index)* of the DAE (2.1).

With $y = \dot{x}$ we have the *underlying ODE* $\dot{x} = \Phi(x, t)$.

In other words, the d-index can be seen as a measurement of how far the DAE is away from an ODE.

Note that the concept of the d-index is only suited for square systems with unique solutions since a DAE with a free solution component cannot lead to an ODE which is uniquely solvable for consistent initial values (see Kunkel/Mehrmann 2006, p. 96).

For linear DAEs with constant coefficients, a simpler characterisation of the d-index can be formulated with the help of the following definition:

Definition 11: 1-full

A block matrix $M \in \mathbb{C}^{kn,ln}$ is called *1-full* (with respect to the block structure built from $n \times n$ matrices) if and only if there exist a nonsingular matrix $R \in \mathbb{C}^{kn,kn}$ and a matrix $H \in \mathbb{C}^{(k-1)n,(l-1)n}$ such that

$$RM = \begin{pmatrix} I_n & 0 \\ 0 & H \end{pmatrix}. \tag{2.40}$$

Lemma 3

Let a matrix pair (E, A) be given with the inflated pair (M_l, N_l). Then, the d-index v_d of the DAE $E\dot{x} = Ax + f$ is the smallest number $v_d \in \mathbb{N}_0$ for which M_{v_d} is 1-full.

Example 10: Differentiation index

In Example 1, we have a linear ODE. Hence, the d-index is 0 since the derivative array equations of order 0 are the ODE itself. Thus, it is uniquely solvable for $y = \dot{x}$:

$$y = \begin{pmatrix} 3 & -1 \\ 4 & -2 \end{pmatrix} x.$$

Consider the differential-algebraic equations from Example 4:

$$\begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{pmatrix}\dot{x} = \begin{pmatrix}1 & 2 & 3\\ 1 & 1 & 2\\ 1 & -1 & 0\end{pmatrix}x + \begin{pmatrix}\sin(t)+\cos(t)\\ \cos(t)\\ \cos(t)\end{pmatrix} \quad \Leftrightarrow \quad E\dot{x} = Ax + f.$$

Obviously, the derivative array equations of order 0 are not uniquely solvable for $y = \dot{x}$ since E is singular.

The derivative array equations of order 1 with $x, y = \dot{x}, w = \ddot{x}$ are given by

$$\begin{pmatrix}1 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ -1 & -2 & -3 & 1 & 0 & 0\\ -1 & -1 & -2 & 0 & 1 & 0\\ -1 & 1 & 0 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}y\\ w\end{pmatrix} = \begin{pmatrix}1 & 2 & 3 & 0 & 0 & 0\\ 1 & 1 & 2 & 0 & 0 & 0\\ 1 & -1 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix} + \begin{pmatrix}\sin(t)+\cos(t)\\ \cos(t)\\ \cos(t)\\ \cos(t)-\sin(t)\\ -\sin(t)\\ -\sin(t)\end{pmatrix}.$$

They are not uniquely solvable for y only dependent on x, t without dependence on w. Thus, consider the derivative array equations of order 2 with $x, y = \dot{x}, w = \begin{pmatrix}\ddot{x}\\ \dddot{x}\end{pmatrix} = \begin{pmatrix}w_1\\ w_2\end{pmatrix}$:

$$\begin{pmatrix}1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ -1 & -2 & -3 & 1 & 0 & 0 & 0 & 0 & 0\\ -1 & -1 & -2 & 0 & 1 & 0 & 0 & 0 & 0\\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & -1 & -2 & -3 & 1 & 0 & 0\\ 0 & 0 & 0 & -1 & -1 & -2 & 0 & 1 & 0\\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}y\\ w_1\\ w_2\end{pmatrix} =$$

$$\begin{pmatrix}1 & 2 & 3 & 0 & 0 & 0 & 0 & 0 & 0\\ 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 & 0\\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}x\\ y\\ w_1\end{pmatrix} + \begin{pmatrix}\sin(t)+\cos(t)\\ \cos(t)\\ \cos(t)\\ \cos(t)-\sin(t)\\ -\sin(t)\\ -\sin(t)\\ -\sin(t)-\cos(t)\\ -\cos(t)\\ -\cos(t)\end{pmatrix}.$$

Multiplying with $R = \begin{pmatrix} -2 & 3 & -1 & 0 & 0 & -3 & 0 & 0 & 0 \\ -2 & 3 & -1 & 0 & 0 & -2 & 0 & 0 & 0 \\ 2 & -3 & 1 & -1 & 1 & 2 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -2 & 3 & -1 & 0 & 0 & -2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$ from the left yields

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -2 & -3 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} y \\ w_1 \\ w_2 \end{pmatrix} =$$

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x \\ y \\ w_1 \end{pmatrix} + \begin{pmatrix} \sin(t) \\ 0 \\ 0 \\ \cos(t) \\ 0 \\ 0 \\ -\sin(t) - \cos(t) \\ -\cos(t) \\ -\cos(t) \end{pmatrix}$$

which contains in the first three rows $y = (\sin(t), 0, 0)^\top$. Therefore, we get the d-index $v_d = 2$.

Note that the resulting system contains the constraint $0 = x_1 - x_2 + \cos(t)$ in the 4th row as well as the hidden constraint $0 = x_2 + x_3$ in the 5th row.

Remark:

In principle, it is possible to determine the hidden constraints from the KCF by back-transformation into the original states of the DAE or from the derivative array. But the determination of the KCF is very technical and involves the computation of JCFs which are very sensitive to perturbations, and thus this approach is numerically not stable. However, the determination of the hidden constraints from the whole derivative array is very intricate, and furthermore it is hard to distinguish the different levels. Therefore, we consider another index concept in the following, the so-called strangeness index.

Strangeness index

Definition 12: Corange, cokernel

Let $E \in \mathbb{C}^{m,n}$. Then, the *corange* and *cokernel* of E are defined as follows:

$$\operatorname{corange}(E) = \ker(E^*), \qquad \operatorname{coker}(E) = \operatorname{range}(E^*). \tag{2.41}$$

Remark:
It holds that $\operatorname{coker}(E)$ is the orthogonal complement of $\ker(E)$, i.e. $\ker(E) \oplus \operatorname{coker}(E) = \mathbb{C}^n$, and $\operatorname{corange}(E)$ is the orthogonal complement of $\operatorname{range}(E)$, i.e. $\operatorname{range}(E) \oplus \operatorname{corange}(E) = \mathbb{C}^m$.

For convenience, the following conventions are used:

1) We say a matrix $T \in \mathbb{C}^{n,k}$ is a basis of a subspace $\mathbb{U} \subseteq \mathbb{C}^n$ if this is valid for its columns, i.e. if $\operatorname{rank}(T) = k$ and $\operatorname{range}(T) = \mathbb{U}$.
2) The empty matrix $\emptyset_{n,0} \in \mathbb{C}^{n,0}$ is the only basis of $\{0\} \in \mathbb{C}^n$ with $\operatorname{rank}(\emptyset_{n,0}) = 0$ and $\det(\emptyset_{0,0}) = 1$.
3) For a given matrix $T \in \mathbb{C}^{n,k}$ with $\operatorname{rank}(T) = k$, we use T' to denote a matrix from $\mathbb{C}^{n,n-k}$ that completes T to a nonsingular matrix $[T \;\; T'] \in \mathbb{C}^{n,n}$.

Theorem 6: Strong canonical form

Let $E, A \in \mathbb{C}^{m,n}$ and introduce the matrices

$$\begin{aligned} T &: \text{basis of } \ker(E), \text{ i.e. } ET = 0,\\ Z &: \text{basis of } \operatorname{corange}(E), \text{ i.e. } Z^*E = 0,\\ V &: \text{basis of } \operatorname{corange}(Z^*AT), \text{ i.e. } V^*Z^*AT = 0. \end{aligned}$$

Then the quantities

$$\begin{array}{llll} r = \operatorname{rank}(E) & \text{(rank)}, & d = r - s & \text{(differential part)},\\ a = \operatorname{rank}(Z^*AT) & \text{(algebraic part)}, & u = n - r - a & \text{(undetermined variables)},\\ s = \operatorname{rank}(V^*Z^*AT') & \text{(strangeness)}, & v = m - r - a - s & \text{(vanishing equations)}, \end{array}$$

called *characteristic values*, are independent of the particular choice of T, V, Z and T' and invariant under strong equivalence.

Furthermore, the matrix pair (E, A) is strongly equivalent to the canonical form

$$\left(\begin{array}{cccc} s & d & a & u \\ I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array} , \begin{array}{ccccc} s & d & a & u & \\ 0 & A_{12} & 0 & A_{14} & s \\ 0 & A_{22} & 0 & A_{24} & d \\ 0 & 0 & I_a & 0 & a \\ I_s & 0 & 0 & 0 & s \\ 0 & 0 & 0 & 0 & v \end{array} \right). \tag{2.42}$$

Proof: The proof is similar to the one of Theorem 3.7 in Kunkel/Mehrmann (2006). In particular, the transformation matrices for the construction of the strong canonical form can be determined with the singular value decomposition (SVD) which is numerically stable. ■

By transforming the matrix pair (E, A) into the strong canonical form (2.42), the corresponding DAE $E\dot{x} = Ax + f$ becomes

$$\dot{x}_1 = A_{12}x_2 + A_{14}x_4 + f_1, \qquad \text{s} \tag{2.43a}$$

$$\dot{x}_2 = A_{22}x_2 + A_{24}x_4 + f_2, \qquad \text{d} \tag{2.43b}$$

$$0 = x_3 + f_3, \qquad \text{a} \tag{2.43c}$$

$$0 = x_1 + f_4, \qquad \text{s} \tag{2.43d}$$

$$0 = f_5. \qquad \text{v} \tag{2.43e}$$

Thus, there is an algebraic equation (2.43c) for x_3 (algebraic part) and a consistency condition (2.43e) for the inhomogeneity f_5 (vanishing equations). Furthermore, there is a differential equation (2.43b) for x_2 (differential part) with a possible free choice in x_4 (undetermined variables). However, there is a problematic coupling, which is called *strangeness*, due to an algebraic equation (2.43d) as well as a differential equation (2.43a) for x_1.

In order to eliminate the strangeness (i.e. to eliminate $\dot{x}_1$), the idea is to differentiate the algebraic equation (2.43d) and to add it to the differential equation (2.43a). Then, this *differentiation-elimination step* (DE-step) corresponds to a conversion into the *modified DAE*

$$0 = A_{12}x_2 + A_{14}x_4 + f_1 + \dot{f}_4, \qquad \text{s} \tag{2.44a}$$

$$\dot{x}_2 = A_{22}x_2 + A_{24}x_4 + f_2, \qquad \text{d} \tag{2.44b}$$

$$0 = x_3 + f_3, \qquad \text{a} \tag{2.44c}$$

$$0 = x_1 + f_4, \qquad \text{s} \tag{2.44d}$$

$$0 = f_5, \qquad \text{v} \tag{2.44e}$$

with corresponding *modified matrix pair*

$$\left(\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & A_{22} & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right) = (E_{\text{mod}}, A_{\text{mod}}). \tag{2.45}$$

Remarks:

- The linear DAE corresponding to (E, A) and the modified DAE corresponding to $(E_{\text{mod}}, A_{\text{mod}})$ have the same set of solutions since we can reverse the procedure by differentiating (2.44d) and subtracting it from (2.44a).
- The rank of the modified matrix E_{mod} is reduced by s, i.e. $\text{rank}(E_{\text{mod}}) = r - s = d$.

Because of the non-uniqueness of the strong canonical form, the following theorem is necessary to ensure that the modified DAE is still characteristic for the original problem.

Theorem 7

Let $E, A, \tilde{E}, \tilde{A} \in \mathbb{C}^{m,n}$ with (E, A) and $(\tilde{E}, \tilde{A})$ being strongly equivalent and in strong canonical form (2.42).

Then, the matrix pairs $(E_{\mathrm{mod}}, A_{\mathrm{mod}})$ and $(\tilde{E}_{\mathrm{mod}}, \tilde{A}_{\mathrm{mod}})$ obtained by the DE-step from (E, A) and $(\tilde{E}, \tilde{A})$, respectively, are also strongly equivalent.

Proof: The proof follows later from the corresponding theorem for linear DAEs with variable coefficients. ■

Now, Theorem 7 allows for the following iterative procedure of finding a strangeness-free formulation:

Procedure 1: Reduction to strangeness-free form

Consider a linear DAE of the form $E\dot{x} = Ax + f$ with $E, A \in \mathbb{C}^{m,n}$.

Starting from $(E_0, A_0) = (E, A)$, $f_0 = f$, iterate for $i = 0, 1, \dots$

1) Transform (E_i, A_i) into the strong canonical form $(\tilde{E}_i, \tilde{A}_i)$, f_i into $\tilde{f}_i$ and x_i into $\tilde{x}_i$ by using the nonsingular matrices $P_i \in \mathbb{C}^{m,m}$ and $Q_i \in \mathbb{C}^{n,n}$ with

$$\tilde{E}_i = P_i E_i Q_i, \qquad \tilde{A}_i = P_i A_i Q_i, \qquad \tilde{f}_i = P_i f_i, \qquad \tilde{x}_i = Q_i^{-1} x_i. \tag{2.46}$$

We get the characteristic values $r_i, a_i, s_i, d_i, u_i, v_i$. If $s_i = 0$, stop the procedure.

2) Transform $(\tilde{E}_i, \tilde{A}_i)$ into (E_{i+1}, A_{i+1}) and $\tilde{f}_i$ into f_{i+1} by the differentiation-elimination step (DE-step), i.e.

$$E_{i+1} = \tilde{E}_i - \begin{pmatrix} I_{s_i} & & & & \\ & 0 & & & \\ & & 0 & & \\ & & & 0 & \\ & & & & 0 \end{pmatrix}, \; f_{i+1} = \tilde{f}_i + \begin{pmatrix} \dot{\tilde{f}}_{i,4} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; A_{i+1} = \tilde{A}_i, \; x_{i+1} = \tilde{x}_i. \tag{2.47}$$

Procedure 1 determines a non-unique sequence of matrix pairs (E_i, A_i) due to the non-uniqueness of the strong canonical form, but a unique sequence of characteristic values (r_i, s_i, a_i) due to Theorem 7. Because of the relation $r_{i+1} = r_i - s_i$, we have $s_i = 0$ after finitely many iterations i. Then, the sequence (r_i, s_i, a_i) becomes stationary and the procedure stops. This gives rise to the definition of a new characteristic quantity of the matrix pair (E, A) and the corresponding DAE as follows:

Definition 13: Strangeness index

Let $E, A \in \mathbb{C}^{m,n}$ and let the sequence (r_i, s_i, a_i), $i \in \mathbb{N}_0$, be determined by Procedure 1.

Then, $v_s = \min\{i \in \mathbb{N}_0 : s_i = 0\}$ is called *strangeness index (s-index)* of the matrix pair (E, A) and of the DAE $E\dot{x} = Ax + f$.

In the case that $v_s = 0$, both the matrix pair (E, A) and the DAE $E\dot{x} = Ax + f$ are called *strangeness-free*. Furthermore, from the sequence (E_i, A_i, f_i) we obtain a *strangenesse-free formulation* $\tilde{E}_{v_s}\dot{x} = \tilde{A}_{v_s}x + \tilde{f}_{v_s}$.

Example 11: Strangeness index

Consider the differential algebraic equations from Example 4:

$$\begin{aligned} \dot{x}_1 &= x_1 + 2x_2 + 3x_3 + \sin(t) + \cos(t) \\ \dot{x}_2 &= x_1 + x_2 + 2x_3 + \cos(t) \\ 0 &= x_1 - x_2 + \cos(t) \end{aligned} \quad \Rightarrow \quad (\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 1 & 2 \\ 1 & -1 & 0 \end{pmatrix}) = (E_0, A_0).$$

1) Transformation into strong canonical form with

$$P_0 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

yields

$$\begin{aligned} \dot{\tilde{x}}_1 &= \tilde{x}_2 + \tilde{x}_3 + \sin(t) \\ \dot{\tilde{x}}_2 &= 2\tilde{x}_2 + 2\tilde{x}_3 \\ 0 &= \tilde{x}_1 + \cos(t) \end{aligned} \quad \Rightarrow \quad (\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 0 & 2 & 2 \\ 1 & 0 & 0 \end{pmatrix}) = (\tilde{E}_0, \tilde{A}_0)$$

with characteristic values $r_0 = 2$, $a_0 = 0$, $s_0 = 1$, $d_0 = 1$, $u_0 = 1$, $v_0 = 0$.

Thus, the DAE is not strangeness-free and we have to proceed.

2) The DE-step yields

$$\begin{aligned} 0 &= \tilde{x}_2 + \tilde{x}_3 \\ \dot{\tilde{x}}_2 &= 2\tilde{x}_2 + 2\tilde{x}_3 \\ 0 &= \tilde{x}_1 + \cos(t) \end{aligned} \quad \Rightarrow \quad (\begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 1 & 1 \\ 0 & 2 & 2 \\ 1 & 0 & 0 \end{pmatrix}) = (E_1, A_1).$$

1) Transformation into strong canonical form with

$$P_1 = \begin{pmatrix} -2 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q_1 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ -1 & 1 & 0 \end{pmatrix}$$

yields

$$\begin{aligned}\dot{\hat{x}}_1 &= 0\\ 0 &= \hat{x}_2\\ 0 &= \hat{x}_3 + \cos(t)\end{aligned} \qquad \Rightarrow \qquad \left(\begin{pmatrix} 1 & 0 & 0\\ 0 & 0 & 0\\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1 \end{pmatrix} \right) = (\tilde{E}_1, \tilde{A}_1)$$

with characteristic values $r_1 = 1$, $a_1 = 2$, $s_1 = 0$, $d_1 = 1$, $u_1 = 0$, $v_1 = 0$.

Therefore, we get the s-index $v_s = 1$.

Note that the state variables x are transformed in step 1). It holds that

$$\hat{x} = \begin{pmatrix} \hat{x}_1\\ \hat{x}_2\\ \hat{x}_3 \end{pmatrix} = Q_1^{-1} Q_0^{-1} x = \begin{pmatrix} 0 & 1 & 0\\ 0 & 1 & 1\\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} x_1\\ x_2\\ x_3 \end{pmatrix} = \begin{pmatrix} x_2\\ x_2 + x_3\\ x_1 - x_2 \end{pmatrix}.$$

The following theorem summarises the obtained results.

Theorem 8

Let v_s be the s-index of the DAE (2.6) and let $f \in \mathcal{C}^{v_s}(\mathbb{I}, \mathbb{C}^m)$.

Then, the DAE (2.6) is equivalent (in the sense that there is a 1-to-1 correspondence between the solution spaces via a nonsingular matrix) to a DAE of the form

$$\dot{\hat{x}}_1 = \hat{A}_{11}\hat{x}_1 + \hat{A}_{13}\hat{x}_3 + \hat{f}_1, \qquad d_{v_s} \tag{2.48a}$$

$$0 = \hat{x}_2 + \hat{f}_2, \qquad a_{v_s} \tag{2.48b}$$

$$0 = \hat{f}_3, \qquad v_{v_s} \tag{2.48c}$$

where $\hat{A}_{11} \in \mathbb{C}^{d_{v_s}, d_{v_s}}$ and $\hat{A}_{13} \in \mathbb{C}^{d_{v_s}, u_{v_s}}$ and the inhomogeneities $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ are determined from (E, A) and f, $f^{(1)}$, ..., $f^{(v_s)}$ using Procedure 1.

Corollary 1

Let v_s be the s-index of the DAE (2.6) and let $f \in \mathcal{C}^{v_s+1}(\mathbb{I}, \mathbb{C}^m)$.

Then we have the following:

1) The DAE (2.6) is solvable if and only if the v_{v_s} consistency conditions $\hat{f}_3 = 0$ are fulfilled.

2) An initial condition (2.7) is consistent if and only if it implies the a_{v_s} conditions $\hat{x}_2(t_0) = -\hat{f}_2(t_0)$.

3) Any IVP (2.6)-(2.7) with consistent initial values is uniquely solvable if and only if $u_{v_s} = 0$ holds.

Example 12

Consider the DAE from Example 8:

$$\begin{aligned} \dot{x}_1 - \dot{x}_2 &= 2x_1 - x_2 \\ \dot{x}_1 &= 2x_1 + x_2 \\ \dot{x}_2 &= -x_1 + 3x_2 + x_3 - x_4 + f_3 \\ 0 &= x_1 - x_2 - x_3 + x_4 + f_4 \end{aligned} \quad \Leftrightarrow \quad E\dot{x} = Ax + f \qquad (*)$$

with

$$E_0 = E = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad A_0 = A = \begin{pmatrix} 2 & -1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 3 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}, \quad f = \begin{pmatrix} 0 \\ 0 \\ f_3 \\ f_4 \end{pmatrix}.$$

Transformation into strong canonical form with

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 1 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

yields

$$\begin{aligned} \dot{\tilde{x}}_1 &= 2\tilde{x}_1 + \tilde{x}_2 \\ \dot{\tilde{x}}_2 &= 2\tilde{x}_2 \\ 0 &= \tilde{x}_3 + f_3 \\ 0 &= f_3 + f_4 \end{aligned} \quad \Rightarrow \quad \left(\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 2 & 1 & 0 & 0 \\ 0 & 2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right) = (\tilde{E}_0, \tilde{A}_0)$$

with characteristic values $r_0 = 2$, $a_0 = 1$, $s_0 = 0$, $d_0 = 2$, $u_0 = 1$, $v_0 = 1$.

Thus, the DAE is strangeness-free.

Solvability statements (see Corollary 1):

1) The DAE $(*)$ is solvable if and only if $0 = f_3 + f_4$.
2) An initial condition $x(t_0) = x_0$ is consistent if and only if $\hat{x}_2(t_0) = -\hat{f}_2(t_0)$. With

$$\tilde{x} = \left(\begin{array}{c} \tilde{x}_1 \\ \tilde{x}_2 \\ \hline \tilde{x}_3 \\ \hline \tilde{x}_4 \end{array}\right) = \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \\ \hat{x}_3 \end{pmatrix} = Q^{-1}x = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 1 & 1 & -1 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ -x_1 + x_2 + x_3 - x_4 \\ -x_2 + x_4 \end{pmatrix}$$

we get $-x_1(t_0) + x_2(t_0) + x_3(t_0) - x_4(t_0) = -f_3(t_0)$.

3) The corresponding IVP is not uniquely solvable since $u_0 = 1$. In particular, $\tilde{x}_4$ can be chosen arbitrarily.

Solving the transformed strangeness-free formulation yields

$$\begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_3 \\ \tilde{x}_4 \end{pmatrix} = \begin{pmatrix} c_1 t e^{2t} + c_2 e^{2t} \\ c_1 e^{2t} \\ -f_3 \\ \tilde{x}_4 \end{pmatrix} \quad \text{with} \quad c_1, c_2 \in \mathbb{R}.$$

Back-transformation yields the general solution

$$x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = Q\tilde{x} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \\ \tilde{x}_1 + \tilde{x}_3 + \tilde{x}_4 \\ \tilde{x}_2 + \tilde{x}_4 \end{pmatrix} = \begin{pmatrix} c_1 t e^{2t} + c_2 e^{2t} \\ c_1 e^{2t} \\ c_1 t e^{2t} + c_2 e^{2t} - f_3 + \tilde{x}_4 \\ c_1 e^{2t} + \tilde{x}_4 \end{pmatrix}.$$

Concluding remarks

The following summarises the obtained results in this section:

1) The index of nilpotency v_n is based on the Kronecker canonical form (KCF) for general matrix pairs or on the Weierstraß canonical form (WCF) for regular matrix pairs. The solvability statements and a solution approach are developed from the different blocks of these canonical forms. But the determination of the KCF is very technical and involves the computation of JCFs which are very sensitive to perturbations, and thus this approach is numerically not stable.
2) The d-index v_d is based on the derivative array equations of the DAE. This index approach is only applicable for square systems which are uniquely solvable, i.e. for regular matrix pairs, because otherwise the d-index does not exist. Furthermore, it is in principle possible to determine the hidden constraints from the derivative array, but this takes a large effort and leads to unnecessary smoothness requirements due to the use of the *whole* derivative array.
3) The s-index v_s is based on an iterative procedure which in each step first determines a strong canonical form of the DAE and subsequently performs a differentiation-elimination step which results in an equivalent strangeness-free formulation. Then, the corresponding solvability statements and a solution approach are developed from this strangeness-free formulation. Therefore, only necessary smoothness requirements are needed since only parts of the DAE are differentiated which results in an efficient procedure compared to the other approaches.
4) Comparing the solvability statements of Theorem 3 and 4 with Corollary 1 results in the following alternative characterisation of regular matrix pairs. A matrix pair (E, A) is regular if and only if the characteristic values satisfy $u_{v_s} = v_{v_s} = 0$ where v_s is the s-index.
Furthermore, regularity of the matrix pair (E, A) implies that the d-index v_d is well-defined and the relation between the two index concepts can be expressed as $v_s = \max\{0, v_d - 1\}$.

2.4 Linear differential-algebraic equations with variable coefficients

In this section, we consider linear DAEs with variable coefficients of the form

$$E(t)\dot{x}(t) = A(t)x(t) + f(t) \tag{2.49}$$

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

$$x(t_0) = x_0 \qquad \text{with} \qquad t_0 \in \mathbb{I},\ x_0 \in \mathbb{C}^n. \tag{2.50}$$

2.4.1 Equivalence and regularity

Considering matrix functions $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$, one might assume that demanding regularity of the matrix pair $(E(t), A(t))$ for all $t \in \mathbb{I}$ leads to unique solvability of the IVP similar to Theorem 3. However, the concept of regularity does not guarantee unique solvability of the IVP as the following two examples show (compare with Examples 3.1 and 3.2 in Kunkel/Mehrmann 2006):

Example 13

Let E, A and f be given by

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

Then it follows that

$$\det(\lambda E(t) - A(t)) = \det \begin{pmatrix} -\lambda t + 1 & -\lambda t^2 \\ -\lambda & \lambda t + 1 \end{pmatrix} = (-\lambda t + 1)(\lambda t + 1) + \lambda^2 t^2 = 1.$$

Thus, $(E(t), A(t))$ is regular for all $t \in \mathbb{I}$.

However, for any $c \in \mathcal{C}(\mathbb{I}, \mathbb{C})$ with $c(t_0) = 0$ the function

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

is a solution of the IVP $E(t)\dot{x}(t) = A(t)x(t)$, $x(t_0) = 0$ because

$$E(t)\dot{x}(t) = \begin{pmatrix} -t & t^2 \\ -1 & t \end{pmatrix} \left(\dot{c}(t) \begin{pmatrix} t \\ 1 \end{pmatrix} + c(t) \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) = \begin{pmatrix} -t \\ -1 \end{pmatrix} c(t) = A(t)x(t).$$

Hence, the IVP is not uniquely solvable despite of regularity.

Example 14

Let E, A and f be given by

$$E(t)=\begin{pmatrix}0 & 0\\ 1 & -t\end{pmatrix},\quad A(t)=\begin{pmatrix}-1 & t\\ 0 & 0\end{pmatrix},\quad f(t)=\begin{pmatrix}f_1(t)\\ f_2(t)\end{pmatrix}\in\mathcal{C}^2(\mathbb{I},\mathbb{C}^2),\quad \mathbb{I}=\mathbb{R}.$$

Then it follows that

$$\det(\lambda E(t)-A(t))=\det\begin{pmatrix}1 & -t\\ \lambda & -\lambda t\end{pmatrix}=-\lambda t+\lambda t=0.$$

Thus, $(E(t),A(t))$ is singular for all $t\in\mathbb{I}$.

However, with $x=\begin{pmatrix}x_1\\ x_2\end{pmatrix}$ we obtain from $E(t)\dot{x}(t)=A(t)x(t)+f(t)$ the two equations

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

The first equation gives $x_1(t)=tx_2(t)+f_1(t)$. Differentiating this equation and inserting it into the second equation yields $x_2(t)=f_2(t)-\dot{f}_1(t)$. Inserting this into the first equation gives $x_1(t)=tf_2(t)-t\dot{f}_1(t)+f_1(t)$. Thus, we have the unique solution

$$x(t)=\begin{pmatrix}x_1(t)\\ x_2(t)\end{pmatrix}=\begin{pmatrix}tf_2(t)-t\dot{f}_1(t)+f_1(t)\\ f_2(t)-\dot{f}_1(t)\end{pmatrix}.$$

Hence, every IVP with consistent initial values is uniquely solvable despite of singularity.

Thus, in derogation from the constant coefficients case, the two properties regularity of the matrix pair $(E(t),A(t))$ for all $t\in\mathbb{I}$ and unique solvability of the corresponding IVP are completely independent of each other. This behaviour can be explained by the inadequacy of the strong equivalence relation (2.10) for linear DAEs with variable coefficients.

Instead of transformations with nonsingular matrices, we should consider pointwise nonsingular matrix functions $P\in\mathcal{C}(\mathbb{I},\mathbb{C}^{m,m})$ and $Q\in\mathcal{C}(\mathbb{I},\mathbb{C}^{n,n})$. Then, the DAE (2.49) can be transformed into another linear DAE with variable coefficients by multiplying with P from the left and a change of variables $x=Q\tilde{x}$ as in the case of constant coefficients. Thus, we get $\dot{x}=Q\dot{\tilde{x}}+\dot{Q}\tilde{x}$, i.e. an additional term $\dot{Q}\tilde{x}$ due to the product rule. Hence, transformation of the DAE (2.49) yields:

$$\begin{aligned}E\dot{x}=Ax+f\quad &\Leftrightarrow\quad PE\dot{x}=PAx+Pf\\ &\Leftrightarrow\quad PEQ\dot{\tilde{x}}=(PAQ-PE\dot{Q})\tilde{x}+Pf.\end{aligned}\tag{2.51}$$

This results in a different kind of equivalence:

Definition 14: Global equivalence

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

$$E_2 = PE_1Q \qquad \text{and} \qquad A_2 = PA_1Q - PE_1\dot{Q}. \tag{2.52}$$

Remarks:

- It can be shown that the relation introduced in Definition 14 is indeed an equivalence relation (see Kunkel/Mehrmann 2006, Lemma 3.4).
- For P and Q constant and therefore $\dot{Q} = 0$, we obtain the strong equivalence.

