Solution of index-2 differential algebraic equations and its application in circuit simulation

Dissertation

zur Erlangung des akademischen Grades

im Fach Mathematik

eingereicht am

Fachbereich Mathematik
der Humboldt-Universität zu Berlin

von

Dipl.-Math. Caren Tischendorf
geb. am 09.06.1969 in Berlin

Präsidentin der Humboldt-Universität zu Berlin:	Prof. Dr. M. Dürkop
Dekan der Mathematisch- Naturwissenschaftlichen Fakultät II:	Prof. Dr. K. Gröger

Gutachter: 1. Prof. Dr. R. März

2. Prof. Dr. L. Petzold

3. Prof. Dr. P. Rentrop

Tag der mündlichen Prüfung: 11.04.1996

Preface

This paper results from my work in the graduate college for "Geometry and nonlinear analysis" at the Humboldt-University of Berlin.

At this point, I gratefully want to thank Prof. R. März for charging me with interesting tasks, for her continuous, committed support and for the inspiring, fruitful cooperation. I would like to express my gratitude to my colleagues at the Humboldt-University for providing a pleasant working environment and numerous helpful conversations.

Further thanks are due to Dr. U. Feldmann from the Central Research and Development of the SIEMENS AG in Munich as well as Dr. M. Günther from the Technical University of Darmstadt for their support and useful discussions in the field of circuit simulation.

Berlin, December 1995 Caren Tischendorf

Contents

Introduction

Subject of this thesis is the numerical analysis of differential algebraic equations (DAEs). DAEs are special implicit ordinary differential equations

$$f(\dot{x}(t), x(t), t) = 0,$$

where the partial Jacobian $f'_{\dot{x}}(\dot{x}, x, t)$ is singular on the domain of f. The dynamic behaviour of numerous problems in physics, in chemistry, and in technical applications can be modelled by differential equations. Additionally, the models often contain implicit nonlinear algebraic equations in order to take into account conservation laws, geometrical or kinematic constraints, Kirchhoff's laws, etc. Hence, DAEs arise in various fields, e.g. in the motion of mechanical systems, the electric circuit analysis, chemical reaction kinetics, control theory, semi-discretization of partial differential equations.

The interest in handling DAEs numerically has increased with the development of the computer technology. Much progress has been made in the analysis of DAEs (see e.g. [GM86], [BCP89], [HLR89], [HW91], [RR91]) during the last ten years.

In contrast to explicit ordinary differential equation, the integration of DAEs may cause essential difficulties. Constraints define a manifold on which solutions of the DAE have to lie. Initial values must be chosen in such a way that they satisfy the constraints. Furthermore, the numerical solution must not drift too far away from the manifold.

For a better understanding of the behaviour of DAEs, they can be characterized by the notion of index. Roughly speaking, the index is a measure for the deviation of a DAE from an (regular) ODE. DAEs of higher index (≥ 2) are ill-posed in the sense that small perturbations in the initial data may cause arbitrarily large changes in the solution data.

The main scope of this thesis is to study the numerical solution of DAEs arising from circuit simulation. The differential algebraic equations in this field are often of lower index (≤ 2).

The solution behaviour of index-1 DAEs is well-understood. There are various codes handling index-1 differential algebraic equations successfully. Common codes like DASSL by L.R. Petzold and LSODI by A.C. Hindmarsh use the BDF method for the integration of DAEs. Recently, also one-step methods such as Runge-Kutta and extrapolation methods have been used successfully. The well-known code RADAU5 by E. Hairer and G. Wanner uses the 3stage Runge-Kutta method Radau IIA. The new code CHORAL by M. Günther was developed for electric circuit simulation and is based on Rosenbrock-Wanner methods.

Regarding these facts we are interested in the numerical solution of index-2 DAEs arising from electric networks. We restrict ourselves to initial value problems. We have chosen the BDF method for our investigations because this method is especially suited for the integration of initial value problems of *stiff* differential equations. The thesis is organized as follows:

- Chapter 1 gives a short introduction to DAEs. We illustrate some properties by examples of circuit simulation. Further, we introduce the notion of the index and explain some index concepts.

- In Chapter 2 we look at modern techniques of electric circuit analysis. We consider the classical and the charge-oriented modified nodal analysis. The DAEs arising from these simulation techniques are the object of our interest in the next chapters.

- Chapter 3 deals with the analysis of perturbed initial value problems. For the numerical solution of DAEs, it is important to study the behaviour of a solution of a perturbed IVP in comparison to a solution of the original IVP.

- The BDF method applied to DAEs is investigated in Chapter 4. The feasibility, i.e., the solvability of the nonlinear equations in each step, is studied. Furthermore, we describe the stability behaviour of the BDF method. It explains the influence of the defects onto the numerical solution in more detail.

- Chapter 5 deals with circuit simulation. The structure of DAEs arising in this field is analyzed. We answer the question of the index of the classic and the charge-oriented modified analysis for some network classes. We present our results of the numerical simulation of two applications, of a NAND-gate and of a ring modulator, by means of our own code DAE2SOL (see [Tis92]) which is based on the BDF method.

- Finally, we state some facts from algebra and analysis in the Appendix. They are useful for the investigations in the Chapters 3 and 4.

Notations and conventions

$\mathbb{R}$	-	set of real numbers
$\mathbb{R}^m$	-	vector space of dimension m over $\mathbb{R}$
$x \in X$	-	x is an element of the set X
$X \subseteq Y$	-	X is a subset of Y
$X \cap Y$	-	intersection of the sets X and Y
$X \cup Y$	-	union of the sets X and Y
$f: X \to Y$	-	f is a mapping of the set X into the set Y
$\exists$	-	there exists
$\forall$	-	for all
$\operatorname{im} A$	-	image space of the operator A
$\ker A$	-	kernel of the operator A
$L(\mathbb{R}^m)$	-	set of linear functions mapping $\mathbb{R}^m$ into $\mathbb{R}^m$
$\dot{x}$	-	$\frac{dx}{dt}$, t - real variable (time), x - function of t
f'	-	$\frac{df}{dx}$, f - function of x
f is smooth	-	f is continuously differentiable
ODE	-	ordinary differential equation
DAE	-	differential algebraic equation
IVP	-	initial value problem
MNA	-	modified nodal analysis
$B(x, \alpha)$	-	sphere around x with the radius α
Q projects onto S	-	$Q^2 = Q$, $\operatorname{im} Q = S$
Q projects along S	-	$Q^2 = Q$, $\ker Q = S$

Chapter 1

Fundamentals of DAEs

Differential algebraic equations are implicit ordinary differential equations of the form

$$f(\dot{x}(t), x(t), t) = 0, \qquad f : \mathbb{R}^m \times \mathcal{D} \times \mathcal{I} \to \mathbb{R}^m, \tag{1.1}$$

where $x : \mathcal{I} \to \mathbb{R}^m$ denotes the unknown function, $\mathcal{I}$ is an open interval of $\mathbb{R}$, and $\mathcal{D}$ is an open subset of $\mathbb{R}^m$. The partial derivative $f'_{\dot{x}}(\dot{x}, x, t)$ is singular and has constant rank on its definition domain $D_f := \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}$.

The behaviour of DAEs differs from that of explicit ODEs in several aspects. We want to describe some of the essential differences here. For a better understanding, let us look at special systems.

Definition 1.1
A DAE is called **semi-explicit** if it is of the form

$$\dot{x}_1(t) + f(x_1(t), x_2(t), t) = 0, \tag{1.2}$$

$$g(x_1(t), x_2(t), t) = 0, \tag{1.3}$$

where $x(t) = (x_1(t), x_2(t)) \in \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ and $m = m_1 + m_2$.

It makes sense to speak of a solution (x_1, x_2) of the system (1.2)-(1.3) if the component x_1 is smooth and x_2 is continuous. In comparison with regular ODEs, it is not necessary to demand smoothness for all components.

Example A. Figure 1.1 shows a small circuit that may be modelled by the following simple semi-explicit system

$$\dot{Q}(t) + I_S(e^{\frac{u(t)}{U_T}} - 1) - I(t) = 0 \tag{1.4}$$
$$u(t) = \mathrm{v}(t) \tag{1.5}$$
$$Q(t) - C \cdot u(t) = 0, \tag{1.6}$$

where I denotes the current through the circuit, u the voltage at the capacitance C, and Q the charge of the capacitance C. The constant I_S denotes the blocked saturation current, the constant U_T represents the voltage in temperature of the diode.

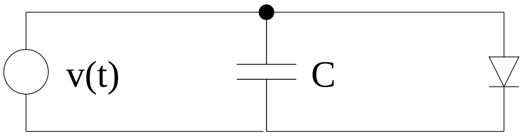

Figure 1.1: Circuit A

This system has a continuous solution

$$\begin{aligned} u(t) &= \mathrm{v}(t), \\ I(t) &= C \cdot \mathrm{v}'(t) + I_S(e^{\frac{\mathrm{v}(t)}{U_T}} - 1), \\ Q(t) &= C \cdot \mathrm{v}(t) \end{aligned}$$

if and only if v is continuously differentiable. For the smoothness of all components of the solution, the input function v has to be twice differentiable.

If we consider autonomous semi-explicit DAEs

$$\dot{x}_1 + f(x_1, x_2) = 0 \tag{1.7}$$
$$g(x_1, x_2) = 0, \tag{1.8}$$

it is obvious that a solution (x_1, x_2) of the system (1.7)-(1.8) has to belong to the submanifold of $\mathbb{R}^m$

$$\mathcal{M}_1 := \{ \left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right) \in \mathbb{R}^m : g(x_1, x_2) = 0 \}.$$

Therefore, initial values for IVPs of such a DAE have to lie in this submanifold. If g depends on its first component x_1 only, one differentiation of the constraint (1.8) implies the relation

$$g'(x_1) f(x_1, x_2) = 0$$

for a solution of the system (1.7)-(1.8), i.e., initial values for IVPs of such a DAE must be elements of the more restricted submanifold of $\mathbb{R}^m$

$$\mathcal{M}_2 := \{ \left(\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right) \in \mathbb{R}^m : \ g(x_1) = 0, \ g'(x_1) f(x_1, x_2) = 0 \}.$$

In the case of Example A above, the initial value (u_0, I_0, Q_0) is even uniquely determined by

$$\begin{aligned} u_0 &= \mathrm{v}(t_0), \\ I_0 &= C \cdot \mathrm{v}'(t_0) + I_S(e^{\frac{\mathrm{v}(t_0)}{U_T}} - 1), \\ Q_0 &= C \cdot \mathrm{v}(t_0). \end{aligned}$$

This example reveals a further aspect. The DAE (1.4)-(1.6) involves a differentiation problem. In the case of regular ODEs, we have to do with integration problems only. Since the differentiation represents an ill-posed problem, such DAEs provide difficulties in handling them numerically.

Example B. Consider a further simple circuit given in Figure 1.2, which may be modelled by the system

$$\dot{Q}(t) - I_S(e^{\frac{u_2(t)-u_1(t)}{U_T}} - 1) = 0 \tag{1.9}$$

$$I_S(e^{\frac{u_2(t)-u_1(t)}{U_T}} - 1) - I(t) = 0 \tag{1.10}$$

$$u_1(t) = \mathrm{v}(t) \tag{1.11}$$

$$Q(t) - C \cdot u_2(t) = 0. \tag{1.12}$$

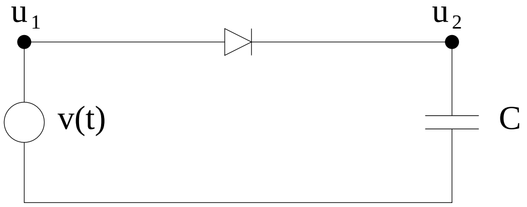

Figure 1.2: Circuit B

This system has a solution

$$u_1(t) = \mathrm{v}(t), \qquad I(t) = I_S(e^{\frac{u_2(t)-\mathrm{v}(t)}{U_T}} - 1), \qquad Q(t) = C u_2(t)$$

if and only if u_2 is a solution of the explicit ordinary differential equation

$$\dot{u}_2(t) = \frac{I_S}{C}(e^{\frac{u_2(t)-\mathrm{v}(t)}{U_T}} - 1).$$

In this case, we do not have to differentiate, we must integrate as in the case of regular ODEs.

For handling DAEs, it is useful to characterize DAEs in such a way that the criteria provide information about their behaviour. Such a characterization is given by the index of a DAE. Roughly speaking, the index of a DAE is a measure of the deviation of a DAE from regular ODEs. The literature provides a number of index concepts. We want to look at some of them in the next sections.

1.1 Linear DAEs with constant coefficients: Solution spaces and index definition

The results obtained in studying linear DAEs with constant coefficients

$$A\dot{x}(t) + Bx(t) = r(t) \tag{1.13}$$

form the basis of all index concepts for DAEs. The matrices A and B are elements of $L(\mathbb{R}^m)$, where A is singular. The solution of this equation system is closely related to the properties of the matrix pencil $\{A, B\}$. Looking at the homogeneous system

$$A\dot{x}(t) + Bx(t) = 0$$

with the initial condition $x(t_0) = 0$, we obtain an infinite-dimensional solution space if the matrix pencil is singular, i.e. if the polynomial

$$p(\lambda) := \det(\lambda A + B)$$

vanishes identically (see e.g. [GM86]). Therefore, it makes sense to consider only nonsingular matrix pencils $\{A, B\}$. Nonsingular matrix pencils $\{A, B\}$ may be transformed into the normal form of Weierstraß (see e.g. [Gan54]) by regular matrices $E, F \in L(\mathbb{R}^m)$

$$EAF = \mathrm{diag}(I, J), \;\; EBF = \mathrm{diag}(W, I),$$

where W lies in $L(\mathbb{R}^{m_1})$, $J \in L(\mathbb{R}^{m_2})$ is a nilpotent block matrix with Jordan blocks of the form

$$\begin{pmatrix} 0 & 1 & & \\ & 0 & \ddots & \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix},$$

and $m = m_1 + m_2$ is true. This transformation leads to a system of the form

$$u'(t) + Wu(t) = \bar{r}_1(t) \tag{1.14}$$

$$Jv'(t) + v(t) = \bar{r}_2(t), \tag{1.15}$$

which is equivalent to equation (1.13). Equation (1.14) represents a regular ODE. The solution of the second equation (1.15) can be put as follows:

$$v(t) = \sum_{k=0}^{\mu-1} (-1)^k (J^k \bar{r}_2(t))^{(k)} \tag{1.16}$$

if $\bar{r}_2(t)$ is differentiable sufficiently often and if μ denotes the nilpotency of the Jordan block matrix J. This μ is independent of the choice of the transformation and is called the *index of the matrix pencil* $\{A, B\}$. Naturally, the **index of the DAE** (1.13) is defined by this number μ.

The system (1.14), (1.15), and its solution (cf. (1.16)) make clear:

(i) DAEs do not only represent integration problems, but differentiation problems, too. Some parts of the right-hand side must be differentiable sufficiently often.

(ii) Some components of the solution are determined algebraically. This implies that the choice of initial values is not free for solutions of IVPs. The initial values must be “consistent” with the DAE.

1.2 The index for nonlinear DAEs

The various index concepts are based on the facts for linear DAEs given in the section above. Here, we want to touch some of them.

The geometrical index (see e.g. [Rhe84], [RR91], [Rei90], [Gri91]) provides useful insights into the geometrical and analytical nature of DAEs. For this approach, DAEs will be regarded as differential equations on manifolds.

The differential index (see e.g. [GP84], [GGL85], [BCP89], [Gea90], [CG93b]) is based on differentiations of the original DAE and often used in the literature.

The perturbation index (see e.g. [HLR89]) orientates on the behaviour of the solution of slightly perturbed DAEs in comparison with that of the solution of the original DAEs. It is a good measure for numerical difficulties.

The tractability index (see e.g. [GM86], [Mär90], [Mär92a]) provides a decomposition of the DAE into its inherent regular ODE as well as algebraic equations and differentiation problems. It is suitable for a detailed analysis of DAEs. Further, it is distinguished by minimal smoothness conditions to the function f. The circuit simulation provides functions with low smoothness properties, hence, we will use the tractability index for our investigations in the next chapters.

Some relations between the tractability index and the geometrical index are given in [Mär94]. In Chapter 3, we present relations between the tractability index and the other index concepts.

In this paper, we will restrict ourselves to the index 2 case. Results on index 1 will be given if this fits into the framework.

1.2.1 Geometrical index

The geometrical index of a DAE describes the behaviour of DAEs as the behaviour of regular ODEs on a constraint manifold. For ease of notation, we consider DAEs in autonomous form

$$f(\dot{x}(t), x(t)) = 0. \tag{1.17}$$

Let f be twice continuously differentiable. Further, let $P_\star(\dot{x}, x)$ be the orthogonal projection onto $\operatorname{im} f'_{\dot{x}}(\dot{x}, x)$ and $Q_\star(\dot{x}, x) := I - P_\star(\dot{x}, x)$. Differentiating equation (1.17) once, applying projections, and dropping the argument t, we obtain a new problem

$$f_1(\dot{x}, x) := P_\star(\dot{x}, x) f(\dot{x}, x) + Q_\star(\dot{x}, x) f'_x(\dot{x}, x)\dot{x} = 0. \tag{1.18}$$

Any C^2-solution of (1.17) solves equation (1.18).

Definition 1.2
The DAE (1.17) has the local **geometrical index 1** around the point (y_0, x_0) $\in f^{-1}(0) \cap f_1^{-1}(0)$ if and only if the partial derivative $(f_1)'_{\dot{x}}(y_0, x_0)$ is invertible.

In this case, problem (1.18) can be transformed, locally around (y_0, x_0), into an explicit ODE.

Consider Example B on page 9. Rewriting the system as an autonomous DAE and dropping the argument t we obtain

$$\begin{aligned} \dot{\tau} - 1 &= 0 \\ \dot{Q} - I_S(e^{\frac{u_2 - u_1}{U_T}} - 1) &= 0 \\ I_S(e^{\frac{u_2 - u_1}{U_T}} - 1) - I &= 0 \\ u_1 - \mathrm{v}(\tau) &= 0 \\ Q - C \cdot u_2 &= 0. \end{aligned}$$

The canonical projector $P_\star(\dot{x}, x)$ is given by

$$P_\star(\dot{x}, x) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

for $x := (\tau, u_1, u_2, I, Q)^T$. Problem (1.18) reads

$$f_1(\dot{x}, x) = \begin{pmatrix} \dot{\tau} - 1 \\ \dot{Q} - I_S(e^{\frac{u_2 - u_1}{U_T}} - 1) \\ \frac{I_S}{U_T} e^{\frac{u_2 - u_1}{U_T}} (\dot{u}_2 - \dot{u}_1) - \dot{I} \\ \dot{u}_1 - \dot{\mathrm{v}}\dot{\tau} \\ \dot{Q} - C\dot{u}_2 \end{pmatrix} = 0.$$

For the partial derivative $(f_1)'_{\dot{x}}(y_0, x_0)$ it holds that

$$(f_1)'_{\dot{x}}(y_0, x_0) = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & -\frac{I_S}{U_T} e^{\frac{u_{20} - u_{10}}{U_T}} & \frac{I_S}{U_T} e^{\frac{u_{20} - u_{10}}{U_T}} & -1 & 0 \\ -\dot{\mathrm{v}} & 1 & 0 & 0 & 0 \\ 0 & 0 & -C & 0 & 1 \end{pmatrix}.$$

Obviously, $(f_1)'_{\dot{x}}(y_0, x_0)$ is invertible for all (y_0, x_0), i.e., the system (1.9)-(1.12) has the geometrical index 1, even globally. In this case, the geometrical index concept requires $\mathrm{v}(t)$ to be continuously differentiable.

Let f be three-times continuously differentiable. Introducing the orthogonal projections $P_{1\star}(\dot{x},x)$ onto $\operatorname{im}(f_1)'_{\dot{x}}(\dot{x},x)$, $Q_{1\star}(\dot{x},x) := I - P_{1\star}(\dot{x},x)$ and differentiating (1.18), we obtain a further problem

$$f_2(\dot{x},x) := P_{1\star}(\dot{x},x)f_1(\dot{x},x) + Q_{1\star}(\dot{x},x)f'_{1x}(\dot{x},x)\dot{x} = 0. \tag{1.19}$$

Any C^1-solution of (1.17) solves the system (1.19).

Definition 1.3
The DAE (1.17) has the local **geometrical index 2** around the point (y_0, x_0) $\in f^{-1}(0) \cap f_1^{-1}(0) \cap f_2^{-1}(0)$ if and only if the partial derivative $(f_1)'_{\dot{x}}(y_0,x_0)$ is singular, the rank $(f_1)'_x(y,x)$ is locally constant near (y_0,x_0) and the partial derivative $(f_2)'_{\dot{x}}(y_0,x_0)$ is invertible.

For such DAEs, problem (1.19) can be transformed, locally around (y_0,x_0), into an explicit ODE.

Consider Example A on page 8. Rewriting the system as an autonomous DAE and dropping the argument t yields

$$\begin{aligned}\dot{\tau} - 1 &= 0\\ \dot{Q} + I_S(e^{\frac{u}{U_T}} - 1) - I &= 0\\ u - \mathrm{v}(\tau) &= 0\\ Q - C \cdot u &= 0.\end{aligned}$$

The canonical projector $P_\star(\dot{x},x)$ is given by

$$P_\star(\dot{x},x) = \begin{pmatrix}1 & 0 & 0 & 0\\ 0 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\end{pmatrix}$$

for $x := (\tau, u, I, Q)^T$. Problem (1.18) reads

$$f_1(\dot{x},x) = \begin{pmatrix}\dot{\tau} - 1\\ \dot{Q} + I_S(e^{\frac{u}{U_T}} - 1) - I\\ \dot{u} - \dot{\mathrm{v}}\dot{\tau}\\ \dot{Q} - C\dot{u}\end{pmatrix} = 0.$$

For the partial derivative $(f_1)'_{\dot{x}}(y_0, x_0)$, we have

$$(f_1)'_{\dot{x}}(y_0, x_0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ -\dot{\mathrm{v}} & 1 & 0 & 0 \\ 0 & -C & 0 & 1 \end{pmatrix}.$$

Obviously, $(f_1)'_{\dot{x}}(y_0, x_0)$ is singular and has rank 3 for all (y_0, x_0). For the projector $P_{1\star}(\dot{x}, x)$, we have

$$P_{1\star}(\dot{x}, x) = \frac{1}{2 + C^2 + C^2\dot{\mathrm{v}}(t)^2} \begin{pmatrix} 2{+}C^2 & C\dot{\mathrm{v}}(t) & -C^2\dot{\mathrm{v}}(t) & -C\dot{\mathrm{v}}(t) \\ C\dot{\mathrm{v}}(t) & 1{+}C^2{+}C^2\dot{\mathrm{v}}(t)^2 & C & 1 \\ -C^2\dot{\mathrm{v}}(t) & C & 2{+}C^2\dot{\mathrm{v}}(t)^2 & -C \\ -C\dot{\mathrm{v}}(t) & 1 & -C & 1{+}C^2{+}C^2\dot{\mathrm{v}}(t)^2 \end{pmatrix}.$$

Then it holds for the partial derivative $(f_2)'_{\dot{x}}(y_0, x_0)$ of problem (1.19) that

$$(f_2)'_{\dot{x}}(y_0, x_0) = \begin{pmatrix} 1 & -\frac{C^2\dot{\mathrm{v}}(\tau_0)}{\alpha_0} - \frac{C\dot{\mathrm{v}}(\tau_0)}{\alpha_0}\beta_0 & \frac{C\dot{\mathrm{v}}(\tau_0)}{\alpha_0} & 0 \\ 0 & \frac{C}{\alpha_0} + \frac{1}{\alpha_0}\beta_0 & -\frac{1}{\alpha_0} & 1 \\ -\dot{\mathrm{v}}(\tau_0) & 1 - \frac{C^2}{\alpha_0} - \frac{C}{\alpha_0}\beta_0 & -\frac{C}{\alpha_0} & 0 \\ 0 & -C - \frac{C}{\alpha_0} - \frac{1}{\alpha_0}\beta_0 & \frac{1}{\alpha_0} & 1 \end{pmatrix}$$

for $\alpha_0 := 2 + C^2 + C^2\dot{\mathrm{v}}(\tau_0)^2$ and $\beta_0 := \frac{I_S}{U_T}e^{\frac{u_0}{U_T}}$. Now, one can compute that this matrix is regular for all τ_0 and u_0, i.e., the system (1.4)-(1.6) has the geometrical index 2. Note that the geometrical index concept requires $\mathrm{v}(t)$ to be twice continuously differentiable.

1.2.2 Differential index

Roughly speaking, the differential index of a DAE is the number of differentiations that are necessary for the transformation of the DAE into an explicit ODE.

Definition 1.4
Let f be once continuously differentiable. The DAE 1.1 has the **differential index 1** if and only if the equation system of the variables t, x, x', x''

$$\begin{aligned} f(x', x, t) & = 0 \\ \frac{d}{dt}f(x', x, t) = \frac{\partial}{\partial x'}f(x', x, t)x'' + \frac{\partial}{\partial x}f(x', x, t)x' + \frac{\partial}{\partial t}f(x', x, t) & = 0 \end{aligned}$$

uniquely determines the variable x' as a continuous function of (x, t).

Looking again at Example B, differentiating the system (1.9)-(1.12) and dropping the argument t, we obtain the system

$$\begin{aligned}
\ddot{Q} - \frac{I_S}{U_T} e^{\frac{u_2-u_1}{U_T}} (\dot{u}_2 - \dot{u}_1) &= 0 \\
\frac{I_S}{U_T} e^{\frac{u_2-u_1}{U_T}} (\dot{u}_2 - \dot{u}_1) - \dot{I} &= 0 \\
\dot{u}_1 &= \dot{\mathrm{v}} \\
\dot{Q} - C\dot{u}_2 &= 0,
\end{aligned}$$

which provides, together with the system (1.9)-(1.12),

$$\begin{aligned}
\dot{u}_1 &= \dot{\mathrm{v}} \\
\dot{u}_2 &= \frac{I_S}{C}(e^{\frac{u_2-u_1}{U_T}} - 1) \\
\dot{I} &= \frac{I_S}{U_T} e^{\frac{u_2-u_1}{U_T}} (\frac{I_S}{C}(e^{\frac{u_2-u_1}{U_T}} - 1) - \dot{\mathrm{v}}) \\
\dot{Q} &= I_S(e^{\frac{u_2-u_1}{U_T}} - 1),
\end{aligned}$$

i.e., the system (1.9)-(1.12) has the differential index 1.

Definition 1.5
Let f be twice continuously differentiable. The DAE 1.1 has the **differential index 2** if and only if it is not of index 1 and the equation system of the variables t, x, x', x'', x'''

$$\begin{aligned}
&f(x', x, t) &&= 0 \\
&\frac{d}{dt} f(x', x, t) = \frac{\partial}{\partial x'} f(x', x, t)x'' + \frac{\partial}{\partial x} f(x', x, t)x' + \frac{\partial}{\partial t} f(x', x, t) &&= 0 \\
&\frac{d^2}{dt^2} f(x', x, t) = \frac{\partial}{\partial x'} f(x', x, t)x''' + ... &&= 0
\end{aligned}$$

uniquely determines the variable x' as a continuous function of (x, t).

For Example A, differentiating the system (1.4)-(1.6) and dropping the argument t, we obtain the system

$$\ddot{Q} + \frac{I_S}{U_T} e^{\frac{u}{U_T}} \dot{u} - \dot{I} = 0 \tag{1.20}$$

$$\dot{u} = \dot{\mathrm{v}} \tag{1.21}$$

$$\dot{Q} - C\dot{u} = 0. \tag{1.22}$$

Trivially, there is no continuous function h such that all solutions $(\dot{I},u,I,Q,t)$ of (1.4)-(1.6) and (1.20)-(1.22), can be described by

$$\dot{I} = h(u, I, Q, t).$$

Consequently, the system (1.4)-(1.6) does not have the differential index 1. Differentiating (1.4)-(1.6) twice with respect to t we obtain

$$\begin{aligned}\dddot{Q} + \frac{I_S}{U_T} e^{\frac{u}{U_T}} \ddot{u} + \frac{I_S}{U_T^2} e^{\frac{u}{U_T}} \dot{u}^2 - \ddot{I} &= 0 \\ \ddot{u} &= \ddot{\mathrm{v}} \\ \ddot{Q} - C\ddot{u} &= 0.\end{aligned}$$

This system, together with the original system (1.4)-(1.6) and the system (1.20)-(1.22), implies

$$\begin{aligned}\dot{u} &= \dot{\mathrm{v}} \\ \dot{I} &= C\ddot{\mathrm{v}} + \frac{I_S}{U_T} e^{\frac{u}{U_T}} \dot{\mathrm{v}} \\ \dot{Q} &= I - I_S(e^{\frac{u}{U_T}} - 1),\end{aligned}$$

i.e., the system (1.4)-(1.6) has the differential index 2.

1.2.3 Perturbation index

The perturbation index interprets the index as a measure of sensitivity of the solutions with respect to perturbations of the given problem.

Definition 1.6
The DAE 1.1 has the **perturbation index 1** along a solution $x_*(t)$ on $\mathcal{I}_0 := [t_0, T]$ if, for all functions $x(t)$ having a defect

$$f(\dot{x}(t), x(t), t) = q(t),$$

there exists an estimate

$$\|x - x_*\| \leq \text{const} \left(\|x(t_0) - x_*(t_0)\| + \max_{t_0 \leq t \leq T} \|q(t)\| \right)$$

whenever the expression on the right-hand side is sufficiently small.

Studying Example B with a perturbed right-hand side $q = (q_1, q_2, q_3, q_4)^T$ and dropping the argument t, we obtain

$$u_1 = \mathrm{v} + q_3, \qquad I = I_S(e^{\frac{u_2 - \mathrm{v} - q_3}{U_T}} - 1) - q_2, \qquad Q = Cu_2 + q_4$$

as the solution, where u_2 is a solution of the explicit ODE

$$\dot{u}_2 = \frac{I_S}{C}(e^{\frac{u_2 - \mathrm{v} - q_3}{U_T}} - 1) + \frac{1}{C}(q_1 - \dot{q}_4)$$

and u_{2*} is a solution of the explicit ODE

$$\dot{u}_{2*} = \frac{I_S}{C}(e^{\frac{u_{2*} - \mathrm{v}}{U_T}} - 1).$$

Hence, the following error estimation holds:

$$\begin{aligned} \left| \begin{pmatrix} u_1 - u_{1*} \\ u_2 - u_{2*} \\ I - I_* \\ Q - Q_* \end{pmatrix} \right| &\leq |u_1 - u_{1*}| + |u_2 - u_{2*}| + |I - I_*| + |Q - Q_*| \\ = |q_3| + |u_2 - u_{2*}| &+ |I_S(e^{\frac{u_2 - \mathrm{v} - q_3}{U_T}} - e^{\frac{u_{2*} - \mathrm{v}}{U_T}}) - q_2| + |C(u_2 - u_{2*}) + q_4| \\ &\leq \mathrm{const}(|u_2(t_0) - u_{2*}(t_0)| + \|q\|) \qquad (1.23) \end{aligned}$$

for sufficiently small defects in the initial value and sufficiently small perturbations q, i.e., the system (1.9)-(1.12) has the perturbation index 1.

