Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation

DISSERTATION

zur Erlangung des akademischen Grades
doctor rerum naturalium
(Dr. rer. nat.)
im Fach Mathematik

eingereicht an der
Mathematisch-Naturwissenschaftlichen Fakultät II
Humboldt-Universität zu Berlin

von

Frau Dipl.-Math. Diana Estévez Schwarz

geboren am 6.4.1972 in Berlin

Präsident der Humboldt-Universität zu Berlin
Prof. Dr. Dr. h.c. Hans Meyer

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät II
Prof. Dr. Bodo Krause

Gutachter/Gutachterin

1. Prof. Dr. R. März
2. Prof. Dr. L. R. Petzold
3. Prof. Dr. P. Rentrop

eingereicht am:	30. Mai 2000
Tag der mündlichen Prüfung	13. Juli 2000

Preface

This thesis results from my work in the BMBF-Project "Untersuchung der speziellen differential-algebraischen Struktur der Netzwerkgleichungen für die Schaltkreissimulation zur Entwicklung zuverlässiger und effizienter Simulationsverfahren" at the Humboldt-University of Berlin.

At this point, I would like to express my gratitude to my advisor Prof. Dr. Roswitha März for her committed support and valuable criticism, as well as for entrusting me with such an interesting task.

I am indebted to Dr. Caren Tischendorf and Dr. René Lamour for many fruitful discussions over the years. Dr. Caren Tischendorf also proof-read carefully this work. Further, I want to thank Dr. Uwe Feldmann of Infineon Technologies AG for the encouragement and suggestions given.

Berlin, July 2000

Diana Estévez Schwarz

Contents

Introduction

Differential algebraic equations (DAEs) are implicit ordinary differential equations (ODEs) of the form

$$f(x'(t), x(t), t) = 0, \tag{1}$$

$f : \mathcal{G}_f \to I\!R^n, \quad \mathcal{G}_f \subset I\!R^n \times I\!R^n \times I\!R$, where the partial Jacobian $f'_y(y, x, t)$ is singular. In fact, this means that (1) consists of coupled systems of differential equations and constraints. DAEs arise in various fields of applications such as the simulation of electric circuits, chemical reactions, vehicle dynamics and optimal control problems. In this thesis, we restrict ourselves to initial value problems (IVP).

The analytical and numerical solutions of (1) depend strongly on its structure and index. Roughly speaking, the index of a DAE is the measure of the deviation of a DAE from regular ODEs, i.e., from equations (1) with nonsingular $f'_y(y, x, t)$. DAEs have, among other things, the following two important properties (see e.g. [5],[25],[31]):

(i) Some components of the solution are determined by constraints. For IVPs, these constraints restrict the choice of initial values, since there is not a solution through every given initial value.

(ii) Higher index (≥ 2) DAEs do not only represent integration problems, but differentiation problems, too. This implies that some parts of the DAE must be differentiable sufficiently often. Moreover, depending on the structure, both differentiations and integrations may be intertwined in a complex manner.

As a consequence, only specific numerical methods should be used to approximate the solutions of DAEs. However, these numerical methods may fail in dependence of the structure of the DAEs, particularly if the index is greater than 2. Thus it is important to recognize the classes of problems for which numerical methods will work. Therefore, various (structural) forms of DAEs, as e.g. the Hessenberg form[1], have been considered.

Property (i) implicates that one of the difficult parts in solving DAEs numerically is to determine a consistent set of initial conditions in order to start the integration. We formulate the problem of computing consistent initial values as follows: Given some user defined guesses about initial values for the DAE, determine values for the variables and the derivatives of variables appearing in the DAE, in such a way that there exists a solution passing through them. In this connection, it has to be emphasized that initialization is important not only for beginning the integration but also for understanding how to handle the solution discontinuities that frequently occur in applications.

For the higher-index cases, the so-called hidden constraints appear if we derive a part of the equations of the DAE. This implies that the consistent values have to be chosen in such a way that not only the explicit equations of the DAE, but additionally these hidden constraints have to be fulfilled. Hence, the task is particularly difficult, since a proper characterization of the hidden constraints becomes necessary.

Up to now, the approaches from the literature to compute consistent initial values require either index or structural assumptions (e.g. the mentioned Hessenberg form), which are not always given in practice, or do not explicitly suppose any assumptions on the index and the structure, trying to cope with a very complex problem. In contrast, here we restrict ourselves to index-2 DAEs and analyse carefully the consequences of some weak structural assumptions. By doing so, we take advantage of the specific index-2 properties. As a consequence, the results are geared to the index-2 DAEs from applications in which we are interested. At this point it has to be emphasized that the mentioned weak structural assumptions comprise a class of DAEs that is much more general than the DAEs in Hessenberg form.