In contrast to strong equivalence (see Lemma 1), regularity of a matrix pair $(E(t), A(t))$ for fixed $t \in \mathbb{I}$ is not invariant under global equivalence which is illustrated in the following example:

Example 15

In Example 13, we have a linear DAE with variable coefficients with corresponding pair of matrix functions

$$(E(t), A(t)) = (\begin{pmatrix} -t & t^2 \\ -1 & t \end{pmatrix}, \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}).$$

We have seen that $(E(t), A(t))$ is regular for all $t \in \mathbb{I}$.

With

$$P(t) = \begin{pmatrix} 0 & -1 \\ -1 & t \end{pmatrix}, \qquad Q(t) = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}$$

we obtain

$$\tilde{E} = PEQ = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \qquad \tilde{A} = PAQ - PE\dot{Q} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} - \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}.$$

Thus, $(E(t), A(t))$ is globally equivalent to the singular matrix pair $(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix})$.

For given matrices $\bar{P} \in \mathbb{C}^{m,m}$ and $\bar{Q}, \bar{R} \in \mathbb{C}^{n,n}$, and for a given $\bar{t} \in \mathbb{I}$, we can define the matrix functions

$$P(t) = \bar{P}, \qquad Q(t) = \bar{Q} + (t - \bar{t})\bar{R} \tag{2.53}$$

satisfying

$$P(\bar{t}) = \bar{P}, \qquad Q(\bar{t}) = \bar{Q}, \qquad \dot{Q}(\bar{t}) = \bar{R}. \tag{2.54}$$

Therefore, we get some freedom in the equivalence relation locally considered at a fixed point $\bar{t} \in \mathbb{I}$. This gives rise to the definition of the following local version of the equivalence relation:

Definition 15: Local equivalence

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

$$E_2 = PE_1Q \qquad \text{and} \qquad A_2 = PA_1Q - PE_1R. \tag{2.55}$$

Remarks:

- Again, it can be shown that the relation introduced in Definition 15 is indeed an equivalence relation (see Kunkel/Mehrmann 2006, Lemma 3.6).
- For $R = 0$ we obtain the strong equivalence. Hence, there are more transformations available for local equivalence compared to strong equivalence in order to simplify the structure of a given matrix pair.
- For $P = I_m$ and $Q = I_n$ we obtain $E_2 = E_1$ and $A_2 = A_1 - E_1R$. Thus, we can subtract multiples of columns of E_1 from arbitrary columns of A_1. So, we can eliminate parts of A_2 with the help of E_1.

2.4.2 Strangeness index approach

With the help of the local equivalence relation, we have the following theorem for matrix pairs in addition to Theorem 6:

Theorem 9: Local canonical form

Let $E, A \in \mathbb{C}^{m,n}$ and let r, a, s, d, u, v be the characteristic values of (E, A) as in Theorem 6.

Then, the matrix pair (E, A) is locally equivalent to the canonical form

$$\left(\begin{array}{c} \begin{array}{cccc} s & d & a & u \end{array} \\ \begin{pmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \begin{array}{c} \begin{array}{cccc} s & d & a & u \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array} \begin{array}{c} \\ s \\ d \\ a \\ s \\ v \end{array} \right). \tag{2.56}$$

Proof: See Kunkel/Mehrmann (2006), Theorem 3.7. ■

Definition 16: Local characteristic values

Let $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$ be matrix functions and let $\bar{t} \in \mathbb{I}$ be fixed. Then, the characteristic values r, a, s, d, u, v from Theorem 6 for $(\bar{E}, \bar{A})$ with $\bar{E} = E(\bar{t})$ and $\bar{A} = A(\bar{t})$ are called *local characteristic values* of (E, A) at $\bar{t}$.

Thus, for a pair of matrix functions $(E(t), A(t))$ we obtain local characteristic values r, a, s for every $t \in \mathbb{I}$, i.e. r, a, s form functions $r, a, s : \mathbb{I} \to \mathbb{N}_0$ of characteristic values.

Now, we want to construct a canonical form under global equivalence by first assuming that these functions are constant on $\mathbb{I}$, i.e. that the block matrices in the strong canonical form as well as in the local canonical form do not depend on $t \in \mathbb{I}$. We will see later what happens if the local characteristic values are not constant and how far the requirement of constancy actually is a loss of generality.

The restriction to constant local characteristic values allows amongst others the application of the following generalisation of the SVD to matrix functions:

Theorem 10: Smooth SVD

Let $E \in \mathcal{C}^l(\mathbb{I}, \mathbb{C}^{m,n})$, $l \in \mathbb{N}_0 \cup \{\infty\}$, with $\operatorname{rank}(E(t)) = r \in \mathbb{N}_0$ for all $t \in \mathbb{I}$.

Then, there exist pointwise unitary functions $U \in \mathcal{C}^l(\mathbb{I}, \mathbb{C}^{m,m})$ and $V \in \mathcal{C}^l(\mathbb{I}, \mathbb{C}^{n,n})$ such that

$$U^* E V = \begin{pmatrix} \Sigma & 0 \\ 0 & 0 \end{pmatrix} \tag{2.57}$$

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

Proof: See Kunkel/Mehrmann (2006), Theorem 3.9; or Steinbrecher (2006), Theorem 2.1.4. ■

Suitable matrix functions U and V in (2.57) can be determined numerically by using the following theorem:

Theorem 11

If $E \in \mathcal{C}^1(\mathbb{I}, \mathbb{C}^{m,n})$, then suitable matrix functions $U = [Z' \;\; Z]$ and $V = [T' \;\; T]$ in (2.57) can be obtained as solutions of the ODEs

$$\begin{aligned} \begin{pmatrix} Z'^* E \\ T^* \end{pmatrix} \dot{T} &= -\begin{pmatrix} Z'^* \dot{E} T \\ 0 \end{pmatrix}, & \begin{pmatrix} T^* \\ T'^* \end{pmatrix} \dot{T}' &= -\begin{pmatrix} \dot{T}^* T' \\ 0 \end{pmatrix}, \\ \begin{pmatrix} T'^* E^* \\ Z^* \end{pmatrix} \dot{Z} &= -\begin{pmatrix} T'^* \dot{E}^* Z \\ 0 \end{pmatrix}, & \begin{pmatrix} Z^* \\ Z'^* \end{pmatrix} \dot{Z}' &= -\begin{pmatrix} \dot{Z}^* Z' \\ 0 \end{pmatrix}, \end{aligned} \tag{2.58}$$

with initial values

$$T(t_0) = T_0, \quad T'(t_0) = T'_0, \quad Z(t_0) = Z_0, \quad Z'(t_0) = Z'_0, \tag{2.59}$$

satisfying

$$[Z'_0 \;\; Z_0]^* E(t_0) [T'_0 \;\; T_0] = \begin{pmatrix} \Sigma_0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \Sigma_0 \text{ nonsingular.} \tag{2.60}$$

Proof: See Kunkel/Mehrmann (2006), Corollary 3.10. ■

By using the smooth SVD, it is possible to construct the following global canonical form for pairs of matrix functions under global equivalence. On the one hand, since $R = \dot{Q}$ has to be satisfied in the local equivalence relation (2.55), we expect a more complicated canonical form than the local canonical form. On the other hand, since we can use time-dependent transformation matrices, we expect a simpler canonical form than the strong canonical form.

Theorem 12: Global canonical form

Let $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$ be matrix functions and suppose that $r(t) \equiv r$, $a(t) \equiv a$, $s(t) \equiv s$ for the local characteristic values $r(t)$, $a(t)$, $s(t)$ of $(E(t), A(t))$.

Then, (E, A) is globally equivalent to the canonical form

$$\left(\begin{array}{c} \begin{array}{cccc} s & d & a & u \end{array} \\ \begin{pmatrix} I_s & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \end{array}, \begin{array}{cc} \begin{array}{cccc} s & d & a & u \end{array} & \\ \begin{pmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} & \begin{array}{c} s \\ d \\ a \\ s \\ v \end{array} \end{array} \right) \tag{2.61}$$

where A_{ij} are (non-unique) matrix functions on $\mathbb{I}$.

Proof: See Kunkel/Mehrmann (2006), Theorem 3.11. ■

Remark:
The global canonical form (2.61) can also be obtained for linear DAEs with constant coefficients if variable transformation matrices are used, in particular in order to eliminate the block A_{22}. In general, the global canonical form (2.61) cannot be obtained with constant transformation matrices.

By transforming the pair of matrix functions (E, A) into the global canonical form (2.61) the corresponding DAE $E(t)\dot{x} = A(t)x + f(t)$ becomes

$$\dot{x}_1 = A_{12}(t)x_2 + A_{14}(t)x_4 + f_1(t), \qquad \text{s} \tag{2.62a}$$

$$\dot{x}_2 = A_{24}(t)x_4 + f_2(t), \qquad \text{d} \tag{2.62b}$$

$$0 = x_3 + f_3(t), \qquad \text{a} \tag{2.62c}$$

$$0 = x_1 + f_4(t), \qquad \text{s} \tag{2.62d}$$

$$0 = f_5(t). \qquad \text{v} \tag{2.62e}$$

Again, we have an algebraic equation (2.62c) for x_3, a consistency condition (2.62e) for the inhomogeneity f_5 and a differential equation (2.62b) for x_2 with a possible free choice in x_4. Furthermore, there is a problematic coupling due to an algebraic equation (2.62d) as well as a differential equation (2.62a) for x_1.

In order to eliminate this problematic coupling, we can apply the differentiation-elimination step in the same way as in the constant coefficient case to obtain the modified DAE

$$
\begin{aligned}
0 &= A_{12}(t)x_2 + A_{14}(t)x_4 + f_1(t) + \dot{f}_4(t), && \text{s} \qquad (2.63a)\\
\dot{x}_2 &= A_{24}(t)x_4 + f_2(t), && \text{d} \qquad (2.63b)\\
0 &= x_3 + f_3(t), && \text{a} \qquad (2.63c)\\
0 &= x_1 + f_4(t), && \text{s} \qquad (2.63d)\\
0 &= f_5(t), && \text{v} \qquad (2.63e)
\end{aligned}
$$

with corresponding modified pair of matrix functions

$$
\left(\begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & I_d & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & A_{12} & 0 & A_{14} \\ 0 & 0 & 0 & A_{24} \\ 0 & 0 & I_a & 0 \\ I_s & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \right) = (E_{\text{mod}}, A_{\text{mod}}). \tag{2.64}
$$

Again, the linear DAEs corresponding to (E, A) and $(E_{\text{mod}}, A_{\text{mod}})$ have the same set of solutions since we can reverse the procedure of differentiating and eliminating.

Because of the non-uniqueness of the global canonical form, the following theorem is necessary to ensure that the modified DAE is still characteristic for the original problem.

Theorem 13

Let $E, A, \tilde{E}, \tilde{A} \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$ with (E, A) and $(\tilde{E}, \tilde{A})$ being globally equivalent and in global canonical form (2.61).

Then, the pairs of matrix functions $(E_{\text{mod}}, A_{\text{mod}})$ and $(\tilde{E}_{\text{mod}}, \tilde{A}_{\text{mod}})$ obtained by the DE-step from (E, A) and $(\tilde{E}, \tilde{A})$, respectively, are also globally equivalent.

Proof: See Kunkel/Mehrmann (2006), Theorem 3.14. ■

Now, Theorem 13 allows an extension of Procedure 1 to linear DAEs with variable coefficients:

Procedure 2: Reduction to strangeness-free form

Consider a linear DAE with variable coefficients of the form $E(t)\dot{x} = A(t)x + f(t)$ with $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$.

Starting from $(E_0, A_0) = (E, A)$, $f_0 = f$, iterate for $i = 0, 1, \ldots$

1) Compute the local characteristic values $r_i(t)$, $a_i(t)$, $s_i(t)$ of (E_i, A_i) for all $t \in \mathbb{I}$.
If $s_i(t) \equiv 0$, stop the procedure.

2) If the local characteristic values are constant on $\mathbb{I}$, transform (E_i, A_i) into the global canonical form $(\tilde{E}_i, \tilde{A}_i)$, f_i into $\tilde{f}_i$ and x_i into $\tilde{x}_i$ by using pointwise nonsingular matrix functions $P_i \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,m})$ and $Q_i \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{n,n})$ with

$$\tilde{E}_i = P_i E_i Q_i, \qquad \tilde{A}_i = P_i A_i Q_i - P_i E_i \dot{Q}_i, \qquad \tilde{f}_i = P_i f_i, \qquad \tilde{x}_i = Q_i^{-1} x_i. \tag{2.65}$$

3) Transform $(\tilde{E}_i, \tilde{A}_i)$ into (E_{i+1}, A_{i+1}) and $\tilde{f}_i$ into f_{i+1} by the differentiation-elimination step (DE-step), i.e.

$$E_{i+1} = \tilde{E}_i - \begin{pmatrix} I_{s_i} & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \\ & & & 0 \end{pmatrix}, \; f_{i+1} = \tilde{f}_i + \begin{pmatrix} \dot{\tilde{f}}_{i,4} \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; A_{i+1} = \tilde{A}_i, \; x_{i+1} = \tilde{x}_i. \tag{2.66}$$

If Procedure 2 does not break down, i.e. all local characteristic values (r_i, s_i, a_i) are constant, then the procedure determines a non-unique sequence of pairs of matrix functions (E_i, A_i) and a unique sequence of characteristic values (r_i, s_i, a_i) which becomes stationary after finitely many iterations i. According to the definition of the s-index for linear DAEs with constant coefficients, the s-index of the pair of matrix functions (E, A) and the corresponding DAE are defined as follows:

Definition 17: Strangeness index

Let $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$ and let the sequence (r_i, s_i, a_i), $i \in \mathbb{N}_0$, be determined by Procedure 2. In particular, let r_i, s_i, a_i be constant on $\mathbb{I}$.

Then, $v_s = \min\{i \in \mathbb{N}_0 : s_i = 0\}$ is called *strangeness index (s-index)* of the pair of matrix functions (E, A) and of the DAE $E(t)\dot{x} = A(t)x + f(t)$.

In the case that $v_s = 0$, both the pair of matrix functions (E, A) and the DAE are called *strangeness-free*. Furthermore, from the sequence (E_i, A_i, f_i) we obtain a *strangenesse-free formulation* $\tilde{E}_{v_s}(t)\dot{x} = \tilde{A}_{v_s}(t)x + \tilde{f}_{v_s}(t)$.

Theorem 14

Let the s-index v_s of the DAE (2.49) be well-defined and let $f \in \mathcal{C}^{v_s}(\mathbb{I}, \mathbb{C}^m)$.

Then, the DAE (2.49) is equivalent (in the sense that there is a 1-to-1 correspondence between the solution spaces via a pointwise nonsingular matrix function) to a DAE of the form

$$\dot{\hat{x}}_1 = \hat{A}_{13}(t)\hat{x}_3 + \hat{f}_1(t), \qquad d_{v_s} \tag{2.67a}$$

$$0 = \hat{x}_2 + \hat{f}_2(t), \qquad a_{v_s} \tag{2.67b}$$

$$0 = \hat{f}_3(t), \qquad v_{v_s} \tag{2.67c}$$

where $\hat{A}_{13} \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{d_{v_s}, u_{v_s}})$ and the inhomogeneities $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ are determined from (E, A) and f, $f^{(1)}$, …, $f^{(v_s)}$ using Procedure 2.

According to the strangeness-free form (2.67), the solvability statements for linear DAEs with variable coefficients are exactly the same as for the constant coefficient case:

Corollary 2

Let the s-index v_s of the DAE (2.49) be well-defined and let $f \in \mathcal{C}^{v_s+1}(\mathbb{I}, \mathbb{C}^m)$.

Then we have the following:

1) The DAE (2.49) is solvable if and only if the v_{v_s} consistency conditions $\hat{f}_3 = 0$ are fulfilled.
2) An initial condition (2.50) is consistent if and only if it implies the a_{v_s} conditions $\hat{x}_2(t_0) = -\hat{f}_2(t_0)$.
3) Any IVP (2.49)-(2.50) with consistent initial values is uniquely solvable if and only if $u_{v_s} = 0$ holds.

Example 16

Consider the DAE $E(t)\dot{x} = A(t)x + f(t)$ from Example 13 with

$$E(t) = \begin{pmatrix} -t & t^2 \\ -1 & t \end{pmatrix}, \quad A(t) = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \quad f(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix}, \quad \mathbb{I} = \mathbb{R}.$$

Following Procedure 2 yields:

- $i = 0: \; E_0 = E, \, A_0 = A, \, f_0 = f.$

  1) **Local characteristic values:**

  $$r_0 = \operatorname{rank}(E_0) = 1 \quad \text{for all } t \in \mathbb{I}.$$

  Choose $T = \begin{pmatrix} t \\ 1 \end{pmatrix}$ as basis of $\ker(E_0)$ and $T' = \begin{pmatrix} 1 \\ -t \end{pmatrix}$. Choose $Z = \begin{pmatrix} 1 \\ -t \end{pmatrix}$ as basis of corange(E_0). Then it follows that:

  $$a_0 = \operatorname{rank}(Z^* A_0 T) = \operatorname{rank}(0) = 0 \quad \text{for all } t \in \mathbb{I}.$$

  Choose $V = 1$ as basis of corange$(Z^* A_0 T)$. Then it follows that:

  $$s_0 = \operatorname{rank}(V^* Z^* A_0 T') = \operatorname{rank}(-1 - t^2) = 1 \quad \text{for all } t \in \mathbb{I}.$$

  Thus, we have $(r_0, a_0, s_0) = (1, 0, 1)$ constant on $\mathbb{I}$. In particular, $s_0 \neq 0$, i.e. the DAE is not strangeness-free.

2) **Global canonical form:**

With

$$P_0 = \begin{pmatrix} 0 & -1 \\ -1 & t \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad \dot{Q}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

we obtain

$$\tilde{E}_0 = P_0 E_0 Q_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{A}_0 = P_0 A_0 Q_0 - P_0 E_0 \dot{Q}_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$\tilde{f}_0 = P_0 f_0 = \begin{pmatrix} -g_2 \\ -g_1 + tg_2 \end{pmatrix}, \quad \tilde{x} = Q_0^{-1} x = \begin{pmatrix} x_1 - tx_2 \\ x_2 \end{pmatrix}.$$

3) **DE-step:**

$$E_1 = \tilde{E}_0 - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \tilde{A}_0 = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

$$f_1 = \tilde{f}_0 + \begin{pmatrix} \dot{\tilde{f}}_{0,2} \\ 0 \end{pmatrix} = \begin{pmatrix} -g_2 - \dot{g}_1 + g_2 + t\dot{g}_2 \\ -g_1 + tg_2 \end{pmatrix} = \begin{pmatrix} -\dot{g}_1 + t\dot{g}_2 \\ -g_1 + tg_2 \end{pmatrix}.$$

- $i = 1$:

1) **Local characteristic values:**

$$r_1 = \operatorname{rank}(E_1) = 0 \quad \text{for all } t \in \mathbb{I}.$$

Choose $T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $T' = \emptyset_{2,0}$ and $Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then it follows that:

$$a_1 = \operatorname{rank}(Z^* A_1 T) = \operatorname{rank}(A_1) = 1 \quad \text{for all } t \in \mathbb{I}.$$

Choose $V = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ as basis of corange$(Z^* A_1 T)$. Then it follows that:

$$s_1 = \operatorname{rank}(V^* Z^* A_1 T') = \operatorname{rank}(\emptyset_{1,0}) = 0 \quad \text{for all } t \in \mathbb{I}.$$

Thus, we have $(r_1, a_1, s_1, d_1, u_1, v_1) = (0, 1, 0, 0, 1, 1)$ constant on $\mathbb{I}$. In particular, $s_1 = 0$, i.e. the procedure stops. Hence, the s-index is $v_s = 1$.

Thus, the corresponding strangeness-free formulation is of the form

$$\begin{aligned} 0 &= -\dot{g}_1 + t\dot{g}_2, \\ 0 &= \tilde{x}_1 - g_1 + tg_2. \end{aligned}$$

Hence, the DAE is solvable if and only if the consistency condition $0 = -\dot{g}_1 + t\dot{g}_2$ is fulfilled. An initial condition $x(t_0) = x_0$ is consistent if and only if $\tilde{x}_1(t_0) = g_1(t_0) - t_0 g_2(t_0)$. With

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = Q_0^{-1} x = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 - t x_2 \\ x_2 \end{pmatrix}$$

we get $x_1(t_0) - t_0 x_2(t_0) = g_1(t_0) - t_0 g_2(t_0)$.

Furthermore, the corresponding IVP is not uniquely solvable since $u_1 = 1$. In particular, $\tilde{x}_2$ can be chosen arbitrarily. Thus, the general solution of the DAE is given by

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = Q_0 \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g_1 - t g_2 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} g_1 - t g_2 + t\tilde{x}_2 \\ \tilde{x}_2 \end{pmatrix}.$$

Example 17

Consider the DAE $E(t)\dot{x} = A(t)x + f(t)$ from Example 14 with

$$E(t) = \begin{pmatrix} 0 & 0 \\ 1 & -t \end{pmatrix}, \quad A(t) = \begin{pmatrix} -1 & t \\ 0 & 0 \end{pmatrix}, \quad f(t) = \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix}, \quad \mathbb{I} = \mathbb{R}.$$

Following Procedure 2 yields:

- $i = 0: \quad E_0 = E,\ A_0 = A,\ f_0 = f.$

  1) **Local characteristic values:**

  $$r_0 = \operatorname{rank}(E_0) = 1 \quad \text{for all } t \in \mathbb{I}.$$

  Choose $T = \begin{pmatrix} t \\ 1 \end{pmatrix}$ as basis of $\ker(E_0)$ and $T' = \begin{pmatrix} 1 \\ -t \end{pmatrix}$. Choose $Z = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ as basis of $\operatorname{corange}(E_0)$. Then it follows that:

  $$a_0 = \operatorname{rank}(Z^* A_0 T) = \operatorname{rank}(0) = 0 \quad \text{for all } t \in \mathbb{I}.$$

  Choose $V = 1$ as basis of $\operatorname{corange}(Z^* A_0 T)$. Then it follows that:

  $$s_0 = \operatorname{rank}(V^* Z^* A_0 T') = \operatorname{rank}(-1 - t^2) = 1 \quad \text{for all } t \in \mathbb{I}.$$

  Thus, we have $(r_0, a_0, s_0) = (1, 0, 1)$ constant on $\mathbb{I}$, i.e. the same local characteristic values as in the previous example. In particular, $s_0 \neq 0$, i.e. the DAE is not strangeness-free.

  2) **Global canonical form:**
  With

  $$P_0 = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}, \quad Q_0 = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix}, \quad \dot{Q}_0 = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

we obtain

$$\tilde{E}_0 = P_0 E_0 Q_0 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{A}_0 = P_0 A_0 Q_0 - P_0 E_0 \dot{Q}_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$
$$\tilde{f}_0 = P_0 f_0 = \begin{pmatrix} g_2 \\ -g_1 \end{pmatrix}, \quad \tilde{x} = Q_0^{-1} x = \begin{pmatrix} x_1 - t x_2 \\ x_2 \end{pmatrix}.$$

3) **DE-step:**

$$E_1 = \tilde{E}_0 - \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad A_1 = \tilde{A}_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix},$$
$$f_1 = \tilde{f}_0 + \begin{pmatrix} \dot{\tilde{f}}_{0,2} \\ 0 \end{pmatrix} = \begin{pmatrix} g_2 - \dot{g}_1 \\ -g_1 \end{pmatrix}.$$

- $i = 1:$

1) **Local characteristic values:**

$$r_1 = \operatorname{rank}(E_1) = 0 \quad \text{for all } t \in \mathbb{I}.$$

Choose $T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, $T' = \emptyset_{2,0}$ and $Z = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Then it follows that:

$$a_1 = \operatorname{rank}(Z^* A_1 T) = \operatorname{rank}(A_1) = 2 \quad \text{for all } t \in \mathbb{I}.$$

Choose $V = \emptyset_{2,0}$ as basis of corange$(Z^* A_1 T)$. Then it follows that:

$$s_1 = \operatorname{rank}(V^* Z^* A_1 T') = \operatorname{rank}(\emptyset_{0,0}) = 0 \quad \text{for all } t \in \mathbb{I}.$$

Thus, we have $(r_1, a_1, s_1, d_1, u_1, v_1) = (0, 2, 0, 0, 0, 0)$ constant on $\mathbb{I}$. In particular, $s_1 = 0$, i.e. the procedure stops. Hence, the s-index is $v_s = 1$.

Thus, the corresponding strangeness-free formulation is of the form

$$0 = -\tilde{x}_2 + g_2 - \dot{g}_1,$$
$$0 = \tilde{x}_1 - g_1.$$

Hence, the DAE is solvable since there is no consistency condition. An initial condition $x(t_0) = x_0$ is consistent if and only if $\tilde{x}_1(t_0) = g_1(t_0)$ and $\tilde{x}_2(t_0) = g_2(t_0) - \dot{g}_1(t_0)$. With

$$\tilde{x} = \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = Q_0^{-1} x = \begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 - t x_2 \\ x_2 \end{pmatrix}$$

we get $x_1(t_0) - t_0 x_2(t_0) = g_1(t_0)$ and $x_2(t_0) = g_2(t_0) - \dot{g}_1(t_0)$.

Furthermore, the corresponding IVP is uniquely solvable since $u_1 = 0$ with the general solution given by

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = Q_0 \begin{pmatrix} \tilde{x}_1 \\ \tilde{x}_2 \end{pmatrix} = \begin{pmatrix} 1 & t \\ 0 & 1 \end{pmatrix} \begin{pmatrix} g_1 \\ g_2 - \dot{g}_1 \end{pmatrix} = \begin{pmatrix} g_1 + tg_2 - t\dot{g}_1 \\ g_2 - \dot{g}_1 \end{pmatrix}.$$

Remark:

The results obtained so far are valid only for constant local characteristic values. The following theorem on the rank of continuous matrix functions gives a characterisation of how restrictive this constancy assumption is.

Theorem 15

Let $\mathbb{I} \subseteq \mathbb{R}$ be a closed interval and $M \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$.

Then, there exist open intervals $\mathbb{I}_j \subseteq \mathbb{I}$, $j \in \mathbb{N}$, with

$$\overline{\bigcup_j \mathbb{I}_j} = \mathbb{I} \qquad \text{and} \qquad \mathbb{I}_j \cap \mathbb{I}_i = \emptyset \quad \text{for } i \neq j, \tag{2.68}$$

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

$$\operatorname{rank}(M(t)) = r_j \qquad \text{for all} \quad t \in \mathbb{I}_j, \tag{2.69}$$

i.e. the rank of $M(t)$ is constant on each subinterval $\mathbb{I}_j$, $j \in \mathbb{N}$.

Proof: See Campbell/Meyer (1979), Theorem 10.5.2. ■

The local characteristic values of a pair of matrix functions correspond to the rank of several matrices (see Theorem 6). Furthermore, the s-index is well-defined if the local characteristic values of all constructed pairs of matrix functions within Procedure 2 are constant. Therefore, as a consequence of Theorem 15, the s-index of a pair of matrix functions (E, A) restricted to the subinterval $\mathbb{I}_j$ is well-defined for all $j \in \mathbb{N}$. Thus, the s-index is well-defined on a dense subset of the given closed interval $\mathbb{I}$. But at an exceptional point, i.e. where any of the ranks changes, we cannot associate an s-index.

2.4.3 Derivative array approach

In the previous subsection, an iterative procedure for linear DAEs was developed in order to determine an equivalent strangeness-free formulation of the considered DAE. In this subsection, we want to investigate an approach which is not iterative and which is based on the derivative array.

The general definition of the derivative array and the corresponding inflated DAEs for nonlinear DAEs (2.39) is also valid here. For linear DAEs with variable coefficients it simplifies to:

Example 18: Derivative array for linear DAEs with variable coefficients

Differentiating a linear DAE with variable coefficients $(E(t)\dot{x} = A(t)x + f(t))$ l times yields

$$\begin{pmatrix} E & 0 & 0 & \cdots & 0 \\ \dot{E} - A & E & 0 & \cdots & 0 \\ \ddot{E} - 2\dot{A} & 2\dot{E} - A & E & \ddots & 0 \\ \vdots & & \ddots & \ddots & 0 \\ E^{(l)} - lA^{(l-1)} & \cdots & \cdots & l\dot{E} - A & E \end{pmatrix} \begin{pmatrix} \dot{x} \\ \ddot{x} \\ \vdots \\ x^{(l+1)} \end{pmatrix} = \begin{pmatrix} A & 0 & \cdots & 0 \\ \dot{A} & 0 & \cdots & 0 \\ \ddot{A} & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ A^{(l)} & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} x \\ \dot{x} \\ \vdots \\ x^{(l)} \end{pmatrix} + \begin{pmatrix} f \\ \dot{f} \\ \vdots \\ f^{(l)} \end{pmatrix},$$

i.e. the derivative array equations of order l for linear DAEs with variable coefficients are given by

$$M_l(t)\dot{z}_l(t) = N_l(t)z_l(t) + g_l(t)$$

where

$$(M_l)_{ij} = \binom{i}{j}E^{(i-j)} - \binom{i}{j+1}A^{(i-j-1)}, \quad i,j = 0,\dots,l, \qquad (z_l)_j = x^{(j)}, \quad j = 0,\dots,l,$$

$$(N_l)_{ij} = \begin{cases} A^{(i)} & \text{for } j = 0 \\ 0 & \text{else} \end{cases}, \qquad i,j = 0,\dots,l, \qquad (g_l)_j = f^{(j)}, \quad j = 0,\dots,l,$$

using the convention that $\binom{i}{j} = 0$ for $j > i$.

The pair of matrix functions (M_l, N_l) with $M_l, N_l \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m(l+1),n(l+1)})$ is called *inflated pair*.