Definition 1.7
The DAE 1.1 has the **perturbation index 2** along a solution $x_*(t)$ on $\mathcal{I}_0 := [t_0, T]$ if it is not of index 1 and, for all functions $x(t)$ having a defect

$$f(\dot{x}(t), x(t), t) = q(t),$$

there exists an estimate

$$\|x - x_*\| \leq \mathrm{const}\left(\|x(t_0) - x_*(t_0)\| + \max_{t_0 \leq t \leq T} \|q(t)\| + \max_{t_0 \leq t \leq T} \|\dot{q}(t)\|\right)$$

whenever the expression on the right-hand side is sufficiently small.

Studying Example A with a perturbed right-hand side $q = (q_1, q_2, q_3)^T$ and dropping the argument t, we obtain the solution

$$\begin{aligned} u &= \mathrm{v} + q_2 \\ I &= C(\dot{\mathrm{v}} + \dot{q}_2) + \dot{q}_3 + I_S(e^{\frac{\mathrm{v} + q_2}{U_T}} - 1) - q_1 \\ Q &= C(\mathrm{v} + q_2) + q_3. \end{aligned}$$

Hence, the following error estimation holds:

$$\begin{aligned}\left|\begin{pmatrix} u-u_* \\ I-I_* \\ Q-Q_* \end{pmatrix}\right| &\leq |u-u_*| + |I-I_*| + |Q-Q_*| \\ &= |q_2| + |\dot{q}_3 + C\dot{q}_2 + I_S(e^{\frac{\mathrm{v}+q_2}{U_T}} - e^{\frac{\mathrm{v}}{U_T}}) - q_1| + |q_3 + Cq_2| \\ &\leq \text{const}(\|q\| + \|\dot{q}\|) \quad (1.24)\end{aligned}$$

for sufficiently small perturbations q. Obviously, we do not find a constant K such that

$$|I - I_*| \leq K(|x(t_0) - x_*(t_0)| + \|q\|)$$

is true. This implies that the system (1.4)-(1.6) has the perturbation index 2.

We want to remark that the perturbation index concept requires information about the solution of the DAE.

1.2.4 Tractability index

The tractability index is suitable for a detailed analysis of DAEs. Further, it distinguishes itself by minimal smoothness conditions to the function f. The circuit simulation provides functions with low smoothness properties. Therefore, we will use the tractability index for the investigations in the next chapters. The function f is assumed to have a constant nullspace of $f'_{\dot{x}}(\dot{x}, x, t)$. It should be mentioned that this is frequently the case in applications, indeed. In particular, this assumption is satisfied for problems arising from charge oriented MNA (see Chapter 5).

Now, there is a projector Q onto the nullspace $\ker f'_{\dot{x}}(\dot{x}, x, t)$. Denoting $P := I - Q$, the relation

$$\begin{aligned}&f(y, x, t) - f(Py, x, t) = \\ &\qquad\qquad \int_0^1 f'_{\dot{x}}(sy + (1-s)Py, x, t)Q\, ds \;=\; 0, \quad (y, x, t) \in D_f,\end{aligned}$$

holds. Hence, equation (1.1) may be written as

$$f(P\dot{x}(t), x(t), t) = 0, \qquad (1.25)$$

and the natural solution space is given by

$$C^1_N(\mathcal{I}, \mathbb{R}^m) := \{x \in C(\mathcal{I}, \mathbb{R}^m) :\; Px \in C^1(\mathcal{I}, \mathbb{R}^m)\}. \qquad (1.26)$$

Let $f : D_f \to \mathbb{R}^m$ be a continuous function and $f_{\dot{x}}$ as well as f_x be continuous. Further, we introduce some useful matrix functions

$$\begin{aligned} A(y,x,t) &:= f'_{\dot{x}}(y,x,t), && (1.27)\\ B(y,x,t) &:= f'_x(y,x,t), && (1.28)\\ G_1(y,x,t) &:= A(y,x,t) + B(y,x,t)Q, && (1.29)\\ B_1(y,x,t) &:= B(y,x,t)P, && (1.30) \end{aligned}$$

and useful spaces

$$\begin{aligned} N &:= \ker A(y,x,t), && (1.31)\\ S(y,x,t) &:= \{z \in \mathbb{R}^m \,:\, B(y,x,t)z \in \operatorname{im} A(y,x,t)\}, && (1.32)\\ N_1(y,x,t) &:= \ker G_1(y,x,t), && (1.33)\\ S_1(y,x,t) &:= \{z \in \mathbb{R}^m \,:\, B_1(y,x,t)z \in \operatorname{im} G_1(y,x,t)\}. && (1.34) \end{aligned}$$

Definition 1.8
The DAE (1.1) is said to be **index-1 tractable** on open $\mathcal{G} \subseteq D_f$ if the relation

$$N \cap S(y,x,t) = \{0\}$$

is true for all $(y,x,t) \in \mathcal{G}$.

Remark 1.9 Regarding Lemma A.1, it is easy to see that $G_1(y,x,t)$ is nonsingular and the relation

$$N \oplus S(y,x,t) = \mathbb{R}^m$$

is satisfied if the DAE (1.1) is index-1 tractable. Note that the regularity of G_1 is independent of the choice of the projector Q.

For Example B we obtain

$$A = \begin{pmatrix} 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} \frac{I_S}{U_T}e^{\frac{u_2-u_1}{U_T}} & -\frac{I_S}{U_T}e^{\frac{u_2-u_1}{U_T}} & 0 & 0\\ -\frac{I_S}{U_T}e^{\frac{u_2-u_1}{U_T}} & \frac{I_S}{U_T}e^{\frac{u_2-u_1}{U_T}} & -1 & 0\\ 1 & 0 & 0 & 0\\ 0 & -C & 0 & 1 \end{pmatrix}$$

for $x := (u_1, u_2, I, Q)^T$. Thus, the relations

$$\begin{aligned} N &= \{(z_1,z_2,z_3,z_4)^T \,:\, z_4 = 0\},\\ S(x) &= \{(z_1,z_2,z_3,z_4)^T \,:\, z_1 = 0,\ z_4 = Cz_2,\ z_3 = \frac{I_S}{U_T}e^{\frac{u_2-u_1}{U_T}} z_2\} \end{aligned}$$

are true, and $N \cap S(x) = \{0\}$ is valid, i.e., the system (1.9)-(1.10) is index-1 tractable.

Definition 1.10
The DAE (1.1) is **index-2 tractable** on $\mathcal{G} \subseteq D_f$ if

- the matrix $G_1(y,x,t)$ is singular on $\mathcal{G}$,
- $\operatorname{rank}(G_1(y,x,t))$ is constant on $\mathcal{G}$ and
- the relation $N_1(y,x,t) \cap S_1(y,x,t) = \{0\}$ is satisfied on $\mathcal{G}$.

Remark 1.11 Applying Lemma A.1 to G_1, B_1, and to a projector Q_1 onto N_1, we conclude the regularity of the matrix

$$G_2(y,x,t) := G_1(y,x,t) + B(y,x,t)PQ_1(y,x,t)$$

and the validity of the relation

$$N_1(y,x,t) \oplus S_1(y,x,t) = \mathbb{R}^m$$

if the DAE (1.1) is index-2 tractable.

In the case of Example A, we now have

$$A = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad B(x) = \begin{pmatrix} \frac{I_S}{U_T} e^{\frac{u}{U_T}} & -1 & 0 \\ 1 & 0 & 0 \\ -C & 0 & 1 \end{pmatrix}$$

for $x := (u, I, Q)^T$. Thus, the relations

$$N = \{(z_1,z_2,z_3)^T : z_3 = 0\}, \quad S = \{(z_1,z_2,z_3)^T : z_1 = z_3 = 0\}$$

are true, and $N \cap S = \{(z_1,z_2,z_3)^T : z_1 = z_3 = 0\} \neq \{0\}$ is valid. Choosing the projector $Q := \left(\begin{smallmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{smallmatrix}\right)$, we obtain

$$G_1(x) = \begin{pmatrix} \frac{I_S}{U_T} e^{\frac{u}{U_T}} & -1 & 1 \\ 1 & 0 & 0 \\ -C & 0 & 0 \end{pmatrix},$$

and the relations

$$N_1 = \{(z_1,z_2,z_3)^T : z_1 = 0, z_2 = z_3\}, \quad S_1 = \{(z_1,z_2,z_3)^T : z_3 = 0\}$$

are satisfied. Hence, the relation $\ker G_1 \cap S_1 = \{0\}$ holds, i.e., the system (1.4)-(1.6) is index-2 tractable.

Chapter 2

Introduction to circuit simulation

The circuits we want to study here are assumed to be modelled by an RLC-network, which can be divided into a dynamic network and a non-dynamic one, that are connected by a b-port. The non-dynamic network consists of linear resistors, nonlinear resistors, independent sources, and controlled sources. The dynamic network contains linear and nonlinear capacitances and inductances. We speak of a nonlinear capacitance if there is a nonlinear differentiable mapping $q_C = \psi(u_C)$ between charge and voltage of the capacitance. Accordingly, we speak of a nonlinear inductance if there is a nonlinear differentiable mapping $\Phi_L = \varphi(I_L)$ between flux and current of the inductance. Such networks may be modelled by differential algebraic equations (cf. [Mat87]).

Many modern circuits consist of a large number of elements. If we want to simulate such networks, the equations have to be generated automatically. We want to study two modern modelling techniques making such an automatic generation possible, namely, the classical approach and the charge-oriented approach of the modified nodal analysis (cf. [BG86], [DR91], [FWZ$^+$92], [FG94]).

Firstly, we will illustrate both of them by means of a little example. We consider a double way rectifier with LC filter. The alternative currents through V_1 and V_2 are being rectified to a directed current through R_1. The diodes D_1 and D_2 rectify the current. The inductances L_1, L_2 and the capacitances C_1, C_2, C_3 filter the remaining oscillations of the current.

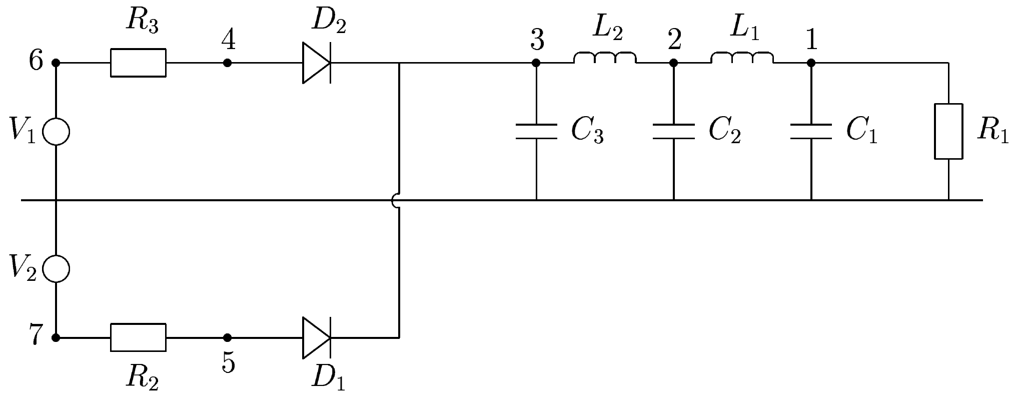

Figure 2.1: Double way rectifier with LC filter

The network constants are given by the following equations:

$$R_1 = 500\,\Omega, \quad R_2 = R_3 = 100\,\Omega,$$
$$C_1 = C_2 = C_3 = 50 \cdot 10^{-6}\,F, \quad L_1 = L_2 = 10\,H.$$

The model functions for the diodes are of the form

$$I_D = G(V_D) := 10^{-9}(e^{39.0\,V_D} - 1).$$

The voltage sources supply the circuit with the voltage

$$V_1(t) = V_2(t) = 30\,\sin(2\pi \cdot 50 \cdot t).$$

2.1 Classical modified nodal analysis

The vector of unknowns x consists of

- the nodal potentials u and
- the currents I of the voltage-controlled elements.

The system contains the equations for each node (except for the node with the zero potential), which are derived by Kirchhoff's nodal law. Additionally, the characteristic equations of the voltage-controlled elements (inductances and sources) belong to the system. The equations of the current-controlled elements are set into the system directly.

In the case of the double way rectifier, we obtain the following system.

$$\begin{aligned}
C_1\dot{u}_1 + \frac{u_1}{R_1} + I_1 &= 0\\
C_2\dot{u}_2 + I_2 - I_1 &= 0\\
C_3\dot{u}_3 + G(u_3 - u_4) + G(u_3 - u_5) - I_2 &= 0\\
\frac{u_4 - u_6}{R_2} - G(u_3 - u_4) &= 0\\
\frac{u_5 - u_7}{R_3} - G(u_3 - u_5) &= 0\\
-\frac{u_4 - u_6}{R_2} + I_3 &= 0\\
\frac{u_5 - u_7}{R_3} - I_4 &= 0\\
L_1\dot{I}_1 - (u_1 - u_2) &= 0\\
L_2\dot{I}_2 - (u_2 - u_3) &= 0\\
u_6 &= V_1(t)\\
u_7 &= V_2(t).
\end{aligned}$$

The variables I_3 and I_4 describe the currents through the voltage sources V_1 and V_2. In general, we obtain a differential algebraic equation system of the form

$$D(x)\dot{x} + f(x) = r(t). \tag{2.1}$$

Let the components of the unknown vector x be ordered in such a way that $x = \binom{u}{I}$ is satisfied – u represents the vector of nodal potentials, and I represents the vector of currents. Further, let the i-th equation of (2.1) represent Kirchhoff's nodal law for the node with the nodal potential u_i, and let the $(n_u + i)$-th equation of (2.1) represent the characteristic equation for the inductance L_i.

2.2 Charge-oriented modified nodal analysis

Here, the vector of unknowns contains

- the nodal potentials u,
- the currents I of the voltage-controlled elements,

- the charge Q of the capacitances and
- the flux Φ of the inductances.

The system contains the equations for each node that are derived by Kirchhoff's nodal law. Additionally, the characteristic equations of the voltage-controlled elements belong to the system. The equations of the current-controlled elements are set into the system directly. Finally, the characteristic equations for charge and flux belong to the system.

This method provides the system

$$
\begin{aligned}
-\dot{Q}_1 + \frac{u_1}{R_1} + I_1 &= 0 \\
-\dot{Q}_2 + I_2 - I_1 &= 0 \\
-\dot{Q}_3 + G(u_3 - u_4) + G(u_3 - u_5) - I_2 &= 0 \\
\frac{u_4 - u_6}{R_2} - G(u_3 - u_4) &= 0 \\
\frac{u_5 - u_7}{R_3} - G(u_3 - u_5) &= 0 \\
-\frac{u_4 - u_6}{R_2} + I_3 &= 0 \\
\frac{u_5 - u_7}{R_3} - I_4 &= 0 \\
\dot{\Phi}_1 - (u_1 - u_2) &= 0 \\
\dot{\Phi}_2 - (u_2 - u_3) &= 0 \\
u_6 &= V_1(t) \\
u_7 &= V_2(t) \\
Q_1 &= -C_1 u_1 \\
Q_2 &= -C_2 u_2 \\
Q_3 &= -C_3 u_3 \\
\Phi_1 &= L_1 \cdot I_1 \\
\Phi_2 &= L_2 \cdot I_2
\end{aligned}
$$

for the double way rectifier. In general, we obtain a differential algebraic equation system of the form

$$A\dot{q} + f(x) = r(t) \tag{2.2}$$

$$q = g(x). \tag{2.3}$$

At this point, let the components of the unknown vector x again be ordered in such a way that $x = \binom{u}{I}$ is satisfied – u represents the vector of nodal potentials, and I the vector of currents. Additionally, let the i-th equation of (2.2) again represent Kirchhoff's nodal law for the node with the nodal potential u_i and let the $(n_u + i)$-th equation of (2.2) represent the characteristic equation for the inductance L_i. Let the components of the vector q be ordered in such a way that $q = \binom{Q}{\Phi}$ is satisfied – Q represents the vector of charges, and Φ the vector of fluxes. Then, the function $g(x)$ is of the form

$$g(x) = \begin{pmatrix} g_1(u) \\ g_2(I) \end{pmatrix}.$$

Remarks 2.1

(1) The coefficient matrix $D(x)$ of equation (2.1) satisfies the relation

$$D(x) = Ag'(x) = A\frac{dg(x)}{dx}. \tag{2.4}$$

(2) If the network is modelled without capacitances, equation (2.3) reads

$$g(x) = g_2(I).$$

Correspondingly, if the network is modelled without inductances, equation (2.3) reads

$$g(x) = g_1(u).$$

(3) If the network contains neither a capacitance nor an inductance, i.e., if the circuit does not have dynamical elements, then equation (2.3) disappears completely. Both modelling techniques lead to the same system

$$f(x) = r(t)$$

in this case. Hence, we may exclude this case when studying the differences between both modelling techniques in the following.

Chapter 3

Solvability of perturbed IVPs of DAEs

We are interested in the numerical solution of initial value problems of differential algebraic equations. Therefore, we study the existence of solutions of perturbed IVPs of DAEs assuming that the original DAE has a solution x_*. Further, we investigate the stability behaviour of the numerical solution, i.e., we present estimations for the deviation of the solution of a perturbed IVP from the solution of the original IVP. For that, we consider compact intervals $\mathcal{I}_0 \subseteq \mathcal{I}$. Let $\mathcal{I}_1$ be the semi-open interval $[a,b)$ if $\mathcal{I}_0$ denotes the closed interval $[a,b]$. Further, let $\mathcal{G}$ be an open neighbourhood of the trajectory $(\dot{x}_*(t), x_*(t), t)$ with $t \in \mathcal{I}_0$. Let us speak of a numerical solution if it is a solution of a perturbed IVP. We speak of an exact solution if it is a solution of the original IVP.

The first two sections of this chapter deal with index-1 problems. We hope that these sections will import the idea of the tractability concept. Furthermore, the analysis of general index-2 problems in later sections will be better understandable if the reader is familiar with the analysis of index-1 DAEs. The results presented for index-1 problems are well-known (see e.g. [GM86], [BCP89], [Mär94], [Tis94]).

3.1 Linear index-1 DAEs

Consider the linear index-1 tractable DAE

$$A(t)\dot{x}(t) + B(t)x(t) = r(t) \tag{3.1}$$

with constant nullspace $\ker A(t)$ (for an explanation of index-1 tractability see Subsection 1.2.4). Let A, B, r be continuous. Then, the matrix

$$G_1(t) = A(t) + B(t)Q$$

is regular (see Remark 1.9 and recall that Q is a projector onto $\ker A(t)$).

For ease of notation, let us drop the argument t in the following. We can transform equation (3.1) equivalently into the form

$$G_1^{-1}A\dot{x} + G_1^{-1}Bx = G_1^{-1}r. \tag{3.2}$$

The relations

$$G_1^{-1}A = P \quad \text{and} \quad G_1^{-1}BQ = Q$$

imply

$$P\dot{x} + G_1^{-1}BPx + Qx = G_1^{-1}r. \tag{3.3}$$

Multiplying (3.3) by P and Q, respectively, we obtain the equivalent system

$$P\dot{x} + PG_1^{-1}BPx = PG_1^{-1}r \tag{3.4}$$

$$QG_1^{-1}BPx + Qx = QG_1^{-1}r. \tag{3.5}$$

The first equation (3.4) represents an explicit ODE for the component Px. The second equation (3.5) allows the algebraical determination of the component Qx depending on Px. Therefore, all solutions $x_*(t)$ of the DAE (3.1) are given by

$$x_* := (I - QG_1^{-1}B)u_* + QG_1^{-1}r \tag{3.6}$$

if u_* is a solution of the IVP

$$\dot{u} + PG_1^{-1}Bu = PG_1^{-1}r \tag{3.7}$$

$$u(t_0) = u_0 \in \operatorname{im} P. \tag{3.8}$$

Hence, one can choose any value in the space $\operatorname{im} P$ as the initial value $u(t_0) = Px(t_0)$. However, the other component $Qx(t_0)$ has to satisfy

$$Qx(t_0) = QG_1^{-1}(t_0)[r(t_0) - B(t_0)u(t_0)].$$

Hence, an IVP of index 1 is appropriately defined if the initial conditions are related to the P-component only. Now, we are in a position to formulate the following theorem.

Theorem 3.1 *Let $x_* \in C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ be a solution of the index-1 tractable DAE (3.1). Then, the perturbed initial value problem*

$$A(t)\dot{x}(t) + B(t)x(t) = r(t) + q(t) \tag{3.9}$$

$$Px(t_0) = u_0 \in \operatorname{im} P \tag{3.10}$$

is uniquely solvable on $C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ for continuous perturbations q. Furthermore, the estimation

$$\|x - x_*\|_\infty + \|P\dot{x} - P\dot{x}_*\|_\infty \le K(\|q\|_\infty + |u_0 - Px_*(t_0)|)$$

is true for a constant K.

The theorem ensures small deviations of the numerical solution from the exact solution if the perturbations q and the deviation in the initial value are sufficiently small.

Proof: Following the way taken before introducing the theorem for a perturbed right-hand side $r + q$ instead of r, the solution of (3.9)-(3.10) is given by

$$x := (I - QG_1^{-1}B)u + QG_1^{-1}(r + q)$$

if u solves the IVP

$$\dot{u} + PG_1^{-1}Bu = PG_1^{-1}(r + q)$$
$$u(t_0) = u_0 \in \operatorname{im} P.$$

The equations (3.6) and (3.7) imply

$$x - x_* = (I - QG_1^{-1}B)(u - u_*) + QG_1^{-1}q \tag{3.11}$$

and

$$\dot{u} - \dot{u}_* = -PG_1^{-1}B(u - u_*) + PG_1^{-1}q.$$

Then, we find a constant K_1 such that

$$|\dot{u}(t) - \dot{u}_*(t)| \le K_1(|u(t) - u_*(t)| + \|q\|_\infty) \qquad \forall\, t \in \mathcal{I}_0$$

is true. Using the Dini derivative

$$D_+m(t) = \liminf_{h>0} \frac{m(t+h) - m(t)}{h}$$

for $m(t) := |u(t) - u_*(t)|$ and following the proof on page 146 in [Tis94], we find a constant K_2 such that

$$\|u - u_*\|_\infty + \|\dot{u} - \dot{u}_*\|_\infty \le K_2(|u(t_0) - u_*(t_0)| + \|q\|_\infty)$$

is satisfied. Regarding equation (3.11), the assertion follows immediately.

□

3.2 Nonlinear index-1 DAEs

We investigate general nonlinear DAEs

$$f(\dot{x}(t), x(t), t) = 0 \tag{3.12}$$

that are index-1 tractable on $\mathcal{G}$.

The existence and the behaviour of numerical solutions for index-1 problems is well-known (see e.g. [BCP89], [GM86]). We formulate the results in the following theorem.

Theorem 3.2 *Let $x_* \in C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ be a solution of the DAE (3.12) which is index-1 tractable on $\mathcal{G}$. If the perturbation q is continuous, then the perturbed initial value problem*

$$f(\dot{x}(t), x(t), t) = q(t) \tag{3.13}$$

$$Px(s) = u_s \in im P, \quad |u_s - Px_*(s)| \leq \tau, \quad \|q\|_\infty \leq \sigma, \tag{3.14}$$

is uniquely solvable on $C^1_N(\mathcal{J}, \mathbb{R}^m)$ for each $s \in \mathcal{I}_1$ and sufficiently small τ and σ, where $\mathcal{J} := [s, S) \subset \mathcal{I}_0$. Further, the inequality

$$\|x - x_*\|_\infty + \|P\dot{x} - P\dot{x}_*\|_\infty \leq K\left(\|q\|_\infty + |u_s - Px_*(s)|\right)$$

holds.

We want to present a proof of this theorem in order to provide some ideas that will be helpful for our investigations in the index-2 case later.

Proof: Firstly, we transform the equation

$$f(\dot{x}(t), x(t), t) = q(t)$$

into the equivalent system

$$A(t)\dot{x}(t) + B(t)x(t) + h(\dot{x}(t), x(t), t) - r_*(t) = q(t) \tag{3.15}$$

where

$$\begin{aligned}
A(t) &:= f'_{\dot{x}}(\dot{x}_*(t), x_*(t), t), \\
B(t) &:= f'_x(\dot{x}_*(t), x_*(t), t), \\
h(y, x, t) &:= f(y, x, t) - A(t)(y - \dot{x}_*(t)) - B(t)(x - x_*(t)), \\
r_*(t) &:= A(t)\dot{x}_*(t) + B(t)x_*(t).
\end{aligned}$$

Then, the function h has the following properties:

$$h(\dot{x}_*(t), x_*(t), t) = 0, \qquad h'_y(\dot{x}_*(t), x_*(t), t) = 0, \qquad h'_x(\dot{x}_*(t), x_*(t), t) = 0.$$

In other words, the functions h, h'_y and h'_x map small neighbourhoods of the trajectory $\{(\dot{x}_*(t), x_*(t), t),\ t \in \mathcal{I}_0\}$ into small spheres around zero.

Since the DAE (3.12) is index-1 tractable, the matrix

$$G_1(t) = A(t) + B(t)Q$$

is regular (see once more Remark 1.9 and recall that Q is a projector onto $\ker A(t)$). Regarding

$$h(y,x,t) - h(Py,x,t) = \int_0^1 f'_{\dot{x}}(\phi y + (1-\phi)Py, x, t)Q\, d\phi \ = \ 0, \quad (y,x,t) \in D_f$$

and dropping the argument t, the system (3.15) is equivalent to

$$G_1^{-1}A\dot{x} + G_1^{-1}Bx + G_1^{-1}h(P\dot{x}, x, \cdot) - G_1^{-1}r_* = G_1^{-1}q.$$

Using the relations $G_1^{-1}A = P$ and $G_1^{-1}BQ = Q$, we obtain

$$P\dot{x} + G_1^{-1}BPx + Qx + G_1^{-1}h(P\dot{x}, x, \cdot) - G_1^{-1}r_* = G_1^{-1}q. \tag{3.16}$$

Multiplying (3.16) by P and Q, respectively, and introducing

$$u := Px, \quad v := Qx, \qquad u_* := Px_*, \quad v_* := Qx_*,$$

we obtain

$$\dot{u} + PG_1^{-1}Bu + PG_1^{-1}h(\dot{u}, u+v, \cdot) - PG_1^{-1}r_* = PG_1^{-1}q \tag{3.17}$$

$$QG_1^{-1}Bu + v + QG_1^{-1}h(\dot{u}, u+v, \cdot) - QG_1^{-1}r_* = QG_1^{-1}q. \tag{3.18}$$

Now, we may solve the algebraic part (3.18) of the DAE with respect to v in a neighbourhood of the solution x_*. We introduce a function $\bar{F} : U_\alpha(x_*) \to \mathbb{R}^m$ defined by

$$\begin{aligned}\bar{F}(v, u', u, q, t) := QG_1^{-1}(t)B(t)u + v - QG_1^{-1}r_* - QG_1^{-1}(t)q \\ + QG_1^{-1}(t)h(u', u+v, t),\end{aligned} \tag{3.19}$$

where

$$U_\alpha(x_*) := \{(v,u',u,q,t):\ |v-v_*(t)|+|u'-\dot{u}_*(t)|+|u-u_*(t)|+|q| < \alpha,\ t\in\mathcal{I}_0\}$$

and the value of α is so small that $(u', u+v, t) \in \mathcal{G}$ is satisfied for

$$(v,u',u,q,t) \in U_\alpha(x_*).$$

Note that the variables v, u', u and q play the role of parameters here. Then,

$$\bar{F}(v_*(t),\dot{u}_*(t),u_*(t),0,t) = 0, \quad \bar{F}'_v(v_*(t),\dot{u}_*(t),u_*(t),0,t) = I \quad \forall\, t\in\mathcal{I}_0$$

is true, and the implicit function theorem provides a function

$$\bar{f}: U_\rho(x_*) \to \mathbb{R}^m,$$

where

$$U_\rho(x_*) := \{(u',u,q,t):\ |u'-\dot{u}_*(t)|+|u-u_*(t)|+|q| < \rho,\ t\in\mathcal{I}_0\},$$

which has continuous partial derivatives $\bar{f}_{u'}$, $\bar{f}_u$, $\bar{f}_q$ and satisfies the relations

$$\begin{aligned}\bar{F}(\bar{f}(u',u,q,t),u',u,q,t) = 0, \quad \bar{f}(u',u,q,t) = Q\bar{f}(u',u,q,t)\\ \bar{f}(\dot{u}_*(t),u_*(t),0,t) = v_*(t), \quad \bar{f}'_{u'}(\dot{u}_*(t),u_*(t),0,t) = 0.\end{aligned}$$

Inserting the function $\bar{f}$ into equation (3.17) we obtain (dropping the argument t)

$$\dot{u} + PG_1^{-1}Bu + PG_1^{-1}h(\dot{u}, u+\bar{f}(\dot{u},u,q,\cdot),\cdot) - PG_1^{-1}r_* = PG_1^{-1}q. \tag{3.20}$$

We introduce a function $\bar{K}: V_\alpha(x_*) \to \mathbb{R}^m$ defined by

$$\begin{aligned}\bar{K}(u',u,q,t) := u' + PG_1^{-1}(t)B(t)u - PG_1^{-1}r_* - PG_1^{-1}(t)q\\ + PG_1^{-1}(t)h(u',u+\bar{f}(u',u,q,t),t),\end{aligned}$$

where

$$V_\alpha(x_*) := \{(u',q,t):\ |u'-\dot{u}_*(t)|+|u-u_*(t)|+|q| \le \alpha,\ t\in\mathcal{I}_0\}.$$

Again, the variables u', u and q play the role of parameters at this point. The function $\bar{K}$ has continuous partial derivatives $\bar{K}_{u'}$, $\bar{K}_u$, $\bar{K}_q$ and satisfies the relations

$$\bar{K}(\dot{u}_*(t),u_*(t),0,t) = 0, \quad \bar{K}'_{u'}(\dot{u}_*(t),u_*(t),0,t) = I \qquad \forall\, t\in\mathcal{I}_0.$$

Applying the implicit function theorem again, we solve the system (3.20) with respect to u' and obtain an explicit ODE of the form

$$u' = \bar{k}(u, q, t), \tag{3.21}$$

where $\bar{k}$ is a function mapping $V_\sigma(x_*)$ into $\mathbb{R}^m$ with

$$V_\sigma(x_*) := \{(u, q, t) : \, |u - u_*(t)| + |q| \leq \sigma, \; t \in \mathcal{I}_0\}.$$

The function $\bar{k}$ has continuous partial derivatives $\bar{k}'_u$, $\bar{k}'_q$ and satisfies the relations

$$\bar{K}(\bar{k}(u,q,t), u, q, t) = 0, \quad \bar{k}(u,q,t) = P\bar{k}(u,q,t), \quad \dot{u}_*(t) = \bar{k}(u_*(t), 0, t).$$

We may solve the regular IVP

$$\dot{u} = \bar{k}(u, q, t), \quad u(s) = u_s \in \operatorname{im} P$$

on $C^1(\mathcal{J}, \mathbb{R}^m)$ with $\mathcal{J} := [s, S) \subset \mathcal{I}_0$ and $s \in \mathcal{I}_1$ if $|u_s - u_*(s)| + \|q\|_\infty$ is sufficiently small. Then, the unique solution of the IVP (3.13)-(3.14) is given by

$$x(t) := u(t) + \bar{f}(\, \bar{k}(u(t), q(t), t), \, u(t), \, q(t), \, t). \tag{3.22}$$

Since $\bar{f}$ has the continuous derivatives $\bar{f}'_{u'}$, $\bar{f}'_u$ and $\bar{f}'_q$, and $\mathcal{I}_0$ is a compact interval, there is a constant $L_{\bar{f}}$ such that

$$\|v - v_*\|_\infty \leq L_{\bar{f}} \left(\|\dot{u} - \dot{u}_*\|_\infty + \|u - u_*\|_\infty + \|q\|_\infty\right) \tag{3.23}$$

is satisfied. Since $\bar{k}$ has the continuous derivatives $\bar{k}'_u$ and $\bar{k}'_q$, and $\mathcal{I}_0$ is a compact interval, there is a constant $L_{\bar{k}}$ such that

$$|\dot{u}(t) - \dot{u}_*(t)| \leq L_{\bar{k}} \left(|u(t) - u_*(t)| + \|q\|_\infty\right) \quad \forall t \in \mathcal{J}.$$

Analogously to the estimations in the proof of Theorem 3.1 on page 31, we obtain

$$|u(t) - u_*(t)| \leq (e^{L_{\bar{k}}(t-s)} - 1)(|u(s) - u_*(s)| + \|q\|_\infty) \quad \forall \, t \in \mathcal{J}.$$

Now, there is a constant K_1 such that the relation

$$\|u - u_*\|_\infty \leq K_1(|u(s) - u_*(s)| + \|q\|_\infty)$$

is satisfied. Further, the inequality

$$\|\dot{u} - \dot{u}_*\|_\infty \leq K_2(|u(s) - u_*(s)| + \|q\|_\infty)$$

is true for a constant K_2. Regarding the estimation (3.23) for v and the description (3.22) of the solution x, we obtain

$$(\|x - x_*\|_\infty + \|P\dot{x} - P\dot{x}_*\|_\infty) \leq K(|u(s) - u_*(s)| + \|q\|_\infty)$$

for a constant K, i.e.,

$$(\|x - x_*\|_\infty + \|P\dot{x} - P\dot{x}_*\|_\infty) \leq K(|u_s - Px_*(s)| + \|q\|_\infty).$$

□

Theorem 3.2 provides a relation between the tractability index and the perturbation index immediately.