One current application, which is also a motivation for our study of systems of DAEs, is the circuit simulation. Due to the fact that the models contain

[1] For a definition see Section 2.6.

functions that are not very smooth, the tractability index, which requires only weak smoothness assumptions on the variables and on the input functions, becomes specially adequate. Another characteristic of circuit simulation is that the equations are generated automatically, because the dimensions are often very large. Since it has turned out that these equations are not in Hessenberg form, hitherto it has not been clear how to identify the pieces of information and model structures that are valuable for the mathematical characterization. In the preliminary work to this thesis [15],[12],[11], it was shown that the nonlinear DAEs obtained by modified nodal analysis (MNA) in circuit simulation present some new interesting structural properties of index-2 DAEs that permit a relatively easy computation of consistent initial values. In practice these results are of special interest, since they allow the location of the mathematically critical model elements by analyzing the network graph. Consequently, in spite of the large dimensions, the relevant properties of the equations can be checked very fast.

In this thesis, a general description of weak and yet helpful structural properties will be given in terms of the spaces and projectors associated with the tractability-index. Starting from these structural properties, a specification of how to take advantage of them for consistent initialization will be presented. In order to achieve a rounded form of the exposition, some of the results for circuit simulation developed in [15],[12],[11] will be reconsidered in connection with the general description.

This thesis is organized as follows:

- Chapter 1 gives a short introduction to DAEs and the notion of their index. Some index concepts from literature are introduced.

- In Chapter 2 some specific structural properties are analysed. Subsequently, an expression for the hidden constraints of index-2 DAEs is deduced, showing that substituting them for a part of the original equations gives place to an index reduction. Based on this expression, it is indicated how to set up a nonlinear system whose solution provides a consistent initial value. Referring to this, careful attention is paid to the simplifications arising from additional structural assumptions.

- Chapter 3 deals with the application of the results from Chapter 2 to circuit simulation (cf. [11]). Thus, a recapitulation of the results that partly have been developed in [15] previously is given. Moreover, a graph-theoretical description of the critical parts of the model (cf. [12]) is presented and some realization specifics are discussed.

- Finally, in the Appendix we state some well-known facts, provide the equations and details of the example discussed in Chapter 2, and give an overview of the assumptions from the Chapters 1 and 2 as well as of some notations from Chapter 3.

Notations and conventions

ODE	–	Ordinary Differential Equation
DAE	–	Differential Algebraic Equation
IVP	–	Initial Value Problem
MNA	–	Modified Nodal Analysis
im A	–	image space of the operator A
$\ker A$	–	kernel of the operator A
Q projects onto R	–	$Q^2 = Q,$ im $Q = R$
W projects along R	–	$W^2 = W,$ $\ker W = R$
$f : X \to Y$ is smooth	–	f is continuously differentiable

Chapter 1

The Index of Differential Algebraic Equations

Differential algebraic equations (DAEs) differ in several aspects from regular ODEs. All the index concepts from the literature precisely give a kind of measure of the deviation of a DAE from regular ODEs. Indeed, the index, in all its variants, measures in some sense how the solution of the DAE depends on the describing equations, initial data, and forcing functions. Moreover, most of these variants coincide when considering linear time-independent DAEs and trace back to the Kronecker canonical form [18].

We briefly introduce two well-known index concepts from the literature, the differential index and the perturbation index, before defining the tractability index, which will be considered in the forthcoming chapters[1].

1.1 The Differential Index

Roughly speaking, the differential index (see e.g. [21],[19],[22],[5],[20],[9],[2])[2] of a DAE is the number of differentiations that are necessary to transform the

[1] Another important index concept is the geometrical index, which describes the behaviour of DAEs as the behaviour of regular ODEs on a constraint manifold (see e.g. [53],[50]).

[2] Actually, there are several slightly different variants of the definition of the differential index.

DAE into a regular ODE. This index concept is often used in the literature.

Definition 1.1.1 *(cf. e.g. [5]) The differential index ν of the general nonlinear, sufficiently smooth DAE*

$$f(x', x, t) = 0 \tag{1.1}$$

is the smallest ν such that

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

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

Fortunately, the structure of the DAEs is frequently such that it will not be necessary to derive the whole function f. Often it suffices to derive the obvious constraints[3] in the index 1 case and, additionally, the hidden constraints in the index 2 case.