Based on the inflated pair (M_l, N_l), $l \in \mathbb{N}_0$, the idea is to construct a DAE with the same solution set as the original DAE, but with better analytical properties, i.e. a strangeness-free DAE with a separated part which explicitly states all constraints of the problem.

By assuming a well-defined s-index v_s, the determination of an equivalent formulation of the DAE $E(t)\dot{x} = A(t)x + f(t)$ with the same solution set and of the form (2.67) by using only information from the inflated pair (M_{v_s}, N_{v_s}) is discussed in Kunkel/Mehrmann (2006, pp. 91-93). In particular, the construction of a matrix function Z_2 of size $(v_s+1)m \times a_{v_s}$ with

$$Z_2^* M_{v_s} = 0 \quad \text{and} \quad \operatorname{rank}\left(Z_2^* N_{v_s} \begin{pmatrix} I_n & 0 & \dots & 0 \end{pmatrix}^*\right) = a_{v_s}, \tag{2.70}$$

a matrix function T_2 of size $n \times (d_{v_s} + u_{v_s})$ with

$$Z_2^* N_{v_s} \begin{pmatrix} I_n & 0 & \dots & 0 \end{pmatrix}^* T_2 = 0 \quad \text{and} \quad \operatorname{rank}(ET_2) = d_{v_s}, \tag{2.71}$$

and a matrix function Z_1 of size $m \times d_{v_s}$ with

$$\operatorname{rank}(Z_1^* E T_2) = d_{v_s} \tag{2.72}$$

is described.

Hence, Z_2 extracts a_{v_s} linearly independent algebraic constraints from the derivative array equations, and Z_1 selects d_{v_s} differential equations such that the constructed pair of matrix functions

$$(\hat{E},\hat{A}) = \left(\begin{pmatrix}\hat{E}_1\\ 0\\ 0\end{pmatrix}, \begin{pmatrix}\hat{A}_1\\ \hat{A}_2\\ 0\end{pmatrix} \right) \begin{array}{l} d_{v_s}\\ a_{v_s}\\ v_{v_s}\end{array} \tag{2.73}$$

with entries

$$\hat{E}_1 = Z_1^* E, \qquad \hat{A}_1 = Z_1^* A, \qquad \hat{A}_2 = Z_2^* N_{v_s} \begin{pmatrix} I_n & 0 & \dots & 0\end{pmatrix}^* \tag{2.74}$$

has the same size as the original pair (E, A). Moreover, it can be shown that this pair of matrix functions is indeed strangeness-free independent of the choice of transformations Z_2, T_2, Z_1:

Theorem 16

Let the s-index v_s of (E, A) with $E, A \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{m,n})$ be well-defined with characteristic values (r_i, a_i, s_i), $i = 1, \dots, v_s$.

Then, every pair of matrix functions $(\hat{E}, \hat{A})$ constructed as in (2.73) has a well-defined s-index $\hat{v}_s = 0$ and the local characteristic values of $(\hat{E}(t), \hat{A}(t))$ are given by

$$(\hat{r}, \hat{a}, \hat{s}) = (d_{v_s}, a_{v_s}, 0) \tag{2.75}$$

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

Proof: See Kunkel/Mehrmann (2006), Theorem 3.32. ■

Remark:
By setting $\hat{f}_1 = Z_1^* f$ and $\hat{f}_2 = Z_2^* g_{v_s}$, we obtain the equations

$$\hat{E}_1(t)\dot{x} = \hat{A}_1(t)x + \hat{f}_1(t) \quad \text{and} \quad 0 = \hat{A}_2(t)x + \hat{f}_2(t) \tag{2.76}$$

from $M_{v_s}(t)\dot{z}_{v_s}(t) = N_{v_s}(t)z_{v_s}(t) + g_{v_s}(t)$.

Furthermore, by setting $\hat{f}_3 = 0$ and

$$\hat{f} = \begin{pmatrix}\hat{f}_1\\ \hat{f}_2\\ \hat{f}_3\end{pmatrix}, \quad \hat{E} = \begin{pmatrix}\hat{E}_1\\ 0\\ 0\end{pmatrix}, \quad \hat{A} = \begin{pmatrix}\hat{A}_1\\ \hat{A}_2\\ 0\end{pmatrix}, \tag{2.77}$$

we get the system

$$\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t). \tag{2.78}$$

Due to the construction of that system, every solution of the original DAE (2.49) is also a solution of (2.78), but by setting $\hat{f}_3 = 0$, we may convert an unsolvable problem into a solvable one.

Example 19

Consider the DAE $E(t)\dot{x} = A(t)x + f(t)$ from Example 13 with

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

In Example 16, we have seen that $v_s = 1$, $d_{v_s} = 0$, $a_{v_s} = 1$ and $v_{v_s} = 1$. Then, the inflated pair of level $v_s = 1$ is given by

$$M_1(t) = \begin{pmatrix} -t & t^2 & 0 & 0 \\ -1 & t & 0 & 0 \\ 0 & 2t & -t & t^2 \\ 0 & 2 & -1 & t \end{pmatrix}, \quad N_1(t) = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad g_1(t) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

We have $\operatorname{rank}(M_1(t)) = 2$ for all $t \in \mathbb{I}$. By choosing $Z_2^* = \begin{pmatrix} 1 & -t & 0 & 0 \end{pmatrix}$ it holds that

$$Z_2^* M_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}, \quad \operatorname{rank}\left(Z_2^* N_1 \begin{pmatrix} I_n & 0 \end{pmatrix}^* \right) = \operatorname{rank}\left(\begin{pmatrix} -1 & t \end{pmatrix} \right) = a_{v_s} = 1.$$

Since $d_{v_s} = 0$, Z_1^* is of size 0×2 and we get

$$\hat{E}(t) = \begin{pmatrix} Z_1^* E \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{A}(t) = \begin{pmatrix} Z_1^* A \\ Z_2^* N_1 \begin{pmatrix} I_n & 0 \end{pmatrix}^* \\ 0 \end{pmatrix} = \begin{pmatrix} -1 & t \\ 0 & 0 \end{pmatrix}, \quad \hat{f}(t) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Thus, the solution $x_1 = tx_2$ with arbitrary x_2 can be read off directly.

Example 20

Consider the DAE $E(t)\dot{x} = A(t)x + f(t)$ from Example 14 with

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

In Example 17, we have seen that $v_s = 1$, $d_{v_s} = 0$, $a_{v_s} = 2$ and $v_{v_s} = 0$. Then, the inflated pair of level $v_s = 1$ is given by

$$M_1(t) = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & -t & 0 & 0 \\ 1 & -t & 0 & 0 \\ 0 & -1 & 1 & -t \end{pmatrix}, \quad N_1(t) = \begin{pmatrix} -1 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad g_1(t) = \begin{pmatrix} f_1 \\ f_2 \\ \dot{f}_1 \\ \dot{f}_2 \end{pmatrix}.$$

We have $\operatorname{rank}(M_1(t)) = 2$ for all $t \in \mathbb{I}$. By choosing $Z_2^* = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \end{pmatrix}$ it holds that

$$Z_2^* M_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \operatorname{rank}\left(Z_2^* N_1 \begin{pmatrix} I_n & 0 \end{pmatrix}^* \right) = \operatorname{rank}\left(\begin{pmatrix} -1 & t \\ 0 & -1 \end{pmatrix} \right) = a_{v_s} = 2.$$

Since $d_{v_s} = 0$, Z_1^* is of size 0×2 and we get

$$\hat{E}(t) = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \quad \hat{A}(t) = Z_2^* N_1 \begin{pmatrix} I_n & 0 \end{pmatrix}^* = \begin{pmatrix} -1 & t \\ 0 & -1 \end{pmatrix}, \quad \hat{f}(t) = Z_2^* g_1 = \begin{pmatrix} f_1 \\ f_2 - \dot{f}_1 \end{pmatrix}.$$

Thus, from $\hat{E}(t)\dot{x} = \hat{A}(t)x + \hat{f}(t)$ it follows that

$$x_2 = f_2 - \dot{f}_1 \qquad \text{and} \qquad x_1 = t x_2 + f_1 = t f_2 - t\dot{f}_1 + f_1.$$

Now, we discuss under which conditions the original DAE (2.49) and the strangeness-free formulation (2.78) have the same solution set, i.e. when such a reformulation is possible without setting $\hat{f}_3 = 0$. Therefore, we formulate the following hypothesis that guarantees that the reformulation can be performed without the presence of a consistency condition for the inhomogeneity f or free solution components:

Hypothesis 1

There exist integers $\hat{v}$, $\hat{a}$ and $\hat{d}$ $(= n - \hat{a})$ such that the inflated pair $(M_{\hat{v}}, N_{\hat{v}})$ of the given pair of matrix functions (E, A) with $E, A \in \mathcal{C}^{\hat{v}}(\mathbb{I}, \mathbb{C}^{n,n})$ satisfies:

1) For all $t \in \mathbb{I}$ we have $\operatorname{rank}(M_{\hat{v}}(t)) = (\hat{v}+1)n - \hat{a}$ such that there exists a smooth matrix function Z_2 of size $(\hat{v}+1)n \times \hat{a}$ and pointwise full rank $\hat{a}$ satisfying $Z_2^* M_{\hat{v}} = 0$.
2) For all $t \in \mathbb{I}$ we have $\operatorname{rank}(\hat{A}_2(t)) = \hat{a}$ where $\hat{A}_2 = Z_2^* N_{\hat{v}} \begin{pmatrix} I_n & 0 & \dots & 0 \end{pmatrix}^*$ such that there exists a smooth matrix function T_2 of size $n \times \hat{d}$ and pointwise full rank $\hat{d}$ satisfying $\hat{A}_2 T_2 = 0$.
3) For all $t \in \mathbb{I}$ we have $\operatorname{rank}(E(t)T_2(t)) = \hat{d}$ such that there exists a smooth matrix function Z_1 of size $n \times \hat{d}$ and pointwise full rank $\hat{d}$ satisfying $\operatorname{rank}(Z_1^* E T_2) = \hat{d}$ for all $t \in \mathbb{I}$.

Remarks:

- If the s-index v_s is well-defined and $u_{v_s} = v_{v_s} = 0$, then Hypothesis 1 is satisfied with $\hat{v} = v_s$, $\hat{a} = a_{v_s}$ and $\hat{d} = d_{v_s}$.
- If Hypothesis 1 is satisfied, then we have a reduction of the derivative array equations $M_{\hat{v}}\dot{z} = N_{\hat{v}} z + g_{\hat{v}}$ to $\hat{E}\dot{x} = \hat{A}x + \hat{f}$ where
$$\hat{E} = \begin{pmatrix} \hat{E}_1 \\ 0 \end{pmatrix} = \begin{pmatrix} Z_1^* E \\ 0 \end{pmatrix}, \; \hat{A} = \begin{pmatrix} \hat{A}_1 \\ \hat{A}_2 \end{pmatrix} = \begin{pmatrix} Z_1^* A \\ Z_2^* N_{\hat{v}} \begin{pmatrix} I_n & 0 & \dots & 0 \end{pmatrix}^* \end{pmatrix}, \; \hat{f} = \begin{pmatrix} \hat{f}_1 \\ \hat{f}_2 \end{pmatrix} = \begin{pmatrix} Z_1^* f \\ Z_2^* g_{\hat{v}} \end{pmatrix}. \tag{2.79}$$
- Note that there is no change of basis in the state space, i.e. only transformations from the left are performed such that the state variables x are still the same.

By construction, it is immediately clear that if Hypothesis 1 is satisfied, any solution x of the original DAE (2.49) also solves the reduced system (2.79). The following theorem shows that also the inverse holds:

Theorem 17

Let (E, A) be a pair of matrix functions satisfying Hypothesis 1.

Then, x solves the DAE (2.49) if and only if x solves the DAE (2.79).

Proof: See Kunkel/Mehrmann (2006), Theorem 3.51. ■

As a consequence, we obtain the following characterisation of consistency of initial values as well as existence and uniqueness of solutions:

Theorem 18

Let (E, A) be a pair of matrix functions satisfying Hypothesis 1 with values $\hat{v}$, $\hat{d}$, $\hat{a}$ and let $E, A \in \mathcal{C}^{\hat{v}+1}(\mathbb{I}, \mathbb{C}^{n,n})$ and $f \in \mathcal{C}^{\hat{v}+1}(\mathbb{I}, \mathbb{C}^{n})$. Then we have the following:

1) An initial condition (2.50) is consistent if and only if it implies the $\hat{a}$ conditions

$$0 = \hat{A}_2(t_0)x_0 + \hat{f}_2(t_0). \tag{2.80}$$

2) Any IVP (2.49)-(2.50) with consistent initial values is uniquely solvable.

Proof: The proof follows from the previous theorem. ■

2.4.4 Differentiation index

The general definition of the d-index for nonlinear DAEs (see subsection 2.3.4) is also valid here.

Similar to linear DAEs with constant coefficients, a simpler characterisation of the d-index for linear DAEs with variable coefficients can be formulated with the help of the following definition:

Definition 18: Smoothly 1-full

A block matrix function $M \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{kn,ln})$ is called *smoothly 1-full* (with respect to the block structure built from $n \times n$ matrix functions) if there exists a pointwise nonsingular matrix function $R \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{kn,kn})$ such that

$$RM = \begin{pmatrix} I_n & 0 \\ 0 & H \end{pmatrix} \tag{2.81}$$

for some $H \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{(k-1)n,(l-1)n})$.

Lemma 4

Let $M \in \mathcal{C}(\mathbb{I}, \mathbb{C}^{kn,ln})$ have constant rank.

Then, M is smoothly 1-full if and only if M is pointwise 1-full.

Proof: See Kunkel/Mehrmann (2006), Lemma 3.36. ■

Note that the constant rank assumption in Lemma 4 cannot be removed (otherwise Theorem 10 cannot be applied).

Lemma 5

Let a pair of matrix functions (E, A) be given with the inflated pair (M_l, N_l), $l \in \mathbb{N}$. Then, the d-index v_d of the DAE $E(t)\dot{x} = A(t)x + f(t)$ is the smallest number $v_d \in \mathbb{N}_0$ for which M_{v_d} is pointwise 1-full and has constant rank.

Now, the next aim is to show that (except for some technical smoothness assumptions) the d-index is well-defined if and only if Hypothesis 1 is satisfied.

Theorem 19

Let (E, A) be a pair of matrix functions with a well-defined d-index v_d.

Then, (E, A) satisfies Hypothesis 1 with

$$\hat{v} = \max\{0, v_d - 1\}, \quad \hat{d} = n - \hat{a}, \quad \hat{a} = \begin{cases} 0 & \text{for } v_d = 0, \\ \operatorname{corank} M_{v_d - 1}(t) & \text{otherwise.} \end{cases} \tag{2.82}$$

Proof: See Kunkel/Mehrmann (2006), Theorem 3.50. ■

Theorem 20

Let (E, A) be a pair of matrix functions that satisfies Hypothesis 1 with characteristic values $\hat{v}$, $\hat{d}$ and $\hat{a}$.

Then, the d-index v_d is well-defined for (E, A). Furthermore, if $\hat{v}$ is chosen minimally, then it holds that

$$v_d = \begin{cases} 0 & \text{for } \hat{a} = 0, \\ \hat{v} + 1 & \text{otherwise.} \end{cases} \tag{2.83}$$

Proof: See Kunkel/Mehrmann (2006), Corollary 3.53. ■

Example 21

Consider the linear DAE

$$\begin{pmatrix} 0 & t \\ 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} f_1(t) \\ f_2(t) \end{pmatrix}, \qquad \mathbb{I} = [-1, 1],$$

(see Example 3.54 in Kunkel/Mehrmann 2006).

Obviously, the matrix function E has a rank drop at $t = 0$. Therefore, the s-index is not well-defined.

Examining the inflated pair of order 1

$$(M_1(t), N_1(t)) = (\begin{pmatrix} 0 & t & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & t \\ 0 & -1 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}),$$

we have that M_1 has a constant corank $\hat{a} = 2$. By choosing

$$Z_2^*(t) = \begin{pmatrix} 1 & 0 & 0 & t \\ 0 & 1 & 0 & 0 \end{pmatrix} \quad \text{with} \quad Z_2^* M_1 = 0,$$

we obtain

$$\hat{A}_2(t) = Z_2^* N_1 \begin{pmatrix} I_n & 0 & \dots & 0 \end{pmatrix}^* = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \hat{f}_2 = Z_2^* g_1 = \begin{pmatrix} f_1(t) + t\dot{f}_2(t) \\ f_2(t) \end{pmatrix}.$$

Thus, (E, A) satisfies Hypothesis 1 with $\hat{v} = 1$, $\hat{a} = 2$ and $\hat{d} = 0$. Hence, the d-index is well-defined with $v_d = 2$.

The reduced system $\hat{E}\dot{x} = \hat{A}x + \hat{f}$ is given by

$$\begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} x + \begin{pmatrix} f_1(t) + t\dot{f}_2(t) \\ f_2(t) \end{pmatrix}.$$

Hence, the DAE has the unique solution

$$x_1(t) = -(f_1(t) + t\dot{f}_2(t)), \qquad x_2(t) = -f_2(t)$$

in the entire interval $\mathbb{I} = [-1, 1]$.

Concluding remarks

The following summarises the obtained results in this section:

1) By assuming constant local characteristic values, a global canonical form of a given pair of matrix functions is derived. Based on this canonical form, an extension of Procedure 1 to linear DAEs with variable coefficients is developed which results in an equivalent strangeness-free formulation of the DAE. Going out from this procedure, the s-index is defined and the corresponding solvability statements and a solution approach are developed.

2) Furthermore, we have seen that the s-index is well-defined on a dense subset of the given closed interval $\mathbb{I}$, but we cannot associate an s-index at an exceptional point where any of the ranks changes. Therefore, an alternative approach is investigated which is not iterative and which is based on the derivative array.

3) Based on the inflated pair, the idea is to construct a DAE with the same solution set as the original DAE, but with better analytical properties, i.e. a strangeness-free DAE with a separated part which explicitly states all constraints of the problem. Therefore, Hypothesis 1 is formulated such that it guarantees that this reformulation can be performed without the presence of a consistency condition for the inhomogeneity or free solution components. In particular, part 1 of Hypothesis 1 ensures that we have $\hat{a}$ constraints and that Z_2 extracts these constraints from the derivative array. Part 2 then requires that these constraints are linearly independent and also excludes consistency conditions for the inhomogeneity f. The matrix function T_2 points to the so-called differential part of the unknown. Finally, part 3 requires that there are $\hat{d}$ equations in the original problem that yield an ODE for the differential part of the unknown if we eliminate the algebraic part, i.e. the other part of the unknown. These equations are selected by Z_1. From the obtained reduced system, a characterisation of consistency of initial values as well as existence and uniqueness of solutions can be read off.

4) Furthermore, we have seen that the d-index is well-defined if and only if Hypothesis 1 is satisfied. Therefore, both the concept of the d-index and the concept of Hypothesis 1 are equivalent approaches except for some smoothness assumptions. The main difference is that the d-index aims at a reformulation of the original problem as an ODE, whereas Hypothesis 1 is constructed with the intention to reformulate it as a strangeness-free DAE with the same solution set.

2.5 Nonlinear differential-algebraic equations

In this section, we consider general nonlinear DAEs, i.e. systems of the form

$$F(\dot{x}, x, t) = 0 \tag{2.84}$$

with $F \in \mathcal{C}(\mathbb{D}_{\dot{x}} \times \mathbb{D}_x \times \mathbb{I}, \mathbb{R}^n)$ where $\mathbb{D}_{\dot{x}}, \mathbb{D}_x \subseteq \mathbb{R}^n$ are open, possibly together with an initial condition

$$x(t_0) = x_0 \qquad \text{with} \qquad t_0 \in \mathbb{I},\ x_0 \in \mathbb{R}^n. \tag{2.85}$$

Note that in this section, we consider only real-valued and square systems $(m = n)$, i.e. the case where the number of equations is equal to the number of unknowns.

2.5.1 Strangeness index approach

Definition 19: Jacobians of the derivative array

Let $\mathcal{F}_l(x^{(l+1)}, \ldots, x^{(1)}, x, t)$ be the derivative array of order $l \in \mathbb{N}_0$ with respect to the DAE $0 = F(\dot{x}, x, t)$, i.e.

$$\mathcal{F}_l(x^{(l+1)}, \ldots, x^{(1)}, x, t) = \begin{pmatrix} F(\dot{x}, x, t) \\ \frac{d}{dt} F(\dot{x}, x, t) \\ \vdots \\ \frac{d^l}{dt^l} F(\dot{x}, x, t) \end{pmatrix}. \tag{2.86}$$

Then, the associated *Jacobians* (M_l, N_l) of the derivative array are defined by

$$\begin{aligned} M_l(x^{(l+1)}, \ldots, x^{(1)}, x, t) &= \mathcal{F}_{l;\dot{x},\ldots,x^{(l+1)}}(x^{(l+1)}, \ldots, x^{(1)}, x, t) \\ &= \begin{pmatrix} \mathcal{F}_{l;\dot{x}} & \mathcal{F}_{l;\ddot{x}} & \ldots & \mathcal{F}_{l;x^{(l+1)}} \end{pmatrix} (x^{(l+1)}, \ldots, x^{(1)}, x, t), \\ N_l(x^{(l+1)}, \ldots, x^{(1)}, x, t) &= \begin{pmatrix} -\mathcal{F}_{l;x} & 0 & \ldots & 0 \end{pmatrix} (x^{(l+1)}, \ldots, x^{(1)}, x, t). \end{aligned} \tag{2.87}$$

Definition 20: Set of solutions of the derivative array

Let $\mathcal{F}_l(x^{(l+1)}, \ldots, x^{(1)}, x, t)$ be the derivative array of order $l \in \mathbb{N}_0$ with respect to the DAE $0 = F(\dot{x}, x, t)$.

Then, the *set of solutions* of the derivative array is defined as

$$\mathbb{L}_l = \{(z_{(l+1)}, \ldots, z_0, t) \in \mathbb{R}^{ln} \times \mathbb{D}_{\dot{x}} \times \mathbb{D}_x \times \mathbb{I} :\ \mathcal{F}_l(z_{(l+1)}, \ldots, z_0, t) = 0\}. \tag{2.88}$$

The following hypothesis generalises Hypothesis 1 to nonlinear DAEs (2.84):

Hypothesis 2

There exist integers v, a and d $(= n - a)$ such that the set of solutions

$$\mathbb{L}_v = \{(x^{(v+1)}, \ldots, \dot{x}, x, t) \in \mathbb{R}^{vn} \times \mathbb{D}_{\dot{x}} \times \mathbb{D}_x \times \mathbb{I} : \mathcal{F}_v(x^{(v+1)}, \ldots, \dot{x}, x, t) = 0\} \tag{2.89}$$

associated with F is nonempty and such that the following holds:

1) $\operatorname{rank}(M_v(x^{(v+1)}, \ldots, \dot{x}, x, t)) = (v+1)n - a$ on $\mathbb{L}_v$ such that there exists a smooth matrix function Z_2 of size $(v+1)n \times a$ and pointwise full rank a satisfying $Z_2^* M_v = 0$ on $\mathbb{L}_v$.
2) $\operatorname{rank}(\hat{A}_2(x^{(v+1)}, \ldots, \dot{x}, x, t)) = a$ on $\mathbb{L}_v$ where $\hat{A}_2 = Z_2^* N_v \begin{pmatrix} I_n & 0 & \ldots & 0 \end{pmatrix}^*$ such that there exists a smooth matrix function T_2 of size $n \times d$ and pointwise full rank d satisfying $\hat{A}_2 T_2 = 0$.
3) $\operatorname{rank}(F_{\dot{x}}(\dot{x}, x, t) T_2(x^{(v+1)}, \ldots, \dot{x}, x, t)) = d$ on $\mathbb{L}_v$ such that there exists a smooth matrix function Z_1 of size $n \times d$ and pointwise full rank d satisfying $\operatorname{rank}(Z_1^* F_{\dot{x}} T_2) = d$.

Definition 21: Strangeness index

Given a nonlinear DAE (2.84), the smallest number v such that F satisfies Hypothesis 2 is called *strangeness index (s-index)* of (2.84).

If $v = 0$, then DAE (2.84) is called *strangeness-free*.

Remarks:

- The variables $x^{(i)}$ in $\mathbb{L}_v$ are treated locally independently as algebraic variables, i.e. the differential relation is ignored.
- The matrix functions Z_2, T_2 and Z_1 always exist locally under the constant rank assumptions. This follows from a generalisation of the smooth SVD (Theorem 10), see Theorem 4.3 in Kunkel/Mehrmann (2006).
- Hypothesis 2 does not require constant ranks in the whole space but only on the set of solutions $\mathbb{L}_v$. This is due to the fact that, in general, away from the solution, the constant rank assumptions in Hypothesis 1 do not hold for the linearisation of the nonlinear DAE (see Example 4.1 in Kunkel/Mehrmann 2006).
- Hypothesis 2 is invariant under the following equivalence transformations:
  1) **Change of variables:**

$$x = Q(t, \tilde{x}) \qquad \text{and} \qquad \tilde{F}(\dot{\tilde{x}}, \tilde{x}, t) = F(\dot{x}, x, t) = F(Q_t(t, \tilde{x}) + Q_{\tilde{x}}(t, \tilde{x})\dot{\tilde{x}}, Q(t, \tilde{x}), t)$$

where $Q \in \mathcal{C}(\mathbb{I} \times \mathbb{R}^n, \mathbb{R}^n)$ is sufficiently smooth, $Q(t, \cdot)$ is bijective for every $t \in \mathbb{I}$ and the Jacobian $Q_{\tilde{x}}(t, \tilde{x})$ is nonsingular for every $(t, \tilde{x}) \in \mathbb{I} \times \mathbb{R}^n$.

2) **Combination of equations:**

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

where $P \in \mathcal{C}(\mathbb{R}^n \times \mathbb{R}^n \times \mathbb{I} \times \mathbb{R}^n, \mathbb{R}^n)$ is sufficiently smooth, $P(\dot{x}, x, t, \cdot)$ is bijective with $P(\dot{x}, x, t, 0) = 0$ and the Jacobian $P_w(\dot{x}, x, t, w)$ is nonsingular for every $(\dot{x}, x, t, w) \in \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{I} \times \mathbb{R}^n$.

3) **Making the system autonomous:**

$$\tilde{F}(\dot{\tilde{x}}, \tilde{x}) = \begin{pmatrix} F(\dot{x}, x, t) \\ \dot{t} - 1 \end{pmatrix}, \qquad \tilde{x} = \begin{pmatrix} x \\ t \end{pmatrix}, \qquad \dot{\tilde{x}} = \begin{pmatrix} \dot{x} \\ \dot{t} \end{pmatrix}.$$

- If F satisfies Hypothesis 2, we obtain the reduced system

$$\hat{F}(\dot{x}, x, t) = \begin{pmatrix} \hat{F}_1(\dot{x}, x, t) \\ \hat{F}_2(x, t) \end{pmatrix} = 0 \tag{2.90}$$

by setting

$$\begin{aligned} \hat{F}_1(\dot{x}, x, t) &= Z_1^T F(\dot{x}, x, t), \\ \hat{F}_2(x, t) &= Z_2^T \mathcal{F}_v(x^{(v+1)}, \dots, \dot{x}, x, t) \end{aligned} \tag{2.91}$$

where $\hat{F}_2 = Z_2^T \mathcal{F}_v$ can be reformulated in such a way that it only depends on t and x (see Kunkel/Mehrmann 2006, p. 161f.). Furthermore, we have:

1) $\hat{F}$ is strangeness-free.
2) Every sufficiently smooth solution x of $F(\dot{x}, x, t)$ also solves the reduced system $\hat{F}(\dot{x}, x, t)$ (see Theorem 4.11 in Kunkel/Mehrmann 2006).

Without any further assumptions, it is not clear whether a solution of the reduced system also solves the original DAE (2.84). The following theorem gives sufficient conditions for this situation (at least locally):

Theorem 21

Consider the nonlinear DAE (2.84) with sufficiently smooth F satisfying Hypothesis 2 with values v, a, d and in addition with $v + 1$, a, d.

Then, for every $z^0_{v+1} \in \mathbb{L}_{v+1}$, the reduced system (2.90) has a unique solution satisfying the initial values given in z^0_{v+1}. Moreover, this solution locally solves the original DAE (2.84).

Proof: See Kunkel/Mehrmann (2006), Theorem 4.13. ■

Remarks:

- The local result can be globalised in the same way as for ODEs (see Theorem I.7.4 in Hairer et al. 1993).
- We have $z^0_{v+1} = (x_0^{(v+2)}, \dots, \dot{x}_0, x_0, t_0)$ where (t_0, x_0) are consistent initial values.

Example 22

Consider the nonlinear DAE

$$F(\dot{x},x,t) = \begin{pmatrix} \dot{x}_2 - x_1 \\ x_2 - x_1\dot{x}_1 + \dot{x}_1\dot{x}_2 \end{pmatrix}, \qquad \mathbb{I} = \mathbb{R}, \mathbb{D}_{\dot{x}} = \mathbb{D}_x = \mathbb{R}^2.$$

Checking Hypothesis 2 with $v = 0$ yields:

$$\begin{aligned}
\mathcal{F}_0 &= F = \begin{pmatrix} \dot{x}_2 - x_1 \\ x_2 - x_1\dot{x}_1 + \dot{x}_1\dot{x}_2 \end{pmatrix}, \\
\mathbb{L}_0 &= \{(\dot{x},x,t) \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R} :\ \dot{x}_2 - x_1 = 0,\ x_2 - x_1\dot{x}_1 + \dot{x}_1\dot{x}_2 = 0\} \\
&= \{(\dot{x},x,t) \in \mathbb{R}^2 \times \mathbb{R}^2 \times \mathbb{R} :\ x_1 = \dot{x}_2,\ x_2 = 0\}, \\
M_0 &= F_{\dot{x}} = \begin{pmatrix} 0 & 1 \\ \dot{x}_2 - x_1 & \dot{x}_1 \end{pmatrix} \overset{\text{on } \mathbb{L}_0}{=} \begin{pmatrix} 0 & 1 \\ 0 & \dot{x}_1 \end{pmatrix}, \qquad N_0 = -F_x = \begin{pmatrix} 1 & 0 \\ \dot{x}_1 & -1 \end{pmatrix}.
\end{aligned}$$

1) $\operatorname{rank}(M_0) = 1 \overset{!}{=} (v+1)n - a$ on $\mathbb{L}_0$. Thus, $a = 1$, $d = n - a = 1$ and we can choose $Z_2^T = \begin{pmatrix} \dot{x}_1 & -1 \end{pmatrix}$ with full rank satisfying $Z_2^T M_0 = 0$ on $\mathbb{L}_0$.