Corollary 3.3 *If the assumptions of Theorem 3.2 are satisfied, then the DAE 3.12 has the perturbation index 1.*

Following the proof of Theorem 4.1 in [Mär95], for the special class of quasilinear systems

$$A\dot{x}(t) + g(x(t)) = r(t), \qquad (3.24)$$

we may additionally derive a relation to the differential index.

Theorem 3.4 *If the assumptions of Theorem 3.2 are satisfied, then the DAE 3.12 has the differential index 1.*

3.3 Properties of projectors and spaces related to index-2 DAEs

Before analyzing index-2 tractable DAEs, we present some useful properties of the projectors and spaces describing the index-2 DAEs in more detail (for definitions see Subsection 1.2.4).

The first two lemmata describe the image-space and the kernel of the matrix function $QQ_1(y, x, t)$. If we look at Example A in Chapter 1, this matrix function has the form

$$QQ_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

if we choose

$$Q = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } Q_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

In other words, QQ_1x represents the current I. This is the variable depending on the derivative of the input function $\mathrm{v}(t)$. Later we will see that the matrix function $QQ_1(y, x, t)$ is useful for describing the differentiation problem involved in the index 2 DAE.

Lemma 3.5 *Independent of the choice of the projectors Q and $Q_1(y, x, t)$ the relation*

$$im\, QQ_1(y, x, t) = N \cap S(y, x, t)$$

is satisfied.

Proof: Let us drop the argument (y, x, t) in the following.

($\subseteq$) For any $z \in \operatorname{im} QQ_1$, we have $z \in \operatorname{im} Q = \ker A = N$. Further, there exists a $w \in \mathbb{R}^m$ such that $z = QQ_1w$ is true. Thus,

$$Bz = BQQ_1w = (G_1 - A)Q_1w = A(-Q_1w) \in \operatorname{im} A$$

is satisfied, i.e., $z \in S$.

($\supseteq$) For any $z \in (N \cap S)$ we find a $v \in \mathbb{R}^m$ such that $Bz = Av$ is valid. Defining $u := z - Pv$ yields

$$\begin{aligned} G_1u &= G_1Qz - Av \qquad (z \in N = \operatorname{im} Q) \\ &= Bz - Av = 0, \end{aligned}$$

i.e., $u \in N_1 = \operatorname{im}\, Q_1$. This implies $z = Qu = QQ_1u$, i.e., $z \in \operatorname{im} QQ_1$.

□

Lemma 3.6 *Independent of the choice of the projectors Q and $Q_1(y, x, t)$, the relation*

$$ker\, QQ_1(y, x, t) = ker\, Q_1(y, x, t)$$

is true.

Proof: Let us drop the argument (y,x,t) again.

($\supseteq$) This is obvious.

($\subseteq$) Let $z \in \ker QQ_1$, hence

$$Q_1 z \in \ker Q. \tag{3.25}$$

On the other hand, $G_1 Q_1 z = 0$, which leads to $AQ_1 z = 0$. That means $Q_1 z \in \operatorname{im} Q$. Together with (3.25) we obtain $Q_1 z = 0$.

□

Taking into account Lemma A.1, we can choose the projector $Q_1(y, x, t)$ along $S_1(y, x, t)$. In this case, we speak of the **canonical projector $\boldsymbol{Q_1}$**(y, x, t). Further, we denote $\boldsymbol{P_1}(y, x, t) := I - Q_1(y, x, t)$.

Then, the relation

$$Q_1(y, x, t)Q = 0 \tag{3.26}$$

is true for the canonical projector $Q_1(y, x, t)$. This follows from the relation

$$Q_1(y, x, t) = Q_1(y, x, t)G_2^{-1}(y, x, t)B(y, x, t)P,$$

which is a trivial conclusion of Lemma A.1.

Next, we define a projector $\boldsymbol{T}(y, x, t)$ onto the space

$$N \cap S(y, x, t)$$

for index-2 DAEs. Let $\boldsymbol{U}(y, x, t) := I - T(y, x, t)$. The space $N \cap S(y, x, t)$ would be equal to $\{0\}$ if the DAE were of index 1. This space may be considered as the space describing the components of a solution involved in the differentiation problem of an index 2 DAE. In the case of Example A in Chapter 1 we have seen that

$$N \cap S(x) = \{(z_1, z_2, z_3)^T : z_1 = z_3 = 0\}.$$

Indeed, the current I of a solution of the system (1.4)-(1.6) depends on the derivative of the second component u.

Remarks 3.7

(1) For semi-explicit systems

$$\begin{aligned} \dot{u} + f(u, v, t) &= 0 \\ g(u, v, t) &= 0, \end{aligned}$$

the space $N \cap S(y,x,t)$ simplifies to

$$\begin{aligned}\{\left(\begin{smallmatrix} z_u \\ z_v \end{smallmatrix}\right) : z_u = 0\} \cap \{\left(\begin{smallmatrix} z_u \\ z_v \end{smallmatrix}\right) : g_u'(u,v,t)z_u + g_v'(u,v,t)z_v = 0\} \\ = \{\left(\begin{smallmatrix} z_u \\ z_v \end{smallmatrix}\right) : z_u = 0;\ g_v'(u,v,t)z_v = 0\}.\end{aligned}$$

In particular, in the case of Hessenberg systems, i.e., if the function g does not depend on v, the space $N \cap S(y,x,t)$ is independent of (y,x,t) and may be written as $\{\left(\begin{smallmatrix} z_u \\ z_v \end{smallmatrix}\right) : z_u = 0\} = N$.

(2) The matrices $T(y,x,t)Q$ and $U(y,x,t)Q$ are projectors. This becomes obvious if we regard

$$QT(y,x,t)Q = T(y,x,t)Q, \tag{3.27}$$

since $\operatorname{im} T(y,x,t) \subseteq \operatorname{im} Q$.

(3) The relation

$$PQ_1(y,x,t)U(y,x,t)Q = 0 \tag{3.28}$$

is satisfied, since $Q_1(y,x,t)Q = 0$ is fulfilled.

For index-2 DAEs, the matrix $G_2(y,x,t)$ is regular (see Remark 1.11), and the following lemma provides a description of $\operatorname{im} A(y,x,t)$ that is closely related to the splitting technique used in the next sections.

Lemma 3.8 *The projector functions and matrix functions defined above satisfy the relation*

$$im\, A(y,x,t) = ker\left([PQ_1(y,x,t) + U(y,x,t)Q]G_2^{-1}(y,x,t)\right).$$

Proof: Let us drop the argument (y,x,t) again.

($\subseteq$) For any $x \in \ker(PQ_1 + UQ)G_2^{-1}$, we obtain

$$PQ_1G_2^{-1}x = 0 \qquad \text{and} \qquad UQG_2^{-1}x = 0.$$

Regarding $\operatorname{im}(TQG_2^{-1}) \subseteq \operatorname{im} T \subseteq \operatorname{im} S$, we obtain

$$x = [A + BQ + BPQ_1]G_2^{-1}x = AG_2^{-1}x + BTQG_2^{-1}x \in \operatorname{im} A.$$

($\supseteq$) For any $x \in \operatorname{im} A$, we find a $y \in \mathbb{R}^m$ such that $x = Ay$ is true. This implies $G_2^{-1}x = P_1Py$, which leads to

$$(PQ_1 + UQ)G_2^{-1}y = UQP_1Py = -UQQ_1y = 0.$$

□

Lemma 3.9 *The image space of $G_1(y,x,t)$ may be described by*

$$im\, G_1(y,x,t) = ker\left(Q_1(y,x,t)G_2^{-1}(y,x,t)\right).$$

Proof: Let us drop the argument (y,x,t) again.

($\subseteq$) For any $x \in \operatorname{im} G_1$, we find a y such that $x = G_1 y$ is true. Hence, $Q_1 G_2^{-1} x = Q_1 G_2^{-1} G_1 y = Q_1 P_1 y = 0$ is fulfilled.

($\supseteq$) For any $x \in \ker\{Q_1 G_2^{-1}\}$, we obtain

$$x = G_2 G_2^{-1} x = G_1 G_2^{-1} x + BPQ_1 G_2^{-1} x = G_1 G_2^{-1} x.$$

□

3.4 Linear index-2 DAEs

We consider the linear index-2 DAE

$$A(t)\dot{x}(t) + B(t)x(t) = r(t). \tag{3.29}$$

Systems of higher index (i.e., index-2 systems, too) are characterized by the fact that there is no algebraic transformation for dividing the DAE into the inherent regular ODE and an algebraic equation. However, we possibly find a splitting into 3 parts,

(a) inherent explicit ODE,

(b) part describing the inherent differentiation problem,

(c) purely algebraic part.

Dropping the argument t we may rewrite the system (3.29) as

$$G_2^{-1} A\dot{x} + G_2^{-1} Bx = G_2^{-1} r.$$

Using the relations

$$\begin{aligned} G_2^{-1} A &= P_1 P \\ G_2^{-1} B &= G_2^{-1} BPP_1 + Q_1 + Q \end{aligned}$$

for any projector Q onto $\ker A$ and the canonical projector Q_1 onto $\ker G_1$ along S_1, we obtain

$$P_1P\dot{x} + G_2^{-1}BPP_1x + Q_1x + Qx = G_2^{-1}r. \tag{3.30}$$

Multiplying the system by PP_1, TQP_1, $UQ + PQ_1$, and taking into account

$$Q_1 = Q_1G_2^{-1}BP, \quad Q_1Q = 0, \quad PP_1Q = 0, \quad TQQ_1 = QQ_1, \quad UQQ_1 = 0,$$

as well as (3.27) and (3.28), the system (3.30) is equivalent to

$$\begin{aligned} PP_1\dot{x} + PP_1G_2^{-1}BPP_1x &= PP_1G_2^{-1}r \\ -QQ_1\dot{x} + TQP_1G_2^{-1}BPP_1x + TQx &= TQP_1G_2^{-1}r \\ UQG_2^{-1}BPP_1x + (UQ + PQ_1)x &= (UQ + PQ_1)G_2^{-1}r. \end{aligned}$$

Note that the equivalence is given, because

$$I = PP_1 + TQP_1 + (UQ + Q_1)(UQ + PQ_1)$$

is true. Supposing the canonical projector Q_1 to be smooth, and defining

$$u := PP_1x, \;\; w := TQx, \;\; y := (PQ_1 + UQ)x,$$

we obtain

$$\dot{\boldsymbol{u}} - P\dot{P}_1(\boldsymbol{u} + Py) + PP_1G_2^{-1}B\boldsymbol{u} = PP_1G_2^{-1}r \tag{3.31}$$

$$-QQ_1(P\dot{y}) + Q\dot{Q}_1u + TQP_1G_2^{-1}Bu + \boldsymbol{w} = TQP_1G_2^{-1}r \tag{3.32}$$

$$UQG_2^{-1}Bu + \boldsymbol{y} = (UQ + PQ_1)G_2^{-1}r. \tag{3.33}$$

This system looks complicated, but it provides the structure of linear index-2 DAEs. Concerning Lemma 3.8 it is obvious that equation (3.33) represents the *inherent algebraic part* of the DAE. This equation makes it possible to determine the component y as a function of the component u. Inserting the term for y into equation (3.31), this equation represents the *inherent explicit ODE* of the DAE determining the component u of the system. Finally, equation (3.32) represents the *part describing the inherent differentiation problem* and makes it possible to determine the component w, the so-called index-2 component, where we have to differentiate a part of equation (3.33) (to be precise, we have to differentiate $PQ_1G_2^{-1}q$).

Note that the splitting technique given here slightly differs from the splitting technique in former papers of März et al. Now, the following theorem is true.

Theorem 3.10 *Let $x_* \in C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ be a solution of the index-2 tractable DAE (3.29). Further, let the canonical projector $Q_1(t)$ and the right-hand side $PQ_1G_2^{-1}r$ be smooth. Then, the perturbed initial value problem*

$$A(t)\dot{x}(t) + B(t)x(t) = r(t) + q(t) \tag{3.34}$$

$$PP_1(t_0)x(t_0) = u_0 \in im\, PP_1(t_0) \tag{3.35}$$

is uniquely solvable on $C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ for continuous perturbations q with the part $PQ_1G_2^{-1}q \in C^1$. Further, the estimation

$$\|x - x_*\|_\infty + \|\frac{d}{dt}(PP_1(x - x_*))\|_\infty \leq$$
$$K\left(\|q\|_\infty + \|\frac{d(PQ_1G_2^{-1}q)}{dt}\|_\infty + |u_0 - PP_1(t_0)x_*(t_0)|\right) \tag{3.36}$$

is fulfilled.

Remark 3.11 If we additionally assume the perturbation-part $UQG_2^{-1}q$ and the matrix function $UQG_2^{-1}B$ to be smooth, then the solution belongs not only to $C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ but the component $U(t)Qx(t)$ is also smooth.

Proof: Following the splitting technique described above for the perturbed right-hand side $r + q$ instead of r, the solution of (3.34)-(3.35) is given by

$$x := u + [(UQ + PQ_1)G_2^{-1}(r + q) - UQG_2^{-1}Bu]$$
$$+ [TQP_1G_2^{-1}(r + q) - TQP_1G_2^{-1}Bu - Q\dot{Q}_1u + QQ_1\frac{d}{dt}(PQ_1G_2^{-1}(r + q))] \tag{3.37}$$

if u is a solution of the IVP

$$\dot{u} - P\dot{P}_1(u + PQ_1G_2^{-1}(r + q)) + PP_1G_2^{-1}Bu = PP_1G_2^{-1}(r + q)$$
$$u(t_0) = u_0 \in \operatorname{im} PP_1(t_0).$$

This implies

$$\|x - x_*\|_\infty \leq K_1\left(\|u - u_*\|_\infty + \|q\|_\infty + \|\frac{d}{dt}(PQ_1G_2^{-1}q)\|_\infty\right)$$

and

$$|\dot{u}(t) - \dot{u}_*(t)| \leq K_2\left(|u(t) - u_*(t)| + \|q\|_\infty\right) \qquad \forall\, t \in \mathcal{I}_0$$

for certain constants K_1 and K_2. Now, the assertion follows with the same arguments as in the proof of Theorem 3.1.

□

3.5 Nonlinear index-2 DAEs

We consider quasilinear index-2 DAEs of the form

$$A(t)\dot{x}(t) + \mathfrak{g}(x(t), t) = 0 \tag{3.38}$$

with constant nullspace $\ker A(t)$. The DAEs arising from charge oriented MNA belong to this class of quasilinear DAEs (see Chapter 5). Let $x_* \in C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ be again a solution of (3.38). In the following, we compute the useful projectors and matrix functions in x_* and denote

$$\begin{aligned} T(t) &:= T(\dot{x}_*(t), x_*(t), t), & U(t) &:= U(\dot{x}_*(t), x_*(t), t)\\ Q_1(t) &:= Q_1(\dot{x}_*(t), x_*(t), t), & P_1(t) &:= P_1(\dot{x}_*(t), x_*(t), t)\\ G_1(t) &:= G_1(\dot{x}_*(t), x_*(t), t), & G_2(t) &:= G_2(\dot{x}_*(t), x_*(t), t). \end{aligned}$$

$Q_1(t)$ is chosen as the canonical projector onto $\ker G_1(t)$ along

$$S_1(t) = \{z \in \mathbb{R}^m : \mathfrak{g}'_x(x_*(t), t)z \in \operatorname{im} A(t)\}.$$

Note that the calculation of these matrix functions and projectors is only of theoretical interest. In praxis, we do not need these special projectors.

Assumption: The function

$$\hat{\mathfrak{g}}(x, t) := [U(t)Q + PQ_1(t)]G_2^{-1}(t)\mathfrak{g}(x, t)$$

is twice continuously differentiable.

This assumption means that the derivative free part of the DAE has to be twice continuously differentiable. For Hessenberg systems of index 2

$$\begin{aligned} \dot{u} + g(u, v, t) &= 0,\\ h(u, t) &= 0, \end{aligned}$$

where $x = \binom{u}{v}$ and $\mathfrak{g} = \binom{g}{h}$, the function $\hat{\mathfrak{g}}$ has the form

$$\hat{\mathfrak{g}}(x, t) = \begin{pmatrix} M(x_*(t), t)h(u, t) \\ 0 \end{pmatrix}$$

with $M := g'_v(h'_u g'_v)^{-1}$. Hence, essentially $h(u, t)$ is required to be twice continuously differentiable. The matrix $M(x_*(t), t)$ can be considered as a certain normalization.

Theorem 3.12 *Let $x_* \in C^1_N(\mathcal{I}_0, \mathbb{R}^m)$ be a solution of the index-2 tractable DAE (3.38). Further, let $UQx_* \in C^1(\mathcal{I}_0, \mathbb{R}^m)$ be fulfilled and let Q_1 as well as $UQG_2^{-1}B$ be of class C^1. If the structural condition*

$$Q_1(t)(I + \hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t))^{-1}T(t)Q = 0 \tag{3.39}$$

for all (x,t) with $|x - x_*(t)| \leq \varrho$ is satisfied, the perturbation q is continuous, and its part

$$\hat{q}(t) := [U(t)Q + PQ_1(t)]G_2^{-1}(t)q(t)$$

is of class C^1, then the perturbed initial value problem

$$A(t)\dot{x}(t) + \mathfrak{g}(x(t),t) = q(t) \tag{3.40}$$

$$PP_1(s)x(s) = u_s \in im PP_1(s), \quad |u_s - PP_1(s)x_*(s)| \leq \tau \tag{3.41}$$

$$\|q\|_\infty + \|\frac{d\hat{q}}{dt}\|_\infty \leq \sigma \tag{3.42}$$

is uniquely solvable on $C^1_N(\mathcal{J}, \mathbb{R}^m)$ for each $s \in \mathcal{I}_1$ and sufficiently small τ and σ, where $\mathcal{J} := [s, S) \subset \mathcal{I}_0$. Moreover, the inequality

$$\|x - x_*\|_\infty + \|\frac{d}{dt}(PP_1(x - x_*))\|_\infty \leq$$
$$K\left(\|q\|_\infty + \|\frac{d\hat{q}}{dt}\|_\infty + |u_s - PP_1(s)x_*(s)|\right)$$

is satisfied for a constant K.

The structural condition (3.39) generalizes the structural condition (3.5) in [MT94]. It ensures that a recursively defined vector field is given on the solution manifold. Hence, Theorem 3.12 is a generalization of Theorem 3.4 in [MT94].

There are networks, e.g. the ring modulator (see Chapter ch.simulation), for which the condition (3.5) given in [MT94] is *not* satisfied. Therefore, we have looked for a generalization. The new structural condition (3.39) seems to be satisfied for electric networks. In any case, all circuit-examples of index 2 we know fulfil (3.39).

We will discuss condition (3.39) after the proof of the theorem in more detail.

Proof: As in Section 3.2 we transform the equation

$$A(t)\dot{x}(t) + \mathfrak{g}(x(t),t) = q(t)$$

into the equivalent system

$$A(t)\dot{x}(t) + B(t)x(t) + h(\dot{x}(t), x(t), t) - r_*(t) = q(t), \tag{3.43}$$

where

$$\begin{aligned} B(t) &:= \mathfrak{g}'_x(x_*(t), t), \\ h(x,t) &:= \mathfrak{g}(x,t) - \mathfrak{g}(x_*(t), t) - B(t)(x - x_*(t)), \\ r_*(t) &:= A(t)\dot{x}_*(t) + B(t)x_*(t). \end{aligned}$$

Then, the function h has the properties

$$h(x_*(t), t) = 0, \quad h'_x(x_*(t), t) = 0 \quad \forall\, t \in \mathcal{I}.$$

Again, h is constructed in such a way that the functions h and h'_x map small neighbourhoods of the trajectory $\{(\dot{x}_*(t), x_*(t), t),\ t \in \mathcal{I}_0\}$ into small spheres around the zero.

Since the DAE (3.38) is index-2 tractable, the matrix $G_2(t)$ is regular (see Remark 1.11).

Dropping the argument t, the system (3.43) is equivalent to

$$G_2^{-1}A\dot{x} + G_2^{-1}Bx + G_2^{-1}h(x,\cdot) - G_2^{-1}r_*(t) = G_2^{-1}q.$$

Using the relations

$$G_2^{-1}A = P_1P, \quad G_2^{-1}BQ = Q, \quad G_2^{-1}BPQ_1 = Q_1$$

we now have

$$P_1P\dot{x} + G_2^{-1}BPP_1x + Q_1x + Qx + G_2^{-1}h(x,\cdot) - G_2^{-1}r_*(t) = G_2^{-1}q. \tag{3.44}$$

Multiplying (3.44) by PP_1, TQP_1, $(UQ + PQ_1)$, and introducing

$$u := PP_1x, \quad w := TQx, \quad y := (UQ + PQ_1)x,$$

we obtain similarly to the linear case (see Section 3.4)

$$\begin{aligned} \dot{\boldsymbol{u}} - P\dot{P}_1(\boldsymbol{u} + Py) + PP_1G_2^{-1}B\boldsymbol{u} - PP_1G_2^{-1}r_* \\ + PP_1G_2^{-1}h(\boldsymbol{u} + w + y, \cdot) = PP_1G_2^{-1}q \end{aligned} \tag{3.45}$$

$$\begin{aligned} - QQ_1(P\dot{y}) + Q\dot{Q}_1u + TQP_1G_2^{-1}Bu + \boldsymbol{w} \\ - TQP_1G_2^{-1}r_*(t) + TQP_1G_2^{-1}h(u + \boldsymbol{w} + y, \cdot) = TQP_1G_2^{-1}q \end{aligned} \tag{3.46}$$

$$\begin{aligned} UQG_2^{-1}Bu + \boldsymbol{y} - (UQ + PQ_1)G_2^{-1}r_*(t) \\ +(UQ + PQ_1)G_2^{-1}h(u + TQw + \boldsymbol{y}, \cdot) = (UQ + PQ_1)G_2^{-1}q. \end{aligned} \tag{3.47}$$

We may solve the algebraic part (3.47) of the DAE with respect to y in a neighbourhood of the solution x_*. We introduce

$$\begin{aligned}\bar{\bar{F}}(y,u,w,\hat{q},t) := {} & U(t)QG_2^{-1}(t)B(t)u + y - (U(t)Q + PQ_1(t))G_2^{-1}(t)r_*(t)\\ & + [U(t)Q + PQ_1(t)]G_2^{-1}(t)h(u + T(t)Qw + y, t) - \hat{q}\\ = {} & U(t)QG_2^{-1}(t)B(t)(u - u_*) + (y - y_*)\\ & + [U(t)Q + PQ_1(t)]G_2^{-1}(t)h(u + T(t)Qw + y, t) - \hat{q}.\end{aligned}$$

Note that the variables y, u, w and $\hat{q}$ are parameters here. Then, the relations

$$\begin{aligned}\bar{\bar{F}}(y_*(t), u_*(t), w_*(t), 0, t) &= 0,\\ \bar{\bar{F}}{}'_y(y_*(t), u_*(t), w_*(t), 0, t) &= I,\\ \bar{\bar{F}}{}'_u(y_*(t), u_*(t), w_*(t), 0, t) &= U(t)QG_2^{-1}(t)B(t),\\ \bar{\bar{F}}{}'_{\hat{q}}(y_*(t), u_*(t), w_*(t), 0, t) &= -I,\\ \bar{\bar{F}}{}'_t(y_*(t), u_*(t), w_*(t), 0, t) &= -U(t)QG_2^{-1}(t)B(t)\dot{u}_*(t) - \dot{y}_*(t)\end{aligned}$$

are true for all $t \in \mathcal{I}_0$, and the relations

$$\begin{aligned}\bar{\bar{F}}{}'_y(y,u,w,\hat{q},t) &= I + \hat{H}(u,w,y,t),\\ \bar{\bar{F}}{}'_w(y,u,w,\hat{q},t) &= \hat{H}(u,w,y,t)T(t)Q,\\ \bar{\bar{F}}{}'_{\hat{q}}(y,u,w,\hat{q},t) &= -I\end{aligned}$$

are satisfied for all $(y,u,w,\hat{q},t)$ in a neighbourhood of the trajectory

$$(y_*(t), u_*(t), w_*(t), 0, t)$$

if the function $\hat{H}$ is defined by

$$\begin{aligned}\hat{H}(u,w,y,t) &:= [PQ_1(t) + U(t)Q]G_2^{-1}(t)h'_x(u + T(t)Qw + y, t)\\ &= \hat{\mathfrak{g}}'_x(u + T(t)Qw + y, t) - \hat{\mathfrak{g}}'_x(x_*(t), t).\end{aligned}$$

Now, the implicit function theorem provides a function

$$\bar{\bar{f}} : U_\rho(x_*) \to \mathbb{R}^m$$

satisfying

$$\bar{\bar{f}}(u_*(t), w_*(t), 0, t) = y_*(t),$$

where

$$U_\rho(x_*) := \{(u,w,\hat{q},t) : \ |u - u_*(t)| + |w - w_*(t)| + |\hat{q}| < \rho, \ t \in \mathcal{I}_0\},$$

This function satisfies the relations

$$\begin{aligned} \bar{\bar{f}}(u,w,\hat{q},t) &= [U(t)Q + PQ_1(t)]\bar{\bar{f}}(u,w,\hat{q},t) \\ QQ_1(t)\bar{\bar{f}}'_w(u,w,\hat{q},t) &= 0 \end{aligned}$$

for $\hat{q} \in \operatorname{im}[U(t)Q + PQ_1(t)]$. The latter relation follows from the structural condition (3.39) since

$$\begin{aligned} QQ_1(t)\bar{\bar{f}}'_w(u,w,\hat{q},t) &= -QQ_1(t)(I + \hat{H}(u,w,y,t))^{-1}\hat{H}(u,w,y,t)T(t)Q \\ &= -QQ_1(t)[I - (I + \hat{H}(u,w,y,t))^{-1}]T(t)Q \\ &= QQ_1(t)(I + \hat{H}(u,w,y,t))^{-1}]T(t)Q = 0 \end{aligned}$$

is true. Furthermore, the relations

$$\begin{aligned} QQ_1(t)\bar{\bar{f}}'_{\hat{q}}(u_*(t), w_*(t), 0, t) &= -QQ_1(t), \\ QQ_1(t)\bar{\bar{f}}'_t(u_*(t), w_*(t), 0, t) &= QQ_1(t)y'_*(t) \end{aligned}$$

are true for all $t \in \mathcal{I}_0$.