The following example illustrates that for nonlinear DAEs the index is a local property.

Example 1.1.2 *[2] Consider*

$$\begin{aligned} x_1' &= x_3, &(1.2)\\ 0 &= x_2(1 - x_2), &(1.3)\\ 0 &= x_1 x_2 + x_3(1 - x_2) - t, &(1.4) \end{aligned}$$

$x_i : \mathcal{I}_f \to I\!R$. Obviously, (1.3) has two solutions $x_2 = 0$ and $x_2 = 1$.

1. *Considering $x_2 = 0$, the third equation leads to $x_3 = t$, and it is easy to see that the differential index is 1.*
2. *Considering $x_2 = 1$, the third equation leads to $x_1 = t$. Then the system has index 2.*

[3] More will be said about constraints in Chapter 2.

1.2 The Perturbation Index

The perturbation index, which was introduced in [30],[31], interprets the index as a measure of sensitivity of the solution with respect to perturbations of the given problem.

Definition 1.2.1 *[30] The equation*

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

has perturbation index m *along a solution* $x_*(t)$ *on a closed interval* $I = [a, b]$, *if* m *is the smallest integer such that, for all functions* $x(t)$ *having a defect*

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

there exists on $I = [a, b]$ *an estimate*

$$\begin{aligned} \| x(t) - x_*(t) \| \leq C \Big(\| x(a) - x_*(a) \| &+ \max_{a \leq \xi \leq t} \| q(\xi) \| + \ldots \\ &+ \max_{a \leq \xi \leq t} \| q^{(m-1)}(\xi) \| \Big) \end{aligned}$$

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

Note that the perturbation index concept requires information about the solution of the DAE. The following example illustrates that the perturbation index may differ from the differential index.

Example 1.2.2 *[30] Consider*

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

$x_i : \mathcal{I}_f \to I\!R$. *If we look at the perturbed system*

$$\begin{aligned} x_1' - x_3 x_2' + x_2 x_3' &= 0, && (1.5) \\ x_2 &= \epsilon \sin \omega t, && (1.6) \\ x_3 &= \epsilon \cos \omega t, && (1.7) \end{aligned}$$

we observe that inserting (1.6) and (1.7) into (1.5) leads to $x_1' = \epsilon^2 \omega$, *i.e., the perturbation index is 2, whereas the differential index is 1.*

1.3 The Tractability Index

The tractability index (see e.g. [25],[39],[40],[41],[44]) orientates on the linearization of a DAE and requires only weak smoothness conditions. Hence, when considering this index, the required smoothness assumptions are specified.

Concretely, we focus on DAEs[4]

$$f(x'(t), x(t), t) = 0, \quad f : I\!R^n \times \mathcal{D}_f \times \mathcal{I}_f \to I\!R^n, \tag{1.8}$$

where $\mathcal{I}_f$ is an open interval of $I\!R$, and $\mathcal{D}_f$ is an open subset of $I\!R^n$. The partial derivative $f'_y(y, x, t)$ is singular and has constant rank for all the triples (y, x, t) of its definition domain $\mathcal{G}_f := I\!R^n \times \mathcal{D}_f \times \mathcal{I}_f$.

For linear time varying DAEs

$$A(t)x'(t) + B(t)x(t) = q(t)$$

with continuous matrix functions $A(\cdot)$, $B(\cdot)$, and continuous functions $q(\cdot)$, the tractability-index is defined considering a matrix chain based on the matrix pencil $(A(\cdot), B(\cdot))$. For nonlinear systems its definition is based on linearization. Roughly speaking, the aim is (cf. [44])

"The DAE (1.8) has index μ if the linearized DAE has it, and vice versa".

In fact, it can be shown that the definition we introduce below for nonlinear systems satisfies this claim (cf. [41]). For a better understanding, we will consider first the definition for linear systems and introduce, afterwards, the definition for nonlinear systems of index 1 and 2. For higher index nonlinear DAEs, many questions concerning an adequate definition remain open.

[4] In the following, we will explicitly write the argument t for $x'(t)$ and $x(t)$ when considering the variables of the DAE, in order to distinguish them from points y and x. For simplicity reasons, this distinction will not be maintained when considering examples and the applications in Chapter 3.