2) $\operatorname{rank}(Z_2^T N_0) = \operatorname{rank}(\begin{pmatrix} 0 & 1 \end{pmatrix}) = 1 = a$. We can choose $T_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ with full rank satisfying $Z_2^T N_0 T_2 = 0$.

3) $\operatorname{rank}(F_{\dot{x}} T_2) = \operatorname{rank}(\begin{pmatrix} 0 \\ 0 \end{pmatrix}) = 0 \neq 1 = d$ on $\mathbb{L}_0$. Thus, Hypothesis 2 is not satisfied for $v = 0$.

Checking Hypothesis 2 with $v = 1$ yields:

$$\mathcal{F}_1 = \begin{pmatrix} \dot{x}_2 - x_1 \\ x_2 - x_1\dot{x}_1 + \dot{x}_1\dot{x}_2 \\ \ddot{x}_2 - \dot{x}_1 \\ \dot{x}_2 - \dot{x}_1^2 - x_1\ddot{x}_1 + \ddot{x}_1\dot{x}_2 + \dot{x}_1\ddot{x}_2 \end{pmatrix}, \qquad N_1 = \begin{pmatrix} -\mathcal{F}_{1;x} & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \dot{x}_1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ \ddot{x}_1 & 0 & 0 & 0 \end{pmatrix},$$

$$\begin{aligned}
\mathbb{L}_1 &= \{(\ddot{x},\dot{x},x,t) \in \mathbb{R}^7 :\ x_1 = \dot{x}_2, x_2 = 0, \ddot{x}_2 - \dot{x}_1 = 0, \dot{x}_2 - \dot{x}_1^2 - x_1\ddot{x}_1 + \ddot{x}_1\dot{x}_2 + \dot{x}_1\ddot{x}_2 = 0\} \\
&= \{(\ddot{x},\dot{x},x,t) \in \mathbb{R}^7 :\ x_1 = 0,\ x_2 = 0,\ \dot{x}_2 = 0,\ \ddot{x}_2 = \dot{x}_1\},
\end{aligned}$$

$$M_1 = \mathcal{F}_{1;\dot{x},\ddot{x}} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ \dot{x}_2 - x_1 & \dot{x}_1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -2\dot{x}_1 + \ddot{x}_2 & 1 + \ddot{x}_1 & \dot{x}_2 - x_1 & \dot{x}_1 \end{pmatrix} \overset{\text{on } \mathbb{L}_1}{=} \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & \dot{x}_1 & 0 & 0 \\ -1 & 0 & 0 & 1 \\ -\dot{x}_1 & 1 + \ddot{x}_1 & 0 & \dot{x}_1 \end{pmatrix}.$$

1) $\operatorname{rank}(M_1) = 2 \overset{!}{=} (v+1)n - a$ on $\mathbb{L}_1$. Thus, $a = 2$, $d = n - a = 0$ and we can choose $Z_2^T = \begin{pmatrix} -\dot{x}_1 & 1 & 0 & 0 \\ -\ddot{x}_1 - 1 & 0 & -\dot{x}_1 & 1 \end{pmatrix}$ with full rank satisfying $Z_2^T M_0 = 0$ on $\mathbb{L}_1$.

2) $\operatorname{rank}(Z_2^T N_1 \begin{pmatrix} I_n & 0 \end{pmatrix}^*) = \operatorname{rank}(\begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix}) = 2 = a$. Thus, $T_2 \in \mathbb{R}^{2\times 0}$.

3) $\operatorname{rank}(F_{\dot{x}} T_2) = 0 = d$ and $Z_1^T \in \mathbb{R}^{0\times 2}$.

Thus, Hypothesis 2 is satisfied for $v = 1$, $a = 2$, $d = 0$. One can show that Hypothesis 2 is also satisfied for $v = 2$, $a = 2$, $d = 0$.

Then, the reduced system is given by

$$\hat{F}(\dot{x}, x, t) = \begin{pmatrix} \hat{F}_1(\dot{x}, x, t) \\ \hat{F}_2(x, t) \end{pmatrix} = \begin{pmatrix} Z_1^T F(\dot{x}, x, t) \\ Z_2^T \mathcal{F}_1(\ddot{x}, \dot{x}, x, t) \end{pmatrix} = Z_2^T \mathcal{F}_1(\ddot{x}, \dot{x}, x, t) = \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} \overset{!}{=} 0.$$

Hence, the given DAE has the unique solution $(x_1, x_2) = (0, 0)$.

2.5.2 Structured problems

In many industrial applications, the arising DAEs are of a special structure. Additional information about the structure of the system can simplify the analysis of such systems, e.g. the application of Hypothesis 2. In the following, we review results obtained from Hypothesis 2 by exploiting the structure of semi-explicit, semi-implicit and quasi-linear DAEs.

Semi-explicit DAEs

Definition 22: Semi-explicit DAE

The DAE (2.84) is called *semi-explicit* if F has the form

$$F(\dot{x}, x, t) = \begin{pmatrix} \dot{y} - f(y, z) \\ -g(y, z) \end{pmatrix} = 0 \tag{2.92}$$

where $x = (y, z)$ with $y \in \mathbb{R}^{n_y}$, $z \in \mathbb{R}^{n_z}$, $n = n_y + n_z$, $f \in \mathcal{C}(\mathbb{R}^{n_y} \times \mathbb{R}^{n_z}, \mathbb{R}^{n_y})$ and $g \in \mathcal{C}(\mathbb{R}^{n_y} \times \mathbb{R}^{n_z}, \mathbb{R}^{m-n_y})$. Then, (2.84) can be written in the form

$$\begin{aligned} \dot{y} &= f(y, z), \\ 0 &= g(y, z). \end{aligned} \tag{2.93}$$

Theorem 22

A uniquely solvable semi-explicit DAE (2.93) with $m = n$ is strangeness-free if and only if $g_z(y, z)$ is nonsingular for all (y, z) satisfying $g(y, z) = 0$.

Remarks:

- The assumption of unique solvability guarantees that the set of solutions $\mathbb{L}_0$ is nonempty.
- The reduced system of the semi-explicit DAE (2.93) is then this DAE itself.
- The d-index of the semi-explicit DAE (2.93) is $v_d = 1$ if $n_z > 0$ and $v_d = 0$ if $n_z = 0$.
- The underlying ODE is obtained by differentiation of the constraints:

$$\begin{aligned} \dot{y} &= f(y,z), \\ \dot{z} &= -g_z^{-1}(y,z)g_y(y,z)f(y,z). \end{aligned}$$

  Note that the underlying ODE does not contain any constraints.

Theorem 23

A uniquely solvable semi-explicit DAE with $m = n$ of the form

$$\begin{aligned} \dot{y} &= f(y,z) \\ 0 &= g(y) \end{aligned} \tag{2.94}$$

has s-index $v_s = 1$ if and only if $g_y(y)f_z(y,z)$ is nonsingular for all (y,z) satisfying $g(y) = 0$ and $g_y(y)f(y,z) = 0$.

Remarks:

- The equations $0 = g_y(y)f(y,z)$ are the hidden constraints of the system.
- The d-index of the semi-explicit DAE (2.94) is $v_d = 2$.
- The DAE obtained by differentiation of the constraints

$$\begin{aligned} \dot{y} &= f(y,z) \\ 0 &= g_y(y)f(y,z) \end{aligned}$$

  is strangeness-free. Thus, the s-index is lowered by one, but the original constraints $0 = g(y)$ are lost.
- In general, one can prove that differentiating all constraints lowers the s-index by one, but these constraints are lost and the set of solutions is increased. In particular, the DAE obtained by differentiation of the constraints has solutions which do not solve the original DAE.

Theorem 24

A uniquely solvable semi-explicit DAE with $m = n$ of the form

$$\begin{aligned} \dot{y} &= f(y,z) \\ 0 &= g(y,z) \end{aligned} \tag{2.95}$$

with $\operatorname{rank}(g_z) = r_z = \text{const.}$ has s-index $v_s = 1$ if and only if $r_z < n_z$ and there exists a smooth pointwise nonsingular matrix function $\begin{pmatrix} Z' \\ Z \end{pmatrix}$ of size $n_z \times n_z$ where Z is of size $(n_z - r_z) \times n_z$ satisfying $Zg_z = 0$ such that $\begin{pmatrix} Z'g_z \\ (Zg_yf)_z \end{pmatrix}$ is nonsingular for all (y,z) satisfying $g(y,z) = 0$ and $(Zg_yf)(y,z) = 0$.

Remarks:

- The equations $0 = (Zg_y f)(y,z)$ are the hidden constraints of the system.
- The d-index of the semi-explicit DAE (2.95) is $v_d = 2$.

Semi-implicit DAEs

Definition 23: Semi-implicit DAE

The DAE (2.84) is called *semi-implicit* if F has the form

$$F(\dot{x},x,t) = \begin{pmatrix} E_1(x,t)\dot{x} - f_1(x,t) \\ -f_2(x,t) \end{pmatrix} = 0 \tag{2.96}$$

where $E_1 \in \mathcal{C}(\mathbb{D}_x \times \mathbb{I}, \mathbb{R}^{m_1,n})$, $f_1 \in \mathcal{C}(\mathbb{D}_x \times \mathbb{I}, \mathbb{R}^{m_1})$, $f_2 \in \mathcal{C}(\mathbb{D}_x \times \mathbb{I}, \mathbb{R}^{m_2})$, $m_1 + m_2 = m$. Then, (2.84) can be written in the form

$$\begin{aligned} E_1(x,t)\dot{x} &= f_1(x,t), \\ 0 &= f_2(x,t). \end{aligned} \tag{2.97}$$

Theorem 25

A uniquely solvable semi-implicit DAE (2.97) with $m = n$ is strangeness-free if and only if $\begin{pmatrix} E_1(x,t) \\ f_{2;x}(x,t) \end{pmatrix}$ is nonsingular for all $(x,t) \in \mathbb{M}$ where $\mathbb{M} = \{(x,t) \in \mathbb{D}_x \times \mathbb{I} :\ 0 = f_2(x,t)\}$.

Remarks:

- The reduced system of the semi-implicit DAE (2.97) is then this DAE itself.
- The d-index of the semi-implicit DAE (2.97) is $v_d = 1$ if $m_2 > 0$ and $v_d = 0$ if $m_2 = 0$.
- The underlying ODE is obtained by differentiation of the constraints:

$$\dot{x} = \begin{pmatrix} E_1 \\ f_{2;x} \end{pmatrix}^{-1} \begin{pmatrix} f_1 \\ -f_{2;t} \end{pmatrix} (x,t).$$

  Note that the underlying ODE does not contain any constraints.

Theorem 26

Let the semi-implicit DAE (2.97) be uniquely solvable with $m = n$ and let $x(t_0) = x_0$ be consistent initial values.

Then, the solution of the IVP is invariant under differentiation of the constraints, i.e.

$$\left.\begin{aligned} E_1(x,t)\dot{x} &= f_1(x,t) \\ 0 &= f_2(x,t) \\ x(t_0) &= x_0 \end{aligned}\right\} \Leftrightarrow \left\{\begin{aligned} E_1(x,t)\dot{x} &= f_1(x,t) \\ f_{2;x}(x,t)\dot{x} &= -f_{2,t}(x,t) \\ x(t_0) &= x_0 \end{aligned}\right. \tag{2.98}$$

Remark:
Without consistent initial values, the dimension of the set of solutions increases by differentiation of constraints due to the loss of these constraints.

Quasi-linear DAEs

Definition 24: Quasi-linear DAE

The DAE (2.84) is called *quasi-linear* if F has the form

$$F(\dot{x},x,t) = E(x,t)\dot{x} - f(x,t) = 0 \tag{2.99}$$

where $E \in \mathcal{C}(\mathbb{D}_x \times \mathbb{I}, \mathbb{R}^{m,n})$ is denoted as *leading matrix* and $f \in \mathcal{C}(\mathbb{D}_x \times \mathbb{I}, \mathbb{R}^m)$ is denoted as *right-hand side*. Then, (2.84) can be written in the form

$$E(x,t)\dot{x} = f(x,t). \tag{2.100}$$

Remark:
Quasi-linear DAEs with a nonsingular leading matrix $E(x,t)$ are called *ODEs in implicit form*.

Theorem 27

A uniquely solvable quasi-linear DAE (2.100) with $m = n$ is strangeness-free if there exists a pointwise nonsingular matrix function $S(x,t) = \begin{pmatrix} S_1(x,t) \\ S_2(x,t) \end{pmatrix}$ with $S_2(x,t)E(x,t) = 0$ and $\begin{pmatrix} S_1(x,t)E(x,t) \\ (S_2(x,t)f(x,t))_x \end{pmatrix}$ is nonsingular for all $(x,t) \in \mathbb{M}$ where

$$\mathbb{M} = \{(x,t) \in \mathbb{D}_x \times \mathbb{I} :\ 0 = S_2(x,t)f(x,t)\}.$$

Remarks:

- Note that the transformation with a pointwise nonsingular $S(x,t)$ from the left corresponds to a combination of equations which does not change the solution set or the s-index.
- The reduced system of the quasi-linear DAE (2.100) is

$$\begin{aligned} S_1(x,t)E(x,t)\dot{x} &= S_1(x,t)f(x,t), \\ 0 &= S_2(x,t)f(x,t). \end{aligned} \tag{2.101}$$

- The d-index of the quasi-linear DAE (2.100) is $v_d = 1$ if $E(x,t)$ is singular and $v_d = 0$ if $E(x,t)$ is nonsingular.
- The underlying ODE is obtained by differentiation of the constraints:

$$\dot{x} = \begin{pmatrix} S_1 E \\ (S_2 f)_x \end{pmatrix}^{-1} \begin{pmatrix} S_1 f \\ -(S_2 f)_t \end{pmatrix} (x,t).$$

Note that the underlying ODE does not contain any constraints.

An alternative approach without using Hypothesis 2 for determining the hidden constraints and the s-index is the following generalisation of Procedure 2 to quasi-linear DAEs:

Procedure 3

Consider a quasi-linear DAE $E(x,t)\dot{x} = f(x,t)$ and assume that E and f are sufficiently smooth. Also, we restrict our focus to uniquely solvable systems with $m = n$.

Starting from $E^0(x,t) = E(x,t)$, $f^0(x,t) = f(x,t)$, $\mathbb{M}_{-1} = \mathbb{D}_x \times \mathbb{I}$, iterate for $i = 0, 1, \ldots$

1) If $E^i(x,t)$ is nonsingular for all $(x,t) \in \mathbb{M}_{i-1}$, stop the procedure with $v = i$.
2) **Transformation matrix function:**

   Suppose there exists a sufficiently smooth matrix function $Z^i(x,t) = \begin{pmatrix} Z_1^i(x,t) \\ Z_2^i(x,t) \end{pmatrix}$ that is nonsingular for all $(x,t) \in \mathbb{M}_{i-1}$ such that

$$Z^i(x,t)E^i(x,t) = \begin{pmatrix} \tilde{E}^i(x,t) \\ 0 \end{pmatrix} \qquad \text{and} \qquad Z^i(x,t)f^i(x,t) = \begin{pmatrix} \tilde{f}_1^i(x,t) \\ \tilde{f}_2^i(x,t) \end{pmatrix} \tag{2.102}$$

   with $r^i = \operatorname{rank}(\tilde{E}^i(x,t)) = \operatorname{rank}(E^i(x,t))$ and $\tilde{E}^i \in \mathcal{C}(\mathbb{M}_i, \mathbb{R}^{r^i,n})$ for all $(x,t) \in \mathbb{M}_i$ and $\mathbb{M}_i = \{(x,t) \in \mathbb{M}_{i-1} : \; 0 = \tilde{f}_2^i(x,t)\}$.
3) **Separation:**

   Multiply the quasi-linear DAE $E^i(x,t)\dot{x} = f^i(x,t)$ with Z^i from the left to obtain the intermediate DAE

$$\begin{pmatrix} \tilde{E}^i(x,t) \\ 0 \end{pmatrix} \dot{x} = \begin{pmatrix} \tilde{f}_1^i(x,t) \\ \tilde{f}_2^i(x,t) \end{pmatrix}. \tag{2.103}$$

   Denote $0 = \tilde{f}_2^i(x,t)$ as *hidden constraints of level* i.
4) **Differentiation of the constraints:**

   Replace the constraints of level i by their derivatives with respect to t in order to obtain

$$\begin{pmatrix} \tilde{E}^i(x,t) \\ \tilde{f}_{2;x}^i(x,t) \end{pmatrix} \dot{x} = \begin{pmatrix} \tilde{f}_1^i(x,t) \\ -\tilde{f}_{2;t}^i(x,t) \end{pmatrix} \qquad \Leftrightarrow \qquad E^{i+1}(x,t)\dot{x} = f^{i+1}(x,t). \tag{2.104}$$

   Increase i by one and continue with 1).

Example 23

Consider the DAE $E(x,t)\dot{x} = f(x,t)$ from Example 14 with

$$E(x,t) = \begin{pmatrix} 0 & 0 \\ 1 & -t \end{pmatrix}, \quad f(x,t) = \begin{pmatrix} -1 & t \\ 0 & 0 \end{pmatrix} x + \begin{pmatrix} g_1(t) \\ g_2(t) \end{pmatrix}, \quad \mathbb{I} = \mathbb{R}, \quad \mathbb{D}_x = \mathbb{D}_{\dot{x}} = \mathbb{R}^2.$$

Following Procedure 3 yields:

- $i = 0:$ $E^0(x,t) = E(x,t)$, $f^0(x,t) = f(x,t)$, $\mathbb{M}_{-1} = \mathbb{D}_x \times \mathbb{I} = \mathbb{R}^2 \times \mathbb{R}$.
  1) $E^0(x,t)$ is singular $\quad\Rightarrow\quad$ continue
  2) **Transformation matrix function:**
  $$Z^0 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
  3) **Separation:**
  Multiplying the quasi-linear DAE $E^0(x,t)\dot{x} = f^0(x,t)$ with Z^0 from the left yields
  $$\begin{pmatrix} 1 & -t \\ 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} g_2 \\ -x_1 + tx_2 + g_1 \end{pmatrix}$$
  and we have $\mathbb{M}_0 = \{(x,t) \in \mathbb{R}^3 : \ x_1 = tx_2 + g_1\}$.
  4) **Differentiation of the constraints:**
  $$\begin{pmatrix} 1 & -t \\ 1 & -t \end{pmatrix} \dot{x} = \begin{pmatrix} g_2(t) \\ x_2 + \dot{g}_1(t) \end{pmatrix} \qquad \Leftrightarrow \qquad E^1(x,t)\dot{x} = f^1(x,t)$$
- $i = 1:$
  1) $E^1(x,t)$ is singular $\quad\Rightarrow\quad$ continue
  2) **Transformation matrix function:**
  $$Z^1 = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$
  3) **Separation:**
  Multiplying the quasi-linear DAE $E^1(x,t)\dot{x} = f^1(x,t)$ with Z^1 from the left yields
  $$\begin{pmatrix} 1 & -t \\ 0 & 0 \end{pmatrix} \dot{x} = \begin{pmatrix} g_2 \\ x_2 + \dot{g}_1 - g_2 \end{pmatrix}$$
  and we have $\mathbb{M}_1 = \{(x,t) \in \mathbb{R}^3 : \ x_1 = tx_2 + g_1, \ x_2 = g_2 - \dot{g}_1\}$.
  4) **Differentiation of the constraints:**
  $$\begin{pmatrix} 1 & -t \\ 0 & 1 \end{pmatrix} \dot{x} = \begin{pmatrix} g_2 \\ \dot{g}_2 - \ddot{g}_1 \end{pmatrix} \qquad \Leftrightarrow \qquad E^2(x,t)\dot{x} = f^2(x,t)$$
- $i = 2:$
  1) $E^2(x,t)$ is nonsingular for all $(x,t) \in \mathbb{M}_1 \quad\Rightarrow\quad$ stop with $v = i = 2$.

Remarks:

- If Procedure 3 stops in step 1) with $v = i$, then ...
  - ... the quasi-linear DAE (2.100) has s-index $v_s = \max\{0, v-1\}$ and d-index $v_d = v$ (this holds only for uniquely solvable and square systems),
  - ... the intermediate DAE $E^v(x,t)\dot{x} = f^v(x,t)$ corresponds to the underlying ODE in implicit form,
  - ... the IVP $E(x,t)\dot{x} = f(x,t)$, $x(t_0) = x_0$ has the same solution as the intermediately obtained IVPs $E^i(x,t)\dot{x} = f^i(x,t)$, $x(t_0) = x_0$, $i = 0, \dots, v$.
- Procedure 3 as well as Hypothesis 2 can be extended to over- and underdetermined systems, i.e. $m \neq n$ (see Procedure 3.5.11 and Hypothesis 3.2.7 in Steinbrecher 2006).

Definition 25: Maximal constraint level

Let Procedure 3 stop with $v = i$. Then, the highest level of existing (hidden) constraints is $v_c = v - 1$ and we call the quasi-linear DAE $E(x,t)\dot{x} = f(x,t)$ of *maximal constraint level* v_c.

Remarks:

- $v_c = -1$ means that there are no constraints, i.e. we have an ODE in implicit form.
- In our case of uniquely solvable systems with $m = n$, we have $v_c = v_s$.
- In a more general case of over- or underdetermined DAEs, this equality no longer holds. In this case, we have $v_c \leq v_s$.

Definition 26: Solution manifold

Let Procedure 3 stop with $v = i$. Then,

$$\mathbb{M} = \mathbb{M}_{v_c} = \{(x,t) \in \mathbb{D}_x \times \mathbb{I} : \; 0 = \tilde{f}_2^0(x,t), \dots, \tilde{f}_2^{v_c}(x,t)\} \tag{2.105}$$

is called *solution manifold* or *set of consistency.*

Remark:

Every solution trajectory x lies in the solution manifold, i.e. $(x(t),t) \in \mathbb{M}$.

Chapter 3

Numerical solution of differential-algebraic equations

This chapter deals with the numerical solution of DAEs. In the first section 3.1, the difficulties which can arise when numerical methods for ODEs are directly used to solve DAEs are discussed. In section 3.2, the idea of index reduction and regularisation of higher index DAEs is presented. Afterwards, in section 3.3, the two main classes of discretisation methods, namely one-step methods and multi-step methods, and their application to strangeness-free DAEs are described. Finally, in section 3.4, the Newton method for the numerical solution of systems of nonlinear equations arising in the integration process is presented.

3.1 Differential-algebraic equations are not ordinary differential equations

In principle, one can easily adapt numerical methods for ODEs in order to apply them directly to DAEs, for example by replacing derivatives by finite differences. But it was observed that, in contrast to ODEs, several difficulties arise when numerical methods are used to solve DAEs, e.g. instabilities, inconsistencies, convergence problems, drift-off effects, order reduction phenomena, inaccurate error estimates or amplifications of perturbations (see Petzold 1982; Brenan et al. 1996; Steinbrecher 2006; Kunkel/Mehrmann 2006).

These difficulties are caused by the algebraic constraints. In particular, additionally to the constraints explicitly occurring in the DAE, the solution of higher index DAEs is restricted by constraints which are hidden in the DAE, i.e. algebraic constraints that are not explicitly given in the system. Thus, a numerical integration of DAEs containing hidden constraints, i.e. not strangeness-free DAEs, is not recommended.

Furthermore, due to algebraic constraints, explicit methods cannot be used to solve DAEs since this would require the solution of a linear system with a typically singular matrix. The following example illustrates that:

Example 24: Explicit Euler method

We consider the DAE $0 = F(\dot{x}(t), x(t), t)$ with initial value $x(t_0) = x_0$ in $\mathbb{I} = [t_0, T] \subseteq \mathbb{R}$. By introducing a grid $t_0 \leq t_1 \leq \ldots \leq t_N = T$ where $h_i = t_i - t_{i-1}$, $i = 1, \ldots, N$, we are interested in a sequence $x_i \in \mathbb{R}^n$, $i = 0, \ldots, N$ which approximates the solution x at the points t_i, i.e. $x_i \approx x(t_i)$.

Similar to the ODE case, the idea of the explicit Euler method is to approximate $\dot{x}(t_i)$ by the forward difference quotient, i.e.

$$\dot{x}(t_i) \approx \frac{x_{i+1} - x_i}{h_{i+1}}.$$

Then, we get from $0 = F(\dot{x}(t), x(t), t)$ the iteration

$$x_0 = x(t_0), \qquad 0 = F\left(\frac{x_{i+1} - x_i}{h_{i+1}}, x_i, t_i\right) \quad \text{for} \quad i = 0, \ldots, N-1.$$

Thus, we have a nonlinear system for x_{i+1} (x_i, t_i, h_{i+1} are known) which has to be solved, e.g. with the Newton method:

$$x_{i+1}^0 = x_i, \qquad x_{i+1}^{j+1} = x_{i+1}^j - J^{-1}(x_{i+1}^j) G(x_{i+1}^j) \quad \text{for} \quad j = 0, \ldots$$

where $G(x_{i+1}^j) = F(\frac{x_{i+1}^j - x_i}{h_{i+1}}, x_i, t_i)$ and $J(x_{i+1}^j) = \dfrac{\partial}{\partial x_{i+1}} G(x_{i+1}^j) = F_{\dot{x}}(\frac{x_{i+1}^j - x_i}{h_{i+1}}, x_i, t_i) \cdot \dfrac{1}{h_{i+1}}$.

In general for DAEs, we have a singular $F_{\dot{x}}$. Therefore, we cannot (uniquely) get x_{i+1} from $0 = F(\frac{x_{i+1} - x_i}{h_{i+1}}, x_i, t_i)$. Thus, the explicit Euler method is not suitable for DAEs.

3.2 Index reduction and regularisation

Due to the difficulties of the numerical integration of higher index DAEs described above, one has to consider an alternative to a direct discretisation of higher index DAEs, namely discretising an equivalent formulation of the problem with s-index zero.

A classical approach for index reduction of higher index DAEs is to use index reduction by differentiating to turn the problem into a strangeness-free DAE. The constraints are replaced by their derivatives and differentiated unknowns are substituted as far as possible. One can prove that differentiating all constraints lowers the index by one. Although this actually was the common approach until the early seventies (see Kunkel/Mehrmann 2006, p. 273), there are major disadvantages to this approach. The differentiated constraints are removed from the DAE, and thus they are not present any more during the numerical integration. Therefore, the reduced system has solutions which do not solve the original DAE, i.e. the set of solutions is increased. Thus, discretisation and round-off errors may lead to a so-called drift-off effect of the solution such that the removed constraints are violated since the solution is no longer restricted into the set of consistency.

Hence, index reduction by differentiation lowers the index, but increases the set of solutions which leads to undesired drift-off effects. Therefore, an index reduction technique which at the same time does not change the set of solutions is preferable. Consequently, a regularisation is defined as follows:

Definition 27: Regularisation

A DAE $0 = \hat{F}(\dot{x}, x, t)$ is called *regularisation* of a DAE $0 = F(\dot{x}, x, t)$ if both have the same set of solutions and the s-index of $0 = \hat{F}(\dot{x}, x, t)$ is smaller than the s-index of the original DAE $0 = F(\dot{x}, x, t)$.

Remarks:

- If the DAE $0 = F(\dot{x}, x, t)$ satisfies Hypothesis 2 with $v_s = v$, then the reduced system (2.90), i.e.
$$0 = \hat{F}(\dot{x}, x, t) = \begin{pmatrix} \hat{F}_1(\dot{x}, x, t) \\ \hat{F}_2(x, t) \end{pmatrix} = \begin{pmatrix} Z_1^T F \\ Z_2^T \mathcal{F}_v \end{pmatrix} \tag{3.1}$$
with Z_1^T and Z_2^T obtained from Hypothesis 2, corresponds to a regularisation of the DAE $0 = F(\dot{x}, x, t)$.
- For numerical integration, strangeness-free regularisations where all constraints are stated explicitly, i.e. as algebraic equations, are to prefer.

Besides getting a regularisation from Hypothesis 2, which is very technical and requires the determination of the whole derivative array, there is another regularisation approach based on Procedure 3 for quasi-linear DAEs. The idea is to determine all constraints and not to replace them by their differentiated versions, but to simply add them to the system. This leads to an overdetermined system, which then has to be lead back to a square system by only selecting certain differential equations.

Regularisation of quasi-linear DAEs

Consider a uniquely solvable quasi-linear DAE

$$E(x, t)\dot{x} = f(x, t) \tag{3.2}$$

with $m = n$ and assume that Procedure 3 applied to (3.2) stops with $v = i \geq 2$. For $v = 0$ or $v = 1$ a regularisation is not necessary.

Then, the maximal constraint level is $v_c = v - 1$ ($v_s = v - 1$, $v_d = v$) and the hidden constraints are determined in step 2) of the procedure as

$$0 = \tilde{f}_2^i(x, t), \qquad i = 0, \ldots, v_c. \tag{3.3}$$

Define the *set of constraints* as

$$0 = g(x,t) = \begin{pmatrix} \tilde{f}_2^0(x,t) \\ \vdots \\ \tilde{f}_2^{v_c}(x,t) \end{pmatrix} \in \mathcal{C}(\mathbb{M}, \mathbb{R}^{m_c}) \tag{3.4}$$

where m_c is the number of all constraints.