Inserting $\bar{\bar{f}}(u,w,\hat{q},t)$ into the equations (3.45) and (3.46), and regarding

$$Q_1P\dot{y} = Q_1\bar{\bar{f}}'_u(u,w,\hat{q},t)\dot{u} + Q_1\bar{\bar{f}}'_{\hat{q}}(u,w,\hat{q},t)\frac{d\hat{q}}{dt} + Q_1\bar{\bar{f}}'_t(u,w,\hat{q},t),$$

we obtain

$$\begin{aligned} \dot{u} - P\dot{P}_1(u + P\bar{\bar{f}}(u,w,\hat{q},\cdot)) + PP_1G_2^{-1}Bu - PP_1G_2^{-1}r_* \\ + PP_1G_2^{-1}h(u + w + \bar{\bar{f}}(u,w,\hat{q},\cdot),\cdot) = PP_1G_2^{-1}q \end{aligned} \tag{3.48}$$

$$\begin{aligned} -QQ_1(\bar{\bar{f}}'_u(u,w,\hat{q},\cdot)\dot{u} + \bar{\bar{f}}'_{\hat{q}}(u,w,\hat{q},\cdot)\frac{d\hat{q}}{dt} + \bar{\bar{f}}'_t(u,w,\hat{q},\cdot)) \\ + QQ'_1u + TQP_1G_2^{-1}Bu + w - TQP_1G_2^{-1}r_* \\ + TQP_1G_2^{-1}h(u + w + \bar{\bar{f}}(u,w,\hat{q},\cdot),\cdot) = TQP_1G_2^{-1}q. \end{aligned} \tag{3.49}$$

We apply the expression for $\dot{u}$ from (3.48) to equation (3.49) and obtain a nonlinear equation that may be solved with respect to w in a neighbourhood of the solution x_*. Now,

$$\begin{aligned} -QQ_1\bar{\bar{f}}'_u(u,w,\hat{q},t)\Big(P\dot{P}_1(u + P\bar{\bar{f}}(u,w,\hat{q},\cdot)) - PP_1G_2^{-1}Bu \\ + PP_1G_2^{-1}r_* - PP_1G_2^{-1}h(u + w + \bar{\bar{f}}(u,w,\hat{q},\cdot),\cdot) + PP_1G_2^{-1}q\Big) \\ - QQ_1\bar{\bar{f}}'_{\hat{q}}(u,w,\hat{q},t)\hat{q}' - QQ_1\bar{\bar{f}}'_t(u,w,\hat{q},t) + Q\dot{Q}_1u - TQP_1G_2^{-1}r_* \\ + TQP_1G_2^{-1}Bu + w + TQP_1G_2^{-1}h(u + w + \bar{\bar{f}}(u,w,\hat{q},\cdot),\cdot) = TQP_1G_2^{-1}q \end{aligned}$$

is satisfied. We denote

$$\hat{p}(t) := \frac{d\hat{q}(t)}{dt}$$

and, dropping the argument t, we define a function

$$\begin{aligned}
&\bar{\bar{G}}(w, u, \hat{p}, \hat{q}, q, \cdot) := \\
&\quad - QQ_1 \bar{\bar{f}}'_u(u, w, \hat{q}, \cdot) \Big(P\dot{P}_1(u + P\bar{\bar{f}}(u, w, \hat{q}, \cdot)) - PP_1 G_2^{-1} Bu \\
&\qquad\quad + PP_1 G_2^{-1} r_* - PP_1 G_2^{-1} h(u + w + \bar{\bar{f}}(u, w, \hat{q}, \cdot), \cdot) + PP_1 G_2^{-1} q \Big) \\
&\qquad - QQ_1 \bar{\bar{f}}'_{\hat{q}}(u, w, \hat{q}, \cdot)\hat{p} - QQ_1 \bar{\bar{f}}'_t(u, w, \hat{q}, \cdot) + Q\dot{Q}_1 u + w - TQP_1 G_2^{-1} r_* \\
&\qquad\quad + TQP_1 G_2^{-1} Bu + TQP_1 G_2^{-1} h(u + w + \bar{\bar{f}}(u, w, \hat{q}, \cdot), \cdot) - TQP_1 G_2^{-1} q.
\end{aligned}$$

Note that the variables w, u, $\hat{p}$, $\hat{q}$ and q play again the role of parameters at this point. Then,

$$\bar{\bar{G}}(w_*(t), u_*(t), 0, 0, 0, t) = 0, \quad \bar{\bar{G}}'_w(w_*(t), u_*(t), 0, 0, 0, t) = I.$$

The implicit-function theorem provides a function

$$\bar{\bar{g}} : \; U_{\rho_2}(x_*) \to \mathbb{R}^m,$$

where

$$U_{\rho_2}(x_*) := \{(u, \hat{p}, \hat{q}, q, t) : \; |u - u_*(t)| + |\hat{p}| + |\hat{q}| + |q|\| < \rho_2, \, t \in \mathcal{I}_0\}.$$

This function satisfies the relation $\bar{\bar{g}}(u, \hat{p}, \hat{q}, q, t) = T(t)Q\bar{\bar{g}}(u, \hat{p}, \hat{q}, q, t)$ for all $(u, \hat{p}, \hat{q}, t)$ in a neighbourhood of the trajectory $(u_*(t), 0, 0, 0, t)$. Furthermore, the relation

$$\bar{\bar{g}}(u_*(t), 0, 0, 0, t) = w_*(t)$$

is fulfilled for all $t \in \mathcal{I}_0$. Applying $\bar{\bar{g}}$ to equation (3.48), we obtain

$$\begin{aligned}
&\dot{u} - P\dot{P}_1(u + P\bar{\bar{f}}(u, w, \hat{q}, \cdot)) + PP_1 G_2^{-1} Bu - PP_1 G_2^{-1} r_* \\
&+ PP_1 G_2^{-1} h(u + \bar{\bar{g}}(u, \hat{p}, \hat{q}, q, \cdot) + \bar{\bar{f}}(u, \bar{\bar{g}}(u, \hat{p}, \hat{q}, q, \cdot), \hat{q}, \cdot) = PP_1 G_2^{-1} q. \quad (3.50)
\end{aligned}$$

Now, we may solve the regular IVP (3.50) together with the initial condition

$$u(s) = u_s, \quad u_s \in \operatorname{im} PP_1(s)$$

on $C^1(\mathcal{J}, \mathbb{R}^m)$ with $\mathcal{J} = [s, S) \subset \mathcal{I}_0$ if $|u_s - u_*(s)| + |\hat{p}| + |\hat{q}| + |q|$ is sufficiently small. Multiplying (3.50) by PP_1 and regarding

$$PP_1(t)\bar{\bar{f}}(u, w, \hat{q}, t) = 0, \quad \forall\, t \in \mathcal{J},$$

we conclude

$$\dot{u} - P\dot{P}_1 u = PP_1\dot{u} - PP_1P\dot{P}_1 u,$$

which implies

$$\frac{d}{dt}((I - PP_1)u) = PP_1P\dot{P}_1(I - PP_1)u.$$

Defining $\hat{u} := (I - PP_1)u$ provides an ODE

$$\frac{d\hat{u}}{dt} = PP_1P\dot{P}_1\hat{u}$$

with

$$\hat{u}(s) = (I - PP_1(s))u(s) = (I - PP_1(s))PP_1(s)u_s = 0,$$

i.e., $\hat{u}(t) = 0$ for all $t \in \mathcal{J}$. This leads to

$$u(t) = PP_1(t)u(t) \quad \forall\, t \in \mathcal{J}.$$

Assuming $\hat{p} = \frac{d\hat{q}}{dt}$ and q to be sufficiently small, which includes $\hat{q}$ to be small, too, the unique solution of the IVP (3.40)-(3.42) is given by

$$x := u + \bar{\bar{g}}(u, \hat{p}, \hat{q}, q, \cdot) + \bar{\bar{g}}(u, \bar{\bar{g}}(u, \hat{p}, \hat{q}, q, \cdot), \hat{q}, \cdot)$$

in a neighbourhood $U_{\rho_3}(x_*)$, where

$$U_{\rho_3}(x_*) := \{(x,t) : \ |x - x_*(t)| < \rho_3, \ t \in \mathcal{J}\}.$$

The relation (3.50) implies

$$|\dot{u}(t) - \dot{u}_*(t)| \leq L_1(|u(t) - u_*(t)| + \|\frac{d\hat{q}}{dt}\|_\infty + \|q\|_\infty), \quad t \geq s, \tag{3.51}$$

for a certain constant L_1. Consequently, there is constant K_1 such that the relation

$$\|u - u_*\|_\infty \leq K_1(|u(s) - u_*(s)| + \|\frac{d\hat{q}}{dt}\|_\infty + \|q\|_\infty) \tag{3.52}$$

is satisfied. Since $\bar{\bar{g}}$ has the continuous partial derivatives $\bar{\bar{g}}'_u$, $\bar{\bar{g}}'_{\hat{p}}$, $\bar{\bar{g}}'_{\hat{q}}$, $\bar{\bar{g}}'_q$, and $\mathcal{I}_0$ is a compact interval, there is a constant $L_{\bar{\bar{g}}}$ satisfying

$$\|w - w_*\|_\infty \leq L_{\bar{\bar{g}}}(\|u - u_*\|_\infty + \|\frac{d\hat{q}}{dt}\|_\infty + \|q\|_\infty). \tag{3.53}$$

Since $\bar{\bar{f}}$ has the continuous partial derivatives $\bar{\bar{f}}'_u$, $\bar{\bar{f}}'_w$, $\bar{\bar{f}}'_{\hat{q}}$, and $\mathcal{I}_0$ is a compact interval, we find a constant $L_{\bar{\bar{f}}}$ such that

$$\|y - y_*\|_\infty \leq L_{\bar{\bar{f}}}(\|u - u_*\|_\infty + \|w - w_*\|_\infty + \|q\|_\infty) \tag{3.54}$$

is true. Now, the estimations (3.51)-(3.54) provide

$$\begin{aligned}\|x-x_*\|_\infty + \|\frac{d}{dt}(PP_1(x-x_*))\|_\infty \\ \leq K(|u(s)-u_*(s)| + \|\frac{d\hat{q}}{dt}\|_\infty + \|q\|_\infty)\end{aligned}$$

for a certain constant K.

□

Let us present some cases in which the structural condition (3.39) is satisfied. We hope that this will make the condition easily accessible for the reader.

1. If the function $\mathfrak{g}$ satisfies the relation

$$[\mathfrak{g}(x,t) - \mathfrak{g}(Px,t)] \in \operatorname{im} A_1(t) \tag{3.55}$$

in a neighbourhood of the trajectory $(x_*(t),t)$ $(t \in \mathcal{I}_0)$, then the structural condition (3.39) is true. This becomes obvious if we regard the following facts. The relation (3.55) implies

$$Q_1(t)G_2^{-1}(t)(\mathfrak{g}'_x(x,t) - \mathfrak{g}'_x(x_*(t),t))Q = 0.$$

Hence,

$$\begin{aligned}Q_1(t)(\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t))T(t)Q = 0,\\ Q_1(t)(\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t))U(t)Q = 0\end{aligned}$$

are fulfilled. Now,

$$\begin{aligned}&Q_1(t)(I + \hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t))^{-1}T(t)Q\\ &= Q_1(t)\sum_{i=0}^{\infty}(-\hat{\mathfrak{g}}'_x(x,t) + \hat{\mathfrak{g}}'_x(x_*(t),t))^iT(t)Q\\ &= Q_1(t)\sum_{i=0}^{\infty}(-PQ_1(t)[\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)])^iT(t)Q\\ &= Q_1(t)\sum_{i=1}^{\infty}(-PQ_1(t)[\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)])^{i-1}\cdot\\ &\qquad\cdot(-PQ_1(t)[\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)]T(t)Q)\\ &= 0.\end{aligned}$$

Condition (3.55) is equivalent to the condition

$$Q^*(t)[\mathfrak{g}(x,t) - \mathfrak{g}(Px,t)] \in \operatorname{im} Q^*(t)\mathfrak{g}'_x(x_*(t),t)Q$$

if $Q^*(t)$ projects along $\operatorname{im} A(t)$. This condition is equivalent to the structural condition (3.5) in [MT94]. Unfortunately, the circuit simulation provides examples for which the condition (3.55) is not satisfied.

2. If the function $\mathfrak{g}$ satisfies the relation

$$[\mathfrak{g}(x,t) - \mathfrak{g}((I - T(t)Q)x,t)] \in \operatorname{im} A(t) \tag{3.56}$$

in a neighbourhood of the trajectory $(x_*(t),t)$ $(t \in \mathcal{I}_0)$, then the structural condition (3.39) is valid. This is easily seen when considering the following facts. The relation (3.56) implies

$$[\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)]T(t)Q = 0.$$

Therefore,

$$\begin{aligned} &Q_1(t)(I + [\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)])^{-1}T(t)Q \\ &\qquad = Q_1(t)(I + [\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)])^{-1} \cdot \\ &\qquad\qquad \cdot [\hat{\mathfrak{g}}'_x(x,t) - \hat{\mathfrak{g}}'_x(x_*(t),t)]T(t)Q = 0 \end{aligned}$$

is true.

3. If the space $N \cap S(x,t)$ does not depend on x in a neighbourhood of the trajectory $(x_*(t),t)$ $(t \in \mathcal{I}_0)$, then the condition (3.39) is true. This follows from (3.56) and

$$[\mathfrak{g}(x,t) - \mathfrak{g}((I - T(t)Q)x,t)] = -\int_0^1 \mathfrak{g}'_x(x - \phi T(t)Qx,t)\, d\phi T(t)Qx,$$

which implies

$$[\mathfrak{g}(x,t) - \mathfrak{g}((I - T(t)Q)x,t)] \in \operatorname{im} A(t)$$

since

$$\operatorname{im}(T(t)Q) \subseteq (N \cap S(x - \phi T(t)Qx,t))$$

is fulfilled. Up to now, we know only examples of circuit simulation (obtained by the charge-oriented MNA) for which the space $N \cap S(x,t)$ is constant. The question whether this space is always independent of (x,t) is still open.

4. For Hessenberg systems

$$\begin{aligned} \dot{x}_1 &= \mathfrak{g}'_1(x_1,x_2,t) \\ 0 &= \mathfrak{g}'_2(x_1,t) \end{aligned}$$

condition (3.39) is obviously satisfied, because

$$N \cap S(x_1, x_2, t) = N$$

is constant in this case.

Theorem 3.12 provides a relation between the tractability index and the perturbation index.

Corollary 3.13 *If the assumptions of Theorem 3.12 are satisfied, then the perturbation index of the DAE (3.38) is not greater than 2.*

Remark 3.14 General differential algebraic equations

$$f(\dot{x}, x, t) = 0 \tag{3.57}$$

with constant nullspace $\ker f_{\dot{x}}$ can be transformed into a quasilinear DAE

$$P\dot{x} - y = 0 \tag{3.58}$$

$$f(y, x, t) = 0, \tag{3.59}$$

equivalently. System (3.57) is index-2 tractable if and only if the system (3.58)-(3.59) is so. Hence, the results for quasilinear systems (3.38) can be generalized to systems of the form (3.57).

Chapter 4

BDF applied to index-2 DAEs

In this Chapter we want to describe the behaviour of the BDF applied to index-2 DAEs. In numerous papers (e.g. [Bre83], [GP84], [GGL85], [LP86], [BE88], [BCP89], [Mär92a], [Tis95]) the BDF method has already been analyzed for several classes of index-2 problems. In this paper, we concentrate on the class arising from MNA in circuit simulation. Since these systems are not of Hessenberg form, we have to generalize the well-known results on this class. Further, we are interested in the question, under which assumptions the BDF method is feasible, i.e., under which assumptions the nonlinear equations arising from the method have a unique solution. The results presented here are a generalization of the results given in [Tis95]. As already mentioned in Chapter 3, there are electric networks, for which the structural condition in [Tis95] is not valid. Therefore, we introduce a new (more general) structural condition (3.39) that is satisfied for all circuit examples (arising from the charge-oriented MNA) that we know. Therefore, we want to assume (3.39) to be satisfied.

We consider quasilinear index-2 tractable DAEs of the form

$$A(t)\dot{x}(t) + \mathfrak{g}(x(t), t) = 0 \tag{4.1}$$

with constant nullspace $\ker A(t)$ as in Section 3.5. Note that the charge-oriented MNA provides DAEs of quasilinear form with constant nullspace $\ker A(t)$ (see Chapter 2).

The splitting technique used in the next sections is closely related to that of Chapter 3. The properties of some projectors and spaces given in Section 3.3 will be frequently applied in the following.

4.1 Analysis

We assume the assumptions of Theorem 3.12 to be fulfilled. We consider a partition π of the closed interval $\mathcal{I}_0$ with the following properties.

$$\begin{aligned}
&\pi : t_0 < t_1 < \cdots < t_N = T, \\
&h_{min} \leq t_\ell - t_{\ell-1} \leq h_{max}, \quad h_{min} > 0, \quad \ell \geq 1, \\
&\kappa_1 \leq \frac{h_{\ell-1}}{h_\ell} \leq \kappa_2, \quad \ell \geq 1,
\end{aligned} \tag{4.2}$$

if κ_1 and κ_2 are suitable constants (cf. [Gri81], [GM86]).

Then, the BDF may be formulated in the following way

$$A(t_\ell)\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}x_{\ell-i} + \mathfrak{g}(x_\ell, t_\ell) = \delta_\ell, \quad \ell \geq k. \tag{4.3}$$

Here, δ_ℓ describes the perturbations in the ℓ-th step for $\ell \geq k$, which is caused by numerical computations including the errors arising from solving the nonlinear equations (e.g. with a Newton-like method). As usually, we denote the stepsize in the ℓ-th step by h_ℓ, i.e., $h_\ell = t_\ell - t_{\ell-1}$. Moreover, we introduce

$$\begin{aligned}
\tilde{x}_\ell &:= x_\ell - x_*(t_\ell), \quad \ell \geq 0, \\
\tau_\ell &:= A(t_\ell)\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}x_*(t_{\ell-i}) + \mathfrak{g}(x_*(t_\ell), t_\ell), \quad \ell \geq k.
\end{aligned}$$

It should be mentioned that the τ_ℓ defined in this way represents the local error of the BDF of order k in the ℓ-th step, and it is of order $O(h_\ell^k)$ if Px_* is sufficiently smooth. This becomes obvious if we regard that $x_*(t_\ell)$ is a solution of the system (4.1), and

$$\begin{aligned}
\tau_\ell &= A(t_\ell)\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}x_*(t_{\ell-i}) + \mathfrak{g}(x_*(t_\ell), t_\ell) \\
&= A(t_\ell)\left(\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}Px_*(t_{\ell-i}) - (Px_*)'(t_\ell)\right)
\end{aligned}$$

is true. At this point, we want to remark that the local error lies in $\operatorname{im} A(t_\ell)$. This will be important later.

Now, equation (4.3) may be written as

$$A(t_\ell)\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{x}_{\ell-i} + \mathfrak{g}(\tilde{x}_\ell + x_*(t_\ell), t_\ell) - \mathfrak{g}(x_*(t_\ell), t_\ell) + \tau_\ell - \delta_\ell = 0,$$

which is equivalent to

$$A_\ell \frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{x}_{\ell-i} + B_\ell \tilde{x}_\ell + \hbar(\tilde{x}_\ell, t_\ell) + \tau_\ell - \delta_\ell = 0, \tag{4.4}$$

where

$$\begin{aligned} A_\ell &:= A(t_\ell) \\ B_\ell &:= \mathfrak{g}'_x(x_*(t_\ell), t_\ell) \\ \hbar(x,t) &:= \mathfrak{g}(x + x_*(t), t) - \mathfrak{g}(x_*(t), t) - \mathfrak{g}'_x(x_*(t), t)x. \end{aligned} \tag{4.5}$$

The function $\hbar(x,t)$ is continuous on $\mathcal{D}_\hbar$, where

$$\mathcal{D}_\hbar := \{(x,t) \in \mathbb{R}^m \times \mathcal{I} \;:\; (x + x_*(t), t) \in D \times \mathcal{I}\}.$$

The quantities introduced above have the following properties:

1. The matrix pencil $\{A_\ell, B_\ell\}$ is index-2 tractable for all ℓ.
2. $\forall\, t \in \mathcal{I}$: $\hbar(0,t) = 0$.
3. The function $\hbar$ is continuously differentiable with respect to x, and $\hbar'_x(0,t) = 0$ is satisfied for all $t \in \mathcal{I}$.

We want to use the splitting technique described in Section 3.5. We define

$$\begin{aligned} U_\ell &:= I - T_\ell = U(t_\ell) \\ G_{1\ell} &:= A_\ell + B_\ell Q = G_1(t_\ell) \\ P_{1\ell} &:= I - Q_{1\ell} = P_1(t_\ell) \\ G_{2\ell} &:= A_{1\ell} + B_\ell P Q_{1\ell} = G_2(t_\ell), \end{aligned}$$

where $T_\ell = T(t_\ell)$ is a projector onto

$$N \cap S_\ell = N \cap S(t_\ell) = \{z \in \mathbb{R}^m \;:\; A_\ell z = 0,\; B_\ell z \in \operatorname{im} A_\ell\},$$

and $Q_{1\ell} = Q_1(t_\ell)$ is the canonical projector onto $\ker G_{1\ell}$ along

$$S_{1\ell} := \{z \in \mathbb{R}^m : B_\ell P z \in \operatorname{im} G_{1\ell}\} = S_1(t_\ell).$$

With Lemma A.1, the relations

$$\begin{aligned} G_{2\ell}^{-1}A_\ell &= P_{1\ell}P, \\ G_{2\ell}^{-1}B_\ell &= G_{2\ell}^{-1}B_\ell PP_{1\ell} + G_{2\ell}^{-1}B_\ell PQ_{1\ell} + G_{2\ell}^{-1}B_\ell Q \\ &= G_{2\ell}^{-1}B_\ell PP_{1\ell} + Q_{1\ell} + Q \end{aligned}$$

are easy to verify. Now, the following lemma provides some further information about the function $\hbar$.

Lemma 4.1 *Let $x_* \in C_N^1$ be a solution of the system (4.1). For each $\epsilon > 0$ there exists a radius $\delta(\epsilon) > 0$ such that the following statements are true for all $z \in \mathbb{R}^m$ with $\|z\| \le \delta(\epsilon)$ and $t \in \mathcal{I}_0$.*

(i) $\|\hbar(z,t)\| \le \epsilon\,\|z\|, \quad \|\hbar_y'(z,t)\| \le \epsilon.$

(ii) The relations

$$\begin{aligned} \|[U(t)Q + PQ_1(t)]G_2^{-1}(t)\hbar(z,t)\| &\le \epsilon\,\|z\|, \\ \|[U(t)Q + PQ_1(t)]G_2^{-1}(t)\hbar_y'(z,t)\| &\le \epsilon \end{aligned}$$

are satisfied.

The correctness of this lemma follows from the continuity properties of the function $\hbar$.

Multiplying the ℓ-th equation of (4.4) by the regular matrix $G_{2\ell}^{-1}$ we obtain the equivalent equation

$$\begin{aligned} P_{1\ell}P\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{x}_{\ell-i} + G_{2\ell}^{-1}B_\ell PP_{1\ell}\tilde{x}_\ell + Q_{1\ell}\tilde{x}_\ell + Q\tilde{x}_\ell \\ + G_{2\ell}^{-1}\hbar(\tilde{x}_\ell, t_\ell) + G_{2\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \qquad (4.6) \end{aligned}$$

for $\ell \ge k$. Multiplying the system by $PP_{1\ell}$, $T_\ell QP_{1\ell}$, $U_\ell QP_{1\ell} + PQ_{1\ell}$, and regarding

$$Q_{1\ell} = Q_{1\ell}G_{2\ell}^{-1}B_\ell P, \quad Q_{1\ell}Q = 0, \quad PP_{1\ell}Q = 0, \quad T_\ell QQ_{1\ell} = QQ_{1\ell},$$

the system (4.6) is equivalent to

$$\begin{aligned} PP_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{x}_{\ell-i} &+ PP_{1\ell}G_{2\ell}^{-1}B_\ell PP_{1\ell}\tilde{x}_\ell \\ &+ PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{x}_\ell,t_\ell) + PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0 \end{aligned} \tag{4.7}$$

$$\begin{aligned} -QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{x}_{\ell-i} &+ T_\ell QP_{1\ell}G_{2\ell}^{-1}B_\ell PP_{1\ell}\tilde{x}_\ell + T_\ell Q\tilde{x}_\ell \\ &+ T_\ell QP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{x}_\ell,t_\ell) + T_\ell QP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0 \end{aligned} \tag{4.8}$$

$$\begin{aligned} U_\ell QG_{2\ell}^{-1}B_\ell PP_{1\ell}\tilde{x}_\ell &+ (U_\ell Q+PQ_{1\ell})\tilde{x}_\ell \\ + (U_\ell Q+PQ_{1\ell})G_{2\ell}^{-1}\hbar(\tilde{x}_\ell,t_\ell) &- (U_\ell Q+PQ_{1\ell})G_{2\ell}^{-1}\delta_\ell = 0. \end{aligned} \tag{4.9}$$

The equivalence is given because

$$I = PP_{1\ell} + T_\ell QP_{1\ell} + (Q_{1\ell}+Q)(U_\ell QP_{1\ell}+PQ_{1\ell})$$

is fulfilled. Fortunately, the influence of the local error vanishes in equation (4.9), since $\tau_\ell \in \operatorname{im} A_\ell$ (as already remarked on page 54) and Lemma 3.8 are valid. Defining

$$\tilde{u}_\ell := PP_{1\ell}\tilde{x}_\ell, \quad \tilde{w}_\ell := T_\ell Q\tilde{x}_\ell, \quad \tilde{y}_\ell := (U_\ell Q+PQ_{1\ell})\tilde{x}_\ell,$$

the system (4.7)–(4.9) is of the form

$$\begin{aligned} \frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{u}_{\ell-i} &+ \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+\tilde{y}_{\ell-j}) \\ + PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell &+ PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\tilde{y}_\ell,t_\ell) + PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0 \end{aligned} \tag{4.10}$$

$$\begin{aligned} -QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{y}_{\ell-i} &- \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}Q(Q_{1\ell}-Q_{1\ell-j})\tilde{u}_{\ell-j} \\ &+ \tilde{w}_\ell + T_\ell QP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell \\ + T_\ell QP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\tilde{y}_\ell,t_\ell) &+ T_\ell QP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0 \end{aligned} \tag{4.11}$$

$$\begin{aligned} \tilde{y}_\ell + U_\ell QG_{2\ell}^{-1}B_\ell\tilde{u}_\ell &+ (U_\ell Q+PQ_{1\ell})G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\tilde{y}_\ell,t_\ell) \\ &- (U_\ell Q+PQ_{1\ell})G_{2\ell}^{-1}\delta_\ell = 0. \end{aligned} \tag{4.12}$$

Equation (4.10) reflects the inherent, discretized, explicit ODE in the u–component. Equation (4.12) is purely algebraic and makes it possible to determine the y–component, which depends on the other components only algebraically. Finally, equation (4.11) represents the differentiation problem of the DAE in discretized form, from which we obtain the w–component.

Now, we want to introduce the following notations

$$\begin{aligned} \hat{\delta}_\ell &:= (U_\ell Q + PQ_{1\ell})G_{2\ell}^{-1}\delta_\ell, \qquad \ell \geq k, && (4.13)\\ \hat{\delta}_\ell &:= (U_\ell Q{+}PQ_{1\ell})G_{2\ell}^{-1}\hbar(\tilde{x}_\ell, t_\ell) + (U_\ell Q{+}PQ_{1\ell})\tilde{x}_\ell, \quad \ell < k, && (4.14)\end{aligned}$$

for making the expressions below somewhat easier. Considering again the system (4.10)-(4.12), we see that $\hat{\delta}_\ell$ represents the defect in the algebraic part for $\ell \geq k$. For the starting values, the corresponding defects are described by (4.14), which will become more transparent when regarding the relation

$$\begin{aligned}(U_\ell Q + PQ_{1\ell})G_{2\ell}^{-1}\hbar(\tilde{x}_\ell, t_\ell) + (U_\ell Q + PQ_{1\ell})\tilde{x}_\ell& \\ = (U_\ell Q + PQ_{1\ell})G_{2\ell}^{-1}(g(x_\ell, t_\ell) - g(x_*(t_\ell), t_\ell)).&\end{aligned}$$

4.2 Feasibility and stability

The following theorem is a generalization of Theorem 3.1 in [Tis95] since the structural condition (3.39) is more general. Further, it contains the implicit Euler case, which was lacking in Theorem 3.7 in [Mär92a]. The results for the implicit Euler method as a special case of the Runge-Kutta method presented in [HLR89] - Theorem 4.1 and 4.2 - are included, because the Hessenberg index-2 systems belong to our class of index-2 DAEs considered in this chapter. Besides the convergence of the BDF method (see also [BCP89] Theorem 3.2.2), we describe the stability behaviour. Furthermore, the theorem provides the feasibility of the BDF method, i.e., the unique solvability of the nonlinear equations arising from the BDF method.

Theorem 4.2 *Let the assumptions of Theorem 3.12 be fulfilled. Supposed there is a constant $C > 0$ such that the starting values satisfy the relation*

$$\|PP_{1\ell}x_\ell - PP_{1\ell}x_*(t_\ell)\| \leq Ch_\ell, \quad \ell < k, \tag{4.15}$$

then the following statements are true:

(i) *There are constants* $\vartheta > 0$ *and* $r > 0$ *such that the BDF with*

$$\|\delta_\ell\| \le \vartheta, \quad \ell \ge k \quad \text{and} \quad \frac{\|\hat{\delta}_\ell\|}{h_\ell} \le \vartheta, \quad \ell \ge 0,$$

is feasible for all partitions (4.2) with sufficiently small stepsize, i.e., the nonlinear equations are solvable with $x_\ell \in B(x_*(t_\ell), r)$.

(ii) *Supposed there is a constant* $C_1 > 0$ *with*

$$\begin{aligned} \|\delta_\ell\| &\le C_1 h_\ell, \quad \ell \ge k, \\ \|\hat{\delta}_\ell\| &\le C_1 h_\ell^2, \quad \ell \ge 0, \end{aligned} \tag{4.16}$$

then we find a constant $C_2 > 0$ *such that the following error estimation holds:*

$$\max_{\ell \ge k} \|x_*(t_\ell) - x_\ell\| \le C_2 \Big[\max_{\ell < k} \|Px_*(t_\ell) - Px_\ell\| + \max_{\ell \ge k} \|\delta_\ell - \tau_\ell\| + \max_{\ell \ge 0} \frac{\|\hat{\delta}_\ell\|}{h_\ell} \Big].$$

Remark 4.3 In general, it is not easy to see which part of the perturbation δ represents the sensitive perturbations $\hat{\delta}$, but, often it is not difficult to determine the image space of $A(t)$. Therefore, the following relation might be useful for controlling the sensitive perturbations $\hat{\delta}$. Separating δ_ℓ into $\delta_{R\ell} + \delta_{N\ell}$ with $\delta_{R\ell} \in \operatorname{im} A_\ell$ and $\delta_{N\ell} \in \operatorname{im} A_\ell{}^\perp$ yields a constant C such that

$$\|\hat{\delta}_\ell\| \le C \, \|\delta_{N\ell}\|$$

is satisfied, since Lemma 3.8 is true.

Proof: Let $\ell \ge k$ be fixed, and let the ℓ-th step of the BDF method be transformed equivalently into the system (see Section 4.2)

$$\begin{aligned} &\frac{1}{h_\ell}\sum_{i=0}^{k} \alpha_{\ell i}\tilde{u}_{\ell-i} + \frac{1}{h_\ell}\sum_{j=1}^{k} \alpha_{\ell j} P(P_{1\ell} - P_{1\ell-j})(\tilde{u}_{\ell-j} + \tilde{y}_{\ell-j}) \\ &+ PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell + PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell + \tilde{w}_\ell + \tilde{y}_\ell, t_\ell) + PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \end{aligned} \tag{4.17}$$

$$\begin{aligned} - QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{y}_{\ell-i} - \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j} Q(Q_{1\ell} - Q_{1\ell-j})\tilde{u}_{\ell-j} \\ + \tilde{w}_\ell + T_\ell Q P_{1\ell} G_{2\ell}^{-1} B_\ell \tilde{u}_\ell \end{aligned}$$

$$+ T_\ell Q P_{1\ell} G_{2\ell}^{-1} \hbar(\tilde{u}_\ell + \tilde{w}_\ell + \tilde{y}_\ell, t_\ell) + T_\ell Q P_{1\ell} G_{2\ell}^{-1}(\tau_\ell - \delta_\ell) = 0 \tag{4.18}$$

$$\tilde{y}_\ell + U_\ell Q G_{2\ell}^{-1} B_\ell \tilde{u}_\ell + (U_\ell Q + P Q_{1\ell}) G_{2\ell}^{-1} \hbar(\tilde{u}_\ell + \tilde{w}_\ell + \tilde{y}_\ell, t_\ell) - \hat{\delta}_\ell = 0, \tag{4.19}$$

where $\hat{\delta}_\ell$ is defined by (4.13) and (4.14). Following the pattern of equation (4.19), we now define the function

$$\begin{aligned}\bar{\bar{F}}(y, u, w, \hat{\delta}, t) &:= y - \hat{\delta} + U(t) Q G_2^{-1}(t) B(t) u \\ &\quad + (U(t)Q + PQ_1(t)) G_2^{-1}(t) \hbar(u + T(t)Qw + y, t).\end{aligned} \tag{4.20}$$

Then $\bar{\bar{F}}$ is is of class C^2. Furthermore,

$$\bar{\bar{F}}(0,0,0,0,t) = 0, \ \bar{\bar{F}}'_v(0,0,0,0,t) = I, \quad t \in \mathcal{I},$$

is fulfilled. Then, the following fact is true.