1.3.1 Linear DAEs

We consider linear time-dependent differential-algebraic equations, i.e. , equations of the form

$$A(t)x'(t) + B(t)x(t) = q(t), \;\; t \in \mathcal{I}_f, \;\; x(t) \in I\!R^n, \tag{1.9}$$

where $A(t)$ is singular and has constant rank on $\mathcal{I}_f$.
Observe that, if $N(t) := \ker A(t)$ depends smoothly[5] on t, $Q(t)$ is a smooth projector onto $N(t)$ and $P(t) := I - Q(t)$, then it holds

$$A(t)x'(t) = A(t)\{(Px)'(t) - P'(t)x(t)\}, \tag{1.10}$$

i.e., (1.9) involves the derivative of $(Px)(t) := P(t)x(t)$, but the derivative of the nullspace component $(Qx)(t)$ is not involved at all. Therefore, solutions of (1.9) lie in

$$C_N^1(\mathcal{I}_f, I\!R^n) := \Big\{ x \in C(\mathcal{I}_f, I\!R^n) : Px \in C^1(\mathcal{I}_f, I\!R^n) \Big\}. \tag{1.11}$$

Notice that, if $W_0(t)$ denotes a projector along im $A(t)$, then all solutions of (1.9) lie in

$$M_0(t) := \{x \in I\!R^n : W_0(t)(B(t)x - q(t)) = 0\}.$$

Hence, let us introduce the space

$$S(t) := \{z \in I\!R^n : W_0(t)B(t)z = 0\},$$

i.e., each solution of the homogeneous equation satisfies $x(t) \in S(t)$, $t \in \mathcal{I}_f$.

Definition 1.3.1 *If $A(t)$ is singular and has constant rank in $\mathcal{I} \subseteq \mathcal{I}_f$, then (1.9) is index-1 tractable on $\mathcal{I}$*

$$\begin{aligned}
&\Longleftrightarrow N(t) \cap S(t) = \{0\},\\
&\Longleftrightarrow N(t) \oplus S(t) = I\!R^n,\\
&\Longleftrightarrow G_1(t) := A(t) + B(t)Q(t) \;\; \textit{is nonsingular},\\
&\Longleftrightarrow A_1(t) := G_1(t)(I - P(t)P'(t)Q(t)) \;\; \textit{is nonsingular},
\end{aligned}$$

for all $t \in \mathcal{I}$.

[5] cf. Definition 4.1.4, Appendix.

Suppose now that the DAE we consider is not index-1 tractable. Then let $W_1(t)$ be a projector along im $G_1(t) = \text{im } A_1(t)$. The relevant spaces on this level are

$$S_1(t) := \{z \in I\!R^n : W_1(t)B(t)P(t)z = 0\}$$

and

$$N_1(t) := \ker A_1(t).$$

We denote by $Q_1(t)$ a projector onto $N_1(t)$ and $P_1(t) := I - Q_1(t)$.

Definition 1.3.2 *If (1.9) is not index-1 tractable, $N_1(t)$ is smooth and $dim N(t) \cap S(t)$ is constant on $\mathcal{I} \subseteq \mathcal{I}_f$, then (1.9) is index-2 tractable on $\mathcal{I}$*

$$\begin{aligned}
&\Longleftrightarrow \quad N_1(t) \cap S_1(t) = \{0\},\\
&\Longleftrightarrow \quad N_1(t) \oplus S_1(t) = I\!R^n,\\
&\Longleftrightarrow \quad G_2(t) := A_1(t) + B(t)P(t)Q_1(t) \ \textit{ is nonsingular,}\\
&\Longleftrightarrow \quad A_2(t) := G_2(t)\left(I - P_1(t)(PP_1)'(t)P(t)Q_1(t)\right) \ \textit{ is nonsingular,}
\end{aligned}$$

for all $t \in \mathcal{I}$.

Remark 1.3.3 *Note that the index definitions introduced above do not depend on the special choice of the different projectors and that the equivalences hold due to Lemma 4.1.3 from the Appendix.*

For index-2 DAEs, $N_1(t) \oplus S_1(t) = I\!R^n$ implies that there exists a projector $Q_{1S}(t)$ fulfilling im $Q_{1S}(t) = N_1(t)$ and $\ker Q_{1S}(t) = S_1(t)$, called the canonical projector. Recall further that this projector is given by $Q_{1S}(t) := Q_1(t)G_2^{-1}(t)B(t)P(t)$ if $Q_1(t)$ is an arbitrary projector onto $N_1(t)$. In the following, we will always consider that $Q_1(t)$ is the canonical projector. Thus, due to $N(t) \subseteq S_1(t)$ it always holds that

$$Q_1(t)Q(t) = 0. \tag{1.12}$$

In order to illustrate the spaces and projectors introduced above, let us consider the following example.