Adding the set of constraints to the quasi-linear DAE (3.2) yields

$$\begin{aligned} E(x,t)\dot{x} &= f(x,t), \\ 0 &= g(x,t). \end{aligned} \tag{3.5}$$

This is an overdetermined system which is uniquely solvable with the same set of solutions as the original DAE (3.2). Furthermore, it is not strangeness-free since there are redundancies between the differential and algebraic equations, i.e. there are redundancies (rank deficiencies) in

$$\begin{pmatrix} (E(x,t)\dot{x} - f(x,t))_{\dot{x}} \\ (-g(x,t))_x \end{pmatrix} = \begin{pmatrix} E(x,t) \\ -g_x(x,t) \end{pmatrix}. \tag{3.6}$$

Eliminating these redundancies by scaling / selecting the differential equations with a so-called *selector matrix function* $S(x,t) \in \mathcal{C}(\mathbb{M}, \mathbb{R}^{m-m_c,m})$ such that

$$\begin{pmatrix} S(x,t)E(x,t) \\ g_x(x,t) \end{pmatrix} \tag{3.7}$$

is nonsingular for all $(x,t) \in \mathbb{M}$ yields a regularisation of the form

$$\begin{aligned} S(x,t)E(x,t)\dot{x} &= S(x,t)f(x,t), \\ 0 &= g(x,t), \end{aligned} \tag{3.8}$$

which is a square system where all redundancies are removed and which has the same set of solutions as the original DAE (3.2). The following theorem summarises the obtained approach:

Theorem 28

Let the quasi-linear DAE $E(x,t)\dot{x} = f(x,t)$ be uniquely solvable with $m = n$. Furthermore, let Procedure 3 applied to the quasi-linear DAE terminate in iteration step $v = i$ with maximal constraint level $v_c = v - 1$ and the constraints $0 = \tilde{f}_2^i(x,t)$, $i = 0, \dots, v_c$.

Then, the DAE

$$\begin{aligned} S(x,t)E(x,t)\dot{x} &= S(x,t)f(x,t), \\ 0 &= g(x,t) \end{aligned} \tag{3.9}$$

with the set of constraints $g(x,t) = \begin{pmatrix} \tilde{f}_2^0(x,t) \\ \vdots \\ \tilde{f}_2^{v_c}(x,t) \end{pmatrix}$ and $\begin{pmatrix} S(x,t)E(x,t) \\ g_x(x,t) \end{pmatrix}$ nonsingular for all $(x,t) \in \mathbb{M} = \{(x,t) \in \mathbb{D}_x \times \mathbb{I} : 0 = g(x,t)\}$ is strangeness-free and has the same set of solutions as the quasi-linear DAE $E(x,t)\dot{x} = f(x,t)$. Thus, (3.9) is a regularisation of the original system.

Proof: See Steinbrecher (2006), Lemma 3.5.47. ■

Remark:
The regularisation is not unique since the selector S is not unique.

Example 25: Mathematical pendulum

The mathematical pendulum can be modeled by the movement of a mass point with mass m around the origin of a Cartesian coordinate system (x_1, x_2) with distance l under the influence of gravity. By introducing the velocity variables $(v_1, v_2) = (\dot{x}_1, \dot{x}_2)$, the system can be described by the following quasi-linear DAE (for more details see section 4.1):

$$\begin{aligned} \dot{x}_1 &= v_1, \\ \dot{x}_2 &= v_2, \\ m\dot{v}_1 &= -2x_1\lambda, \\ m\dot{v}_2 &= -2x_2\lambda - mg, \\ 0 &= x_1^2 + x_2^2 - l^2. \end{aligned}$$

From Procedure 3 we obtain the maximal constraint level $v_c = 2$ and the constraints

$$\left.\begin{aligned} 0 &= \tilde{f}_2^0(x,t) = x_1^2 + x_2^2 - l^2 \\ 0 &= \tilde{f}_2^1(x,t) = x_1 v_1 + x_2 v_2 \\ 0 &= \tilde{f}_2^2(x,t) = v_1^2 + v_2^2 - \frac{2}{m}\left(x_1^2 + x_2^2\right)\lambda - g x_2 \end{aligned}\right. \qquad \Leftrightarrow \qquad 0 = g(x,t).$$

Therefore, we have

$$\begin{pmatrix} E(x,t) \\ g_x(x,t) \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & m & & \\ & & & m & \\ & & & & 0 \\ 2x_1 & 2x_2 & 0 & 0 & 0 \\ v_1 & v_2 & x_1 & x_2 & 0 \\ -\frac{4}{m}x_1\lambda & -\frac{4}{m}x_2\lambda - g & 2v_1 & 2v_2 & -\frac{2}{m}\left(x_1^2 + x_2^2\right) \end{pmatrix}.$$

With the selector

$$S(x,t) = \begin{pmatrix} x_2 & -x_1 & 0 & 0 & 0 \\ 0 & 0 & x_2 & -x_1 & 0 \end{pmatrix}$$

we obtain

$$\begin{pmatrix} S(x,t)E(x,t) \\ g_x(x,t) \end{pmatrix} = \begin{pmatrix} x_2 & -x_1 & 0 & 0 & 0 \\ 0 & 0 & mx_2 & -mx_1 & 0 \\ 2x_1 & 2x_2 & 0 & 0 & 0 \\ v_1 & v_2 & x_1 & x_2 & 0 \\ -\dfrac{4}{m}x_1\lambda & -\dfrac{4}{m}x_2\lambda - g & 2v_1 & 2v_2 & -\dfrac{2}{m}\left(x_1^2 + x_2^2\right) \end{pmatrix},$$

which is nonsingular since $x_1^2 + x_2^2 = l^2 \neq 0$.

Thus, we obtain the strangeness-free regularisation $\begin{pmatrix} S(x,t)E(x,t) \\ 0 \end{pmatrix} \dot{x} = \begin{pmatrix} S(x,t)f(x,t) \\ g(x,t) \end{pmatrix}$:

$$\begin{aligned} x_2\dot{x}_1 - x_1\dot{x}_2 &= x_2v_1 - x_1v_2, \\ mx_2\dot{v}_1 - mx_1\dot{v}_2 &= -2x_1x_2\lambda + 2x_1x_2\lambda + mx_1g, \\ 0 &= x_1^2 + x_2^2 - l^2, \\ 0 &= x_1v_1 + x_2v_2, \\ 0 &= v_1^2 + v_2^2 - \frac{2}{m}\left(x_1^2 + x_2^2\right)\lambda - gx_2. \end{aligned}$$

Remark:

Note that if we use Procedure 3 for the determination of a strangeness-free regularisation, it is only necessary to differentiate the algebraic constraints. In contrast, if we use Hypothesis 2, it is necessary to determine the derivative array, i.e. to determine the derivatives of the whole DAE. Therefore, the use of Hypothesis 2 for the determination of a regularisation of a quasi-linear DAE is more involved than the use of Procedure 3. However, Procedure 3 is only applicable for quasi-linear DAEs, while Hypothesis 2 is suited for general nonlinear DAEs.

3.3 Methods for strangeness-free differential-algebraic equations

In this section, the two main classes of discretisation methods, namely one-step methods and linear multi-step methods, and their application to strangeness-free DAEs are discussed.

In general, we are interested in a numerical solution of an initial value problem of a general nonlinear DAE of the form

$$F(\dot{x}, x, t) = 0, \qquad x(t_0) = x_0 \tag{3.10}$$

in the interval $\mathbb{I} = [t_0, T] \subseteq \mathbb{R}$.

We denote by $t_0 \leq t_1 \leq \ldots \leq t_N = T$ grid points in the interval $\mathbb{I}$ and by x_i approximations to the solution $x(t_i)$. For convenience, we only consider a fixed step size h, i.e. it holds $t_i = t_0 + ih$, $i = 0, \ldots, N$ and $N = \frac{T - t_0}{h}$.

3.3.1 One-step methods

A one-step method for the numerical solution of (3.10) is defined as follows:

Definition 28: One-step method

A *one-step method* for the determination of discrete approximations x_i, $i = 0, \dots, N$ to the values $x(t_i)$ of a solution x of (3.10) is given by an iteration

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h) \qquad \text{for} \quad i = 0, \dots, N-1, \tag{3.11}$$

where Φ is called *increment function*.

We are then interested in conditions that guarantee convergence of the methods, i.e. that x_N tends to $x(t_N)$ when h tends to zero. Therefore, we need the following definitions:

Definition 29: Consistent of order p

A one-step method (3.11) is called *consistent of order* p, $p \in \mathbb{N} \setminus \{0\}$, if

$$\|x(t_{i+1}) - x(t_i) - h\Phi(t_i, x(t_i), h)\| \leq Ch^{p+1} \tag{3.12}$$

for all $i = 0, \dots, N-1$ with a constant C independent of h.

Definition 30: Convergent of order p

A one-step method (3.11) is called *convergent of order* p, $p \in \mathbb{N} \setminus \{0\}$, if

$$\|x(t_N) - x_N\| \leq Ch^p \tag{3.13}$$

with a constant C independent of h.

Theorem 29: Convergence of one-step methods

Let a one-step method (3.11) with order of consistency p be given. Furthermore, let Φ be Lipschitz continuous with respect to its second argument, i.e.

$$\|\Phi(t, x, h) - \Phi(t, y, h)\| \leq L \|x - y\| \tag{3.14}$$

for all $t \in [t_0, T]$, $0 < h < T - t$, $x, y \in \mathbb{C}^n$ with Lipschitz constant $L > 0$.

Then, the one-step method is also convergent of order p.

Proof: See Bollhöfer/Mehrmann (2004), Satz 3.10. ■

In order to develop appropriate one-step methods for DAEs, the idea is to generalise Runge-Kutta methods for ODEs of the form $\dot{x} = f(t, x)$ to DAEs of the form $0 = F(\dot{x}, x, t)$. The concept of Runge–Kutta methods for ODEs consists in using an increment function $\Phi(t_i, x_i, h)$ that is a linear combination of the values of the right-hand side f in discrete points. A general implicit Runge-Kutta method is defined as follows:

Definition 31: s-stage Runge-Kutta method for ODEs

Let $s \in \mathbb{N} \setminus \{0\}$ and $a_{ij}, b_i, c_i \in \mathbb{R}$, $i, j = 1, \ldots, s$.

Then, an *s-stage Runge-Kutta method* for the solution of $\dot{x} = f(t, x)$, $x(t_0) = x_0$ is given by

$$x_{i+1} = x_i + h \sum_{j=1}^{s} b_j k_j \qquad \text{for} \quad i = 0, \ldots, N-1 \tag{3.15}$$

with

$$k_j = f(t_i + c_j h, x_i + h \sum_{l=1}^{s} a_{jl} k_l), \qquad j = 1, \ldots, s, \tag{3.16}$$

where s is the *number of stages*, k_j are called *stages*, $A = [a_{ij}]_{i,j=1,\ldots,s} \in \mathbb{R}^{s,s}$ denotes the *Runge-Kutta matrix*, $b = [b_i]_{i=1,\ldots,s} \in \mathbb{R}^s$ denotes the *weight vector* and $c = [c_i]_{i=1,\ldots,s} \in \mathbb{R}^s$ denotes the *node vector*.

The coefficients $a_{ij}, b_i, c_i \in \mathbb{R}$, $i, j = 1, \ldots, s$, determine the particular Runge-Kutta method and are conveniently expressed in a tableau, a so-called *Butcher tableau*:

$$\begin{array}{c|ccc} c_1 & a_{11} & \ldots & a_{1s} \\ \vdots & \vdots & \ddots & \vdots \\ c_s & a_{s1} & \ldots & a_{ss} \\ \hline & b_1 & \ldots & b_s \end{array} \qquad \Leftrightarrow \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array} \tag{3.17}$$

Note that the stages k_j approximate the derivative $\dot{x}$ at $t_i + c_j h$, and furthermore $x_i + h \sum\limits_{l=1}^{s} a_{jl} k_l$ corresponds to an approximation for x at $t_i + c_j h$. For DAEs, an explicit approximation for $\dot{x}$ at $t_i + c_j h$ cannot be provided since $F_{\dot{x}}$ is in general singular. Therefore, set $X_j' = k_j$ and $X_j = x_i + h \sum\limits_{l=1}^{s} a_{jl} X_l'$. Then, we obtain an implicit Runge-Kutta method for DAEs as follows:

Definition 32: s-stage Runge-Kutta method for DAEs

Let $s \in \mathbb{N} \setminus \{0\}$ and $a_{ij}, b_i, c_i \in \mathbb{R}$, $i, j = 1, \ldots, s$.

Then, an *s-stage Runge-Kutta method* for the solution of $0 = F(\dot{x}, x, t)$, $x(t_0) = x_0$ is given by

$$x_{i+1} = x_i + h \sum_{j=1}^{s} b_j X_j' \qquad \text{for} \quad i = 0, \ldots, N-1 \tag{3.18}$$

with

$$\left. \begin{aligned} X_j &= x_i + h \sum_{l=1}^{s} a_{jl} X_l' \\ 0 &= F(X_j', X_j, t_i + c_j h) \end{aligned} \right\} \qquad \text{for} \quad j = 1, \ldots, s. \tag{3.19}$$

Note that (3.19) forms a system of nonlinear equations of size $2ns \times 2ns$ for the determination of X_j' and X_j for $j = 1, \dots, s$. The numerical solution of such systems of nonlinear equations is discussed in the next section.

Example 26: Implicit Euler method

The implicit Euler method is given by the Butcher tableau $\begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}$.

Thus, we get from (3.18) and (3.19):

$$\begin{aligned} x_{i+1} &= x_i + hX_1', \\ X_1 &= x_i + hX_1', \\ 0 &= F(X_1', X_1, t_i + h). \end{aligned}$$

The first equation yields $X_1' = \dfrac{x_{i+1} - x_i}{h}$. Then, from the second equation it follows that $X_1 = x_i + hX_1' = x_{i+1}$. Thus, by inserting X_1' and X_1 into the last equation we obtain

$$0 = F\left(\frac{x_{i+1} - x_i}{h}, x_{i+1}, t_{i+1}\right).$$

Example 27: Midpoint method

The midpoint method is given by the Butcher tableau $\begin{array}{c|c} \frac{1}{2} & \frac{1}{2} \\ \hline & 1 \end{array}$.

Thus, we get from (3.18) and (3.19):

$$\begin{aligned} x_{i+1} &= x_i + hX_1', \\ X_1 &= x_i + \frac{1}{2}hX_1', \\ 0 &= F(X_1', X_1, t_i + \frac{1}{2}h). \end{aligned}$$

The first equation yields $X_1' = \dfrac{x_{i+1} - x_i}{h}$. Then, from the second equation it follows that $X_1 = x_i + \dfrac{1}{2}hX_1' = x_i + \dfrac{x_{i+1} - x_i}{2} = \dfrac{x_i + x_{i+1}}{2}$. Thus, by inserting X_1' and X_1 into the last equation we obtain

$$0 = F\left(\frac{x_{i+1} - x_i}{h}, \frac{x_i + x_{i+1}}{2}, t_i + \frac{1}{2}h\right).$$

Note that equations (3.18) and (3.19) only describe a well-defined discretisation method if they define a unique x_{i+1} at least for sufficiently small step sizes h. Furthermore, another important aspect in the numerical treatment of DAEs is the consistency of numerical solution x_{i+1}. In addition, we are interested in the convergence properties of the resulting methods compared to the classical order of convergence for ODEs.

Concerning the uniqueness of x_{i+1}, Kunkel/Mehrmann (2006, pp. 227-228) show that already for the case of linear DAEs with constant coefficient it is necessary that the Runge-Kutta matrix A is nonsingular. Therefore, we are restricted to implicit Runge-Kutta methods since explicit methods have a singular A.

If we consider the consistency of x_{i+1}, it is clear that in the case of strangeness-free DAEs where all constraints are given in an explicit way, the stages X_j, $j = 1, \ldots, s$, are consistent due to equation (3.19), i.e. they satisfy all constraints. Unfortunately, this does not guarantee the consistency of the numerical solution x_{i+1} since x_{i+1} is given by equation (3.18). Hence, Runge-Kutta methods which automatically guarantee the consistency of x_{i+1} are to be preferred. In this context, an important class of Runge-Kutta methods are the so-called *stiffly accurate* Runge-Kutta methods. These are defined to satisfy the condition $a_{sj} = b_j$, $j = 1, \ldots, s$ (see Prothero/ Robinson 1974). From this it follows that the numerical solution x_{i+1} coincides with the last stage X_s, i.e. $x_{i+1} = X_s$. Therefore, the consistency of x_{i+1} can be guaranteed for strangeness-free DAEs where all constraints are given in an explicit way. However, for DAEs of higher index or for methods which are not stiffly accurate, the consistency of x_{i+1} cannot be guaranteed.

Concerning the convergence of Runge-Kutta methods, it has been shown that the obtained order of implicit Runge-Kutta methods applied to DAEs differs from the obtained order of these methods applied to ODEs, the so-called *classical order*. Besides an order reduction, it even may happen that the convergence is completely lost. For more details see e.g. Petzold (1986); Hairer et al. (1989); Steinbrecher (2006), section 3.5.4.5; and Kunkel/Mehrmann (2006), section 5.2. In Hairer et al. (1989), results on the order of implicit Runge-Kutta methods applied to semi-explicit DAEs of d-index one, two and three are investigated. Steinbrecher (2006) generalises these results to strangeness-free quasi-linear DAEs. His results for certain important classes of implicit Runge-Kutta methods are listed in Table 3.1.

Method	Stages	Classical order	Order of convergence	Stability properties
Gauß	$s \begin{cases} \text{odd} \\ \text{even} \end{cases}$	$2s$	$\begin{cases} s+1 \\ s \end{cases}$	A-stable
RadauIa	s	$2s-1$	s	A-, L-stable
RadauIIa	s	$2s-1$	$2s-1$	A-, L-stable, stiffly accurate
LobattoIIIa	s	$2s-2$	$2s-2$*	A-stable, stiffly accurate
LobattoIIIb	s	$2s-2$	**	A-stable
LobattoIIIc	s	$2s-2$	$2s-2$	A-, L-stable, stiffly accurate

*Results for LobattoIIIa are proven for $s = 2, 3$ and conjectured for larger s (see Steinbrecher 2006, p. 105)

**LobattoIIIb methods are not considered by Hairer et al. (1989) and consequently not generalised by Steinbrecher (2006)

Table 3.1: Properties of Runge-Kutta methods applied to strangeness-free quasi-linear DAEs

For the numerical solution of DAEs, there are several important classes of implicit Runge-Kutta methods which are well investigated and very popular, namely Gauß methods, Radau methods and Lobatto methods.

Gauß methods are collocation methods based on the Gauß quadrature formulas. The s-stage Gauß method is of (maximally reachable) classical order $2s$, while the numerical integration of strangeness-free quasi-linear DAEs in general leads to an order of convergence of only $s+1$ for odd s and an order of convergence of s for even s. Furthermore, all Gauß methods are A-stable, but they are not L-stable and not stiffly accurate. Due to the choice of nodes, i.e. the parameters c_j, there is no stage X_j, $j = 1, \dots, s$, which coincides with the numerical solution x_{i+1}. Instead of that, the numerical solution x_{i+1} is just a linear combination of the stages X_j. Therefore, the consistency of x_{i+1} cannot be assured. Hence, due to the order reduction and the possible inconsistency of x_{i+1}, the Gauß methods are impracticable for the numerical solution of DAEs.

$$
\begin{array}{c}
s=1 \\[4pt]
\begin{array}{c|c}
\frac{1}{2} & \frac{1}{2} \\
\hline
 & 1
\end{array}
\end{array}
\qquad
\begin{array}{c}
s=2 \\[4pt]
\begin{array}{c|cc}
\frac{1}{2}-\frac{\sqrt{3}}{6} & \frac{1}{4} & \frac{1}{4}-\frac{\sqrt{3}}{6} \\
\frac{1}{2}+\frac{\sqrt{3}}{6} & \frac{1}{4}+\frac{\sqrt{3}}{6} & \frac{1}{4} \\
\hline
 & \frac{1}{2} & \frac{1}{2}
\end{array}
\end{array}
\qquad
\begin{array}{c}
s=3 \\[4pt]
\begin{array}{c|ccc}
\frac{1}{2}-\frac{\sqrt{15}}{10} & \frac{5}{36} & \frac{2}{9}-\frac{\sqrt{15}}{15} & \frac{5}{36}-\frac{\sqrt{15}}{30} \\
\frac{1}{2} & \frac{5}{36}+\frac{\sqrt{15}}{24} & \frac{2}{9} & \frac{5}{36}-\frac{\sqrt{15}}{24} \\
\frac{1}{2}+\frac{\sqrt{15}}{10} & \frac{5}{36}+\frac{\sqrt{15}}{30} & \frac{2}{9}+\frac{\sqrt{15}}{15} & \frac{5}{36} \\
\hline
 & \frac{5}{18} & \frac{4}{9} & \frac{5}{18}
\end{array}
\end{array}
$$

Table 3.2: Butcher tableaus of the first three Gauß methods

The s-stage Radau methods are of classical order $2s-1$ which is slightly lower than the classical order of the Gauß methods. But they are designed in such a way that they have excellent stability properties, i.e. Radau methods are A-stable as well as L-stable. The node vector c of the RadauIa methods satisfies the condition $c_1 = 0$, whereas the node vector c of the RadauIIa methods satisfies the condition $c_s = 1$ and furthermore it holds that $a_{sj} = b_j$ for $j = 1, \dots, s$, i.e. RadauIIa methods are stiffly accurate. Thus, the numerical solution x_{i+1} coincides with the last stage X_s for RadauIIa methods. Therefore, the consistency of x_{i+1} obtained from a RadauIIa method applied to strangeness-free DAEs with all constraints stated in an explicit way follows from the consistency of the stages. The numerical integration of strangeness-free quasi-linear DAEs with RadauIa methods in general leads to an order of convergence of only s, whereas the order of convergence of RadauIIa methods applied to strangeness-free quasi-linear DAEs is still $2s-1$.

The great stability properties, the guaranteed consistency of x_{i+1} and the absence of order reduction make RadauIIa methods excellent candidates for the numerical solution of initial value problems for strangeness-free quasi-linear DAEs.

$$
\begin{array}{c}
s=1 \\[4pt]
\begin{array}{c|c}
0 & 1 \\
\hline
 & 1
\end{array}
\end{array}
\qquad
\begin{array}{c}
s=2 \\[4pt]
\begin{array}{c|cc}
0 & \frac{1}{4} & -\frac{1}{4} \\
\frac{2}{3} & \frac{1}{4} & \frac{5}{12} \\
\hline
 & \frac{1}{4} & \frac{3}{4}
\end{array}
\end{array}
\qquad
\begin{array}{c}
s=3 \\[4pt]
\begin{array}{c|ccc}
0 & \frac{1}{9} & \frac{-1-\sqrt{6}}{18} & \frac{-1+\sqrt{6}}{18} \\
\frac{6-\sqrt{6}}{10} & \frac{1}{9} & \frac{88+7\sqrt{6}}{360} & \frac{88-43\sqrt{6}}{360} \\
\frac{6+\sqrt{6}}{10} & \frac{1}{9} & \frac{88+43\sqrt{6}}{360} & \frac{88-7\sqrt{6}}{360} \\
\hline
 & \frac{1}{9} & \frac{16+\sqrt{6}}{36} & \frac{16-\sqrt{6}}{36}
\end{array}
\end{array}
$$

Table 3.3: Butcher tableaus of the first three RadauIa methods

$$s=1 \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array} \qquad\qquad s=2 \qquad \begin{array}{c|cc} \frac{1}{3} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{3}{4} & \frac{1}{4} \\ \hline & \frac{3}{4} & \frac{1}{4} \end{array}$$

$$s=3 \qquad \begin{array}{c|ccc} \frac{4-\sqrt{6}}{10} & \frac{88-7\sqrt{6}}{360} & \frac{296-169\sqrt{6}}{1800} & \frac{-2+3\sqrt{6}}{225} \\ \frac{4+\sqrt{6}}{10} & \frac{296+169\sqrt{6}}{1800} & \frac{88+7\sqrt{6}}{360} & \frac{-2-3\sqrt{6}}{225} \\ 1 & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \\ \hline & \frac{16-\sqrt{6}}{36} & \frac{16+\sqrt{6}}{36} & \frac{1}{9} \end{array}$$

Table 3.4: Butcher tableaus of the first three RadauIIa methods

The s-stage Lobatto methods are of classical order $2s-2$. The numerical integration of strangeness-free quasi-linear DAEs with Lobatto methods in general leads to the same order of convergence of $2s-2$. There are three families of Lobatto methods, called LobattoIIIa, LobattoIIIb and LobattoIIIc. The node vector c of the Lobatto methods satisfies the conditions $c_1 = 0$ and $c_s = 1$. Furthermore, the Lobatto methods are A-stable but only the LobattoIIIc methods are L-stable. In addition, LobattoIIIa and LobattoIIIc methods are stiffly accurate. For LobattoIIIa methods, the first row of the Runge-Kutta matrix A is identical to 0, so that A is singular. Similarly, for LobattoIIIb methods, the last column of the Runge-Kutta matrix A is identical to 0, so that A is singular. Thus, LobattoIIIa and LobattoIIIb methods are not suited for the numerical solution of DAEs because they are not well-defined, i.e. the uniqueness of x_{i+1} is not assured, while LobattoIIIc methods are good candidates for the numerical solution of strangeness-free DAEs.

$$s=2 \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad\qquad s=3 \qquad \begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{5}{24} & \frac{1}{3} & -\frac{1}{24} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Table 3.5: Butcher tableaus of the first two LobattoIIIa methods

$$s=2 \qquad \begin{array}{c|cc} 0 & \frac{1}{2} & 0 \\ 1 & \frac{1}{2} & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad\qquad s=3 \qquad \begin{array}{c|ccc} 0 & \frac{1}{6} & -\frac{1}{6} & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 \\ 1 & \frac{1}{6} & \frac{5}{6} & 0 \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Table 3.6: Butcher tableaus of the first two LobattoIIIb methods

$$s=2 \qquad \begin{array}{c|cc} 0 & \frac{1}{2} & -\frac{1}{2} \\ 1 & \frac{1}{2} & \frac{1}{2} \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array} \qquad\qquad s=3 \qquad \begin{array}{c|ccc} 0 & \frac{1}{6} & -\frac{1}{3} & \frac{1}{6} \\ \frac{1}{2} & \frac{1}{6} & \frac{5}{12} & -\frac{1}{12} \\ 1 & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \\ \hline & \frac{1}{6} & \frac{2}{3} & \frac{1}{6} \end{array}$$

Table 3.7: Butcher tableaus of the first two LobattoIIIc methods

3.3.2 Multi-step methods

The idea of multi-step methods is to not only use the previous approximation x_{i-1} for the determination of the approximation x_i to the solution $x(t_i)$, but also to use several of the previous approximations $x_{i-1}, \dots, x_{i-k}$ for the computation of x_i. In the case of linear multi-step methods, a linear combination of the previous approximations x_j and of the right-hand sides $f(t_j, x_j)$ is used. Thus, a linear multi-step method for the numerical solution of ODEs is defined as follows:

Definition 33: Linear multi-step method for ODEs

Let $k \in \mathbb{N} \setminus \{0\}$ and $\alpha_i, \beta_i \in \mathbb{R}$, $i = 0, \dots, k$.

Then, a *linear multi-step method* for the solution of $\dot{x} = f(t, x)$, $x(t_0) = x_0$ is given by

$$\sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = h \sum_{l=0}^{k} \beta_{k-l} f(t_{i-l}, x_{i-l}) \qquad \text{for} \quad i = k, \dots, N \tag{3.20}$$

with $\alpha_k \neq 0$ and $\alpha_0^2 + \beta_0^2 \neq 0$. The method (3.20) is also called a *k-step method.*

Remarks:

- In order to initialise the iteration, the values $x_0, \dots, x_{k-1}$ must be provided. The computation of these values is usually done by an appropriate one-step method.
- The linear multi-step method is called *explicit* if $\beta_k = 0$, and *implicit* if $\beta_k \neq 0$.
- By multiplying (3.20) with a nonzero scalar, we may assume that $\alpha_k = 1$ or, in the case of implicit methods, that $\beta_k = 1$.

Definition 34: Consistent of order p

A multi-step method (3.20) is called *consistent of order p*, $p \in \mathbb{N} \setminus \{0\}$, if

$$\left\| \sum_{l=0}^{k} \alpha_{k-l} x(t_{i-l}) - h \sum_{l=0}^{k} \beta_{k-l} \dot{x}(t_{i-l}) \right\| \leq C h^{p+1} \tag{3.21}$$

for all sufficiently smooth functions x with a constant C independent of h.

Theorem 30: Consistency of multi-step methods

A multi-step method (3.20) is consistent if and only if

$$\sum_{l=0}^{k} \alpha_l = 0 \qquad \text{and} \qquad \sum_{l=0}^{k} l\alpha_l = \sum_{l=0}^{k} \beta_l. \tag{3.22}$$

Furthermore, (3.20) has order of consistency p if

$$\sum_{l=0}^{k} l^i \alpha_l = \sum_{l=0}^{k} i l^{i-1} \beta_l \qquad \text{for} \quad i = 0, \dots, p. \tag{3.23}$$

Proof: See Kunkel/Mehrmann (2006), Theorem 5.20. ■

Definition 35: Stability

A multi-step method (3.20) is called *stable* if

1) all roots $\lambda \in \mathbb{C}$ of $\varrho(\lambda) = \sum\limits_{l=0}^{k} \alpha_l \lambda^l$ satisfy $|\lambda| \leq 1$ and

2) all roots $\lambda \in \mathbb{C}$ of $\varrho(\lambda) = \sum\limits_{l=0}^{k} \alpha_l \lambda^l$ with $|\lambda| = 1$ are simple.

Definition 36: Convergent of order p

A multi-step method (3.20) is called *convergent of order* p, $p \in \mathbb{N} \setminus \{0\}$, if for all $t_N \in \mathbb{I}$ and all $h = \dfrac{t_N - t_0}{N}$ it holds that

$$\|x(t_N) - x_N\| \leq Ch^p \tag{3.24}$$

with a constant C independent of h.

Theorem 31: Convergence of multi-step methods

If the multi-step method (3.20) is stable and consistent of order p, then it is convergent of order p.