There exist a radius α and a unique C^2-function

$$\bar{\bar{f}}(u, w, \hat{\delta}, t) \ : \ B(0, \alpha) \times \mathcal{I}_0 \ \rightarrow \ B(0, \rho)$$

with the properties

(i) $\bar{\bar{F}}(\bar{\bar{f}}(u, w, \hat{\delta}, t), u, w, \hat{\delta}, t) = 0$

(ii) $\bar{\bar{f}}(0,0,0,t) = 0, \ \bar{\bar{f}}'_u(0,0,0,t) = 0, \ \bar{\bar{f}}'_w(0,0,0,t) = 0, \ \bar{\bar{f}}'_{\hat{\delta}}(0,0,0,t) = I$

(iii)
$$\|\bar{\bar{f}}(u, w, \hat{\delta}, t)\| \le \|u\| + \|w\| + 3\|\hat{\delta}\| \tag{4.21}$$

(iv) If $\hat{\delta} \in \operatorname{im}[U(t)Q + PQ_1(t)]$ is fulfilled, then

$$\bar{\bar{f}}'(u, w, \hat{\delta}, t) = [U(t)Q + PQ_1(t)] \bar{\bar{f}}(u, w, \hat{\delta}, t). \tag{4.22}$$

is satisfied.

(v)
$$Q_1(t) \bar{\bar{f}}'_w(u, w, \hat{\delta}, t) = 0. \tag{4.23}$$

Lemma A.2 and Lemma A.3 (see Appendix) imply the correctness of the assertions (i)-(iii). Multiplying equation (i) by $[U(t)Q + PQ_1(t)]$, we obtain the relation (iv). The assertion (v) follows from the structural condition (3.39) in the same way as on page 47.

Now, we have to solve the following system

$$\begin{aligned}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{u}_{\ell-i} + PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell \\ + \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+P\bar{\bar{f}}(\tilde{u}_{\ell-j},\tilde{w}_{\ell-j},\hat{\delta}_{\ell-j},t_{\ell-j})) \\ + PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell) + PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0\end{aligned} \tag{4.24}$$

$$\begin{aligned}-QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}(\tilde{u}_{\ell-i},\tilde{w}_{\ell-i},\hat{\delta}_{\ell-i},t_{\ell-i}) + T_\ell QP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell \\ -\frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}Q(Q_{1\ell}-Q_{1\ell-j})\tilde{u}_{\ell-j}+\tilde{w}_\ell \\ + T_\ell QP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell) + T_\ell QP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell) = 0\end{aligned} \tag{4.25}$$

$$\tilde{y}_\ell - \bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell) = 0. \tag{4.26}$$

Regarding Lemma 3.6 and (4.23),

$$Q_1(t)\bar{\bar{f}}'_w(u,w,\hat{\delta},t) = 0$$

is true for $u, w, \hat{\delta} \in B_\alpha(0)$. Using (4.22), we obtain

$$\begin{aligned}P\bar{\bar{f}}(\tilde{u}_{\ell-i},\tilde{w}_{\ell-i},\hat{\delta}_{\ell-i},t_{\ell-i}) &= PQ_{1\ell-i}\bar{\bar{f}}(\tilde{u}_{\ell-i},\tilde{w}_{\ell-i},\hat{\delta}_{\ell-i},t_{\ell-i}) \\ &= PQ_{1\ell-i}\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i}) \\ &\quad + P\int_0^1 Q_{1\ell-i}\bar{\bar{f}}'_w(\tilde{u}_{\ell-i},s\tilde{w}_{\ell-i},\hat{\delta}_{\ell-i},t_{\ell-i})\,ds\,\tilde{w}_{\ell-i} \\ &= P\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i})\end{aligned} \tag{4.27}$$

for $i = 0, ..., k$ and $\tilde{u}_{\ell-i}, \tilde{w}_{\ell-i}, \hat{\delta}_{\ell-i} \in B_\alpha(0)$ (note that $\hat{\delta}_{\ell-i} \in \operatorname{im}[U(t_{\ell-i})Q + PQ_1(t_{\ell-i})]$). Now, considering the relations

$$Q_{1\ell} = Q_{1\ell}P, \qquad PP_{1\ell-i} = PP_{1\ell-i}P \quad \text{for} \quad i = 0, ..., k,$$

the system (4.24)-(4.26) is equivalent to

$$\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{u}_{\ell-i} + PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell$$

$$\begin{aligned}
&+\frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+P\bar{\bar{f}}(\tilde{u}_{\ell-j},0,\hat{\delta}_{\ell-j},t_{\ell-j}))\\
&+PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell)+PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell)=0
\end{aligned}\tag{4.28}$$

$$\begin{aligned}
&-QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i})+T_\ell QP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell\\
&\qquad-\frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}Q(Q_{1\ell}-Q_{1\ell-j})\tilde{u}_{\ell-j}+\tilde{w}_\ell\\
&+T_\ell QP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell)+T_\ell QP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell)=0
\end{aligned}\tag{4.29}$$

$$\tilde{y}_\ell-\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell)=0.\tag{4.30}$$

In order to be able to solve equation (4.29) with respect to $\tilde{w}_\ell$, it is obviously necessary to investigate the term

$$QQ_{1\ell}\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i})$$

with respect to its relationship to the stepsize h_ℓ in more detail. Therefore, we write this term in the following equivalent form:

$$\begin{aligned}
&QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i})\\
&\quad=QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}[\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i})-\bar{\bar{f}}(0,0,0,t_{\ell-i})]\\
&\quad=QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\left[\int_0^1\bar{\bar{f}}'_u(s\tilde{u}_{\ell-i},0,0,t_{\ell-i})\,ds\,\tilde{u}_{\ell-i}\right.\\
&\qquad\qquad\qquad\left.+\int_0^1\bar{\bar{f}}'_{\hat{\delta}}(\tilde{u}_{\ell-i},0,s\hat{\delta}_{\ell-i},t_{\ell-i})\,ds\,\hat{\delta}_{\ell-i}\right]\\
&\quad=QQ_{1\ell}\int_0^1\bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)\,ds\,\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{u}_{\ell-i}\\
&\quad-QQ_{1\ell}\int_0^1\sum_{j=1}^{k}\alpha_{\ell j}\frac{\bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)-\bar{\bar{f}}'_u(s\tilde{u}_{\ell-j},0,0,t_{\ell-j})}{h_\ell}\,ds\,\tilde{u}_{\ell-j}\\
&\quad+QQ_{1\ell}\int_0^1\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}'_{\hat{\delta}}(\tilde{u}_{\ell-i},0,s\hat{\delta}_{\ell-i},t_{\ell-i})\,ds\,\frac{\hat{\delta}_{\ell-i}}{h_\ell}.
\end{aligned}$$

Due to (4.28), the term $\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\tilde{u}_{\ell-i}$ may be replaced by

$$\begin{aligned} &-PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell - \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+\bar{\bar{f}}(\tilde{u}_{\ell-j},0,\hat{\delta}_{\ell-j},t_{\ell-j})) \\ &\qquad - PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell) - PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell). \end{aligned}$$

Then, we obtain the following, long description for the most sensitive term of the index-2 DAE

$$\begin{aligned} &QQ_{1\ell}\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}(\tilde{u}_{\ell-i},0,\hat{\delta}_{\ell-i},t_{\ell-i}) = \\ &\qquad QQ_{1\ell}\int_0^1 \bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)\,ds\,\Big(-PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell \\ &\qquad - \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+\bar{\bar{f}}(\tilde{u}_{\ell-j},0,\hat{\delta}_{\ell-j},t_{\ell-j})) \\ &\qquad - PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+\tilde{w}_\ell+\bar{\bar{f}}(\tilde{u}_\ell,\tilde{w}_\ell,\hat{\delta}_\ell,t_\ell),t_\ell) - PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell)\Big) \\ &\qquad - QQ_{1\ell}\int_0^1\sum_{j=1}^{k}\alpha_{\ell j}\frac{\bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)-\bar{\bar{f}}'_u(s\tilde{u}_{\ell-j},0,0,t_{\ell-j})}{h_\ell}\,ds\,\tilde{u}_{\ell-j} \\ &\qquad\qquad + QQ_{1\ell}\int_0^1\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}'_{\hat{\delta}}(\tilde{u}_{\ell-i},0,s\hat{\delta}_{\ell-i},t_{\ell-i})\,ds\,\frac{\hat{\delta}_{\ell-i}}{h_\ell}. \end{aligned}$$

Now, we may solve (4.29) with respect to $\tilde{w}_\ell$. We define a function

$$\begin{aligned} &\bar{\bar{G}}_\ell(w,\tilde{u}_{\ell-i},\hat{\delta}_{\ell-i},\tau_\ell-\delta_\ell) := \\ &\qquad - QQ_{1\ell}\int_0^1 \bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)\,ds\,\Big(-PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell \\ &\qquad - \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}P(P_{1\ell}-P_{1\ell-j})(\tilde{u}_{\ell-j}+\bar{\bar{f}}(\tilde{u}_{\ell-j},0,\hat{\delta}_{\ell-j},t_{\ell-j})) \\ &\qquad - PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell+w+\bar{\bar{f}}(\tilde{u}_\ell,w,\hat{\delta}_\ell,t_\ell),t_\ell) - PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell-\delta_\ell)\Big) \\ &\qquad + QQ_{1\ell}\int_0^1\sum_{j=1}^{k}\alpha_{\ell j}\frac{\bar{\bar{f}}'_u(s\tilde{u}_\ell,0,0,t_\ell)-\bar{\bar{f}}'_u(s\tilde{u}_{\ell-j},0,0,t_{\ell-j})}{h_\ell}\,ds\,\tilde{u}_{\ell-j} \\ &\qquad - QQ_{1\ell}\int_0^1\sum_{i=0}^{k}\alpha_{\ell i}\bar{\bar{f}}'_{\hat{\delta}}(\tilde{u}_{\ell-i},0,s\hat{\delta}_{\ell-i},t_{\ell-i})\,ds\,\frac{\hat{\delta}_{\ell-i}}{h_\ell} \\ &\qquad - \frac{1}{h_\ell}\sum_{j=1}^{k}\alpha_{\ell j}Q(Q_{1\ell}-Q_{1\ell-j})\tilde{u}_{\ell-j} + w \end{aligned}$$

$$+ T_\ell Q P_{1\ell} G_{2\ell}^{-1} \hbar(\tilde{u}_\ell + w + \bar{\bar{f}}(\tilde{u}_\ell, w, \hat{\delta}_\ell, t_\ell), t_\ell) + T_\ell Q P_{1\ell} G_{2\ell}^{-1}(\tau_\ell - \delta_\ell).$$

Writing $\bar{\bar{G}}_\ell(w, \tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell)$ we mean

$$\bar{\bar{G}}_\ell(w, \tilde{u}_\ell, ..., \tilde{u}_{\ell-k}, \hat{\delta}_\ell, ..., \hat{\delta}_{\ell-k}, \tau_\ell - \delta_\ell).$$

Further, the index i goes from 0 to k and the index j goes from 1 to k in the following, provided that no other information is given.

Then, the function $\bar{\bar{G}}_\ell$ is continuously differentiable,

$$\bar{\bar{G}}_\ell(0) = 0, \quad \bar{\bar{G}}_{\ell_w}'(0) = I,$$

and we may conclude:

There is a radius σ (independent of the partition) and a unique C^1-function

$$\bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell) \ : \ B(0,\sigma) \ \rightarrow \ \mathbb{R}^m$$

with the following properties.

(i) $\bar{\bar{G}}_\ell(\bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell), \tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell) = 0, \quad \bar{\bar{g}}_\ell(0) = 0$

(ii)
$$\begin{aligned} \bar{\bar{g}}_\ell'(0) =& (-QP_{1\ell}G_{2\ell}^{-1}B_\ell, \frac{\alpha_{\ell 1}}{h_\ell} Q(Q_{1\ell} - Q_{1\ell-1}), ..., \\ & \frac{\alpha_{\ell k}}{h_\ell} Q(Q_{1\ell} - Q_{1\ell-k}), \frac{\alpha_{\ell 0}}{h_\ell} QQ_{1\ell}, ..., \frac{\alpha_{\ell k}}{h_\ell} QQ_{1\ell}, -T_\ell QP_{1\ell}G_{2\ell}^{-1}) \end{aligned}$$

(iii)
$$\begin{aligned} \|\bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell)\| \ \leq& \ \{1 + 2\|QP_{1\ell}G_{2\ell}^{-1}B_\ell\|\}\|\tilde{u}_\ell\| \\ & + \sum_{j=1}^{k}\{1 + 2\frac{|\alpha_{\ell j}|}{h_\ell}\|Q(Q_{1\ell} - Q_{1\ell-j})\|\}\|\tilde{u}_{\ell-j}\| \\ + \sum_{i=0}^{k}\{1 + 2\frac{|\alpha_{\ell i}|}{h_\ell}\|QQ_{1\ell}\|\}\|\hat{\delta}_{\ell-i}\| & + \{1 + 2\|T_\ell QP_{1\ell}G_{2\ell}^{-1}\|\}\|\tau_\ell - \delta_\ell\|. \end{aligned} \tag{4.32}$$

For the proof of this assertion use Lemma A.2 and Lemma A.3, or modify the proof of Lemma 3.3 in [Tis95].

It remains to solve equation (4.28) with respect to $\tilde{u}_\ell$, which has the following form

$$\begin{aligned} & \sum_{i=0}^{k} \alpha_{\ell i}\tilde{u}_{\ell-i} + h_\ell PP_{1\ell}G_{2\ell}^{-1}B_\ell\tilde{u}_\ell + h_\ell PP_{1\ell}G_{2\ell}^{-1}(\tau_\ell - \delta_\ell) \\ & + \sum_{j=1}^{k} \alpha_{\ell j}P(P_{1\ell} - P_{1\ell-j})(\tilde{u}_{\ell-j} + \bar{\bar{f}}(\tilde{u}_{\ell-j}, 0, \hat{\delta}_{\ell-j}, t_{\ell-j})) \\ & \qquad + h_\ell PP_{1\ell}G_{2\ell}^{-1}\hbar(\tilde{u}_\ell + \bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell) \\ & \qquad\quad + \bar{\bar{f}}(\tilde{u}_\ell, \bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell), \hat{\delta}_\ell, t_\ell), t_\ell) = 0 \end{aligned} \tag{4.33}$$

We define

$$
\begin{aligned}
&\bar{\bar{R}}_\ell(u, \tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell) := \\
&\qquad \alpha_{\ell 0} u + \sum_{j=1}^{k} \alpha_{\ell j} \tilde{u}_{\ell-j} + h_\ell P P_{1\ell} G_{2\ell}^{-1} B_\ell u \\
&\qquad + h_\ell P P_{1\ell} G_{2\ell}^{-1} \hbar(u + \bar{\bar{f}}(u, \bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell), \hat{\delta}_\ell, t_\ell) \\
&\qquad\qquad\qquad + \bar{\bar{g}}_\ell(\tilde{u}_{\ell-i}, \hat{\delta}_{\ell-i}, \tau_\ell - \delta_\ell), t_\ell) \\
&\qquad + \sum_{j=1}^{k} \alpha_{\ell j} P(P_{1\ell} - P_{1\ell-j})(\tilde{u}_{\ell-j} + \bar{\bar{f}}(\tilde{u}_{\ell-j}, 0, \hat{\delta}_{\ell-j}, t_{\ell-j})) \\
&\qquad + h_\ell P P_{1\ell} G_{2\ell}^{-1}(\tau_\ell - \delta_\ell),
\end{aligned}
$$

for $i = 0, ..., k$ and $j = 1, ..., k$. Then, R_ℓ is continuously differentiable,

$$\bar{\bar{R}}_\ell(0) = 0, \ \bar{\bar{R}}'_{\ell_u}(0) = \alpha_{\ell 0} I + h_\ell P P_{1\ell} G_{2\ell}^{-1} B_\ell.$$

Taking into account the properties of $\hbar$, $\bar{\bar{f}}$, and $\bar{\bar{g}}_\ell$, we obtain by the same arguments as above:
There is a radius χ (independent of the partition) and a unique C^1-function

$$\bar{\bar{r}}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell) \ : \ B((0), \chi) \ \to \ B(0, \chi')$$

for sufficiently small h_ℓ with the properties

$$\bar{\bar{R}}_\ell(\bar{\bar{r}}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell), \tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell) = 0, \quad \bar{\bar{r}}_\ell(0) = 0.$$

The function determined in this way

$$
\begin{aligned}
&\tilde{x}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell) := \\
&\qquad \bar{\bar{r}}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell) + \bar{\bar{f}}(\bar{\bar{r}}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell), \hat{\delta}_\ell, t_\ell) \\
&\qquad\qquad + \bar{\bar{g}}_\ell(\bar{\bar{r}}_\ell(\tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell), \tilde{u}_{\ell-j}, \hat{\delta}_{\ell-i}, \delta_\ell - \tau_\ell)
\end{aligned}
$$

is a solution of the system (4.17)–(4.19), i.e., the BDF method provides a solution x_ℓ at the time point t_ℓ.

Considering again equation (4.33), using the estimations (4.21) and (4.32), and taking standard arguments for explicit ODEs, we find a constant $c' > 0$ for suitable constants κ_1 and κ_2 (see (4.2)) such that

$$\tilde{u}_\ell \leq c' \, h_\ell \quad \text{for } \ell \geq k$$

is true, provided that the assumption (4.15) is satisfied. Finally, the assertions of the theorem follow immediately.

□

Remarks 4.4

(1) For the problems given in Theorem 4.2, a week instability is present in all components of the solution. In the proof we have seen that the inherent differential ODE is influenced by the defects in the algebraic part of the DAE. However, from the theory for linear index-2 problems (see e.g. [Mär92a]) we know that instability occurs only in the algebraic components (Qx). Unfortunately, this is not correct for all DAEs. There are even DAEs in Hessenberg form (cf. Examples 1 and 2 in Section 3.2 of [Tis95]) for which the instability occurs also in the differential components (Px).

(2) For a successful integration of quasilinear index-2 DAEs with the variable order variable stepsize BDF method the following aspects should be taken into consideration.

- If the nullspace of the leading coefficient matrix is not constant, the BDF need not converge, even need not work at all (see [GP84]).
- In general, all components of the solution will be influenced by a week instability arising from defects in the derivative-free part of the DAE. That's why, for obtaining a solution, it is necessary to ensure that these defects remain smaller than the stepsize used. Additionally, for the BDF of an order greater than 1 the components of the starting values in the involved inherent ODE must be sufficiently exact in relation to the stepsize.
- If the algebraic components go into the DAE only in a linear way, the integration will work better when the stepsize and order control is based on the differential components only (see [Tis95]). These components are not affected by the week instability. In order to improve the other variables a smaller stepsize won't be helpful. Therefore, the defects in the algebraic part of the DAE have to be made smaller.

Remark 4.5 Since general index-2 differential algebraic equations

$$f(\dot{x}, x, t) = 0 \tag{4.34}$$

with constant nullspace $\ker f_{\dot{x}}$ can be equivalently transformed into a quasilinear index-2 DAE

$$P\dot{x} - y = 0 \tag{4.35}$$

$$f(y, x, t) = 0, \tag{4.36}$$

the results for quasilinear systems (4.1) can be generalized to systems of the form (4.34). For the class of quasilinear DAEs satisfying the structural condition (2.2) in [Tis95], this was already done in [Fre95].

Chapter 5

Simulation of electrical circuits

In this chapter, we investigate the problem of the index of the systems arising from modified nodal analysis. This problem was already studied in [FG94] for some examples. We present some results for all models whose capacitances are described by a one-port. In this case, each capacitance of the network has two uniquely determined nodals (including the node with the zero potential), enclosing this capacitance. That means, for each capacitance of the network the voltage through this capacitance may be expressed by the difference of the nodal potentials of these two uniquely determined nodals. Note that a wide class of electric networks can be modelled in such a way since general capacitances can often be modelled by controlled current sources in such a way that

$$I = \dot{q} \quad \text{with} \quad q = f(u).$$

For an example, see the nonlinear current source in the MOSFET model on page 88.

5.1 Structure of the circuit systems

From Chapter 2 we know that the classical modified nodal analysis leads to systems of the form

$$D(x)\dot{x} + f(x) = r(t). \tag{5.1}$$

The charge-oriented modified nodal analysis implies systems of the form

$$\begin{aligned} A\dot{q} + f(x) &= r(t) && (5.2)\\ q &= g(x). && (5.3) \end{aligned}$$

Recall the relation $D(x) = Ag'(x)$ for the systems above.

Let us denote the number of nodals by $n_u \geq 0$ and the number of inductances in the network by $n_L \geq 0$. Further, we sort the vector x and the vector q in such a way that

$$x = \begin{pmatrix} u \\ I_L \\ I_s \end{pmatrix}, \quad q = \begin{pmatrix} Q \\ \Phi \end{pmatrix},$$

where u represents the vector of nodal voltages, I_L the vector of inductances, I_s the vector of voltage-controlled sources, Q the vector of charges, and Φ the vector of fluxes.

Let all capacitances of the network have a certain direction. Then, each capacitance has a uniquely determined "left" node and a uniquely determined "right" node (if we regard also the node with the zero potential). The voltage through the capacitance may be expressed by the difference of the nodal potential of the right and of the left node.

We numerate the capacitances and the inductances in the network in such a way that Q_j represents the charge of the capacitance C_j for $j = 1, ..., n_C$ and Φ_j represents the flux of the inductance L_j for $j = 1, ..., n_L$. Now, the entries a_{ij} of the rectangular matrix A in equation (5.2) satisfy the following relations.

- If the i-th node is the right node of the capacitance C_j, then a_{ij} is equal to 1.
- If the i-th node is the left node of the capacitance C_j, then a_{ij} is equal to -1.
- If the current through the inductance L_i is denoted by $x_j = I_{j-n_u}$, then a_{ij} is equal to 1.
- Otherwise, a_{ij} is equal to 0.

For more clarity, we look again at the double way rectifier on page 23. In that case, the incidence matrix A has the following structure:

$$A = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where

$$Ag'(x) = \begin{pmatrix} C_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & C_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & L_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & L_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The matrix A describes the topology of the dynamical network elements. Are there any relations to the entries of $g'(x)$ (see equation (2.4))? We are able to answer this question if we analyze the expressions for the capacitances and inductances in the network in more detail.

For each capacitance in the network, there is a positive, differentiable function ψ_j such that

$$Q_j = \psi_j(v_j)$$

is satisfied for the voltage v_j of the capacitance C_j. Therefore,

$$\dot{Q}_j = \psi_j'(v_j)\dot{v}_j$$

is valid. Furthermore, the voltage v_j is a linear function of u. More precisely, it is equivalent to $u_k - u_l$ (u_k denotes the nodal potential of the right node, u_l denotes the nodal potential of the left node of the capacitance C_j). Hence, we may write

$$\dot{Q} = \begin{pmatrix} \psi_1'(v_1) & 0 & \dots & 0 \\ 0 & \psi_2'(v_2) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_{n_C}'(v_{n_C}) \end{pmatrix} Y\dot{u}.$$

The entries y_{ij} of the matrix Y satisfy the following relations.

- If the j-th node is the right node of the capacitance C_i and the nodal potential of the j-th node is denoted by u_j, then y_{ij} is equal to 1.

- If the j-th node is the left node of the capacitance C_i and the nodal potential of the j-th node is denoted by u_j, then y_{ij} is equal to -1.

- Otherwise, y_{ij} is equal to 0.

In the case of the double way rectifier, the matrix Y has the form

$$Y = \begin{pmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

For each inductance, there is a positive, differentiable function φ_j such that

$$\Phi_j = \varphi_j(I_i)$$

is satisfied for the current I_i through the inductance L_j. Now,

$$\dot{\Phi}_j = \varphi_j'(I_i)\dot{I}_i$$

is fulfilled. Hence, we may write

$$\dot{\Phi} = \begin{pmatrix} \varphi_1'(x) & 0 & \dots & 0 \\ 0 & \varphi_2'(x) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varphi_{n_L}'(x) \end{pmatrix} Z\dot{I}$$

if the entries z_{ij} of Z satisfy the following relations.

- If the current through the inductance L_i is denoted by $x_j = I_{j-n_u}$, then z_{ij} is equal to 1.
- Otherwise, z_{ij} is equal to 0.

For the double way rectifier, we have

$$Z = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

These facts lead to the following property of the function $g(x)$ (see the equations (5.3) and (2.4)).

$$g'(x) = R(x)A^T, \tag{5.4}$$

where

$$R(x) = \begin{pmatrix} \psi_1'(x) & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & \psi_2'(x) & \dots & 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & \dots & \psi_{n_C}'(x) & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \varphi_1'(x) & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 0 & \varphi_2'(x) & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & \varphi_{n_L}'(x) \end{pmatrix}$$

is symmetric and positive-definite.

Remarks 5.1

(1) If the network is modelled without capacitances, the matrix $R(x)$ reads

$$R(x) = \begin{pmatrix} \varphi_1'(x) & 0 & \dots & 0 \\ 0 & \varphi_2'(x) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \varphi_{n_L}'(x) \end{pmatrix}.$$

(2) If the network is modelled without inductances, the matrix $R(x)$ reads

$$R(x) = \begin{pmatrix} \psi_1'(x) & 0 & \dots & 0 \\ 0 & \psi_2'(x) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \psi_{n_C}'(x) \end{pmatrix}.$$

(3) More precisely, the matrix A may be written as

$$A \;=\; \begin{array}{cc} \left(\begin{array}{cc} M & 0 \\ 0 & I \\ 0 & 0 \end{array}\right) & \begin{array}{c} n_u \\ n_L \ , \\ n_s \end{array} \\ \begin{array}{cc} n_C & n_L \end{array} & \end{array} \tag{5.5}$$

where M is a matrix with entries -1, 0, 1 only. It describes the occurrence of the capacitors in the network. I represents the identity matrix. The dimension n_s represents the number of voltage-controlled sources of the circuit. Note that if the circuit does not contain all kinds of elements, some dimensions (e.g. n_L) may be zero. Then, obviously, some rows or columns of the description (5.5) will disappear.

Now, the following lemma is true.

Lemma 5.2 *For the systems (5.1) and (5.2)-(5.3), the relations*

$$im\, A = im\, D(x) \quad and \quad ker\, D(x) = ker\, g'(x),$$

are satisfied.

Proof: Since

$$g'(x) = R(x)A^T$$

is valid for a symmetric, positive-definite matrix $R(x)$, we find a symmetric regular matrix $R_s(x)$ such that $R(x) = R_s(x)R_s(x)$ is satisfied. Hence,

$$\operatorname{rank} AR(x)A^T = \operatorname{rank} AR_s(x)(AR_s(x))^T = \operatorname{rank} AR_s(x) = \operatorname{rank} A.$$

This implies

$$\operatorname{im} D(x) = \operatorname{im} AR(x)A^T = \operatorname{im} A.$$

Secondly,

$$\begin{aligned}\operatorname{rank} AR(x)A^T &= \operatorname{rank} AR_s(x)(AR_s(x))^T \\ &= \operatorname{rank} (AR_s(x))^T = \operatorname{rank} R(x)A^T.\end{aligned}$$

Now,

$$\ker D(x) = \ker AR(x)A^T = \ker g'(x)$$

is valid.

□

Corollary 5.3 *The relation $ker D(x) = ker A^T$ is valid.*

Proof: It holds that $\ker D(x) = \ker g'(x) = \ker R(x)A^T$. Since, $R(x)$ is regular, we may conclude $\ker D(x) = \ker A^T$.

□

5.2 Solution spaces adapted to circuit systems

Due to Corollary 5.3, equation (5.1) can be rewritten more precisely as

$$D(x)P\dot{x} + f(x) - r(t) = 0, \tag{5.6}$$

where P denotes any constant projector matrix projecting along the constant nullspace $N := \ker D(x) = \ker A^T$. This reformulation (5.6) provides information on what kind of functions we should accept to be solutions of the DAE (5.1), in fact. Namely, such a solution has to be a continuous function with a continuously differentiable P-component. However, the other component should not be expected to belong to C^1 in general.

Analogously, by introducing the projector P_A along the nullspace of A, the system (5.2)-(5.3) reads

$$\begin{aligned}AP_A\dot{q} + f(x) &= r(t), \\ q - g(x) &= 0.\end{aligned}$$

Thus, the function spaces

$$\begin{aligned} C^1_N &:= \{x \in C(\mathcal{I}, \mathbb{R}^m) : Px \in C^1(\mathcal{I}, \mathbb{R}^m)\}, \\ C^1_{\tilde{N}} &:= \{\tilde{x} = (q, x) \in C(\mathcal{I}, \mathbb{R}^{n+m}) : P_A q \in C^1(\mathcal{I}, \mathbb{R}^n)\} \end{aligned}$$

with $m = n_u + n_L + n_s$ and $n = n_C + n_L$ result to be natural ones, where the solutions of (5.1) resp. (5.2)-(5.3) should belong to.

Theorem 5.4 *$x \in C^1_N$ solves (5.1) and $q = g(x)$ if and only if $(q, x) \in C^1_{\tilde{N}}$ is a solution of (5.2)-(5.3).*

Proof:
($\rightarrow$) Denoting $Q := I - P$, the projector Q projects onto $\ker A^T$. Equation (5.4) leads to the relation

$$g(x) - g(Px) = \int_0^1 R(Px + sQx)A^T Q \, ds = 0,$$

i.e., $g(x) = g(Px)$, $x \in \mathcal{D}$. Hence, if $x \in C^1_N$ is a solution of (5.1) and $q = g(x)$ is satisfied, then

$$q = g(Px)$$

is true, i.e., the function q is continuously differentiable. Thus, $P_A q$ is also continuously differentiable, i.e., the pair (q, x) belongs to $C^1_{\tilde{N}}$. Trivially, (q, x) solves the equation system (5.2)-(5.3).