Proof: See Kunkel/Mehrmann (2006), Theorem 5.22. ■

A very popular class of implicit multi-step methods are the so-called BDF methods. The abbreviation BDF stands for backward differentiation formula. BDF methods are constructed by setting $\beta_0 = \beta_1 = \ldots = \beta_{k-1} = 0$, $\beta_k = 1$ and determining the coefficients α_i, $i = 0, \ldots, k$, by backward differences, such that the order of consistency is as large as possible. Thus, BDF methods for ODEs have the following form:

$$\frac{1}{h} \sum_{l=0}^{k} \alpha_{k-l} x_{i-l} = f(t_i, x_i). \tag{3.25}$$

The resulting coefficients α_i, $i = 0, \ldots, k$, are stated in Table 3.8:

$k = 1$	$\alpha_0 = -1,$	$\alpha_1 = 1$					
$k = 2$	$\alpha_0 = \frac{1}{2},$	$\alpha_1 = -2,$	$\alpha_2 = \frac{3}{2}$				
$k = 3$	$\alpha_0 = -\frac{1}{3},$	$\alpha_1 = \frac{3}{2},$	$\alpha_2 = -3,$	$\alpha_3 = \frac{11}{6}$			
$k = 4$	$\alpha_0 = \frac{1}{4},$	$\alpha_1 = -\frac{4}{3},$	$\alpha_2 = 3,$	$\alpha_3 = -4,$	$\alpha_4 = \frac{25}{12}$		
$k = 5$	$\alpha_0 = -\frac{1}{5},$	$\alpha_1 = \frac{5}{4},$	$\alpha_2 = -\frac{10}{3},$	$\alpha_3 = 5,$	$\alpha_4 = -5,$	$\alpha_5 = \frac{137}{60}$	
$k = 6$	$\alpha_0 = \frac{1}{6},$	$\alpha_1 = -\frac{6}{5},$	$\alpha_2 = \frac{15}{4},$	$\alpha_3 = -\frac{20}{3},$	$\alpha_4 = \frac{15}{2},$	$\alpha_5 = -6,$	$\alpha_6 = \frac{49}{20}$

Table 3.8: Coefficients of the BDF methods

By construction, the BDF methods are consistent of order $p = k$. However, in order to achieve convergence, stability of the methods is required, see Theorem 31. In Hairer/Wanner 1983, it is shown that the BDF methods are stable and thus convergent for $1 \leq k \leq 6$ and unstable for $k \geq 7$. Furthermore, BDF methods are A-stable only for $k \leq 2$ and A(α)-stable for $3 \leq k \leq 6$.

In order to generalise the BDF methods for ODEs of the form $\dot{x} = f(t, x)$ to DAEs of the form $0 = F(\dot{x}, x, t)$, it should be noted that in (3.25), the discretisation is yielded by simply replacing $x(t_i)$ by the approximation x_i and $\dot{x}(t_i) = f(t_i, x(t_i))$ by the approximation $\frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} x_{i-l}$. Therefore, by using this discretisation in the same way we obtain the BDF methods for DAEs as follows:

$$0 = F\left(\frac{1}{h}\sum_{l=0}^{k} \alpha_{k-l} x_{i-l}, x_i, t_i\right). \tag{3.26}$$

Example 28: BDF method $k = 1$

For $k = 1$ we have $\alpha_0 = -1$ and $\alpha_1 = 1$. Thus, we get from (3.26):

$$0 = F\left(\frac{x_i - x_{i-1}}{h}, x_i, t_i\right).$$

This is the implicit Euler method.

Example 29: BDF method $k = 2$

For $k = 2$ we have $\alpha_0 = \frac{1}{2}$, $\alpha_1 = -2$, $\alpha_2 = \frac{3}{2}$. Thus, we get from (3.26):

$$0 = F\left(\frac{1}{h}\left(\frac{3}{2}x_i - 2x_{i-1} + \frac{1}{2}x_{i-2}\right), x_i, t_i\right).$$

Note that in the case of strangeness-free DAEs where all constraints are given in an explicit way, the consistency of x_i is obviously guaranteed due to equation (3.26).

Concerning the convergence of BDF methods, it is shown that the obtained order of BDF methods applied to strangeness-free DAEs with all constraints given in an explicit way is equal to the order of these methods applied to ODEs, i.e. BDF methods applied to those DAEs are convergent of order $p = k$ for $1 \leq k \leq 6$ (see Kunkel/Mehrmann 2006, Theorem 5.27). However, for higher index DAEs, the order of convergence is reduced. Furthermore, according to Petzold (1986, p. 838), it is shown by März (1981) that general linear multi-step methods applied to strangeness-free DAEs have to satisfy an extra set of conditions for the method to be convergent of the classical order, i.e. not to suffer from order reduction. Fortunately, these conditions are satisfied for BDF methods, which makes them excellent candidates for the numerical solution of initial value problems for strangeness-free DAEs.

3.4 Numerical solution of systems of nonlinear equations

In this section, we consider systems of nonlinear equations of the form $f(x) = 0$ where $f \in \mathcal{C}(\mathbb{R}^m, \mathbb{R}^m)$ is a nonlinear system of functions mapping a vector $x \in \mathbb{R}^m$ to a vector $y = f(x) \in \mathbb{R}^m$, defined by a system of m nonlinear equations

$$y = f(x) = \begin{pmatrix} f_1(x_1, \dots, x_m) \\ \vdots \\ f_m(x_1, \dots, x_m) \end{pmatrix}. \tag{3.27}$$

In the previous section, we have seen that the numerical solution of DAEs always yields such systems of nonlinear equations. In particular, in each iteration step of a Runge-Kutta method a nonlinear system of size $2ns \times 2ns$ for the determination of X_j' and X_j for $j = 1, \dots, s$ has to be solved where s is the number of stages, see equation (3.19). The size of the system can be reduced to $ns \times ns$ by simply substituting the expression of X_j into the second part of (3.19). On the other hand, in each iteration step of a BDF method a nonlinear system of size $n \times n$ has to be solved, see equation (3.26). Note that the size of the system does not depend on the order k.

Thus, an efficient algorithm for the numerical solution of systems of nonlinear equations is necessary for the solution of DAEs. The numerical solution of systems of nonlinear equations has been investigated extensively in the literature, e.g. see Deuflhard (2011), but it is not the main topic of this thesis. Therefore, only the ordinary Newton method for the numerical solution of systems of nonlinear equations is presented here:

Definition 37: Newton method

Let $f : X \to \mathbb{R}^n$ be a continuously differentiable function with $X \subset \mathbb{R}^n$ open and convex. Let $x^0 \in \mathbb{R}^n$ be given. Suppose that the Jacobian $J_f(x)$ is nonsingular for all $x \in X$.

Then, the *Newton method* for the determination of a sequence $\{x^k\}$ in order to find a solution of $f(x) = 0$ is given by the iteration

$$J_f(x^k)\Delta x^k = -f(x^k), \qquad x^{k+1} = x^k + \Delta x^k. \tag{3.28}$$

Remarks:

- Obviously, the Newton method treats the numerical solution of a nonlinear problem by solving a sequence of linear problems with the Jacobian as system matrix.
- Under certain conditions, the existence and uniqueness of a solution x as well as quadratic convergence of the Newton iterates x^k can be shown (see e.g. Deuflhard 2011, Theorem 2.3).
- One of these conditions requires a *sufficiently good* initial guess x^0 of the solution, otherwise convergence cannot be guaranteed.
- The Newton method will be carried out until a certain termination criterion is satisfied. There exist many different termination criteria in the literature. In the present work, the Newton iteration is terminated if the residual norm $\|f(x^k)\|$ or the correction norm $\|\Delta x^k\|$ fall below a prescribed tolerance, or if a maximum number of iterations is exceeded.

Chapter 4

Test problems

4.1 Mathematical pendulum

One of the most famous elementary examples of DAEs is the mathematical pendulum which is modeled by the movement of a mass point with mass m around the origin of a Cartesian coordinate system (x_1, x_2) with distance l under the influence of gravity, see Figure 4.1.

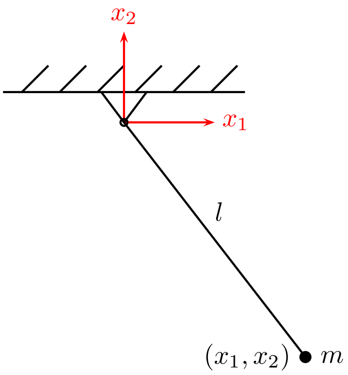


Figure 4.1: The mathematical pendulum

The kinetic energy of the system is given by

$$E_k = \frac{1}{2}m(\dot{x}_1^2 + \dot{x}_2^2). \tag{4.1}$$

Furthermore, the potential energy is given by

$$E_p = mgx_2 \tag{4.2}$$

where g is the gravitational acceleration.

Together with the constraint equation $0 = x_1^2 + x_2^2 - l^2$, the Lagrange function is defined by

$$\mathcal{L} = E_k - E_p - \lambda(x_1^2 + x_2^2 - l^2) = \frac{1}{2}m(\dot{x}_1^2 + \dot{x}_2^2) - mgx_2 - \lambda(x_1^2 + x_2^2 - l^2) \tag{4.3}$$

where λ is the Lagrange multiplier.

Then, the equations of motion are of the form

$$0 = \frac{d}{dt}\left(\frac{\partial \mathcal{L}}{\partial \dot{q}}\right) - \frac{\partial \mathcal{L}}{\partial q} \tag{4.4}$$

for the variables $q = x_1, x_2, \lambda$, i.e. we obtain

$$\begin{aligned} 0 &= m\ddot{x}_1 + 2x_1\lambda, \\ 0 &= m\ddot{x}_2 + 2x_2\lambda + mg, \\ 0 &= x_1^2 + x_2^2 - l^2. \end{aligned} \tag{4.5}$$

This system can be transformed into a quasi-linear DAE by introducing the velocity variables $(v_1, v_2) = (\dot{x}_1, \dot{x}_2)$ in order to eliminate the second time derivatives

$$\begin{aligned} \dot{x}_1 &= v_1, \\ \dot{x}_2 &= v_2, \\ m\dot{v}_1 &= -2x_1\lambda, \\ m\dot{v}_2 &= -2x_2\lambda - mg, \\ 0 &= x_1^2 + x_2^2 - l^2. \end{aligned} \tag{4.6}$$

Following Procedure 3 in order to determine the s-index, the d-index and the hidden constraints of the DAE (4.6) yields after the first differentiation of the constraint

$$\begin{aligned} \dot{x}_1 &= v_1, \\ \dot{x}_2 &= v_2, \\ m\dot{v}_1 &= -2x_1\lambda, \\ m\dot{v}_2 &= -2x_2\lambda - mg, \\ 0 &= x_1v_1 + x_2v_2. \end{aligned} \tag{4.7}$$

By differentiating the constraint once more we obtain

$$\begin{aligned} \dot{x}_1 &= v_1, \\ \dot{x}_2 &= v_2, \\ m\dot{v}_1 &= -2x_1\lambda, \\ m\dot{v}_2 &= -2x_2\lambda - mg, \\ 0 &= v_1^2 + v_2^2 - \frac{2}{m}\left(x_1^2 + x_2^2\right)\lambda - gx_2. \end{aligned} \tag{4.8}$$

Further differentiating of the constraint of (4.8) would lead to an ODE in implicit form. Thus, the s-index of the mathematical pendulum (4.6) is 2, and the d-index is 3. Additionally, (4.7) has s-index 1 and d-index 2, (4.8) has s-index 0 and d-index 1.

A strangeness-free regularisation of the mathematical pendulum is given by (see Example 25):

$$\begin{aligned} x_2\dot{x}_1 - x_1\dot{x}_2 &= x_2 v_1 - x_1 v_2, \\ m x_2\dot{v}_1 - m x_1\dot{v}_2 &= -2x_1x_2\lambda + 2x_1x_2\lambda + m x_1 g, \\ 0 &= x_1^2 + x_2^2 - l^2, \\ 0 &= x_1 v_1 + x_2 v_2, \\ 0 &= v_1^2 + v_2^2 - \frac{2}{m}\left(x_1^2 + x_2^2\right)\lambda - g x_2. \end{aligned} \tag{4.9}$$

4.2 Reheat furnace model

Introduction

In the steel production industry, many steel products are hot rolled in a hot mill. The products entering the hot mill are relatively cold and need to be heated up before they can be rolled which is done in a so-called reheat furnace. Figure 4.2 shows a schematic side view of a slab reheat furnace for a hot mill. Burners are located at the sides, top and bottom in order to heat the steel slabs. The intensity of the burner flames can be adjusted by fuel flows where the air-to-fuel ratio is controlled automatically. The cold steel slabs enter the furnace on the left hand side, move slowly through the furnace and are pushed out hot on the right-hand side. The flow direction of the waste gases is from the right to the left, where they are conveyed away through a chimney.

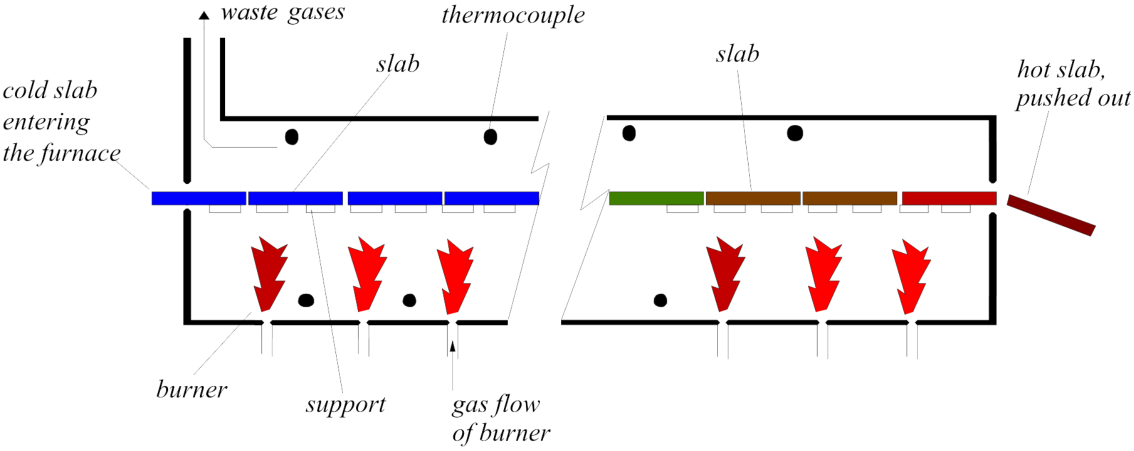


Figure 4.2: Schematic view of a reheat furnace (Source: DotX Control Solutions BV 2014b)

Since such reheat furnaces consume considerably large amounts of fuel, the objective is to heat up the steel products in the furnace in such a way that the consumption of fuel is minimized. Therefore, a simplified model of the reheat furnace is developed which is described in the following. For other and more detailed models see e.g. Pike/Citron (1970), Ko et al. (2000), Zhang et al. (2002) and Chen et al. (2003).

Model equations

The simplified model of the temperatures in a reheat furnace is based on a single reference frame fixed to the furnace. Variations across the width of the furnace are ignored since the furnace is loaded with slabs of a uniform length L_s which are pushed sideways. In order to discretise the furnace in space, it is divided into N sections in the length direction where each section has the same length Δx as shown in Figure 4.3. It is assumed that in each section, the temperature of the waste gases as well as the temperature of the steel products are uniformly distributed.

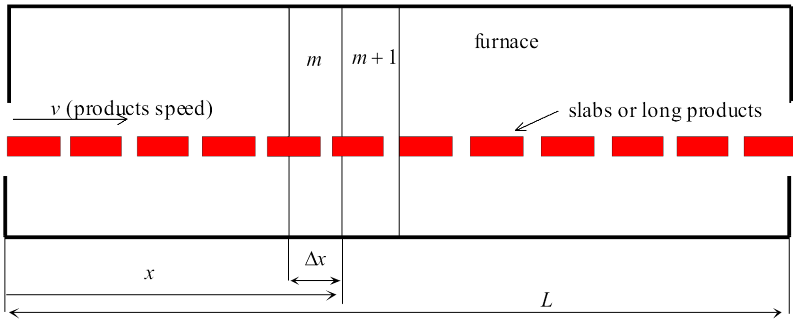
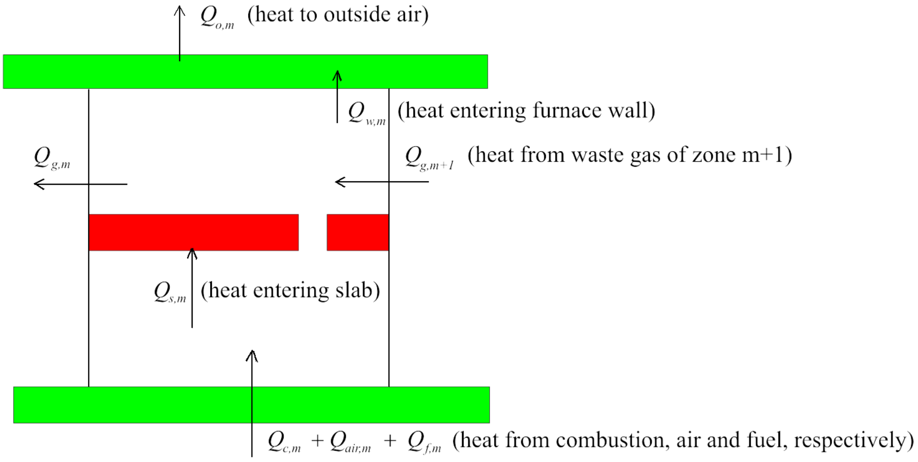


Figure 4.3: Space discretisation of the furnace (Source: DotX Control Solutions BV 2014b)

Figure 4.4 shows the total heat transfer in one section m where the following amounts of heat are considered:

$Q_{g,m+1}$ .. heat brought in by waste gas of section $m+1$,
$Q_{c,m}$ .. heat brought in by combustion in section m,
$Q_{air,m}$ .. heat brought in by air in section m,
$Q_{f,m}$ .. heat brought in by fuel in section m,
$Q_{s,m}$ .. heat entering steel products in section m,
$Q_{w,m}$ .. heat entering the furnace wall in section m,
$Q_{o,m}$ .. heat leaving the furnace wall to outside air in section m.



$Q_{o,m}$ (heat to outside air)
$Q_{w,m}$ (heat entering furnace wall)
$Q_{g,m}$
$Q_{g,m+1}$ (heat from waste gas of zone m+1)
$Q_{s,m}$ (heat entering slab)
$Q_{c,m}+Q_{air,m}+Q_{f,m}$ (heat from combustion, air and fuel, respectively)

Figure 4.4: Heat balance in section m of the furnace (Source: DotX Control Solutions BV 2014b)

Then, the heat balance of a section is given by

$$Q_{g,m+1} - Q_{g,m} + Q_{c,m} + Q_{air,m} + Q_{f,m} - Q_{w,m} - Q_{s,m} = 0 \qquad \text{for} \quad m = 1, \ldots, N \tag{4.10}$$

where $Q_{g,N+1}$ is set to be zero, i.e. there is no heat brought in by waste gas in section N.

The steel slabs are modeled as a flow of steel through the furnace with speed v where the gaps between the single slabs are neglected. Furthermore, it is assumed that the temperature of the steel products T_s only varies in the direction of the length of the furnace x. Hence, the temperature variation of the steel products can be described by the convection equation

$$\rho c_s \left(\frac{\partial T_s}{\partial t} + v \frac{\partial T_s}{\partial x} \right) = \frac{Q_s}{V_s} \tag{4.11}$$

where c_s is the heat capacity of the slabs, ρ is the density of steel, v is the speed of the products travelling through the furnace, Q_s is the heat entering the steel products and V_s is the volume of the steel products.

Using the implicit Euler method for discretising equation (4.11) in space yields

$$\dot{T}_{s,m} = v \, \frac{T_{s,m-1} - T_{s,m}}{\Delta x} + \frac{Q_{s,m}}{\rho c_s V_{s,m}} \qquad \text{for} \quad m = 1, \ldots, N \tag{4.12}$$

where $T_{s,0} = 293$ K, i.e. the steel products entering the furnace are cold.

The heat entering the steel products is assumed to be only due to radiation, and thus is given by

$$Q_{s,m} = A_{s,m} \sigma \varepsilon_s \left(T_{g,m}^4 - T_{s,m}^4 \right) \qquad \text{for} \quad m = 1, \ldots, N \tag{4.13}$$

where $A_{s,m}$ is the surface area of the steel products in section m, σ is the Stefan-Boltzmann constant, ε_s is the emissivity of the steel products, $T_{g,m}$ is the temperature of the waste gas in section m and $T_{s,m}$ is the temperature of the steel products in section m.

The temperature variation of the furnace wall is modeled by

$$c_w \rho_w D_w A_{w,m} \dot{T}_{w,m} = Q_{w,m} - Q_{o,m} \qquad \text{for} \quad m = 1, \ldots, N \tag{4.14}$$

where c_w is the heat capacity of the material of the wall, ρ_w is the density of the wall, D_w is the thickness of the wall, $A_{w,m}$ is the surface area of the wall in section m and $Q_{o,m}$ is the heat leaving the furnace wall to outside air in section m.

Again, the heat is assumed to be only due to radiation, and thus, the heat entering and leaving the furnace wall, respectively, is given by

$$Q_{w,m} = A_{w,m} \sigma \varepsilon_w \left(T_{g,m}^4 - T_{w,m}^4 \right) \qquad \text{for} \quad m = 1, \ldots, N, \tag{4.15}$$

$$Q_{o,m} = A_{w,m} \sigma \varepsilon_w \left(T_{w,m}^4 - T_o^4 \right) \qquad \text{for} \quad m = 1, \ldots, N \tag{4.16}$$

where ε_w is the emissivity of the wall and T_o is the temperature of the air outside the furnace.

Furthermore, the heat produced by combustion in section m is given by

$$Q_{c,m} = \phi_{f,m} H_0 \qquad \text{for} \quad m = 1, \ldots, N \tag{4.17}$$

where $\phi_{f,m}$ is the fuel flow into section m and H_0 is the lower calorific value of fuel.

The specific amounts of heat of fuel, air and waste gas are computed by

$$\begin{aligned} Q_{f,m} &= \phi_{f,m} c_f T_f && \text{for} \quad m = 1, \ldots, N, && (4.18) \\ Q_{air,m} &= \phi_{f,m} R_{af} c_{air} T_{air} && \text{for} \quad m = 1, \ldots, N, && (4.19) \\ Q_{g,m} &= \phi_{g,m} c_g T_{g,m} && \text{for} \quad m = 1, \ldots, N && (4.20) \end{aligned}$$

where c_f is the heat capacity of fuel, c_{air} is the heat capacity of air, c_g is the heat capacity of waste gas, R_{af} is the ratio of air to fuel, $\phi_{g,m}$ is the waste gas flow from section m into $m-1$, T_f is the temperature of the fuel and T_{air} is the temperature of the air used in combustion.

Under the assumption of incompressibility, the waste gas flow from section m follows from the conservation of volume

$$\phi_{g,m} = \phi_{g,m+1} + \phi_{f,m} + \phi_{f,m} R_{af} \qquad \text{for} \quad m = 1, \ldots, N \tag{4.21}$$

where $\phi_{g,N+1}$ is set to be zero, i.e. there is no waste gas flow entering section N.

Finally, it is assumed that the furnace consists of three zones, a preheat zone from $x = 0\,\text{m}$ to $15\,\text{m}$, a heating zone from $x = 15\,\text{m}$ to $22.8\,\text{m}$ and a soaking zone from $x = 22.8\,\text{m}$ to $30\,\text{m}$ where $L = 30\,\text{m}$ is the length of the furnace. In each zone, there is one burner and the heat of each burner is assumed to be equally distributed over all sections of the corresponding zone such that for $N = 50$, we have

$$\phi_{f,m} = \begin{cases} \frac{1}{25} u_1 & \text{for } m = 1, \ldots, 25 \\ \frac{1}{13} u_2 & \text{for } m = 26, \ldots, 38 \\ \frac{1}{12} u_3 & \text{for } m = 39, \ldots, 50 \end{cases} \tag{4.22}$$

where u_1, u_2, u_3 are the fuel flows to burners 1, 2, 3, respectively.

Control criteria

The objective of the reheat furnace model is to maximise the production speed at minimal energy cost. This is to be achieved by adjusting the fuel flows u_1, u_2, u_3 and the furnace speed v.

Thus, the objective function J is defined as

$$J = \int_0^T c_{prod} v \rho L_s D_s - c_{fuel}(u_1 + u_2 + u_3)\, dt \tag{4.23}$$

where c_{prod} is the production profit, c_{fuel} is the cost of fuel and T is the time horizon.

The inputs are constrained as follows:

$$0 < u_1 < 4\,\mathrm{Nm}^3/\mathrm{s}, \quad 0 < u_2 < 2\,\mathrm{Nm}^3/\mathrm{s}, \quad 0 < u_3 < 1.5\,\mathrm{Nm}^3/\mathrm{s}, \quad 0 < v < v_{\max}. \tag{4.24}$$

Furthermore, the dropout temperature of each steel slab must remain within a predefined window such that we have the following output constraints:

$$1478\,\mathrm{K} < T_{s,N} < 1523\,\mathrm{K}. \tag{4.25}$$

In addition, the temperature of the waste gas is limited such that we have

$$T_{g,m} < 1573\,\mathrm{K} \qquad \text{for} \quad m = 1, \ldots, N. \tag{4.26}$$

As initial condition it is supposed that the furnace starts up cold, i.e.

$$T_{g,m}(0) = T_{s,m}(0) = T_{w,m}(0) = 293\,\mathrm{K} \qquad \text{for} \quad m = 1, \ldots, N. \tag{4.27}$$

Model parameters

The values of the parameters used in the model are given in the following table:

Parameter	Meaning	Value	Unit
L	length of the furnace	30	m
N	number of sections	50	-
B	width of the furnace	10	m
H	height of the furnace	2	m
ρ	density of steel products	7800	kg/m^3
ρ_w	density of furnace wall	1000	kg/m^3
c_s	heat capacity of steel products	650	J/(kg K)
c_w	heat capacity of furnace wall	840	J/(kg K)
c_f	heat capacity of fuel	1000	J/(Nm^{3}K)
c_g	heat capacity of waste gas	1700	J/(Nm^{3}K)
c_{air}	heat capacity of air	500	J/(Nm^{3}K)
L_s	length of steel products	8	m
D_s	thickness of steel products	0.2	m
D_w	thickness of furnace wall	0.4	m
T_o	temperature of air outside the furnace	373	K
T_f	temperature of fuel	773	K
T_{air}	temperature of air used in combustion	773	K
σ	Stefan-Boltzmann constant	$5.67 \cdot 10^{-8}$	W/(m^{2}K^4)
H_0	lower calorific value of fuel	30	MJ/(Nm^3)
ε_w	emissivity of furnace wall	0.8	-
ε_s	emissivity of steel products	0.8	-
R_{af}	ratio of air to fuel	10	-
T	time horizon	15	h
c_{prod}	production profit	0.5	€/kg
c_{fuel}	cost of fuel	0.2	€/(Nm^3)

Table 4.1: Parameter values of the reheat furnace model

Chapter 5

Numerical results

In this chapter, the numerical results of several implementations of BDF methods and RadauIIa methods applied to the two test problems are presented. Section 5.1 deals with the efficiency of different approaches for solving systems of nonlinear equations. In section 5.2, the results of the different formulations of the mathematical pendulum are compared and the advantages of a strangeness-free regularisation are shown. Finally, in section 5.3, a comparison of the implemented methods for the two test problems with respect to computing time and accuracy is presented and a best-practice method based on these results is suggested.

In order to compare the calculated numerical solutions in section 5.2 and 5.3 with respect to accuracy, the following parameters for determining the base solutions are used:

	Method	Tolerance	Step size	End point
Pendulum (regularised s-index 0 formulation)	RadauIIa $s = 3$	$tol = 10^{-10}$	$h = 0.0001\,s$	$T = 100\,s$
Furnace	RadauIIa $s = 3$	$tol = 10^{-12}$	$h = 3\,s$	$T = 5\,h$

Table 5.1: Parameter values of the base solutions

Then, the accuracy is determined by the maximum norm of the difference between the calculated numerical solution and the base solution restricted to the evaluated points of the numerical solution.

All calculations are done with MATLAB 8.3 (R2014a) on a 32-bit Windows 7 system with an Intel® Core™ i3 CPU (2.26 GHz) and 4 GB RAM.

5.1 Numerical solution of systems of nonlinear equations

In this section, various approaches for solving systems of nonlinear equations are analysed. Therefore, the intrinsic MATLAB function `fsolve` and several different versions of the Newton method, where the required Jacobian is either given by the user, estimated at each discretisation point by central differences or determined with the MATLAB Symbolic Math Toolbox, are compared. For all approaches, the same prescribed tolerance is used.

Figure 5.1 shows the resulting computing times in dependence of several end points T for the different solvers for systems of nonlinear equations for the mathematical pendulum. The implicit Euler method with step size $h = 0.001\,s$ is used where the arising system of nonlinear equations of size 5×5 in each time step is solved by the different approaches, all with prescribed tolerance $tol = 10^{-10}$. As anticipated, the Newton method with given Jacobian yields the best results, while the MATLAB intrinsic function `fsolve` is the slowest. The Newton method with estimated Jacobian lies in between. Unexpectedly, the Newton method with symbolic Jacobian is almost as good as the Newton method with given Jacobian. Since the given Jacobian must be determined by the user for every single time stepping method and for every test problem, this approach is associated with a large effort for the user. Therefore, the usage of the Newton method with symbolic Jacobian is preferred.