($\leftarrow$) We define an auxiliary function

$$F_h(y, z, P_A q) := \begin{pmatrix} P_A q - P_A g(Py) + Q_A z \\ Qy + Az \end{pmatrix}.$$

This function is continuously differentiable, and the relations

$$\begin{aligned} (F_h)'_{(y,z)}(y, z, P_A q) &= \begin{pmatrix} P_A R(Py) A^T & Q_A \\ Q & A \end{pmatrix} \\ (F_h)'_{P_A q}(y, z, P_A q) &= \begin{pmatrix} I \\ 0 \end{pmatrix} \end{aligned}$$

are satisfied for $Py \in \mathcal{D}$ and $q \in \mathbb{R}^m$. The matrix $(F_h)'_{(y,z)}(y, z, P_A q)$ is regular, and the implicit-function theorem provides a continuously differentiable function $f = (f_{h1}, f_{h2})^T$ satisfying

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} f_{h1}(P_A q) \\ f_{h2}(P_A q) \end{pmatrix}, \quad F_h(f_{h1}(P_A q), f_{h2}(P_A q), q) = 0.$$

Since $F_h(Px, 0, P_A q) = 0$ and $P_A q \in C^1$ are satisfied for the solution (q, x) of the system (5.2)-(5.3), the relation $Px \in C^1$ is true, i.e., $x \in C^1_N$ is fulfilled.

□

Corollary 5.5 *If $(x, q) \in C^1_{\tilde{N}}$ solves (5.2)-(5.3), then we have $Px \in C^1$.*

5.3 Index of the circuit systems

One of the important questions is the question of the index of the equivalent systems (5.1) and (5.2)-(5.3). We want to investigate the tractability index of both systems. Regarding Section 1.2.4, we have to study the relevant spaces N, S, N_1, and S_1 of (5.1) and (5.2)-(5.3), respectively. In order to be able to distinguish the relevant spaces for both systems, we will mark the corresponding spaces of the system (5.2)-(5.3) by a tilde. As in Section 1.2.4 we introduce

$$N := \ker Ag'(x), \quad \tilde{N} := \{ \left(\begin{smallmatrix} \gamma \\ \tilde{z} \end{smallmatrix}\right) : A\gamma = 0\}$$

and

$$\begin{aligned} S(x) &:= \{z : f'(x)z \in \operatorname{im} A\} \\ \tilde{S}(x) &:= \{ \left(\begin{smallmatrix} \gamma \\ \tilde{z} \end{smallmatrix}\right) : \ f'(x)\tilde{z} \in \operatorname{im} A, \ \gamma = g'(x)\tilde{z}\} \end{aligned}$$

as well as

$$\begin{aligned} N_1(x) &:= \ker (Ag'(x) + f'(x)Q) \\ \tilde{N}_1(x) &:= \{ \left(\begin{smallmatrix} \gamma \\ \tilde{z} \end{smallmatrix}\right) : \ A\gamma + f'(x)\tilde{z} = 0, \ Q_A\gamma = g'(x)\tilde{z}\} \end{aligned}$$

and

$$\begin{aligned} S_1(x) &:= \{z : f'(x)Pz \in \operatorname{im} (Ag'(x) + f'(x)Q)\} \\ \tilde{S}_1(x) &:= \{ \left(\begin{smallmatrix} \gamma \\ \tilde{z} \end{smallmatrix}\right) : \ \exists \ \alpha, \beta : \ 0 = A\alpha + f'(x)\beta, \ P_A\gamma = Q_A\alpha - g'(x)\beta\}. \end{aligned}$$

Note that if g is twice differentiable, the spaces introduced correspond to the spaces introduced in Section 1.2.4 since

$$Ag''(x)yQ = \lim_{\tau \to 0} \frac{Ag'(x + \tau y)Q - Ag'(x)Q}{\tau} = 0$$

is true for all $(y, x) \in \mathbb{R}^m \times \mathcal{D}$.

The special structure of the circuits implies the hoped-for results:

Theorem 5.6 *System (5.1) is index-1 tractable if and only if system (5.2)-(5.3) is so.*

Proof:
($\rightarrow$) For any $\binom{\gamma}{\tilde{z}} \in \tilde{N} \cap \tilde{S}(x)$, we get

$$A\gamma = 0, \quad \gamma = g'(x)\tilde{z}, \quad f'(x)\tilde{z} \in \operatorname{im} A.$$

Thus, $Ag'(x)\tilde{z} = 0$ is satisfied. Using the condition

$$N \cap S(x) = \{0\},$$

we obtain $\tilde{z} = 0$. The relation $\gamma = g'(x)\tilde{z}$ implies $\gamma = 0$, i.e.,

$$\tilde{N} \cap \ \tilde{S}(x) = \{0\}.$$

($\leftarrow$) For any $z \in N \cap S(x)$ we compute $\gamma := g'(x)z$ and $\tilde{z} := z$. Then, the relations

$$A\gamma = 0, \quad \gamma = g'(x)\tilde{z}, \quad f'(x)\tilde{z} \in \operatorname{im} A$$

are satisfied, i.e., $\binom{\gamma}{\tilde{z}} \in \tilde{N} \cap \tilde{S}(x)$. Therefore, $\tilde{z} = 0$ is valid, i.e., $z = 0$.

□

Theorem 5.7 *System (5.1) is index-2 tractable if and only if system (5.2)-(5.3) is so.*

Proof: Firstly, we will show that the relation

$$\dim N_1(x) = \dim \tilde{N}_1(x) \tag{5.7}$$

is satisfied.

(1) Let $\{z^1, ..., z^r\}$ be a basis of $N_1(x)$. For any $i \in \{1, ..., r\}$ we compute $\tilde{z}^i := Qz^i$. Then, $\tilde{z}^i$ lies in $\operatorname{im} Q = \ker g'(x)$, i.e., $g'(x)\tilde{z}^i = 0$. Setting

$$\gamma^i := P_A g'(x) z^i,$$

we may follow $\begin{pmatrix} \gamma^i \\ \tilde{z}^i \end{pmatrix} \in \tilde{N}_1(x)$. Assuming $\begin{pmatrix} \gamma^1 \\ \tilde{z}^1 \end{pmatrix}, ..., \begin{pmatrix} \gamma^r \\ \tilde{z}^r \end{pmatrix}$ to be linearly dependent, we find $\lambda_1, ..., \lambda_r \in \mathbb{R}$ such that

$$\sum_{i=1}^{r} \lambda_i \begin{pmatrix} \gamma^i \\ \tilde{z}^i \end{pmatrix} = 0.$$

Hence, the relation $\sum\limits_{i=1}^{r} \lambda_i \tilde{z}^i = 0$ is satisfied. Therefore,

$$0 = \sum_{i=1}^{r} \lambda_i Q z^i = Q \sum_{i=1}^{r} \lambda_i z^i, \quad \text{i.e.,} \quad \sum_{i=1}^{r} \lambda_i z^i \in \ker Q. \tag{5.8}$$

Now, we know

$$0 = \sum_{i=1}^{r} \lambda_i (Ag'(x) + f'(x)Q) z^i = Ag'(x) \sum_{i=1}^{r} \lambda_i z^i,$$

i.e., $\sum\limits_{i=1}^{r} \lambda_i z^i \in \ker Ag'(x) = \operatorname{im} Q$. Considering (5.8), we may follow

$$\sum_{i=1}^{r} \lambda_i z^i = 0,$$

which is a contradiction to the choice of z^1,..., z^r. Therefore, we are able to conclude that the set $\binom{\gamma^1}{\tilde{z}^1}$,..., $\binom{\gamma^r}{\tilde{z}^r}$ is a linearly independent subset of $\tilde{N}_1(x)$, i.e.,

$$\dim N_1(x) \leq \dim \tilde{N}_1(x).$$

(2) Let $\binom{\gamma^1}{\tilde{z}^1}$,..., $\binom{\gamma^s}{\tilde{z}^s}$ be a basis of $\tilde{N}_1(x)$. Regarding the relation

$$g'(x)\tilde{z}^i = Q_A \gamma^i,$$

we obtain

$$Ag'(x)\tilde{z}^i = 0, \text{ i.e., } \tilde{z}^i \in \operatorname{im} Q. \tag{5.9}$$

Further, the relation $A\gamma^i + f'(x)\tilde{z}^i = 0$ is fulfilled for each $i \in \{1, ..., s\}$. This implies

$$f'(x)\tilde{z}^i \in \operatorname{im} A = \operatorname{im} Ag'(x),$$

consequently, there is a unique $w^i \in \operatorname{im} P$ such that

$$f'(x)\tilde{z}^i = Ag'(x) w^i.$$

Computing $z^i := \tilde{z}^i - w^i$ and regarding (5.9), we see

$$(Ag'(x) + f'(x)Q) z^i = -Ag'(x) w^i + f'(x)\tilde{z}^i = 0,$$

i.e., $z^i \in N_1(x)$. Assuming z^1,..., z^s to be linearly dependent, we find λ_1,..., $\lambda_s \in \mathbb{R}$ such that

$$\sum_{i=1}^{s} \lambda_i z^i = 0$$

is fulfilled. Then, $Q \sum_{i=1}^{s} \lambda_i z^i = 0$ is valid, i.e., $\sum_{i=1}^{s} \lambda_i \tilde{z}^i = 0$. Further,

$$A \sum_{i=1}^{s} \lambda_i \gamma^i = -\sum_{i=1}^{s} f'(x) \lambda_i \tilde{z}^i = 0, \quad Q_A \sum_{i=1}^{s} \lambda_i \gamma^i = \sum_{i=1}^{s} g'(x) \lambda_i \tilde{z}^i = 0,$$

implying

$$\sum_{i=1}^{s} \lambda_i \begin{pmatrix} \gamma^i \\ \tilde{z}^i \end{pmatrix} = 0,$$

which is a contradiction to the choice of $\binom{\gamma^1}{\tilde{z}^1}$,..., $\binom{\gamma^s}{\tilde{z}^s}$. Finally, the set $\{z^1, ..., z^s\}$ is a linearly independent subset of $N_1(x)$, i.e.,

$$\dim N_1(x) \geq \dim \tilde{N}_1(x).$$

($\rightarrow$) Regarding the proof of Theorem 5.6 we obtain

$$\tilde{N} \cap \tilde{S}(x) \neq \{0\}.$$

For any $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x) \cap \tilde{S}_1(x)$, there exist α, β such that

$$P_A \gamma = Q_A \alpha - g'(x)\beta \tag{5.10}$$

$$0 = A\alpha + f'(x)\beta \tag{5.11}$$

hold. Because of $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x)$, we may conclude that

$$\tilde{z} \in \ker Ag'(x) = \operatorname{im} Q$$

if we regard $Q_A \gamma = g'(x)\tilde{z}$. Now, we compute $z := \tilde{z} - P\beta$. Using (5.10) we obtain

$$Ag'(x)z = Ag'(x)\tilde{z} - Ag'(x)\beta = -Ag'(x)\beta = A\gamma.$$

Since $\binom{\gamma}{\tilde{z}} \in \tilde{N}_1(x)$, the relation

$$-A\gamma = f'(x)\tilde{z} = f'(x)Q\tilde{z} = f'(x)Qz$$

is fulfilled. The latter two equations lead to

$$z \in N_1(x). \tag{5.12}$$

Further, we conclude

$$\begin{aligned}f'(x)Pz &= -f'(x)P\beta = -f'(x)\beta + f'(x)Q\beta \\ &= A\alpha + f'(x)Q\beta \qquad \text{(see (5.11))} \\ &= Ag'(x)\alpha_1 + f'(x)Q\beta \quad (\text{for an } \alpha_1, \text{ since } \operatorname{im} A = \operatorname{im} Ag') \\ &= (Ag'(x) + f'(x)Q)(P\alpha_1 + Q\beta),\end{aligned}$$

that means

$$z \in S_1(x). \tag{5.13}$$

Now, (5.12) and (5.13) imply $z = 0$. Because of $\tilde{z} = Q\tilde{z}$, the relation $\tilde{z} = Qz = 0$ is valid. Further, we obtain from (5.10) that

$$A\gamma = -Ag'(x)\beta = Ag'(x)Pz = 0$$

is satisfied. Finally, $\gamma = Q_A\gamma = g'(x)\tilde{z} = 0$, i.e.,

$$\tilde{N}_1(x) \cap \tilde{S}_1(x) = \{0\}.$$

($\leftarrow$) For any $z \in N_1(x) \cap S_1(x)$, we find an α_1 such that

$$f'(x)Pz = Ag'(x)\alpha_1 + f'(x)Q\alpha_1. \tag{5.14}$$

We consider

$$\begin{aligned}&\gamma := P_A g'(x)z, \quad \tilde{z} := Qz, \\ &\alpha := P_A g'(x)\alpha_1 - g'(x)z, \quad \beta := Q\alpha_1 - z.\end{aligned}$$

Then, we obtain

$$\begin{aligned}A\gamma + f'(x)\tilde{z} &= Ag'(x)z + f'(x)Qz = 0 \\ Q_A\gamma = 0 &= g'(x)Qz = g'(x)\tilde{z} \qquad (\text{note } \operatorname{im} Q = \ker g'(x)).\end{aligned}$$

Hence,

$$\left(\begin{smallmatrix}\gamma \\ \tilde{z}\end{smallmatrix}\right) \in \tilde{N}_1(x) \tag{5.15}$$

is satisfied. Further,

$$\begin{aligned}0 = (Ag'(x) + f'(x)Q)z &= -A\alpha - f'(x)\beta \quad \text{(see (5.14))} \\ P_A\gamma = P_Ag'(x)z &= -Q_Ag'(x)z + g'(x)z = Q_A\alpha - g'(x)\beta,\end{aligned}$$

i.e., $\left(\begin{smallmatrix}\gamma \\ \tilde{z}\end{smallmatrix}\right) \in \tilde{S}_1(x)$. Together with (5.15) this leads to

$$\left(\begin{smallmatrix}\gamma \\ \tilde{z}\end{smallmatrix}\right) \in \tilde{N}_1(x) \cap \tilde{S}_1(x),$$

i.e., $\gamma = \tilde{z} = 0$. Now, we know $P_Ag'(x)z = 0, \quad Qz = 0$. The first relation implies $z \in \ker Ag'(x)$, i.e., $z \in \operatorname{im} Q$. Together with the second relation, $z = 0$ is valid, i.e.,

$$N_1(x) \cap S_1(x) = 0.$$

□

5.4 Calculation of projectors and index criteria

The charge-oriented modified nodal analysis is preferred in modern circuit simulation, because it makes it possible to control the charge and flux conservation of the circuit. Further, system (5.2)-(5.3) obtained by the charge-oriented model is trivially one degree smoother than the system (5.1) obtained by the classical model if we regard (2.4). Moreover, we have seen that both modelling techniques lead to the same index (in the lower case). Therefore, we want to analyze the charge-oriented system (5.2)-(5.3) in more detail now.

Firstly, we rewrite the system (5.2)-(5.3) as

$$\begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \dot{q} \\ \dot{x} \end{pmatrix} + \begin{pmatrix} f(x) \\ q - g(x) \end{pmatrix} = \begin{pmatrix} r(t) \\ 0 \end{pmatrix}. \tag{5.16}$$

In the following, we denote the corresponding matrices, projectors, and spaces of this large system by a tilde, i.e.,

$$\tilde{A} = \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix}, \quad \tilde{\mathfrak{g}}(\tilde{x}) = \begin{pmatrix} f(x) \\ q - g(x) \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} q \\ x \end{pmatrix}, \quad \ldots$$

Since A is an incidence matrix of easy structure, it is not difficult to calculate a projector Q_A onto its nullspace. Then, the matrix

$$\tilde{Q} := \begin{pmatrix} Q_A & 0 \\ 0 & I \end{pmatrix}$$

represents a projector onto the nullspace of $\tilde{A}$. We have

$$\tilde{B}(\tilde{x}) = \begin{pmatrix} 0 & f'(x) \\ I & -g'(x) \end{pmatrix}, \quad \tilde{G}_1(\tilde{x}) = \begin{pmatrix} A & f'(x) \\ Q_A & -g'(x) \end{pmatrix}.$$

Considering the relation (5.4) and Lemma 5.2, the interesting space $\tilde{N} \cap \tilde{S}(\tilde{x})$ may be expressed by

$$\begin{aligned} \tilde{N} \cap \tilde{S}(\tilde{x}) &= \{ \left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : A\gamma = 0,\ f'(x)z \in \operatorname{im} A,\ \gamma = g'(x)z \} \\ &= \{ \left(\begin{smallmatrix} \gamma \\ z \end{smallmatrix}\right) : \gamma = 0,\ f'(x)z \in \operatorname{im} A,\ z \in \ker A^T \}. \end{aligned}$$

This information implies the following criteria for the index-1 case.

Theorem 5.8 *The system (5.2)-(5.3) is index-1 tractable if and only if the relation*

$$\{x : f'(x)z \in im\, A,\ z \in ker\, A^T\} = \{0\}$$

is satisfied.

Introducing a projector $\boldsymbol{Q}_*(x) \in L(\mathbb{R}^m)$ onto

$$S(x) = \{z \in \mathbb{R}^m : f'(x)z \in \operatorname{im} A\},$$

we may also give a criterion for the index-2 case.

Theorem 5.9 *Let the subspace*

$$N \cap S(x) = \{z :\ z \in ker\, A^T, f'(x)z \in im\, A\}$$

have constant dimension for $x \in \mathcal{D}$. Then, system (5.2)-(5.3) is index-2 tractable if and only if, for $x \in \mathcal{D}$,

$$z \in N \cap S(x),\ f'(x)z \in im\, Ag'(x)Q_*(x) \qquad imply \qquad z = 0 \tag{5.17}$$

is satisfied.

Proof: The nullspace of $\tilde{G}_1(\tilde{x})$ satisfies the relation

$$\tilde{N}_1(\tilde{x}) = \ker \tilde{G}_1(\tilde{x}) = \{(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}) :\ \gamma \in \ker Q_A,\ z \in \ker A^T,\ A\gamma + f'(x)z = 0\}. \tag{5.18}$$

Since the space $N \cap S(x)$ has constant dimension, the nullspace of $G_1(x)$ has also constant dimension

$$\dim \tilde{G}_1(\tilde{x}) = \dim\ N \cap S(x).$$

Now,

$$\begin{aligned}\tilde{S}_1(\tilde{x}) &:= \{(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}) : \begin{pmatrix}0\\ P_A\gamma\end{pmatrix} \in \operatorname{im} \begin{pmatrix}A & f'(x)\\ Q_A & -g'(x)\end{pmatrix}\} \\ &= \{(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}) : A\gamma \in \operatorname{im} Ag'(x)Q_*(x)\}.\end{aligned} \tag{5.19}$$

In particular, $(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}) \in \ker \tilde{G}_1(\tilde{x}) \cap \tilde{S}_1(\tilde{x})$ implies $z \in S(x)$, hence

$$A\gamma + f'(x)z = 0.$$

Now, the assertion follows immediately.

□

Introduce a continuous matrix function $\boldsymbol{R}_*(x)$ satisfying

$$AR(x)A^T R_*(x) = A, \quad R_*(x) = R_*(x)P_A. \tag{5.20}$$

Such a matrix function does exist, e.g. we can choose

$$R_*(x) = (AR(x)A^T)^+ A.$$

The following theorem describes the relation between the canonical projector $Q_1(x)$ of the system (5.1) and the canonical projector $\tilde{Q}_1(\tilde{x})$ of the system (5.2)-(5.3).

Theorem 5.10 *Let Q be a projector onto N and $Q_1(x)$ be the canonical projector onto $N_1(x)$ along $S_1(x)$. Then,*

$$\tilde{Q}_1(\tilde{x}) := \begin{pmatrix} P_A g'(x)Q_1(x)R_*(x) & 0 \\ QQ_1(x)R_*(x) & 0 \end{pmatrix}$$

represents the canonical projector onto $\tilde{N}_1(\tilde{x})$ along $\tilde{S}_1(\tilde{x})$.

Proof: Regarding (5.20), and (5.4), the relation

$$PR_*(x)P_A g'(x) = P$$

is satisfied, which yields

$$Q_1(x)R_*(x)P_A g'(x) = Q_1(x). \tag{5.21}$$

By this relation, it is easy to see that the matrix $\tilde{Q}_1(\tilde{x})$ defined in the theorem above is a projector for all $\tilde{x}$. Next, we show that

$$\operatorname{im} \tilde{Q}_1(\tilde{x}) = \tilde{N}_1(\tilde{x}).$$

($\subseteq$) Let $\binom{\gamma}{z}$ be an element of $\operatorname{im} \tilde{Q}_1(\tilde{x})$. Then, we find an α such that

$$\begin{aligned} \gamma &= P_A g'(x)Q_1(x)R_*(x)\alpha \\ z &= QQ_1(x)R_*(x)\alpha. \end{aligned}$$

Obviously, γ lies in $\ker Q_A$, and z lies in $\ker A^T$. Furthermore,

$$\begin{aligned} A\gamma + f'(x)z &= Ag'(x)Q_1(x)R_*(x)\alpha + f'(x)QQ_1(x)R_*(x)\alpha \\ &= [Ag'(x) + f'(x)Q]Q_1(x)R_*(x)\alpha = 0 \end{aligned}$$

is true, since $Q_1(x)$ projects onto $N_1(x)$. Regarding (5.18), we obtain $\binom{\gamma}{z}$ to be an element of $\tilde{N}_1(\tilde{x})$.

($\supseteq$) Let $\binom{\gamma}{z}$ be an element of $\tilde{N}_1(\tilde{x})$. Then, the relations

$$\gamma = P_A\gamma, \quad z = Qz, \quad A\gamma + f'(x)z = 0$$

are satisfied. Thus,

$$[Ag'(x) + f'(x)Q][PR_*(x)\gamma + Qz] = A\gamma + f'(x)z = 0,$$

i.e., $PR_*(x)\gamma + Qz$ lies in $\operatorname{im} Q_1(x)$. Let β be chosen in such a way that

$$PR_*(x)\gamma + Qz = Q_1(x)\beta$$

is satisfied. Regarding (5.21), there is an α such that

$$Q_1(x)\beta = Q_1(x)R_*(x)\alpha$$

is true. Then,

$$\begin{aligned} QQ_1(x)R_*(x)\alpha &= Qz = z \\ P_Ag'(x)Q_1(x)R_*(x)\alpha &= P_Ag'(x)R_*(x)\gamma = P_A\gamma = \gamma, \end{aligned}$$

i.e., $\binom{\gamma}{z}$ is an element of $\operatorname{im} \tilde{Q}_1(\tilde{x})$.

Finally, we show

$$\ker \tilde{Q}_1(\tilde{x}) = \tilde{S}_1(\tilde{x}).$$

($\subseteq$) Let $\binom{\gamma}{z}$ be an element of $\ker \tilde{Q}_1(\tilde{x})$. Then, the relations

$$P_Ag'(x)Q_1(x)R_*(x)\gamma = 0, \quad QQ_1(x)R_*(x)\gamma = 0$$

are satisfied. The first relation implies $PQ_1(x)R_*(x)\gamma = 0$, i.e., together with the second equation we obtain that $R_*(x)\gamma$ lies in $\ker Q_1(x) = S_1(x)$. Now, we find an α_1 such that

$$f'(x)PR_*(x)\gamma = [Ag'(x) + f'(x)Q]\alpha_1$$

is fulfilled. Computing

$$\alpha := P_Ag'(x)\alpha_1 - Q_Ag'(x)R_*(x)\gamma, \quad \beta := Q\alpha_1 - PR_*(x)\gamma,$$

we obtain

$$\begin{aligned} A\alpha + f'(x)\beta &= Ag'(x)\alpha_1 + f'(x)Q\alpha_1 - f'(x)PR_*(x)\gamma \\ Q_A\alpha - g'(x)\beta &= P_Ag'(x)R_*(x)\gamma - g'(x)Q\alpha_1, \end{aligned}$$

which yields

$$\begin{pmatrix} 0 \\ P_A\gamma \end{pmatrix} = \begin{pmatrix} A & f'(x) \\ Q_A & -g'(x) \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

i.e., $\binom{\gamma}{z}$ is an element of $\tilde{S}_1(\tilde{x})$.

($\supseteq$) Let $\left(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}\right)$ be an element of $\tilde{S}_1(\tilde{x})$. Then, we find α and β such that

$$P_A\gamma = Q_A\alpha - g'(x)\beta$$

is true. Regarding (5.20), we have

$$\begin{aligned} f'(x)PR_*(x)\gamma &= -f'(x)P\beta = f'(x)Q\beta - f'(x)\beta \\ &= f'(x)Q\beta + A\alpha = [Ag'(x) + f'(x)Q][PR_*(x)\alpha + Q\beta], \end{aligned}$$

i.e., $R_*(x)\gamma$ lies in $S_1(x) = \ker Q_1(x)$. Therefore $Q_1(x)R_*(x)\gamma = 0$ is true, and hence $\left(\begin{smallmatrix}\gamma\\ z\end{smallmatrix}\right)$ is an element of $\ker \tilde{Q}_1(\tilde{x})$.

□

5.5 Numerical specifics

Applying the BDF to the charge-oriented modelling, we obtain an equation system of the following form

$$A\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}q_{\ell-i} + f(x_\ell) = r(t_\ell) \tag{5.22}$$

$$q_\ell = g(x_\ell). \tag{5.23}$$

In order to obtain a lower-dimensional system, one puts the second equation into the first one, solves

$$A\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}g(x_{\ell-i}) + f(x_\ell) = r(t_\ell)$$

by a Newton-like method, and calculates $q_\ell = g(x_\ell)$ then.

Remark 5.11 Chapter 4 deals with the BDF applied to quasilinear index-2 DAEs of the form (4.3). Since, in general, the charge-oriented modified nodal analysis provides systems belonging to this class of DAEs, we may successfully apply the BDF method to general models (i.e., also to models with non-reciprocal capacitances) if the assumptions of Theorem 4.2 are satisfied.

The investigations in Chapter 4 make clear that all components of the solution are influenced by a week instability arising from defects in the derivative-free part of the DAE. Hence, for obtaining a solution, it is necessary to ensure that these defects remain smaller than the stepsize used. Obviously, the defects in equation (5.23) vanish. Dividing equation (5.22) by a $Q_{im\,A}$ onto $\operatorname{im} A$ into the system

$$A\frac{1}{h_\ell}\sum_{i=0}^{k}\alpha_{\ell i}q_{\ell-i} + Q_{im\,A}f(x_\ell) = Q_{im\,A}r(t_\ell) \tag{5.24}$$

$$(I - Q_{im\,A})f(x_\ell) = (I - Q_{im\,A})r(t_\ell), \tag{5.25}$$

it is only necessary to observe the defects arising from the solution of equation (5.25).

The values of the components q are very small ($\approx 10^{-12}$) in comparison to the values of x ($\approx 10^{-3}$). We tested the variable stepsize variable order BDF method by controlling the "reliable" non-nullspace components only successfully in [Tis92]. In our case here, the non-nullspace components are represented by $P_A q$, i.e., the tolerance requirements must be adapted to absolute values of $\approx 10^{-12}$.

5.6 NAND-gate

5.6.1 Model

Most of the industrially integrated circuits contain NAND- and NOR-gates as basic elements. These types of gates may be economically produced and universally used. Figure 5.1 displays a circuit simulating a NAND-gate (see [GR94]). It consists of two n-channel enhancement MOSFETs (ME), one n-channel depletion MOSFET (MD), and a load capacitor C (cf. [SH68]).

Digital MOS-circuits contain no other elements besides the MOSFETs as a rule. MOSFETs also take the function of controlled resistors. In our example, gate and source of the depletion transistor MD are connected, i.e., this MOSFET works as a controlled resistor here.

The drain voltage of MD is constant at $V_{DD} = 5V$. The bulk voltages are not at ground, $V_{BB} = -2.5V$. The source voltages of both MEs are at ground. The gate voltages are controlled by the voltage sources V_1 and V_2. The response at node 1 is only LOW (FALSE) if both, the input signal V_1

and the input signal V_2, are HIGH (TRUE).

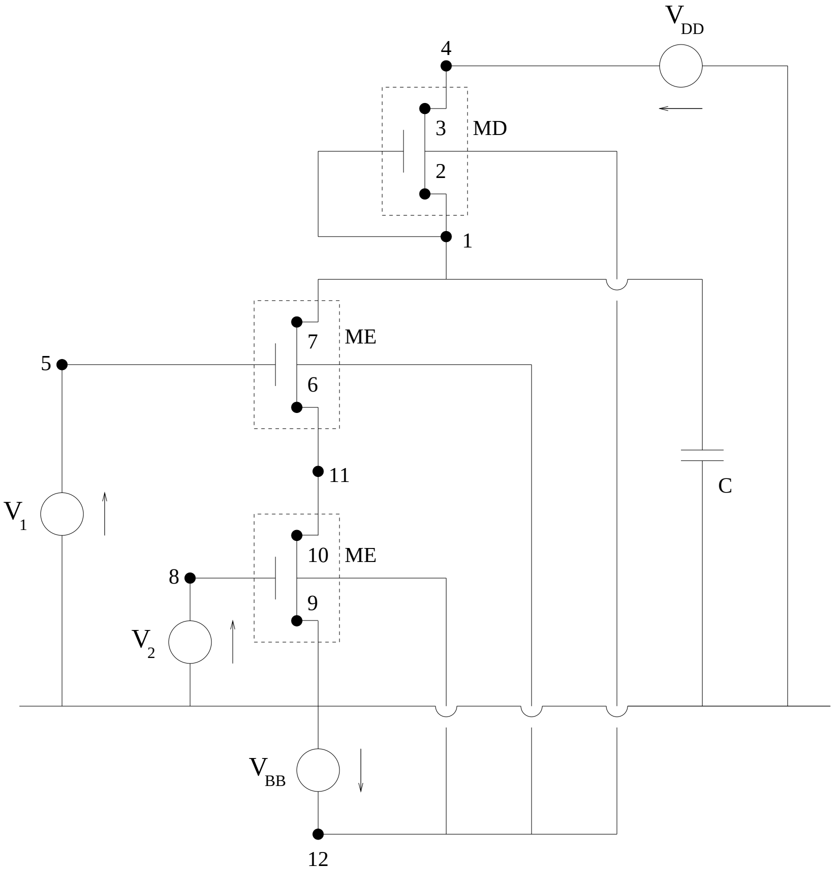

Figure 5.1: NAND-gate model

The circuit model for the MOSFETs MD and ME is given in Figure 5.2 (see [SH68]). Later, we will show that the model leads to an index-2 DAE for the NAND-gate. The model used in [FG94] and [Gün95] is a regularization of this model and of index-1. The transistors MD and ME differ only in parameter values (see Table 5.1).

The current i_{ds} flows from drain to source if and only if the controlling voltage U_{gs} between gate and source is larger than a technology dependent threshold voltage U_T. The gate is isolated from the channel DS by a thin SiO_2-layer, i.e., the resistance R_{sd} between source and drain is almost infinitely high ($\sim 10^{15}\Omega$).