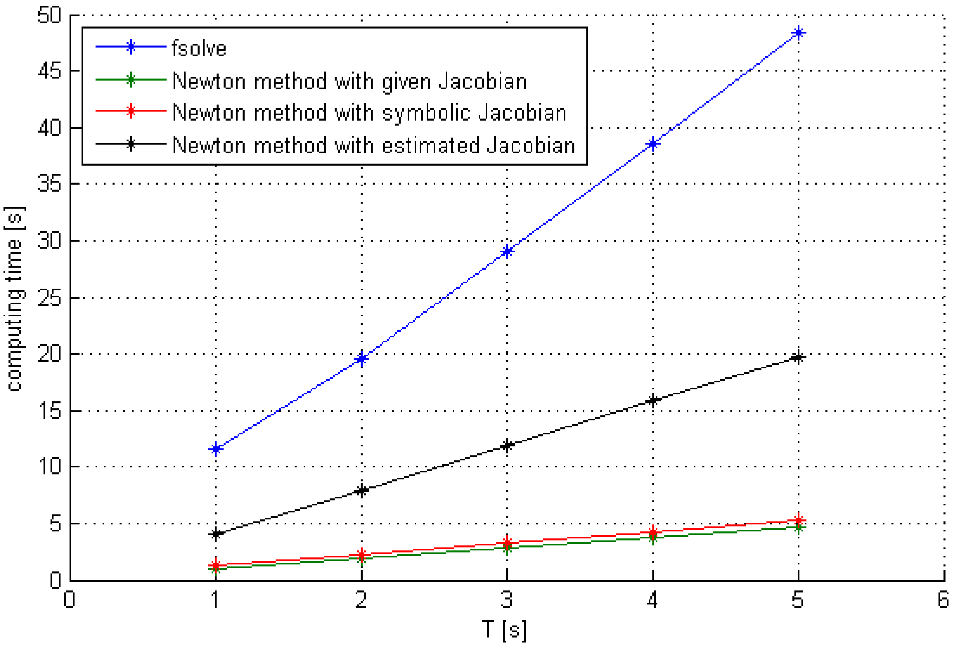


Figure 5.1: Different solvers for systems of nonlinear equations for the pendulum (implicit Euler method, $h = 0.001\,s$, $tol = 10^{-10}$)

In Figure 5.2, the resulting computing times in dependence of several end points T for the different solvers for systems of nonlinear equations for the reheat furnace model are depicted. The implicit Euler method with step size $h = 60\,s$ is used where the arising system of nonlinear equations in each time step is solved by different approaches, all with prescribed tolerance $tol = 10^{-10}$. For this problem, the Newton method with estimated Jacobian is as slow as the MATLAB intrinsic function `fsolve`. Furthermore, the Newton method with given Jacobian is not calculated due to the big effort for such a large system of size 150×150. Thus, the Newton method with symbolic Jacobian yields clearly the best results.

In consequence of these results, the Newton method in combination with the symbolic Jacobian is used in the following.

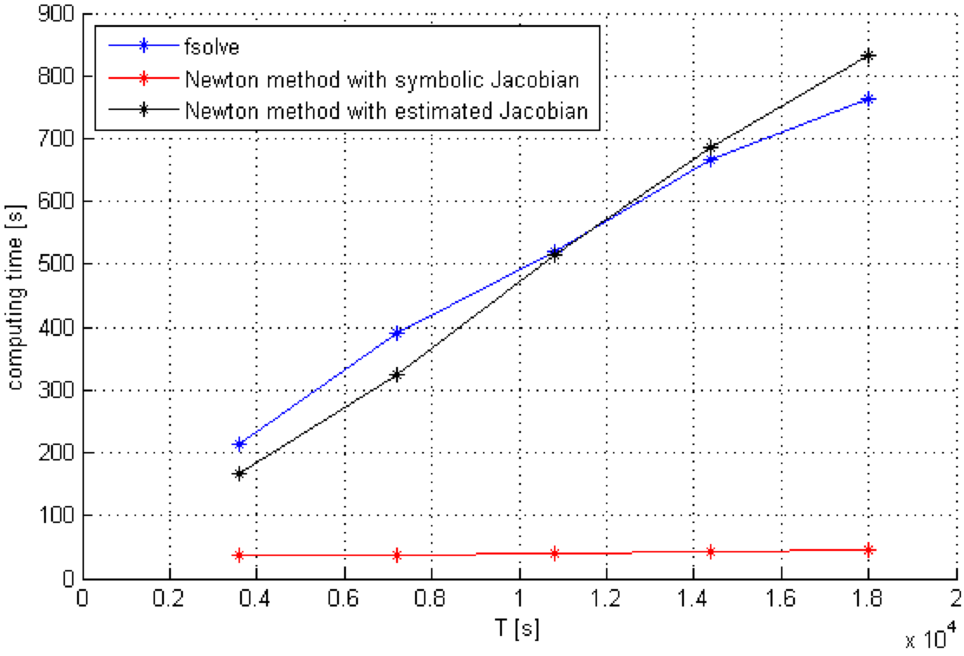


Figure 5.2: Different solvers for systems of nonlinear equations for the furnace (implicit Euler method, $h = 60\,s$, $tol = 10^{-10}$)

5.2 Index reduction and regularisation

In this section, the influence of the different formulations of the mathematical pendulum is analysed. Therefore, several implementations of BDF methods and RadauIIa methods are applied to the model equations, and the reached accuracy with respect to the step size is determined for all formulations of the pendulum. Furthermore, the residuals of the constraints are calculated for the BDF $k = 6$ method. Four formulations of the mathematical pendulum are considered: the original s-index 2 formulation (4.6), the regularised s-index 0 formulation (4.9) as well as the s-index 1 and s-index 0 formulations (4.7) and (4.8) obtained by index reduction by differentiation of the constraints.

In Figures 5.3-5.6, the accuracy of the different methods with respect to the step size for the four formulations is depicted. A closer look at Figure 5.4 shows that the original s-index 2 formulation of the pendulum seems to have a lower bound of 10^{-4} for the accuracy, although the prescribed tolerance is 10^{-10}. So, this formulation is not suitable for numerical integration. The reached accuracy for the s-index 1 formulation (Figure 5.5) gives similar results as the regularised s-index 0 formulation (Figure 5.3), except that the accuracy for the RadauIIa $s = 3$ method does not fall below 10^{-8}. Furthermore, the s-index 0 formulation (Figure 5.6) also gives similar results as the regularised s-index 0 formulation (Figure 5.3), except that for the implicit Euler method (RadauIIa $s = 1$ and BDF $k = 1$) and relatively big step sizes h the value of the accuracy is much higher. To explain this behaviour, an analysis of the residuals of the constraints is in order.

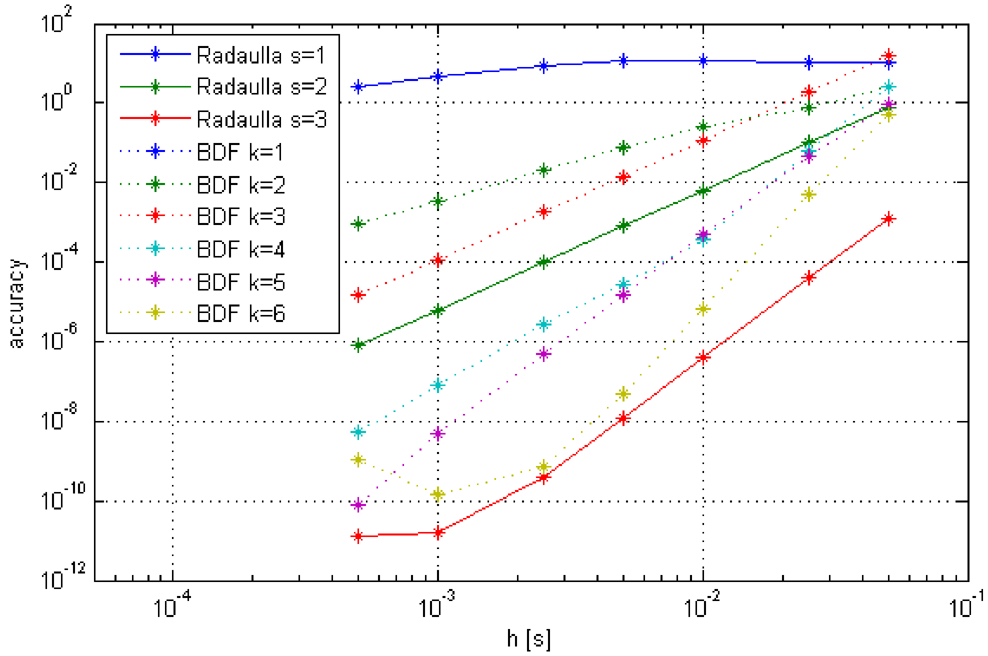


Figure 5.3: Accuracy of different methods with respect to the step size h for the regularised s-index 0 formulation ($tol = 10^{-10}$, $T = 10\,s$)

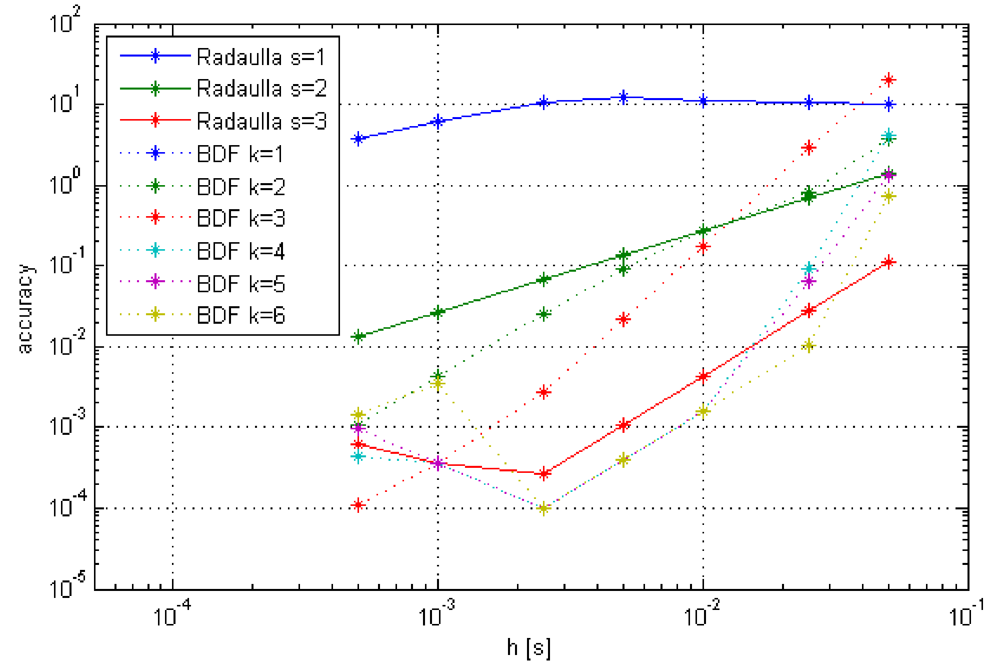


Figure 5.4: Accuracy of different methods with respect to the step size h for the s-index 2 formulation ($tol = 10^{-10}$, $T = 10\,s$)

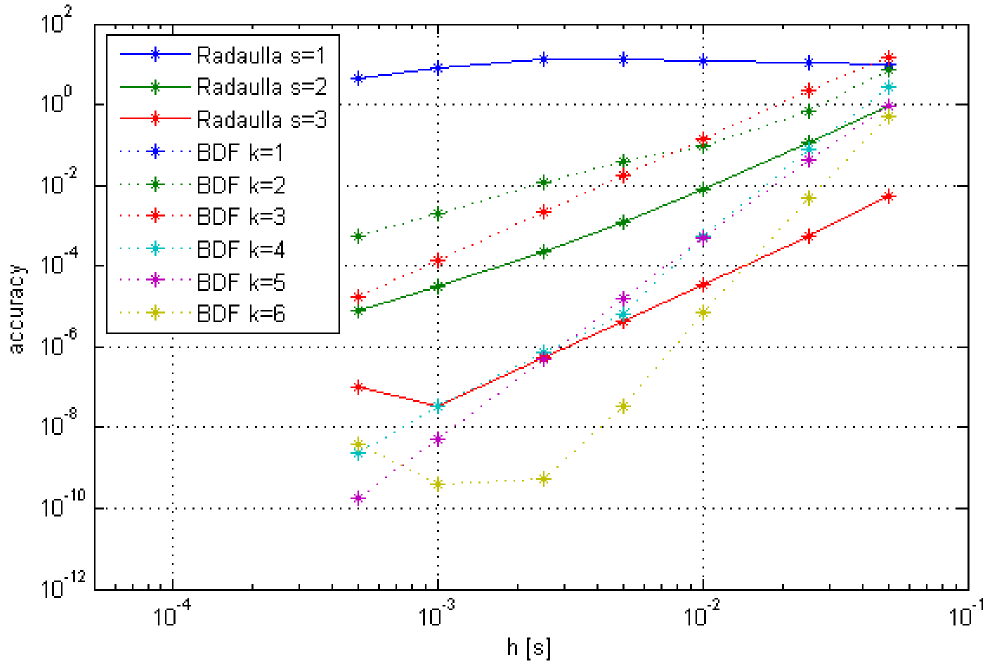


Figure 5.5: Accuracy of different methods with respect to the step size h for the s-index 1 formulation ($tol = 10^{-10}$, $T = 10\,s$)

Figures 5.7-5.10 show the residuals of the constraints for the BDF $k = 6$ method with $h = 0.005\,s$, $tol = 10^{-10}$ and $T = 1000\,s$. So, compared to the accuracy analysis, a much higher end point T is chosen in order to show the numerical effects more clearly. All three constraints are considered: the constraint of level 0 ($0 = x_1^2 + x_2^2 - l^2$), the constraint of level 1 ($0 = x_1v_1 + x_2v_2$) and the constraint of level 2 ($0 = v_1^2 + v_2^2 - \frac{2}{m}(x_1^2 + x_2^2)\lambda - gx_2$).

For the regularised s-index 0 formulation (Figure 5.7) the residuals of the constraints are all below the prescribed tolerance as desired, whereas for the original s-index 2 formulation (Figure 5.8) only the residual of the constraint of level 0 is satisfactory. The residuals of the constraints of levels 1 and 2 are oscillating around zero with an amplitude higher than the prescribed tolerance $tol = 10^{-10}$. This is caused by the fact that both constraints are hidden in the DAE and not explicitly given as in the regularised s-index 0 formulation. Therefore, the numerical method notices that the hidden constraints are violated and tries to correct this, but overreacts, leading to the occurrence of these oscillations. Furthermore, the deeper the constraints are hidden, the stronger are the overreactions and therefore the amplitude of the oscillations. In the s-index 1 formulation (Figure 5.9), the constraint of level 0 is no longer contained in the system of equations and the constraint of level 1 is explicitly given, whereas the constraint of level 2 is hidden. Therefore, a drift of the residual of the constraint of level 0 can be observed due to the discretisation error and the residual of the constraint of level 2 is oscillating with an amplitude higher than 10^{-10}. At last, in the s-index 0 formulation (Figure 5.10) both constraints of level 0 and 1 are not contained in the DAE and the constraint of level 2 is explicitly given. Thus, a drift of the residuals of the constraints of level 0 and 1 can be observed, whereas the residual of the constraint of level 2 is below the prescribed tolerance.

In summary, formulations which are not strangeness-free contain hidden constraints which are causing oscillations of the residuals with amplitudes higher than the prescribed tolerance. Furthermore, index reduction by differentiation of the constraints lowers the index, but also removes the constraints from the system such that a drift of the residuals can be observed. Therefore, the regularised s-index 0 formulation is the most suitable formulation since all constraints are explicitly given in the system.

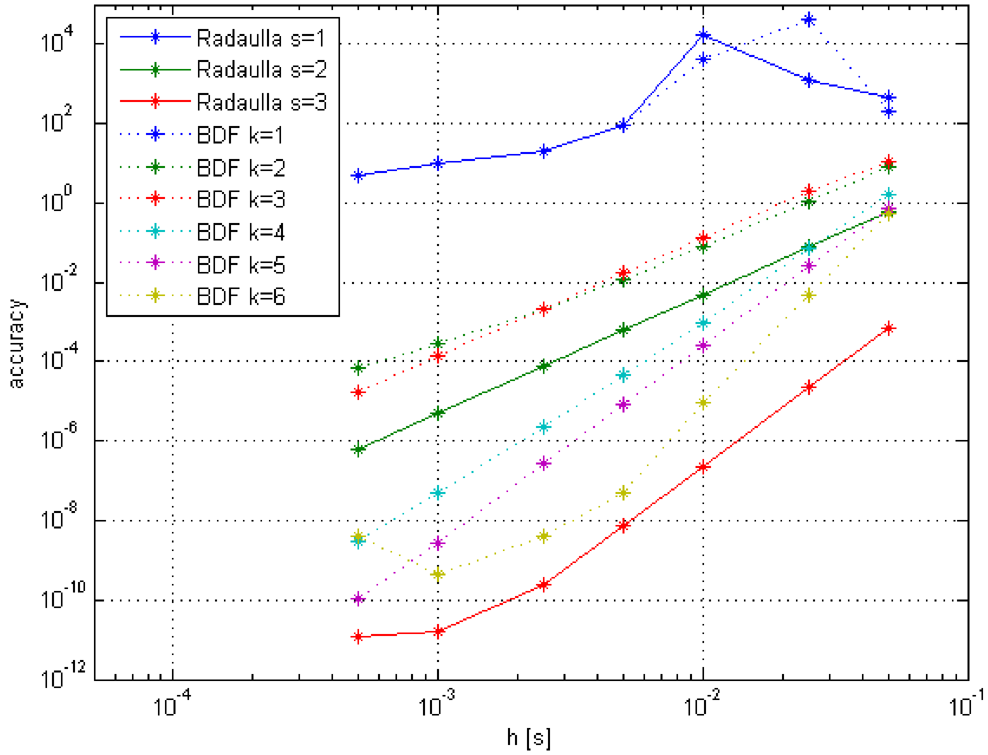


Figure 5.6: Accuracy of different methods with respect to the step size h for the s-index 0 formulation ($tol = 10^{-10}$, $T = 10\,s$)

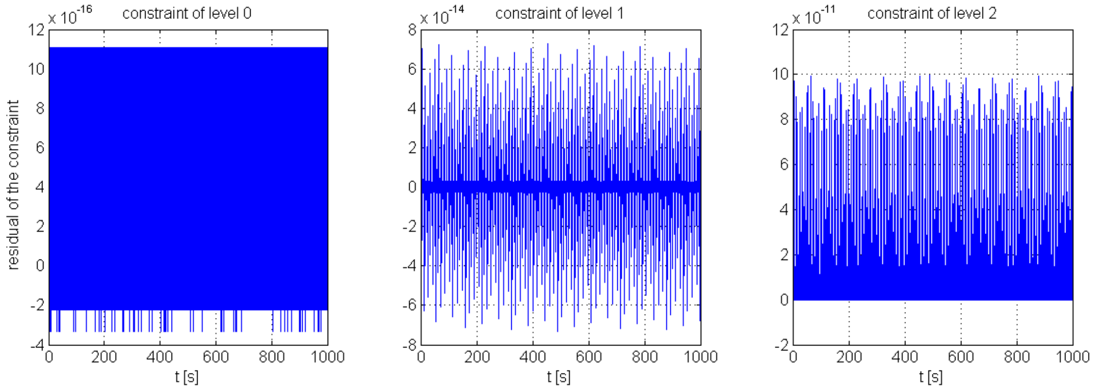


Figure 5.7: Residuals of the constraints for the regularised s-index 0 formulation (BDF $k = 6$ method, $h = 0.005\,s$, $tol = 10^{-10}$, $T = 1000\,s$)

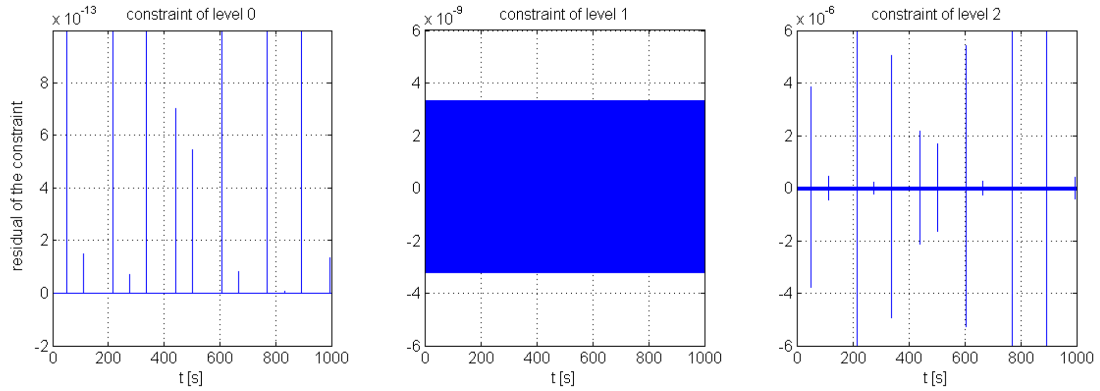


Figure 5.8: Residuals of the constraints for the s-index 2 formulation (BDF $k = 6$ method, $h = 0.005\,s$, $tol = 10^{-10}$, $T = 1000\,s$)

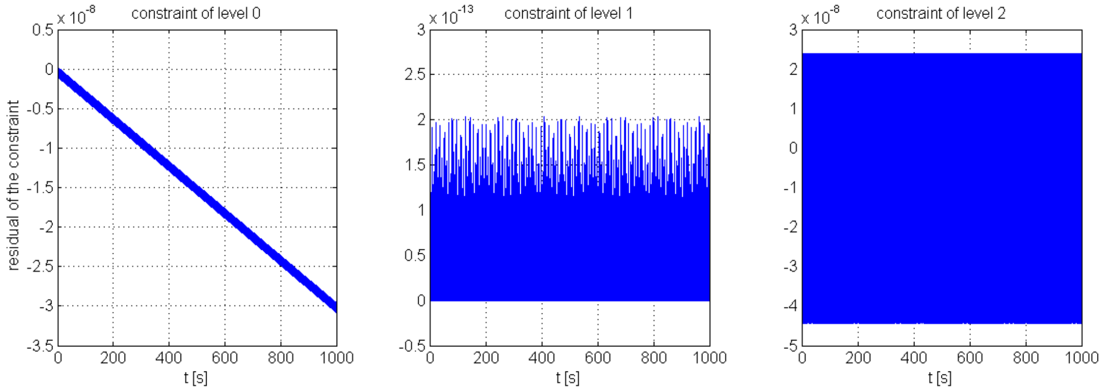


Figure 5.9: Residuals of the constraints for the s-index 1 formulation (BDF $k = 6$ method, $h = 0.005\,s$, $tol = 10^{-10}$, $T = 1000\,s$)

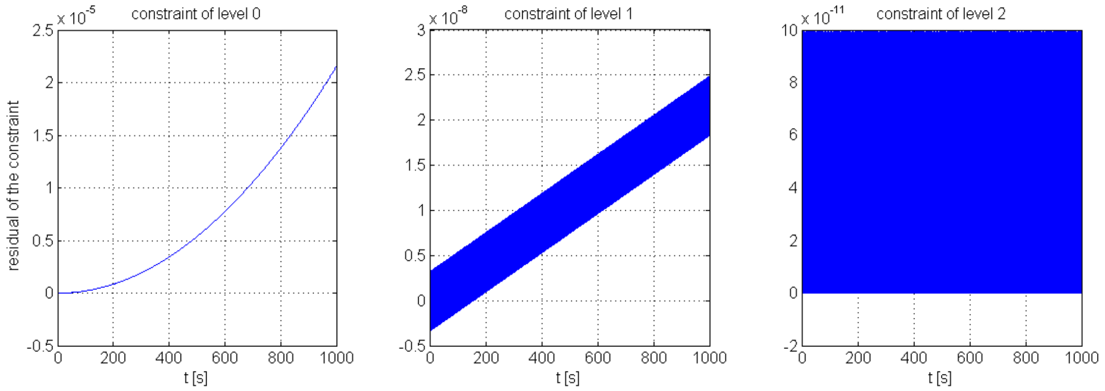


Figure 5.10: Residuals of the constraints for the s-index 0 formulation (BDF $k = 6$ method, $h = 0.005\,s$, $tol = 10^{-10}$, $T = 1000\,s$)

5.3 Computing time and accuracy

In this section, a comparison of the implemented methods for the two test problems with respect to computing time and accuracy is presented. The first subsection deals with the mathematical pendulum and the second subsection addresses the reheat furnace model.

5.3.1 Mathematical pendulum

In Figure 5.3, we have already seen the accuracy of the different methods with respect to the step size h for the regularised s-index 0 formulation. As expected, methods with a higher order achieve a good accuracy for relatively large step sizes h, whereas methods with a lower order require relatively small step sizes h in order to achieve a reasonable accuracy.

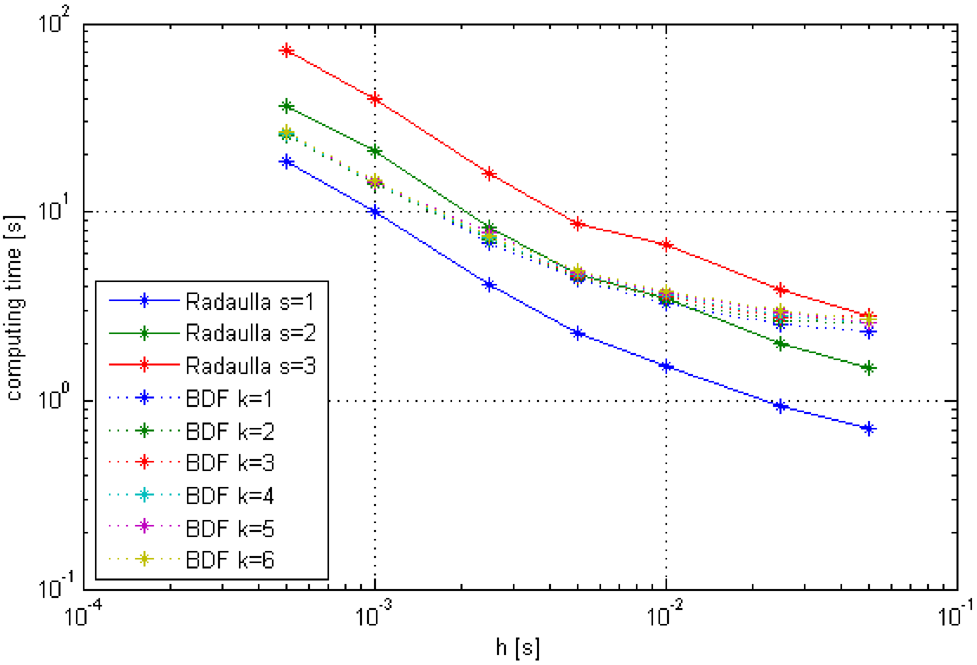


Figure 5.11: Computing time of each of different methods with respect to the step size h for the regularised s-index 0 formulation ($tol = 10^{-10}$, $T = 10\,s$)

Figure 5.11 shows the computing time of each of the different methods with respect to the step size h for the regularised s-index 0 formulation. It can be observed that the RadauIIa methods require more computing time with increasing order due to the larger system of nonlinear equations which has to be solved in each time step, i.e. a system of size 5×5 for RadauIIa $s = 1$, a system of size 10×10 for RadauIIa $s = 2$ and a system of size 15×15 for RadauIIa $s = 3$. In contrast, the BDF methods do not require significantly more computing time with increasing order since the systems of nonlinear equations have the same size of 5×5.

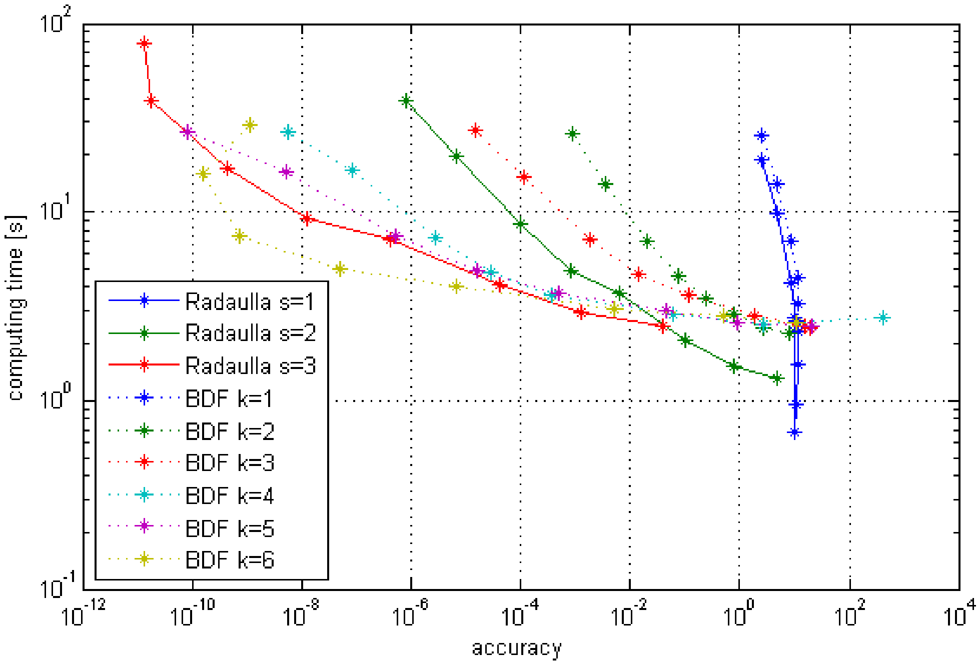


Figure 5.12: Computing time of each of different methods with respect to the reached accuracy for the regularised s-index 0 formulation ($tol = 10^{-10}$, $T = 10\,s$)

In Figure 5.12, the computing time of each of the different methods with respect to the reached accuracy for the regularised s-index 0 formulation is depicted. It shows that the BDF $k = 6$ method is preferable for a desired accuracy better than 10^{-4}, and that the RadauIIa $s = 3$ method should be chosen for a desired accuracy between 10^{-4} and 10^{-2}.

5.3.2 Reheat furnace model

In Figure 5.13, the accuracy of the different methods with respect to the step size h for the furnace is depicted. It is striking that the RadauIIa $s = 3$ method yields the best accuracy results. Furthermore, it can be observed that the BDF $k = 6$ method has two runaway values for the step sizes $h = 120\,s$ and $h = 240\,s$. For these large step sizes, the Newton method does not converge within the maximum of 100 iterations. Therefore, for these cases the reached accuracy is extremely bad.