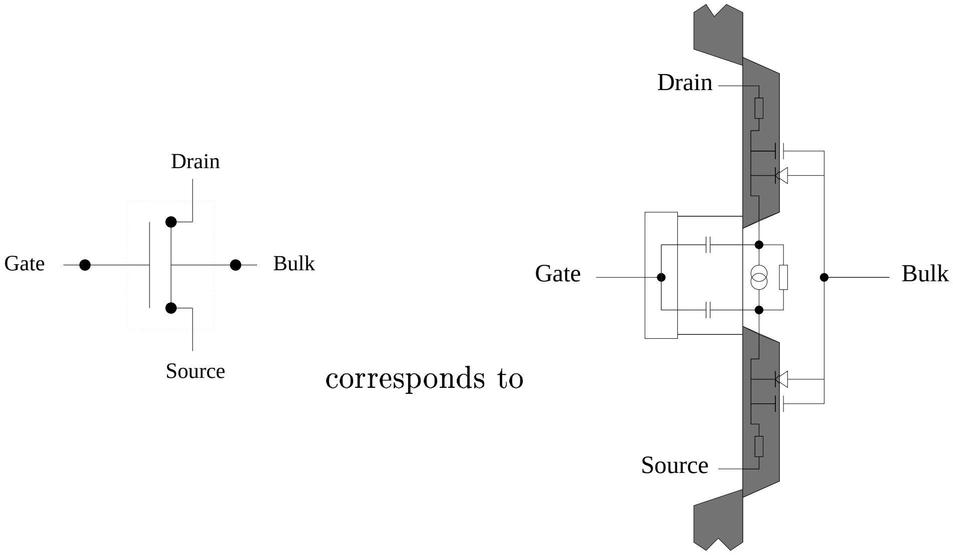

Figure 5.2: MOSFET model

Using the charge-oriented modified nodal analysis described in Chapter 2, we obtain the following DAE system

$$\frac{u_1 - u_2}{R_s} - \frac{u_7 - u_1}{R_d} + \dot{Q} + \dot{Q}_{1gd} + \dot{Q}_{1gs} = 0 \tag{5.26}$$

$$\begin{aligned}-\dot{Q}_{1gs} + \dot{Q}_{1sb} + \frac{u_2 - u_1}{R_s} + \frac{u_2 - u_3}{R_{sd}} + i_{bs}^D(u_{12} - u_2)&\\ + i_{ds}^D(u_3 - u_2, u_1 - u_2, u_{12} - u_2) &= 0\end{aligned} \tag{5.27}$$

$$\begin{aligned}-\dot{Q}_{1gd} + \dot{Q}_{1db} + \frac{u_3 - u_4}{R_d} - \frac{u_2 - u_3}{R_{sd}} + i_{bd}^D(u_{12} - u_3)&\\ - i_{ds}^D(u_3 - u_2, u_1 - u_2, u_{12} - u_2) &= 0\end{aligned} \tag{5.28}$$

$$\frac{u_4 - u_3}{R_d} + I_{DD} = 0 \tag{5.29}$$

$$+\dot{Q}_{2gd} + \dot{Q}_{2gs} + I_1 = 0 \tag{5.30}$$

$$\begin{aligned}-\dot{Q}_{2gs} + \dot{Q}_{2sb} + \frac{u_6 - u_{11}}{R_s} + \frac{u_6 - u_7}{R_{sd}} + i_{bs}^E(u_{12} - u_6)&\\ + i_{ds}^E(u_7 - u_6, u_5 - u_6, u_{12} - u_6) &= 0\end{aligned} \tag{5.31}$$

$$\begin{aligned}-\dot{Q}_{2gd} + \dot{Q}_{2db} + \frac{u_7 - u_1}{R_d} - \frac{u_6 - u_7}{R_{sd}} + i_{bd}^E(u_{12} - u_7)&\\ - i_{ds}^E(u_7 - u_6, u_5 - u_6, u_{12} - u_6) &= 0\end{aligned} \tag{5.32}$$

$$+\dot{Q}_{3gd} + \dot{Q}_{3gs} + I_2 = 0 \tag{5.33}$$

$$\begin{aligned} -\dot{Q}_{3gs} + \dot{Q}_{3sb} + \frac{u_9}{R_s} + \frac{u_9 - u_{10}}{R_{sd}} + i^E_{bs}(u_{12} - u_9) \\ + i^E_{ds}(u_{10} - u_9, u_8 - u_9, u_{12} - u_9) = 0 \end{aligned} \tag{5.34}$$

$$\begin{aligned} -\dot{Q}_{3gd} + \dot{Q}_{3db} + \frac{u_{10} - u_{11}}{R_d} - \frac{u_9 - u_{10}}{R_{sd}} + i^E_{bd}(u_{12} - u_{10}) \\ - i^E_{ds}(u_{10} - u_9, u_8 - u_9, u_{12} - u_9) = 0 \end{aligned} \tag{5.35}$$

$$\frac{u_{11} - u_6}{R_s} - \frac{u_{10} - u_{11}}{R_d} = 0 \tag{5.36}$$

$$\begin{aligned} -\dot{Q}_{1db} - \dot{Q}_{1sb} - i^D_{bs}(u_{12} - u_2) - i^D_{bd}(u_{12} - u_3) \\ -\dot{Q}_{2db} - \dot{Q}_{2sb} - i^E_{bs}(u_{12} - u_6) - i^E_{bd}(u_{12} - u_7) \\ -\dot{Q}_{3db} - \dot{Q}_{3sb} - i^E_{bs}(u_{12} - u_9) - i^E_{bd}(u_{12} - u_{10}) + I_{BB} = 0 \end{aligned} \tag{5.37}$$

$$Q - C\,u_1 = 0 \tag{5.38}$$

$$Q_{1gd} - q_{gd}(u_1 - u_3) = 0 \tag{5.39}$$

$$Q_{1gs} - q_{gs}(u_1 - u_2) = 0 \tag{5.40}$$

$$Q_{1db} - q_{db}(u_3 - u_{12}) = 0 \tag{5.41}$$

$$Q_{1sb} - q_{sb}(u_2 - u_{12}) = 0 \tag{5.42}$$

$$Q_{2gd} - q_{gd}(u_5 - u_7) = 0 \tag{5.43}$$

$$Q_{2gs} - q_{gs}(u_5 - u_6) = 0 \tag{5.44}$$

$$Q_{2db} - q_{db}(u_7 - u_{12}) = 0 \tag{5.45}$$

$$Q_{2sb} - q_{sb}(u_6 - u_{12}) = 0 \tag{5.46}$$

$$Q_{3gd} - q_{gd}(u_8 - u_{10}) = 0 \tag{5.47}$$

$$Q_{3gs} - q_{gs}(u_8 - u_9) = 0 \tag{5.48}$$

$$Q_{3db} - q_{db}(u_{10} - u_{12}) = 0 \tag{5.49}$$

$$Q_{3sb} - q_{sb}(u_9 - u_{12}) = 0 \tag{5.50}$$

$$u_4 - V_{DD} = 0 \tag{5.51}$$

$$u_{12} - V_{BB} = 0 \tag{5.52}$$

$$u_5 - V_1 = 0 \tag{5.53}$$

$$u_8 - V_2 = 0 \tag{5.54}$$

The current through the diode between bulk and source as well as the current through the diode between bulk and drain is given by the function

$$i_{bs}(U) = i_{bd}(U) = \begin{cases} -i_s \cdot \left(\exp(\frac{U}{U_T}) - 1\right) & \text{for } U \leq 0 \\ 0 & \text{for } U > 0 \end{cases}. \tag{5.55}$$

The current through the controlled current source between drain and source is modelled by the function

$$i_{ds}(U_{ds}, U_{gs}, U_{bs}) =$$
$$\begin{cases} 0 & \text{for } U_{gs} - U_{TE} \leq 0 \\ -\beta \cdot (1 + \delta \cdot U_{ds}) \cdot (U_{gs} - U_{TE}) & \text{for } 0 < U_{gs} - U_{TE} \leq U_{ds} \\ -\beta \cdot U_{ds} \cdot (1 + \delta \cdot U_{ds}) \cdot [2(U_{gs} - U_{TE}) - U_{ds}] & \text{for } 0 < U_{ds} < U_{gs} - U_{TE} \end{cases}$$

with $U_{TE} = U_{T0} + \gamma \cdot \left(\sqrt{\Phi - U_{bs}} - \sqrt{\Phi}\right)$. The technical parameters for the MOSFETs MD and ME are given in Table 5.1.

	ME	MD
i_s	10^{-14}A	10^{-14}A
U_T	25.85V	25.85V
U_{T0}	0.8V	-2.43V
β	$1.748 \cdot 10^{-3}$A/V^2	$5.35 \cdot 10^{-4}$A/V^2
γ	$0.0\sqrt{\text{V}}$	$0.2\sqrt{\text{V}}$
δ	0.02V^{-1}	0.02V^{-1}
Φ	1.01V	1.28V

Table 5.1: Technical parameters

The values for the resistances are chosen for all MOSFETs as

$$R_s = R_d = 4\Omega, \quad R_{sd} = 10^{15}\Omega.$$

The load capacitance is constant of size $C = 0.5 \cdot 10^{-13}F$. The capacitance between gate and source as well as that between gate and drain are modelled as linear capacitors, i.e.,

$$q_{gs}(u) = q_{gd}(u) = C_1 \cdot u \text{ with } C_1 = 0.6 \cdot 10^{-13}F.$$

The capacitance between bulk and drain as well as that between bulk and source may be modelled on two levels (see [GR94]):

- Level A: Linear capacitances

$$q_{db}(u) = q_{sb}(u) = -C_0 \cdot u \text{ with } C_0 = 0.24 \cdot 10^{-13}F.$$

- Level B: Nonlinear capacitances

$$q_{db}(u) = q_{sb}(u) = \begin{cases} -C_0 \cdot \Phi_B \cdot \left(1 - \sqrt{1 - \frac{u}{\Phi_B}}\right) & \text{for } u \leq 0 \\ -C_0 \cdot \left(1 + \frac{u}{4\Phi_B}\right) \cdot u & \text{for } u > 0 \end{cases}$$

with

$$C_0 = 0.24 \cdot 10^{-13} F \text{ and } \Phi_B = 0.87V.$$

5.6.2 Index analysis

Let us study the related spaces for the NAND-gate equation system (5.26)-(5.54) using the vector

$$(q,x) = (Q, Q_{1gd}, Q_{1gs}, Q_{1db}, Q_{1sb}, Q_{2gd}, Q_{2gs}, Q_{2db}, Q_{2sb}, Q_{3gd}, Q_{3gs},\\ Q_{3db}, Q_{3sb}, u_1, u_2, u_3, u_4, u_5, u_6, u_7, u_8, u_9, u_{10}, u_{11}, u_{12}, I_1, I_2, I_{BB}, I_{DD}).$$

It is easy to verify that

$$\begin{aligned} \ker A = N = \{z:\ & z_1 = 0,\ z_2 = -z_3 = z_4 = -z_5, \\ & z_6 = -z_7 = z_8 = -z_9, \\ & z_{10} = -z_{11} = z_{12} = -z_{13}\}. \end{aligned}$$

The image space of A may be written as

$$\operatorname{im} A = \{y:\ y_4 = y_{11} = 0, y_{13} = y_{14} = ... = y_{29} = 0\}. \tag{5.56}$$

This leads to

$$\begin{aligned} S(u) = \{z: & \ z_{17} = z_{25} = z_{18} = z_{21} = 0,\ z_{29} = \frac{1}{R_d} z_{16}, z_1 = C z_{14}, \\ & -\frac{1}{R_s} z_{19} - \frac{1}{R_d} z_{23} + (\frac{1}{R_s} + \frac{1}{R_d}) z_{24} = 0, \\ & z_2 = C_1[z_{14} - z_{16}],\ z_3 = C_1[z_{14} - z_{15}],\ z_4 = C_2(u_3 - u_{12}) z_{16}, \\ & z_5 = C_2(u_2 - u_{12}) z_{15},\ z_6 = -C_1 z_{20},\ z_7 = -C_1 z_{10}, \\ & z_8 = C_2(u_7 - u_{12}) z_{20},\ z_9 = C_2(u_6 - u_{12}) z_{19},\ z_{10} = -C_1 z_{23}, \\ & z_{11} = -C_1 z_{22},\ z_{12} = C_2(u_{10} - u_{12}) z_{23},\ z_{13} = C_2(u_9 - u_{12}) z_{22}\} \end{aligned}$$

if the capacitance function $C_2(u)$ is defined by

- $C_2(u) = C_0$ in case of Level A

- $C_2(u) = \begin{cases} C_0 \cdot \left(1 - (1 - \frac{u}{\Phi_B})^{-\frac{1}{2}}\right) & \text{for } u \leq 0 \\ C_0 \cdot \left(1 + \frac{u}{2\Phi_B}\right) & \text{for } u > 0 \end{cases}$ in case of Level B.

Hence,

$$N \cap S = \{z : \ z_1 = z_2 = ... = z_{25} = 0, \ z_{29} = 0\}, \tag{5.57}$$

i.e., the space $N \cap S$ is constant, and the structural condition (3.39) is satisfied. Further, we see that the components z_{26}, z_{27} and z_{28}, i.e., I_1, I_2 and I_{BB}, represent the critical variables. A differentiation is necessary for computing these variables.

As a projector Q onto N we choose $Q = (q_{ij})$ for $i, j = 1, ..., 29$ with

$$\begin{aligned} &q_{2\,2} = -q_{3\,2} = q_{4\,2} = -q_{5\,2} = 1, \quad q_{6\,6} = -q_{7\,6} = q_{4\,6} = -q_{8\,6} = 1 \\ &q_{10\,10} = -q_{11\,10} = q_{12\,10} = -q_{13\,10} = 1, \quad q_{i\,i} = 1 \quad \text{for } i = 14, ..., 29 \\ &q_{i\,j} = 0 \quad \text{in all the other cases.} \end{aligned}$$

Calculating

$$G_1(x) := A + B(x)Q$$

we obtain that

$$\begin{aligned} \ker G_1(x) = N_1 = \{z : \ & z_2 = z_4 = z_6 = z_8 = z_{10} = z_{12} = 0, \ z_{29} = 0, \\ & z_{14} = z_{15} = ... = z_{25} = 0, \ z_{28} = -z_5 - z_9 - z_{13} \\ & z_5 = z_3 = -z_1, \ z_{26} = z_9 = z_7, \ z_{27} = z_{13} = z_{11}\} \end{aligned}$$

is constant. For the image space of $G_1(x)$ we obtain

$$\begin{aligned} \operatorname{im} G_1(u) = \{y : \ & y_{22} + y_{23} - 2C_1 y_{29} \\ & + \frac{C_1}{C_1 + C_{2\,5}(u)} [y_{24} - y_{22} + C_1 y_{29} + C_{2\,5}(u) y_{27}] \\ & + \frac{C_1}{C_1 + C_{2\,6}(u)} [y_{25} - y_{23} + C_1 y_{29} + C_{2\,6}(u) y_{27}] = 0, \\ & y_{18} + y_{19} - 2C_1 y_{28} \\ & + \frac{C_1}{C_1 + C_{2\,3}(u)} [y_{20} - y_{18} + C_1 y_{28} + C_{2\,3}(u) y_{27}] \\ & + \frac{C_1}{C_1 + C_{2\,4}(u)} [y_{21} - y_{19} + C_1 y_{28} + C_{2\,4}(u) y_{27}] = 0, \\ & \frac{C}{C_1 \cdot [C_{2\,1}(u) + C_{2\,2}] + 2C_{2\,1}(u) C_2(u_2 - u_{12})} [C_{2\,1}(u) y_{14} \end{aligned}$$

$$
\begin{aligned}
&+ C_2(u_7 - u_{12})y_{15} + \frac{C_{21}(u)C_{22}}{C_1}[y_{14} + y_{15}] \\
&+ [C_1 + C_{22}][y_{16} + C_{21}(u)y_{27}] + [C_1 + C_{21}(u)][y_{17} + C_{22}y_{27}]] = y_{13}\}
\end{aligned}
$$

if

$$
\begin{aligned}
C_{21}(u) &:= C_2(u_3 - u_{12}), & C_{22}(u) &:= C_2(u_2 - u_{12}) \\
C_{23}(u) &:= C_2(u_7 - u_{12}), & C_{24}(u) &:= C_2(u_6 - u_{12}) \\
C_{25}(u) &:= C_2(u_{10} - u_{12}), & C_{26}(u) &:= C_2(u_9 - u_{12}).
\end{aligned}
$$

Thus,

$$
\begin{aligned}
S_1(u) = \{z :\ &\frac{C_{25}(u)}{C_1 + C_{25}(u)}[z_{10} + z_{12}] + \frac{C_{26}(u)}{C_1 + C_{26}(u)}[z_{11} + z_{13}] = 0, \\
&\frac{C_{23}(u)}{C_1 + C_{23}(u)}[z_6 + z_8] + \frac{C_{24}(u)}{C_1 + C_{24}(u)}[z_7 + z_9] = 0, \\
&\frac{C}{C_1 \cdot [C_{21}(u) + C_{22}(u)] + 2C_{21}(u)C_{22}(u)}\Big[[C_1 + C_{22}(u)]z_4 \\
&\quad + [C_1 + C_{21}(u)]z_5 + C_{22}(u)[1 + \frac{C_{21}(u)}{C_1}][z_2 + z_3]\Big] = z_1\},
\end{aligned}
$$

which implies

$$N_1 \cap S_1(u) = \{0\},$$

i.e., the NAND-gate equation system (5.26)-(5.54) is index-2 tractable.

5.6.3 Numerical results

Taking into account Remark 4.3 and relation (5.56), the sensitive perturbations, which one has to control in relation to the stepsize, are those arising when solving the equations (5.29), (5.36), and (5.38)-(5.54). Considering Section 5.5, the equations (5.38)-(5.50) are solved exactly. Further, the defects of the other equations (5.29), (5.36), and (5.51)-(5.54) are small since these equations are linear. Consequently, the assumptions of Theorem 4.2 are satisfied.

For the numerical integration, we used the Level B approach. The absolute tolerance and the relative tolerance were chosen as 10^{-12} and 10^{-3}, respectively. We started at the inconsistent starting value $(q(0), x(0)) = 0$ without any problems. The integration required 602 steps over the interval $[0ns, 80ns]$.

The input signals are given in Figure 5.3.

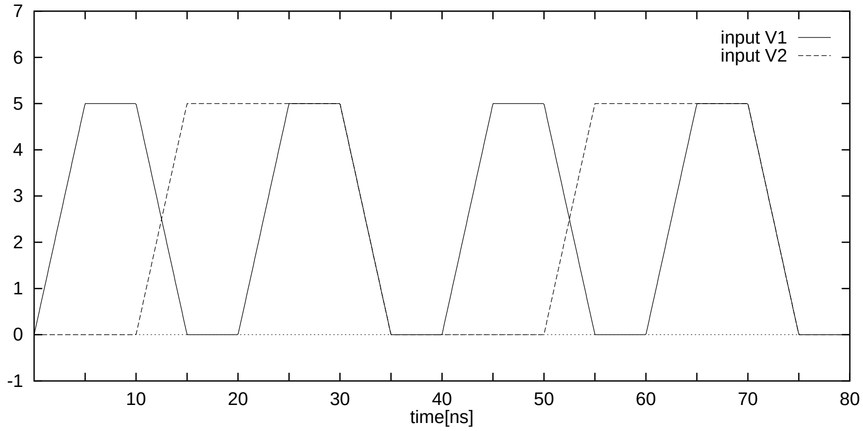

Figure 5.3: Input signals V_1 and V_2

The simulation results reflect the real output of the NAND-gate (see Figure 5.4). The voltage at nodal 1 is low if and only if the input voltages V_1 and V_2 are high. The regions $[10ns, 15ns]$ and [50ns,55ns] are critical. Both signals, V_1 and V_2, are relatively high around the points of time 12.5ns and 52.5ns.

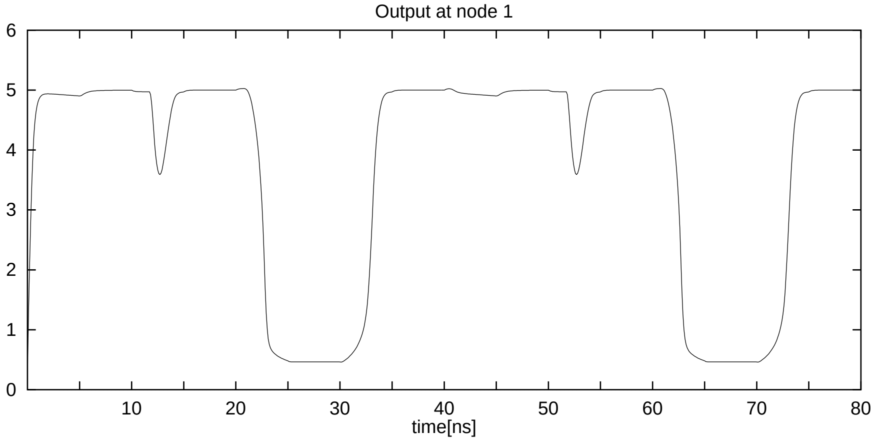

Figure 5.4: Response at node 1

Regarding (5.57), the currents I_1 through V_1, I_2 through V_2, and I_{BB} through

V_{BB} represent the so-called index-2 variables. Figure 5.5 shows the result for I_1 and Figure 5.6 for I_2.

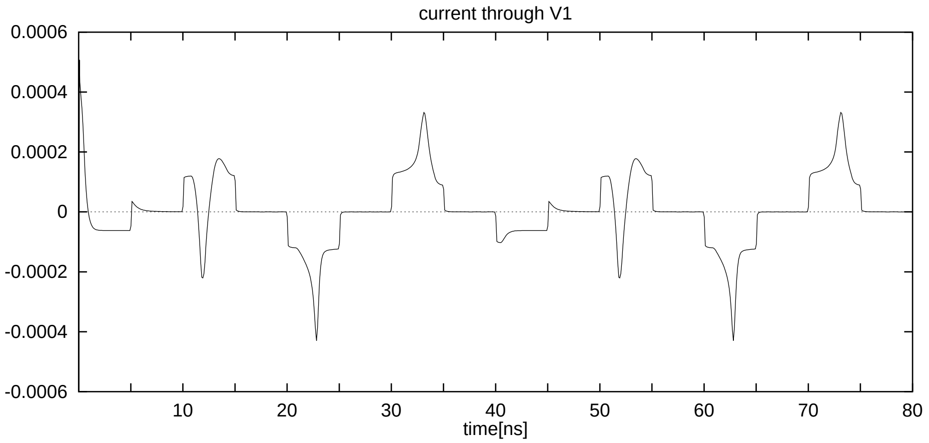

Figure 5.5: Current I_1

The currents I_1 and I_2 vanish in the intervals [5ns, 10ns], [15ns, 20ns], [25ns, 30ns], [35ns, 40ns], [45ns, 50ns], [55ns, 60ns], [65ns, 70ns], [75ns, 80ns] since the input signals V_1 and V_2 are constant in these intervals. This gives us the possibility to determine the global error in these intervals.

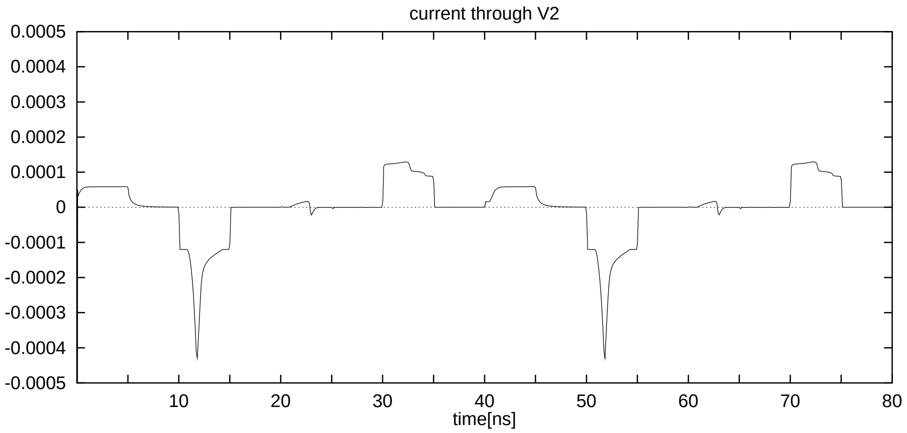

Figure 5.6: Current I_2

All calculations were carried out by the BDF code DAE2SOL ([Tis92]) with controlled order and stepsize via the smooth component

$$\tilde{P}\tilde{x}_\ell = \begin{pmatrix} P_A & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} q_\ell \\ x_\ell \end{pmatrix} = \begin{pmatrix} P_A q_\ell \\ 0 \end{pmatrix},$$

where

$$P_A = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 1
\end{pmatrix}.$$

From our experience, this control works essentially more effective than that of the complete value $\tilde{x}_\ell$.

As far as the weak instability term $\frac{1}{h_\ell} Q_{1\ell} G_{2\ell}^{-1} \delta_\ell$ involved in the error estimation of Theorem 4.2 is concerned, this effect is even typical of index-2 DAEs. Besides the usual error propagation expected from the index-1 case, a certain defect component amplified by h_ℓ^{-1} influences the computation strongly.

By the following table we realize those instability effects. The table shows the values computed by the constant stepsize backward Euler method with different stepsizes for approximating the currents $I_1(T) = 0$, $I_2(T) = 0$, $I_{DD}(T) = 0$, and $I_{BB}(T) = 0$, which have to vanish at the final point $T = 80 \cdot 10^{-9}$. The produced values reflect the theoretical results as expected. If we decrease the stepsize, the error becomes smaller up to the stepsize 2e-10. The error increases for stepsizes smaller than $2e - 10$. This clearly reflects the weak instability.

stepsize	I_1	I_2	I_{DD}	I_{BB}
8e-10	5.41e-10	1.25e-15	2.30e-09	3.29e-09
5e-10	2.36e-10	1.20e-15	1.00e-09	1.44e-09
2e-10	1.51e-10	1.19e-15	6.23e-10	9.00e-10
1e-10	1.88e-10	1.22e-15	4.91e-10	1.53e-10
5e-11	1.24e-09	2.25e-15	2.72e-09	5.34e-10

5.7 Ring modulator

5.7.1 Model

The ring modulator considered here represents a small circuit that interferes a high frequency signal $e_1(t)$ with a low frequency signal $e_2(t)$.

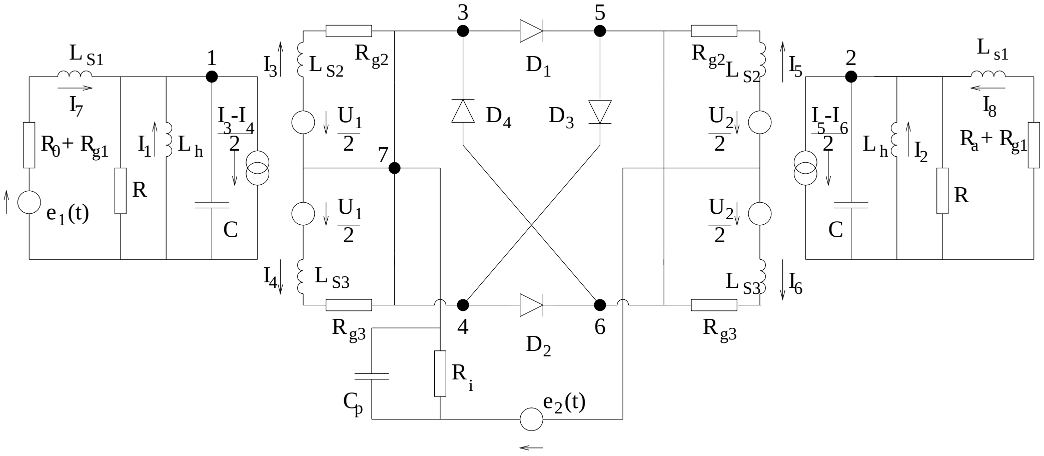

Figure 5.7: Ring modulator

This circuit was modelled by [Hor76]. If we discard the very small capacitances of diodes, then we obtain the following DAE by applying the classical modified nodal analysis.

$$C\dot{u}_1 = I_1 - I_3 \cdot 0.5 + I_4 \cdot 0.5 + I_7 - u_1/R \tag{5.58}$$

$$C\dot{u}_2 = I_2 - I_5 \cdot 0.5 + I_6 \cdot 0.5 + I_8 - u_2/R \tag{5.59}$$

$$0 = I_3 - G(UD_1) + G(UD_4) \tag{5.60}$$

$$0 = -I_4 + G(UD_2) - G(UD_3) \tag{5.61}$$

$$0 = I_5 + G(UD_1) - G(UD_3) \tag{5.62}$$

$$0 = -I_6 - G(UD_2) + G(UD_4) \tag{5.63}$$

$$C_P\dot{u}_7 = u_7/R_i + G(UD_1) + G(UD_2) - G(UD_3) - G(UD_4) \tag{5.64}$$

$$L_h\dot{I}_1 = -u_1 \tag{5.65}$$

$$L_h\dot{I}_2 = -u_2 \tag{5.66}$$

$$L_{S2}\dot{I}_3 = u_1 \cdot 0.5 - u_3 - R_{g2} \cdot I_3 \tag{5.67}$$
$$L_{S3}\dot{I}_4 = -u_1 \cdot 0.5 + u_4 - R_{g3} \cdot I_4 \tag{5.68}$$
$$L_{S2}\dot{I}_5 = u_2 \cdot 0.5 - u_5 - R_{g2} \cdot I_5 \tag{5.69}$$
$$L_{S3}\dot{I}_6 = -u_2 \cdot 0.5 + u_6 - R_{g3} \cdot I_6 \tag{5.70}$$
$$L_{S1}\dot{I}_7 = -u_1 + e_1(t) - (R_0 + R_{g1}) \cdot I_7 \tag{5.71}$$
$$L_{S1}\dot{I}_8 = -u_2 - (R_a + R_{g1}) \cdot I_8, \tag{5.72}$$

the diode-functions are given by

$$G(UD) = 40.67286402 \cdot 10^{-9} \cdot [exp(17.7493332 \cdot UD) - 1],$$

the voltages at the different diodes are expressed by

$$\begin{aligned} UD_1 &= u_3 - u_5 - u_7 - e_2(t) \\ UD_2 &= -u_4 + u_6 - u_7 - e_2(t) \\ UD_3 &= u_4 + u_5 + u_7 + e_2(t) \\ UD_4 &= -u_3 - u_6 + u_7 + e_2(t). \end{aligned}$$

For the technical parameters, it holds that

$$\begin{aligned} R_{g1} &= 36.3\Omega, \quad R_{g2} = R_{g3} = 17.3\Omega \\ R_0 &= R_i = 50\Omega \\ R_a &= 600\Omega, \quad R = 25000\Omega \\ C &= 16 \cdot 10^{-9}F, \quad C_P = 10 \cdot 10^{-9}F \\ L_h &= 4.45H, \quad L_{S1} = 2 \cdot 10^{-3}H, \quad L_{S2} = L_{S3} = 0.5 \cdot 10^{-3}H \end{aligned}$$

The input signals are as follows:

$$\begin{aligned} e_2(t) &= 2 \cdot \sin(2\pi \cdot 10^4 \cdot t) \\ e_1(t) &= 0.5 \cdot \sin(2\pi \cdot 10^3 \cdot t). \end{aligned}$$

The considered capacitances of the diodes C_S are of small order $10^{-12}F$. Therefore, the system is extremely stiff. Recent investigations by [DR89] and [HLR89] have shown the undesirable oscillations in the diode-voltages to be the smaller the smaller the parameter C_S is. In case of $C_S = 0$, the simulation result is most adapted to the curve measured. In this case, the system (5.58)-(5.72) becomes an index-2 tractable DAE, which we will see in the next subsection.