Figure 5.14 shows the computing time of each of the different methods with respect to the step size h for the furnace. It can be observed that the RadauIIa methods require more computing time with increasing order due to the larger system of nonlinear equations which has to be solved in each time step, i.e. a system of size 150×150 for RadauIIa $s = 1$, a system of size 300×300 for RadauIIa $s = 2$ and a system of size 450×450 for RadauIIa $s = 3$. In contrast, the BDF methods do not require significantly more computing time with increasing order since the systems of nonlinear equations have the same size of 150×150. Unexpectedly, the BDF methods require much more computing time than the RadauIIa $s = 1$ method, although the system of nonlinear equations has the same size. We will investigate this further in the next paragraphs.

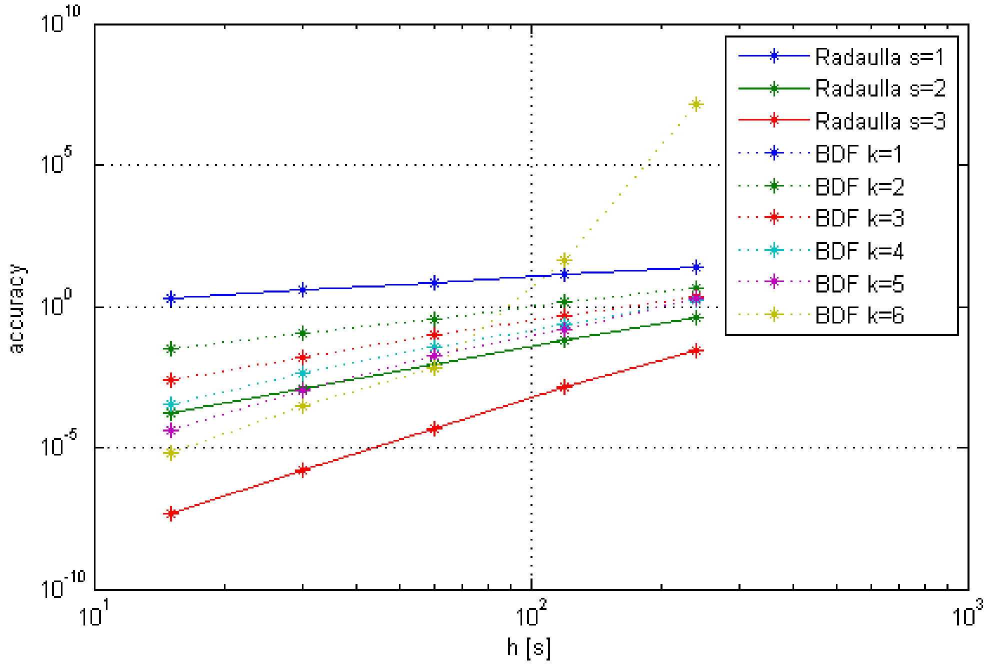


Figure 5.13: Accuracy of different methods with respect to the step size h for the furnace ($tol = 10^{-10}$, $T = 5\,h$)

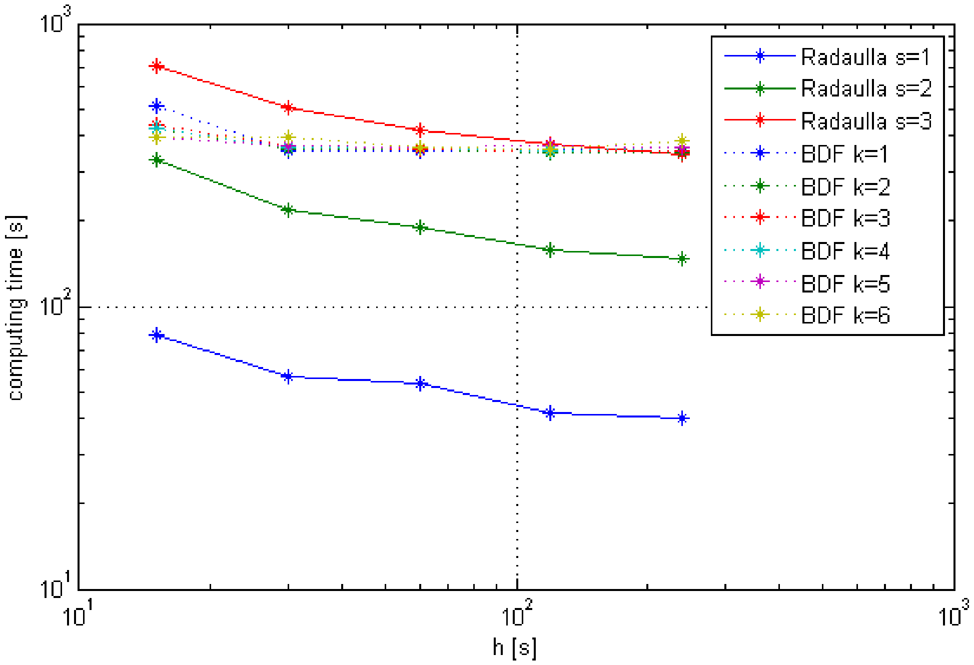


Figure 5.14: Computing time of each of different methods with respect to the step size h for the furnace ($tol = 10^{-10}$, $T = 5\,h$)

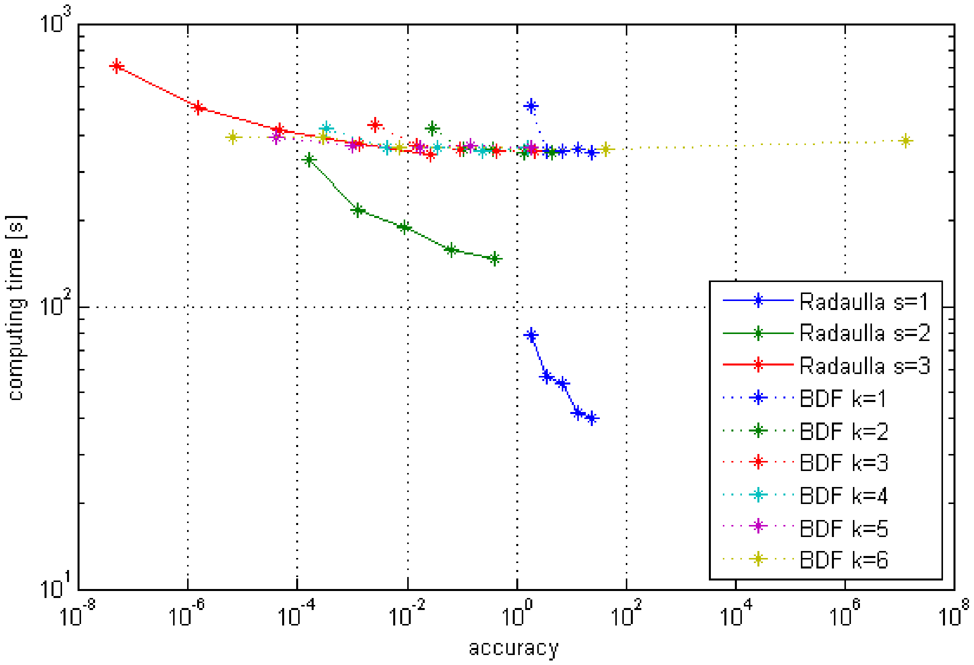


Figure 5.15: Computing time of each of different methods with respect to the reached accuracy for the furnace ($tol = 10^{-10}$, $T = 5\,h$)

In Figure 5.15, the computing time of each of the different methods with respect to the reached accuracy for the furnace is depicted. It shows that the RadauIIa $s = 2$ method is preferable for a desired accuracy above 10^{-4}. For a desired accuracy better than 10^{-4} either the RadauIIa $s = 3$ method or one of the higher order BDF methods ($k = 5$ or $k = 6$) should be chosen.

It is remarkable that the RadauIIa $s = 2$ method, and not one of the BDF methods, is the best method for the relevant accuracy range between 10^{-4} and 10^{-2}, although most of the BDF methods have a better order and a smaller system of nonlinear equations which has to be solved in each time step. Furthermore, it is striking that for the BDF methods the slope of the curves in Figure 5.14 is close to zero. This suggests that the time spent for the time integration is insignificant compared to the total computing time. Therefore, a detailed look at the different parts contributing to the computing time is in order.

In Table 5.2, the analysis of the computing time for the BDF methods in absolute values and in percent is shown. It becomes apparent that the time spent for the time integration only contributes 3 % to the total time, while determining the symbolic Jacobian takes around 10 % to 11 % of the total time, and most time is spent for determining the initial values, i.e. around 86 % to 87 % of the total time. For all BDF methods, the missing initial values are determined by the RadauIIa $s = 3$ method since it has the highest order of the implemented one-step methods. Therefore, the analysis of the computing time for the RadauIIa methods in absolute values and in percent is shown in Table 5.3. For all RadauIIa methods, the major part of the total computing time with 77 % to 79 % is spent for determining the symbolic Jacobians, while the time integration only takes around 21 % to 23 % of the total time. In particular, determining the symbolic Jacobians for the RadauIIa $s = 3$ method takes 301 s, which explains the high contribution of the time spent for determining the initial values for the BDF methods.

	BDF $k = 1$	BDF $k = 2$	BDF $k = 3$
Time for determining the symbolic Jacobian	38 s (11%)	35 s (10%)	36 s (10%)
Time for determining the initial values	300 s (86%)	307 s (87%)	302 s (87%)
Time for the time integration	12 s (3%)	11 s (3%)	10 s (3%)
Total time	350 s	353 s	348 s

	BDF $k = 4$	BDF $k = 5$	BDF $k = 6$
Time for determining the symbolic Jacobian	38 s (11%)	40 s (11%)	41 s (11%)
Time for determining the initial values	305 s (86%)	303 s (86%)	306 s (86%)
Time for the time integration	11 s (3%)	12 s (3%)	11 s (3%)
Total time	354 s	354 s	358 s

Table 5.2: Analysis of the computing time of the BDF methods for the furnace ($tol = 10^{-10}$, $h = 60\,s$, $T = 5\,h$)

	RadauIIa $s = 1$	RadauIIa $s = 2$	RadauIIa $s = 3$
Time for determining the symbolic Jacobian	36 s (79%)	130 s (77%)	301 s (77%)
Time for the time integration	9 s (21%)	38 s (23%)	88 s (23%)
Total time	45 s	168 s	390 s

Table 5.3: Analysis of the computing time of the RadauIIa methods for the furnace ($tol = 10^{-10}$, $h = 60\,s$, $T = 5\,h$)

In consequence of the above results, the computing time for the determination of the Jacobians should be considerably decreased in order to have an applicable and efficient numerical solver for DAEs. Therefore, the transformation of the symbolic functions into MATLAB function handles by the function `matlabFunction` is changed by the additional option `'file'`, which causes that instead of a function handle, a function file with optimized code is created (see MathWorks 2014). This procedure only has to be carried out once for every method and problem, such that in this case nine files for the functions themselves and nine files for the Jacobians are necessary. Hence, the resulting computing time of a simulation by using the generated function files is caused mainly by the time integration, i.e. the time integration takes more than 99 % of the total time.

It is remarkable that the generation of the function files takes much more time than creating the MATLAB function handles. In Table 5.4 and 5.5, the computing times for generating the function files for the BDF methods and for the RadauIIa methods, respectively, are shown. For the BDF methods, the computing time for generating the function file for G grows slowly with the order, while the computing time for generating the function files for the Jacobians J_G remains constant with the order. Remember that the size of the system of nonlinear equations is constant for all BDF methods, i.e. a system of size 150 × 150. In contrast, the size of the system of nonlinear equations increases with the order for the RadauIIa methods. Therefore, the computing time for generating the function files increases with the order, but the increase seems to be growing exponentially with the size of the systems. In particular, generating the function file for the Jacobian J_G for the RadauIIa $s = 3$ method takes more than one day, while creating the corresponding function handle with the old procedure took only 301 s. Since the generation of the files has to be done only once, this amount of time is still acceptable. Particularly with regard to the planned usage of the DAE solver for NMPC where the simulation has to be carried out every time a new measurement is available, the generation of the function files should be preferred to the function handles since the function handles are created over and over again for every new simulation. Additionally, the computing time for generating the function files can possibly be improved by using other symbolic differentiation codes which are more efficient than MATLAB.

	BDF $k = 1$	BDF $k = 2$	BDF $k = 3$	BDF $k = 4$	BDF $k = 5$	BDF $k = 6$
Function G	24 s	27 s	32 s	34 s	36 s	40 s
Jacobian J_G	211 s	214 s	217 s	214 s	213 s	212 s

Table 5.4: Computing times of generating the function files and corresponding files for the Jacobians for the BDF methods

	RadauIIa $s = 1$	RadauIIa $s = 2$	RadauIIa $s = 3$
Function G	53 s	7 min	21 min
Jacobian J_G	316 s	203 min	37 h

Table 5.5: Computing times of generating the function files and corresponding files for the Jacobians for the RadauIIa methods

By using the generated function files, new results regarding the computing time with respect to the step size and with respect to the reached accuracy can be obtained, while the results of the accuracy with respect to the step size remain the same (see Figure 5.13).

Figure 5.16 shows the new computing time of each of the different methods with respect to the step size h for the furnace. Compared to the old computing times (see Figure 5.14), now all values of the computing time are much lower for all step sizes and methods. Furthermore, it can be observed that the BDF methods require almost the same amount of computing time as the RadauIIa $s = 1$ method, because the systems of nonlinear equations have the same size. In addition, the computing time of the BDF $k = 6$ method for the step size $h = 240\,s$ is much higher than expected due to the divergence of the Newton method, i.e. the maximum of 100 iterations is exhausted.

In Figure 5.17 the new computing time of each of the different methods with respect to the reached accuracy for the furnace is depicted. It shows that the BDF $k = 6$ method is preferable for a desired accuracy better than 10^{-2}, and that the BDF $k = 5$ method should be chosen for a desired accuracy between 10^{-2} and 10^{-1}.

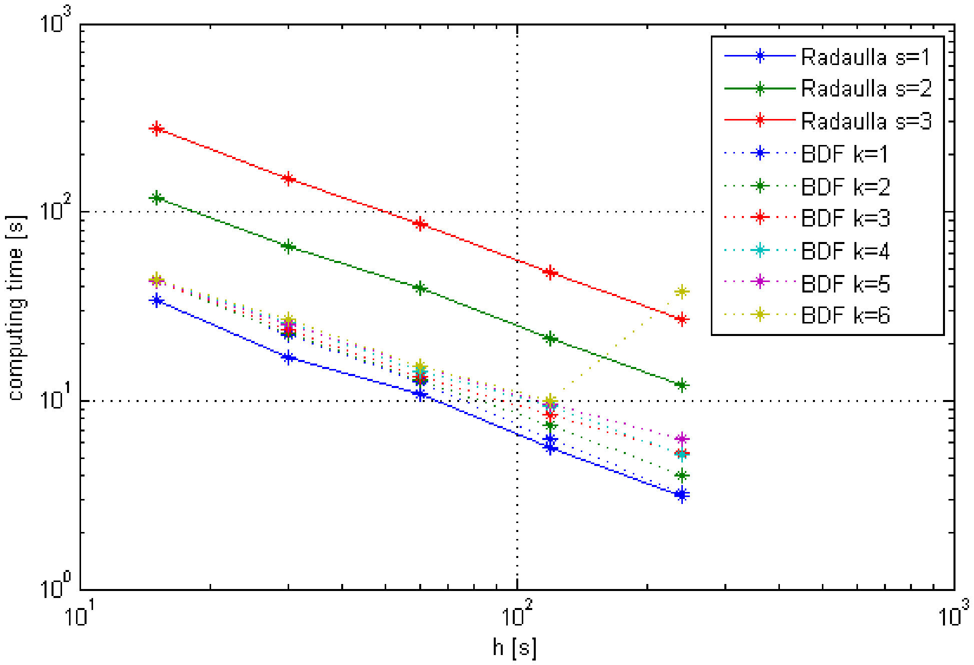


Figure 5.16: Computing time of each of different methods with respect to the step size h for the furnace ($tol = 10^{-10}$, $T = 5\,h$)

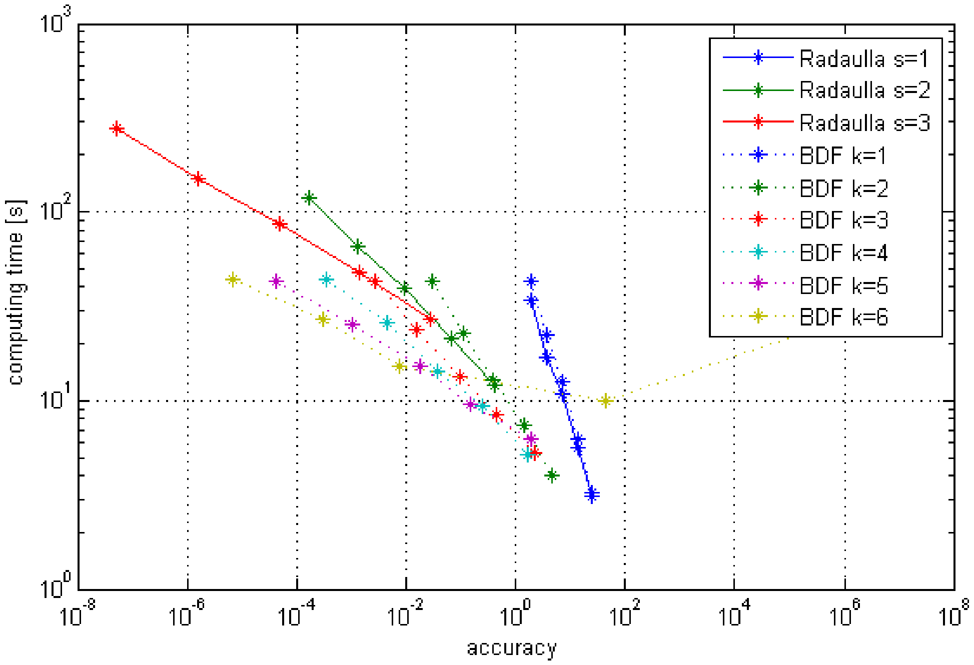


Figure 5.17: Computing time of each of different methods with respect to the reached accuracy for the furnace ($tol = 10^{-10}$, $T = 5\,h$)

Chapter 6

Conclusion and future research

In this thesis, we have investigated three main topics. First, we have intensively focused on the analysis of DAEs. Second, we have investigated the numerical treatment of DAEs. And as a third topic, we have introduced two test examples and applied several implementations of numerical methods to them.

Concerning the analysis of DAEs, we considered the range of DAEs from simple types to more general ones in chapter 2. Starting with linear DAEs with constant coefficient, we successively treated linear DAEs with variable coefficients and finally general nonlinear DAEs. In particular, we investigated different concepts of the so-called index with a main focus on the strangeness index. We have seen that DAEs are not just simple combinations of differential and algebraic equations. Further algebraic constraints can be hidden in the system. These hidden constraints are the main challenge of DAEs since they impose additional consistency conditions on the initial values and cause stronger smoothness requirements of the problem in order to obtain existence and uniqueness results. Therefore, they provoke several difficulties in the analysis as well as in the numerical treatment of DAEs of higher index.

Chapter 3 was concerned with the numerical solution of DAEs. Due to the difficulties which can arise when numerical methods for ODEs are directly used to solve higher index DAEs, we first presented two regularisation techniques which transform the original DAE to a strangeness-free DAE with the same set of solutions. The first approach was based on Hypothesis 2 which was very technical and furthermore requires the determination of the whole derivative array. In contrast, the second approach is based on Procedure 3, which means that it is only necessary to differentiate the algebraic constraints and not the whole DAE. However, Procedure 3 is only applicable for quasi-linear DAEs, while Hypothesis 2 is suited for general nonlinear DAEs. Afterwards, the two main classes of discretisation methods, namely one-step methods and linear multi-step methods, and their application to strangeness-free DAEs were described, and important aspects like the uniqueness of the numerical solution x_{i+1}, the consistency of x_{i+1} as well as the convergence and stability properties of certain methods were discussed. As a result, we have seen that RadauIIa methods and BDF methods are excellent candidates for the numerical solution of initial value problems for strangeness-free DAEs.

In chapter 5, the numerical results of several implementations of BDF methods and RadauIIa methods applied to the two test problems introduced in chapter 4, the mathematical pendulum and a reheat furnace model, were presented and discussed. First, the efficiency of different approaches for solving systems of nonlinear equations was analysed, leading to the conclusion that the usage of the Newton method where the required Jacobians are determined with the MATLAB Symbolic Math Toolbox is to prefer. Second, the different formulations of the mathematical pendulum were compared with respect to the accuracy of their results and the advantages of a strangeness-free regularisation were shown. In particular, it was illustrated that formulations which are not strangeness-free lead to oscillations of the residuals of the constraints. Furthermore, it was shown that index reduction by differentiation of the constraints lowers the index, but also removes the constraints from the system such that a drift of the residuals was observed. Therefore, strangeness-free regularisations where all constraints are stated explicitly are to prefer. Third, a comparison of the implemented methods for the two test problems with respect to computing time and accuracy was presented. For the mathematical pendulum, it turned out that the best method for a desired accuracy better than 10^{-4} is the BDF $k = 6$ method. For the reheat furnace model, the BDF $k = 6$ method is also best for a desired accuracy better than 10^{-2}. But one should be careful with the conjecture that BDF methods are in general better than RadauIIa methods. The reader should keep in mind that RadauIIa methods have much better stability properties, i.e. they are A- and L-stable, such that for stiff problems they could be the better choice. We have already seen this in Figure 5.13 where the BDF $k = 6$ method has two runaway values for the step sizes $h = 120\,s$ and $h = 240\,s$ which can be caused by stability problems. Another conclusion of this chapter was that the generation of the Jacobians during the numerical integration is very time consuming compared to the time spent for the time integration, especially for the reheat furnace model. Thus, in order to reduce the computing time of the numerical methods, the required Jacobians for every method and problem were generated and stored in a function file in advance.

Finally, some directions for future research that might build up on this thesis work. In particular, there are many possibilities to improve the current implementation of the DAE solver:

- It would be desirable to implement an automated detection that recognises if the provided DAE is not strangeness-free. Furthermore, an automated determination of the strangeness index as well as of a strangeness-free regularisation of the considered DAE would be desirable. Until now, this effort has to be done by the user in advance, i.e. the user has to determine the index of the considered problem and if the index is not zero, the user has to provide a regularisation of the problem in order to get reliable results of the numerical integration. Thus, an implementation of Procedure 3 and subsequently an automated regularisation of the DAE would be desirable.
- The performance of the DAE solver could be improved by implementing a step size control, i.e. not to use a fixed step size h anymore.

- Furthermore, the used method for the numerical solution of systems of nonlinear equations could be improved. The implemented ordinary Newton method belongs to the so-called local Newton methods, which require sufficiently good initial guesses for convergence. In contrast, global Newton methods are able to compensate bad initial guesses for instance by damping or trust region strategies (see Deuflhard 2011).
- Another approach for the numerical solution of systems of nonlinear equations could be the implementation of the simplified Newton method which uses a fixed approximation of the Jacobian for several or all steps of the Newton iteration process. Especially for large systems where the determination of the Jacobian at discrete values is costly, this could improve the overall computing time, although the method is only linearly convergent and not quadratic convergent anymore.
- The determined Jacobians for the Newton iteration are so far not considered to be sparse, although in many applications this would be the case. By implementing sparsity of the Jacobians, the efficiency of the solution of the arising linear systems in the Newton iteration could be improved.
- In addition, for extremely large systems of nonlinear equations, it would be desirable to solve the arising linear systems in the Newton iteration with an iterative solver, e.g. GMRES. Until now, the backslash operator of MATLAB is used for the solution of the linear system. Thus, the linear system is solved by a direct solver, namely the LU decomposition.
- We have seen that the required computing time for the determination of the Jacobians and the following generation of the function files increases extremely with the size of the system. For a system of size 450×450, it already takes more than one day (see Table 5.5). Therefore, it is considerable to use another more efficient programme than MATLAB for the determination of the symbolic derivatives in order to be able to treat problems with even larger system size.

In conclusion, we are confident that we have developed efficient numerical methods for solving DAEs in order to enable a prospectively follow-up thesis project to develop an applicable NMPC algorithm with DAEs as system models.

Bibliography

Bollhöfer, Matthias / Mehrmann, Volker (2004): *Numerische Mathematik.* Vieweg, Wiesbaden.

Brenan, K. E. / Campbell, Stephen L. / Petzold, Linda R. (1996): *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* Society for Industrial and Applied Mathematics, Philadelphia.

Campbell, Stephen L. (1987): *A general form for solvable linear time varying singular systems of differential equations.* SIAM Journal on Mathematical Analysis, Volume 18, Issue 4, pp. 1101–1115.

Campbell, Stephen L. / Gear, C. William (1995): *The index of general nonlinear DAEs.* Numerische Mathematik, Volume 72, Issue 2, pp. 173–196.

Campbell, Stephen L. / Griepentrog, Eberhard (1995): *Solvability of general differential algebraic equations.* SIAM Journal on Scientific Computing, Volume 16, Issue 2, pp. 257–270.

Campbell, Stephen L. / Meyer, Carl D. (1979): *Generalized Inverses of Linear Transformations.* Pitman, San Francisco.

Chen, Zhigang / Xu, Chao / Zhang, Bin / Shao, Huihe / Zhang, Jianmin (2003): *Advanced control of walking-beam reheating furnace.* Journal of University of Science and Technology Beijing, Volume 10, Issue 4, pp. 69–74.

Deuflhard, Peter (2011): *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms.* Springer Series in Computational Mathematics, Volume 35, Springer-Verlag, Berlin.

DotX Control Solutions BV (2014a): *Process control projects.* ⟨URL: `http://www.dotxcontrol.com/en/projects/process-control.html`⟩.

DotX Control Solutions BV (2014b): *Reheat Furnace Model.* Internal project notes.

DotX Control Solutions BV (2014c): *Wind turbine control projects.* ⟨URL: `http://www.dotxcontrol.com/en/projects/wind-turbine-control.html`⟩.

Gantmacher, F.R. (1959): *The Theory of Matrices, Volume 2.* Chelsea Publishing Company, New York.

GEAR, C. WILLIAM (1988): *Differential-algebraic equation index transformations.* SIAM Journal on Scientific and Statistical Computing, Volume 9, Issue 1, pp. 39–47.

GRIEPENTROG, EBERHARD / MÄRZ, ROSWITHA (1986): *Differential-algebraic equations and their numerical treatment.* Teubner Verlag, Leipzig.

HAIRER, ERNST / LUBICH, CHRISTIAN / ROCHE, MICHEL (1989): *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods.* Springer-Verlag, Berlin.

HAIRER, ERNST / NØRSETT, SYVERT P. / WANNER, GERHARD (1993): *Solving Ordinary Differential Equations I.* Springer-Verlag, Berlin.

HAIRER, ERNST / WANNER, GERHARD (1983): *On the Instability of the BDF Formulas.* SIAM Journal on Numerical Analysis, Volume 20, Issue 6, pp. 1206–1209.

KO, HYUN SUK / KIM, JUNG-SU / YOON, TAE-WOONG / LIM, MOKEUN / YANG, DAE RYUK / JUN, IK SOO (2000): *Modeling and Predictive Control of a Reheating Furnace.* Proceedings of the American Control Conference, Chicago, USA, pp. 2725–2729.

KRONECKER, L. (1890): *Algebraische Reduction der Schaaren bilinearer Formen.* Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, zweiter Halbband, pp. 1225–1237.

KUNKEL, PETER / MEHRMANN, VOLKER (2006): *Differential-Algebraic Equations - Analysis and Numerical Solution.* European Mathematical Society, Zürich.

MATHWORKS (2014): *MATLAB Documentation.* ⟨URL: `http://www.mathworks.de/help/symbolic/matlabfunction.html`⟩.

MÄRZ, ROSWITHA (1981): *Multistep methods for initial value problems in implicit differential-algebraic equations.* Humboldt-Universität zu Berlin, Sektion Mathematik, Preprint No. 22.

PANTELIDES, CONSTANTINOS C. (1988): *The Consistent Initialization of Differential-Algebraic Systems.* SIAM Journal on Scientific and Statistical Computing, Volume 9, Issue 2, p. 213–231.

PETZOLD, LINDA R. (1982): *Differential/Algebraic Equations are not ODE's.* Journal on Scientific and Statistical Computing, Volume 3, Issue 3, pp. 367–384.

PETZOLD, LINDA R. (1986): *Order Results for Implicit Runge-Kutta Methods Applied to Differential/Algebraic Systems.* SIAM Journal on Numerical Analysis, Volume 23, Issue 4, pp. 837–852.

PIKE, H. E. JR. / CITRON, S. J. (1970): *Optimization Studies of a Slab Reheating Furnace.* Automatica, Volume 6, Issue 1, pp. 41–50.

PROTHERO, A. / ROBINSON, A. (1974): *On the Stability and Accuracy of One-Step Methods for Solving Stiff Systems of Ordinary Differential Equations.* Mathematics of Computation, Volume 28, Issue 125, pp. 145–162.

RHEINBOLDT, WERNER C. (1984): *Differential-Algebraic Systems as Differential Equations on Manifolds*. Mathematics of Computation, Volume 43, Issue 168, pp. 473–482.

STEINBRECHER, ANDREAS (2006): *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems*. Ph. D thesis, Technische Universität Berlin.

WEIERSTRASS, K. (1858): *Über ein die homogenen Functionen betreffendes Theorem, nebst Anwendung desselben auf die Theorie der kleinen Schwingungen*. Monatsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin, pp. 207–220.

WEINTRAUB, STEVEN H. (2009): *Jordan Canonical Form: Theory and Practice*. Morgan & Claypool.

ZHANG, BIN / CHEN, ZHIGANG / XU, LIYUN / WANG, JINGCHENG / ZHANG, JIANMIN / SHAO, HUIHE (2002): *The Modeling and Control of A Reheating Furnace*. Proceedings of the American Control Conference, Anchorage, USA, pp. 3823–3828.