5.7.2 Index analysis

Let us study the related spaces for the ring modulator equation system (5.58)-(5.72). It is easy to verify that

$$\ker A = N = \{z: \ z_1 = z_2 = 0, \ z_7 = z_8 = ... = z_{15} = 0\}.$$

The image space of A may be written as

$$\operatorname{im} A = \{y: \ y_3 = y_4 = y_5 = y_6 = 0\}. \tag{5.73}$$

This leads to

$$\begin{aligned} S(u,t) = \{z: \ & [g_1(u,t) + g_4(u,t)]z_3 - g_1(u,t)z_5 + g_4(u,t)z_6 \\ & \qquad - [g_1(u,t) + g_4(u,t)]z_7 - z_{10} = 0, \\ & [g_2(u,t) + g_3(u,t)]z_4 + g_3(u,t)z_5 - g_2(u,t)z_6 \\ & \qquad + [g_2(u,t) + g_3(u,t)]z_7 + z_{11} = 0, \\ & - g_1(u,t)z_3 + g_3(u,t)z_4 + [g_1(u,t) + g_3(u,t)]z_5 \\ & \qquad + [g_1(u,t) + g_3(u,t)]z_7 - z_{12} = 0, \\ & g_4(u,t)z_3 - g_2(u,t)z_4 + [g_2(u,t) + g_4(u,t)]z_6 \\ & \qquad - [g_2(u,t) + g_4(u,t)]z_7 + z_{13} = 0\} \end{aligned}$$

if the diode-functions $g_1(u)$, $g_2(u)$, $g_3(u)$, and $g_4(u)$ are defined by

$$\begin{aligned} g_1(u,t) &:= G'(u_3 - u_5 - u_7 - e_2(t)) \\ g_2(u,t) &:= G'(-u_4 + u_6 - u_7 - e_2(t)) \\ g_3(u,t) &:= G'(u_4 + u_5 + u_7 + e_2(t)) \\ g_4(u,t) &:= G'(-u_3 - u_6 + u_7 + e_2(t)) \end{aligned}$$

with

$$G'(u) = 17.7493332 \cdot 40.67286402 \cdot 10^{-9} \cdot exp(17.7493332 \cdot u).$$

Hence,

$$\begin{aligned} N \cap S(u,t) = \{z: \ z_1 = z_2 = 0, \ z_7 = z_8 = ... = z_{15} = 0, \\ z_3 = -z_4 = z_5 = -z_6\}, \end{aligned} \tag{5.74}$$

i.e., the space $N \cap S$ is constant. Thus, the structural condition (3.39) is satisfied. Further, we see that the components z_3, z_4, z_5, and z_6, i.e., u_3, u_4, u_5, and u_6 represent the critical variables. For computing them, a differentiation is necessary.

As a projector Q onto N we choose $Q = (q_{ij})$ for $i, j = 1, ..., 15$ with

$$\begin{aligned} &q_{3\,3} = q_{4\,4} = q_{5\,5} = q_{6\,6} = 1 \\ &q_{i\,j} = 0 \quad \text{in all other cases.} \end{aligned}$$

Calculating

$$G_1(x) := A + B(x)Q$$

we obtain that

$$\begin{aligned} \ker G_1(x) = N_1 = \{z:\ &z_1 = z_2 = 0,\ z_7 = z_8 = z_9 = 0,\ z_{14} = z_{15} = 0, \\ &z_3 = -z_4 = z_5 = -z_6,\ z_3 = -L_{S2}z_{10}, \\ &z_4 = L_{S3}z_{11} = -z_1,\ z_5 = -L_{S2}z_{12},\ z_6 = L_{S3}z_{13}\} \end{aligned}$$

is constant. For the image space of $G_1(x)$ we obtain

$$\operatorname{im} G_1(x) = \{y:\ y_3 - y_4 + y_5 - y_6 = 0\}.$$

Thus, we obtain

$$S_1 = \{z:\ z_{10} + z_{11} + z_{12} + z_{13} = 0\},$$

which implies

$$N_1 \cap S_1 = \{0\},$$

i.e., the ring modulator equation system (5.58)-(5.72) is index-2 tractable.

5.7.3 Numerical results

Due to Remark 4.3 and the relation (5.73), the sensitive perturbations, which one has to control in relation to the stepsize, are the perturbations resulting from solving the equations (5.60)-(5.63). These equations contain exponential functions. Therefore, it is important to keep the defects in these equations small in relation to the stepsize in order to make the BDF feasible over the integration interval.

The integration required 944 steps for an absolute tolerance and a relative tolerance of 10^{-6}.

Comparing the simulation result for the output voltage u_2 presented in Figure 5.8 with the measured voltage (see e.g. [KRS92]), the simulation results reflect the real output of the ring modulator.

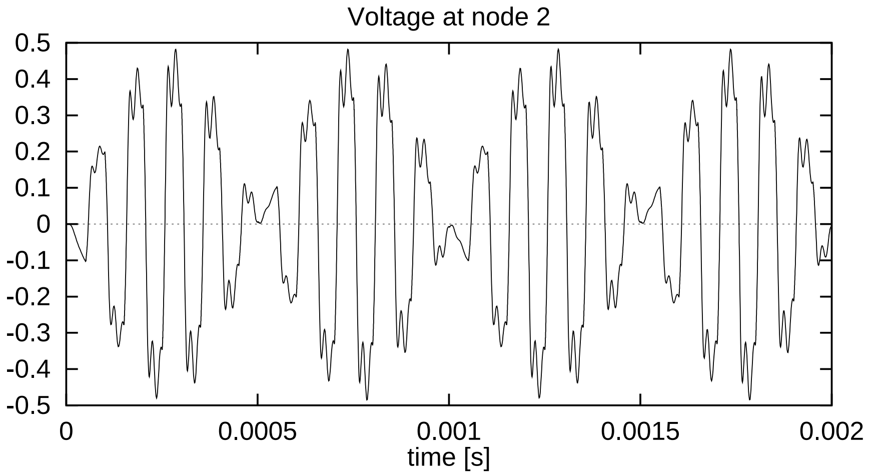

Figure 5.8: Voltage u_2

Considering (5.74), the voltages u_3, u_4, u_5, and u_6 represent the so-called index-2 variables. Figure 5.9 shows the result for u_3. The time points $5 \cdot 10^{-5}$, $10 \cdot 10^{-5}$, $15 \cdot 10^{-5}$, $20 \cdot 10^{-5}$, $25 \cdot 10^{-5}$, ... are critical, but the periodical behaviour over the long integration interval is kept. Figure 5.10 shows the result for a current, for I_2. Unfortunately, we do not know the exact solution. Therefore, we cannot investigate the errors in more detail.

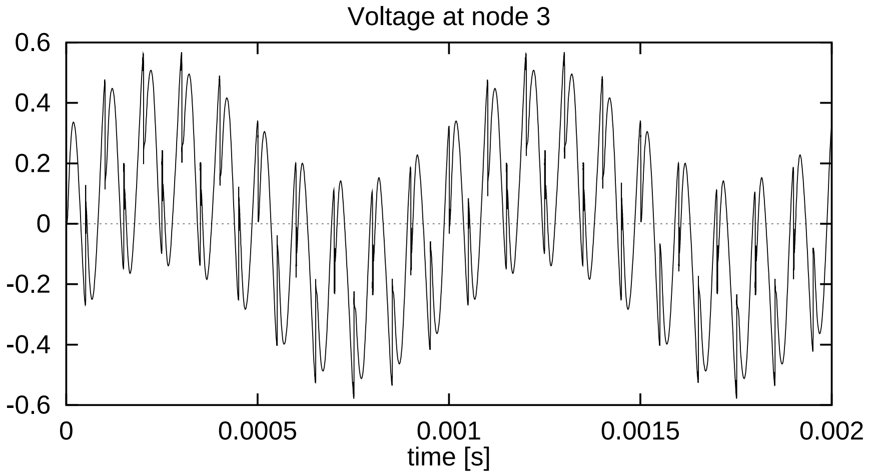

Figure 5.9: Voltage u_3

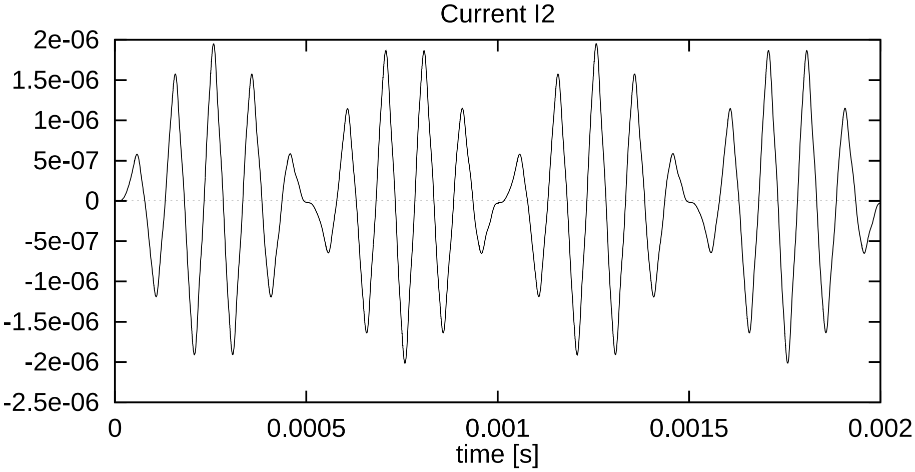

Figure 5.10: Current I_2

Summary

Differential-algebraic equations arise in numerous fields. In particular, modern modelling techniques in the circuit simulation like the modified nodal analysis lead to DAEs. The index of such DAEs is often low (≤ 2). The solution behaviour of differential algebraic equations depends essentially on the index of the system. Index-1 DAEs contain integration problems and algebraic problems. General index-2 DAEs include not only integration and algebraic problems, but also differentiation problems. Differentiation problems are ill-posed, i.e., small defects in the initial data may cause arbitrarily large defects in the solution data. Hence, a careful analysis of the behaviour of numerical solutions of index-2 DAEs is necessary. Primarily, different modelling techniques can provide DAEs of different indices.

We investigate two modern modelling techniques, the charge-oriented and the classical MNA. Both techniques are shown to lead to the same index if the capacitances of the circuit are all one-port capacitances. Often, this restriction can be fulfilled, since general capacitances can be interpreted as current sources whose element equations are formulated in a charge-oriented way (see e.g. the model of the MOSFET in the NAND-gate).

In Chapter 3, the numerical solvability is proved for a large class of index-2 differential algebraic equations satisfying the structural condition (3.39). All the circuit-examples with index-2 we know, e.g. the NAND-gate and the ring modulator, fulfil this condition. The question, whether all index-2 DAEs in this field have this property, is still open.

The BDF method is shown to be feasible (i.e., the nonlinear equations are uniquely solvable), and weakly unstable for index-2 DAEs if the defects arising from solving the nonlinear equations are sufficiently small. Thereby, the defects in the derivative-free part of the DAE have to be sufficiently small in relation to the stepsize. In the case of electric circuits, the derivative-free part consists of the characteristic equations of voltage sources, charges and fluxes as well as some linear combinations of the nodal equations, that can be

obtained by following all loops of capacitances of the network. The numerical results confirm the assertion that an error control based on the differential components only, i.e., on the "reliable" components only, works well for the network equation systems of index-2.

The presented simulations of the NAND-gate and the ring modulator demonstrate the possibility of a successful integration of index-2 circuit DAEs. They can be considered as model cases for handling other electric networks.

Appendix A

Basics from algebra and analysis

A fundamental relation between the spaces appearing at the tractability index and the choice of the corresponding projectors is given by the following lemma, which may be obtained directly from Theorem A.13. and Lemma A.14. in [GM86].

Lemma A.1 *Let $A_*, B_*, Q_* \in L(\mathbb{R}^m)$ be given, let Q_* be a projector onto $ker A_*$, i.e., $Q_*^2 = Q_*$, $im\, Q_* = ker\, A_*$. Denote*

$$S_* := \{z \in \mathbb{R}^m : B_* z \in im\, A_*\}.$$

Then the following conditions are equivalent:

(i) *The matrix $G_* := A_* + B_* Q_*$ is regular.*

(ii) $\mathbb{R}^m = S_* \oplus ker A_*$

(iii) $S_* \cap ker A_* = \{0\}$

If G_* is regular, then the relation

$$Q_{*s} = Q_* {G_*}^{-1} B_*$$

holds for the canonical projector Q_{*s} (canonical means: Q_{*s} projects onto $ker\, A_*$ along S_*).

Proof:
$(i) \rightarrow (ii)$ First, the space $\mathbb{R}^m$ can be described as $S_* + \ker A_*$ because

$$z = (I - Q_* {G_*}^{-1} B_*) z + Q_* {G_*}^{-1} B_* z =: z_1 + z_2 \qquad \text{(A.1)}$$

is satisfied for any $z \in \mathbb{R}^m$. Now, z_2 obviously lies in $\ker A_*$ since Q_* is a projector onto $\ker A_*$. For z_1 we obtain

$$B_* z_1 = (I - B_* Q_* {G_*}^{-1}) B_* z = A_* {G_*}^{-1} B_* z \in \operatorname{im} A_*,$$

i.e., $z_1 \in S_*$. It remains to show that

$$S_* \cap \ker A_* = \{0\}.$$

For that, let $x \in S_* \cap \ker A_*$. Then, $x = Q_* x$ holds and there is a $z \in \mathbb{R}^m$ such that

$$A_* z = B_* x = B_* Q_* x \quad \text{and , hence,} \quad G_*^{-1} A_* z = G_*^{-1} B_* Q_* x,$$

i.e.,

$$(I - Q_*) z = Q_* x, \quad \text{thus,} \quad 0 = Q_* x = x.$$

$(ii) \rightarrow (iii)$ This trivially holds per definition.

$(iii) \rightarrow (i)$ Let $x \in \mathbb{R}^m$ be chosen such that $G_* x = 0$, i.e.,

$$B_* Q_* x = -A_* x.$$

Hence, $Q_* x \in S_*$ is satisfied. On the other hand, $Q_* x$ lies in $\ker A_*$. Thus $x \in kerQ_*$ holds because of the assumption. That means $A_* x = 0$, consequently $x \in \operatorname{im} Q_*$. Now, $x = 0$ must be fulfilled, and G_* is regular.

Since partition (A.1) is unique, the latter assertion follows immediately.

□

The following two lemmas describe known facts from functional analysis (see [Die85]). We use them in the Chapters 3 and 4.

Lemma A.2 *Let E_1, E_2,..., E_n, F, ($n \in \mathcal{N}$, $n \geq 2$) be Banach spaces and f a C^1-function mapping an open subset A of*

$$E_1 \times E_2 \times \cdots \times E_n$$

into F. At the point

$$x^0 := (x_1^0, x_2^0, \ldots, x_n^0)$$

of A let $f(x^0) = 0$ be fulfilled, and let the partial derivative $D_1 f(x^0)$ be a linear homeomorphism of E_1 onto F. Then, there is a connected open neighbourhood U of

$$(x_2^0, \dots, x_n^0) \in E_2 \times \dots \times E_n$$

and a unique C^1-function u of U into E_1 in such a way that the equation

$$u(x_2^0, \dots, x_n^0) = x_1^0$$

is true, and the relations

$$(u(x_2, \dots, x_n), x_2, \dots, x_n) \in A, \quad f(u(x_2, \dots, x_n), x_2, \dots, x_n) = 0$$

are satisfied for all $(x_2, \dots, x_n) \in U$. Further, the inequality

$$\begin{aligned} &\|u(x_2, \dots, x_n) - u(x_2^0, \dots, x_n^0)\| \\ &\qquad \leq (2\|D_{x_2} u(x_2^0, \dots, x_n^0)\| + 1)\|x_2 - x_2^0\| \quad + \quad \dots \\ &\qquad\qquad + (2\|D_{x_n} u(x_2^0, \dots, x_n^0)\| + 1)\|x_n - x_n^0\| \end{aligned}$$

is valid for each $(x_2, \dots, x_n) \in U$.

Lemma A.3 *Let E, F be Banach spaces, U and V be open balls in E and F, respectively, with the center 0 and the radii ρ and α, respectively. Further, let v be a continuous function mapping $U \times V$ into F of such a kind that*

$$\|v(x, y_1) - v(x, y_2)\| \leq k \, \|y_1 - y_2\|$$

is satisfied for $x \in U$, $y_1 \in V$, $y_2 \in V$, $0 \leq k < 1$. Then, there exists a unique function f of U into V such that

$$f(x) = v(x, f(x)), \quad x \in U,$$

is true if

$$\|v(x, 0)\| \leq \rho(1 - k), \quad x \in U,$$

is fulfilled. The mapping f is continuous on U.

Bibliography

[AP91a] U. Ascher and L. Petzold. Stability of computational methods for constrained dynamics systems. *SIAM J. Sci. Stat. Comput.*, (1):95–120, 1991.

[AP91b] U. M. Ascher and L. R. Petzold. Projected implicit Runge-Kutta methods for differential-algebraic equations. *SIAM J. Numer. Anal.*, 28:1097–1120, 1991.

[AP92] U. M. Ascher and L. R. Petzold. Projected collocation methods for higher-order higher-index differential-algebraic equations. *J. Comput. Appl. Math.*, 43:243–259, 1992.

[Arn93] M. Arnold. Stability of numerical methods for differential-algebraic equations of higher index. *Appl. Numer. Math.*, 13:5–15, 1993.

[Arn95] M. Arnold. Applying BDF to quasilinear differential-algebraic equations of index 2: Perturbation analysis. In preparation, 1995.

[ASW87] M. Arnold, K. Strehmel, and R. Weiner. Small perturbations in differential-algebraic systems of index 2. Preprint 93/1, Univ. Rostock, Germany, 1987.

[BCP89] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *The Numerical Solution of Initial Value Problems in Ordinary Differential-Algebraic Equations.* North Holland Publishing Co., 1989.

[BE88] K. E. Brenan and B. E. Engquist. Backward differentiation approximations of nonlinear differential/algebraic systems. *Math. Comp.*, (51):659–676, S7–S16, 1988.

[BG86] R. Bulirsch and A. Gilg. Effiziente numerische Verfahren für die Simulation elektrischer Schaltungen. In H. Schwärtzel, editor, *Informatik in der Praxis*, pages 3–12. Springer Verlag, 1986.

[BP89] K.E. Brenan and L.R. Petzold. The numerical solution of higher index differential/algebraic equations by implicit methods. *SIAM J. Numer. Anal.*, 26:976–996, 1989.

[Bre83] K. E. Brenan. *Stability and Convergence of Difference Approximations for Higher Index Differential-Algebraic Systems with Applications in Trajectory Control.* PhD thesis, University of California, Los Angeles, 1983.

[Cam85] S. L. Campbell. The numerical solution of higher index linear time varying singular systems of differential equations. *SIAM J. Sci. Stat. Comput.*, (6):334–348, 1985.

[Cam86] S. L. Campbell. Index two linear time–varying system of differential equations. *Circuits Systems Signal Process.*, (5):97–107, 1986.

[CG93a] S. L. Campbell and C. W. Gear. The index of general nonlinear daes. Report, North Carolina State Univ., NC, U.S.A., 1993.

[CG93b] S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. Report, North Carolina State Univ., NC, U.S.A., 1993.

[CL75] L. O. Chua and Pen-Min Lin. *Computer-Aided Analysis of Electronic Circuits.* Prentice Hall, Englewood Cliffs, 1975.

[CL90] S.L. Campbell and B. Leimkuhler. Differentiation of constraints in differential-algebraic equations. In R. Deyo and E. Haug, editors, *NATO Advanced Research Workshop on Real – Time Integration Methods for Mechanical System Simulation.* Springer, Heidelberg, to appear 1990.

[Cla88] K. D. Clark. A structural form for higher index semistate equations I: Theory and application to circuit and control theory. *Linear Alg. Appl.*, 98:169–197, 1988.

[Den88] G. Denk. Die numerische Integration von Algebro-Differentialgleichungen bei der Simulation elektrischer Schaltkreise mit SPICE2. Technical Report TUM-M8809, Mathematisches Institut, TU München, 1988.

[Die85] J. Dieudonné. *Grundzüge der modernen Analysis*, volume 1. Deutscher Verlag der Wissenschaften, Berlin, 1985.

[DR89] G. Denk and P. Rentrop. Mathematical models in electric circuit simulation and their numerical treatment. Technical Report TUM-M8903, Mathematisches Institut, TU München, 1989.

[DR91] G. Denk and P. Rentrop. Mathematical models in electric circuit simulation and their numerical treatment. In *Teubner-Texte zur Mathematik*, volume 121, pages 305–316, 1991.

[FG94] U. Feldmann and M. Günther. The dae-index in electric circuit simulation. In I. Troch and F. Breitenecker, editors, *Proc. IMACS Symposium on Mathematical Modelling*, 4, pages 695–702, 1994.

[Fre95] S. Freude. Projizierende Defektkorrektur für Algebro-Differentialgleichungen mit dem Index 2, 1995. Humboldt-Univ. zu Berlin, Diplomarbeit.

[Füh88] C. Führer. *Differential-algebraische Gleichungssysteme in mechanischen Mehrkörpersystemen.* PhD thesis, Mathematisches Institut, Technische Universität München, 1988.

[FWZ+92] U. Feldmann, U. A. Wever, Q. Zheng, R. Schultz, and H. Wriedt. Algorithm for modern circuit simulation. *Archiv für Elektronik und Übertragungstechnik*, 46:274–285, 1992.

[Gan54] F. R. Gantmacher. *Teorija matrits.* Gosudarstv. Izdat. Techn.-Teor. Lit., Moskva, 1954.

[Gea90] C. W. Gear. Differential algebraic equations, indices, and integral algebraic equations. *SIAM J. Numer. Anal.*, 27(6):1527–1534, 1990.

[GGL85] C. W. Gear, G. K. Gupta, and B. J. Leimkuhler. Automatic integration of the Euler-Lagrange equations with constraints. *J. Comp. Appl. Math.*, 12,13:77–90, 1985.

[GHM92] E. Griepentrog, M. Hanke, and R. März. Towards a better understanding of differential algebraic equations. In E. Griepentrog, M. Hanke, and R. März, editors, *Berlin Seminar on Differential-Algebraic Equations, Fachbereich Mathematik, Humboldt-Univ. Berlin*, pages 2–13, 1992.

[GHP81] C. W. Gear, H. H. Hsu, and L. Petzold. Differential-algebraic equations revisited. In *Proc. Oberwolfach Conf. on Stiff Equations*, Bericht des Instituts für Geom. und Prakt. Math.; 9, Aachen, 1981. Rhein.-Westfälische Techn. Hochschule.

[GM86] E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment.* Teubner-Texte zur Mathematik No. 88. BSB B.G. Teubner Verlagsgesellschaft, Leipzig, 1986.

[GP84] C. W. Gear and L. R. Petzold. Ode methods for the solution of differential/algebraic systems. *SIAM J. Numer. Anal.*, 21:716–728, 1984.

[GR93] M. Günther and P. Rentrop. Multirate row methods and latency of electric circuits. *Appl. Numer. Math.*, 13:83–102, 1993.

[GR94] M. Günther and P. Rentrop. Suitable one-step methods for quasilinear-implicit odes. Technical Report TUM-M9405, Mathematisches Institut, TU München, 1994.

[Gri81] R. D. Grigorieff. Stabilität von Mehrschrittverfahren auf variablen Gittern. Preprint Reihe Mathematik 89, Techn. Univ. Berlin, 1981.

[Gri90] E. Griepentrog. The index of differential-algebraic equations and its significance for the circuit simulation. In R. E. Bank, R. Bulirsch, and K. Merten, editors, *Mathem. Modelling and Simulation of Electrical Circuits and Semiconductor Devices*, pages 11–26. Birkhäuser, Basel, 1990. ISNM 93.

[Gri91] E. Griepentrog. Index reduction methods for differential algebraic equations. Technical Report 91-12, Fachbereich Mathematik, Humboldt-Univ. zu Berlin, 1991.

[Gün95] M. Günther. *Ladungsorientierte Rosenbrock-Wanner-Methoden zur numerischen Simulation digitaler Schaltungen.* PhD thesis, Techn. Univ. München, 1995.

[Han89] B. Hansen. Comparing different concepts to treat differential-algebraic equations. Technical Report 220, Fachbereich Mathematik, Humboldt-Univ. zu Berlin, 1989.

[Han90] M. Hanke. On the asymptotic representation of a regularization approach to nonlinear semiexplicit higher index differential-algebraic equations. *IMA J. Appl. Math.*, 1990.

[Hin80] A. C. Hindmarsh. LSODE and LSODI, two new initial value ordinary differential equation solvers. *ACM–SIGNUM Newsletters*, 15:10–11, 1980.

[HLR89] E. Hairer, Ch. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods.* Lecture Notes in Mathematics Vol. 1409. Springer, Heidelberg, 1989.

[HNW87] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems.* Springer Series in Computational Mathematics 8. Springer-Verlag, Berlin, Heidelberg, 1987.

[Hor76] E.-H. Horneber. *Analyse nichtlinearer RLCÜ–Netzwerke mit Hilfe der gemischten Potentialfunktion mit einer systematischen Darstellung der Analyse nichtlinearer dynamischer Netzwerke.* PhD thesis, FB: Elektrotechnik, Univ. Kaiserslautern, 1976.

[HW91] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and differential-algebraic problems.* Springer Series in Computational Mathematics 14. Springer-Verlag, Berlin, Heidelberg, 1991.

[KRS92] W. Kampowsky, P. Rentrop, and W. Schmidt. Classification and numerical solution of electric circuits. *Mathematics for Industry*, pages 23–65, 1992.

[Lam91] R. Lamour. A well-posed shooting method for transferable dae's. *Numer. Math.*, 59(8):815–829, 1991.

[Lei86] B. J. Leimkuhler. Error estimates for differential-algebraic equations. Technical Report UIUCDCD-R-86-1287, Dept. of Computer Science Univ. of Illinois, 1986.

[Lei89] B. J. Leimkuhler. Some notes on perturbations of differential-algebraic equations. Technical report, Institute of Mathematics, Helsinki University of Technology, 1989.

[LP86] P. Lötstedt and L. Petzold. Numerical solution of nonlinear differential equations with algebraic constraints i: Convergence results for backward differentiation formulas. *Math. Comp.*, 46:491–516, 1986.

[LPG91] B. J. Leimkuhler, L. R. Petzold, and C. W. Gear. Approximation methods for the consistent initialization for differential-algebraic equations. *SIAM J. Numer. Anal.*, 28:205–226, 1991.

[Lub88] Ch. Lubich. h^2 extrapolation methods for differential-algebraic equations of index-2. Technical report, Universität Insbruck, Institut für Mathematik und Geometrie, A-6020 Insbruck, 1988.

[Mär89] R. März. Index-2 differential–algebraic equations. *Results in Mathematics*, pages 148–171, 1989.

[Mär90] R. März. Higher-index differential-algebraic equations: Analysis and numerical treatment. *Banach Center Publ.*, 24:199–222, 1990.

[Mär92a] R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, pages 141–198, 1992.

[Mär92b] R. März. On quasilinear index 2 differential algebraic equations. In E. Griepentrog, M. Hanke, and R. März, editors, *Berlin Seminar on Differential-Algebraic Equations, Fachbereich Mathematik, Humboldt-Univ. Berlin*, 1992.

[Mär94] R. März. Progress in handling differential algebraic equations. *Annals of Numerical Mathematics*, 1:279–292, 1994.

[Mär95] R. März. On linear differential-algebraic equations and linearizations. *APNUM*, 18:267–292, 1995.

[Mat87] W. Mathis. *Theorie nichtlinearer Netzwerke.* Springer Verlag Berlin Heidelberg NewYork, 1987.

[MT94] R. März and C. Tischendorf. Solving more general index 2 differential algebraic equations. *Comp. and Math. with Appl.*, 28(10-12):77–105, 1994.

[Pet82a] L. R. Petzold. Automatic selection of methods for solving stiff and nonstiff systems of ordinary differential equations. Technical Report SAND80-8230, Sandia National Laboratories, Livermore, 1982.

[Pet82b] L. R. Petzold. A description of DASSL: A Differential/Algebraic system solver. In *Proc. 10th IMACS World Congress, August 8-13 Montreal 1982*, 1982.

[PL86] L. R. Petzold and Lötstedt. Numerical solution of nonlinear differential equations with algebraic constraints ii: Practical implications. *SIAM J. Sci. Stat. Comput.*, (7):720–733, 1986.

[Rei90] S. Reich. *Beitrag zur Theorie der Algebrodifferentialgleichungen*. PhD thesis, Techn. Univ. Dresden, 1990.

[Rhe84] W. C. Rheinboldt. Differential-algebraic systems as differential equations on manifolds. *Math. Comp.*, 43:473–482, 1984.

[RR91] P. Rabier and W. Rheinboldt. A general existence and uniqueness theory for implicit differential-algebraic equations. *Differential and Integral Equations*, 4(3):563–582, 1991.

[SEYE81] R. F. Sincovec, A. M. Erisman, E. L. Yip, and M. A. Epton. Analysis of descriptor systems using numerical algorithms. *IEEE Trans. Automatic Control*, 26:139–147, 1981.

[SFR91] B. Simeon, C. Führer, and P. Rentrop. Differential-algebraic equations in vehicle system dynamics. *Surv. Math. Ind.*, 1:1–37, 1991.

[SH68] H. Shichman and D. A. Hodges. Insulated-gate field-effect transistor switching circuits. *IEEE J. Solid State Circuits*, SC-3:285–289, 1968.

[Soe89] G. Soederlind. Stiff differential equations. Technical Report V1.18, Carl-Cranz-Gesellschaft, D-8031 Wessling, 1989.

[Tis92] C. Tischendorf. Die BDF für nichtlineare Algebro-Differentialgleichungen vom Index 2, 1992. Humboldt-Univ. zu Berlin, Diplomarbeit.

[Tis94] C. Tischendorf. On the stability of solutions of autonomous index-1 tractable and quasilinear index-2 tractable DAEs. *Circuits Systems Signal Process.*, 13(2-3):139–154, 1994.

[Tis95] C. Tischendorf. Feasibility and stability behaviour of the BDF applied to index-2 differential algebraic equations. *ZAMM*, 75(12):927–946, 1995.

[Win94] R. Winkler. On simple impasse points and their numerical computations. Technical Report 94-15, Fachbereich Mathematik, Humboldt-Univ. zu Berlin, 1994.

[Wri88] H. Wriedt. Über Theorie und Numerik von Algebro-Differentialgleichungssystemen, 1988. Universität Hamburg, Diplomarbeit.