Example 1.3.4 *Consider*

$$\begin{aligned} x_1' + x_1 + x_2 &= q_1, \\ x_2' + x_3 + x_4 &= q_2, \\ x_2 &= q_3, \\ x_4 &= q_4, \end{aligned}$$

$x_i, q_i : \mathcal{I}_f \to I\!R$. *Straight forward computation shows:*

$$N = \operatorname{im} Q = \operatorname{im} \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}, \quad \operatorname{im} A = \ker W_0 = \ker \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 1 \end{pmatrix},$$

$$N \cap S = \operatorname{im} \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 0 \end{pmatrix}, \; PQ_1 = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}, \; PP_1 = \begin{pmatrix} 1 & & & \\ & 0 & & \\ & & 0 & \\ & & & 0 \end{pmatrix}.$$

Thus, we recognize that PP_1x, corresponding to x_1, represents the variable that is determined by the inherent regular ODE. PQ_1x, corresponding to x_2, represents the component that appears in dynamic form but is determined by a derivative-free equation. The $N \cap S$-component, corresponding to x_3, is determined by an inherent differentiation. Finally, x_4 is simply determined by a derivative-free equation, whereas it does not appear in dynamic form.

Let us observe that the definitions can be continued for higher index DAEs. In [39],[41] the following matrix chain is introduced in order to characterize the index of the DAE (1.9):

$$\begin{aligned} A_0 &:= A, \\ B_0 &:= B - AP_0', \\ A_{i+1} &:= A_i + B_iQ_i, \\ B_{i+1} &:= (B_i - A_{i+1}(P_0P_1 \dots P_{i+1})'P_0P_1 \dots P_{i-1})P_i, \end{aligned}$$

where Q_i is defined to be a projector onto $N_i := \ker A_i$, and $P_i := I - Q_i$ and the arguments are dropped for the sake of simplicity[6]. Further, the projectors Q_i are chosen in such a way that $Q_jQ_i = 0$ is true for $j > i$.

[6] Note that this definition means $Q_0 = Q$ and $P_0 = P$.

Definition 1.3.5 *The DAE (1.9) is said to be index-μ tractable on $\mathcal{I}$ if all matrices $A_j(t)$, $t \in \mathcal{I}$, $j = 0, \ldots, \mu - 1$, within the above chain are singular with smooth null spaces, and $A_\mu(t)$ remains nonsingular on $\mathcal{I}$.*

1.3.2 Nonlinear DAEs

The definition of the tractability-index for nonlinear systems

$$f(x'(t), x(t), t) = 0 \tag{1.13}$$

is based on an analogous chain of subspaces, projectors and matrices, using the Jacobians $f'_y(y, x, t)$ and $f'_x(y, x, t)$ point-wise instead of $A(t)$ and $B(t)$ (cf. [25],[41],[44]). To this end, we suppose that $f : \mathcal{G}_f \to I\!R^n$, $\mathcal{G}_f = I\!R^n \times \mathcal{D}_f \times \mathcal{I}_f$, is a continuous function and that $f'_y(y, x, t)$, $f'_x(y, x, t) \in L(I\!R^n)$ exist for all $(y, x, t) \in \mathcal{G}_f$, and f'_y, $f'_x \in C(\mathcal{G}_f, I\!R^n)$.

We focus on the quasilinear DAEs[7]

$$A(x(t), t)x'(t) + b(x(t), t) = 0. \tag{1.14}$$

Note that if the coefficient matrix $A(x, t)$ is nonsingular, (1.14) represents an implicitly regular ODE. But we are interested in the case of $A(x, t)$ remaining singular and assume that[8]

$$\mathbf{A1}: \quad N(t) := \ker A(x, t), \quad \text{im } A(x, t) \quad \text{depend smoothly on } t,$$

and do not depend on x for $(x, t) \in \mathcal{D}_f \times \mathcal{I}_f$. For a proper analysis of these systems we define the smooth projectors[9] $Q(t)$ onto $N(t)$, $P(t) := I - Q(t)$, and $W_0(t)$ along im $A(x, t)$.

Since

$$A(x, t) = A(x, t)P(t), \qquad (x, t) \in \mathcal{D}_f \times \mathcal{I}_f, \tag{1.15}$$

[7] Note that index-1 tractable DAEs can also be defined even if they do not present a quasilinear structure [25]. Since we will restrict our considerations to quasilinear DAEs in the following, for reasons of uniformity we preferred to introduce the index-1 concept also for quasilinear DAEs only.

[8] Observe that, by this assumption, we precisely exclude Example 1.2.2. To consider problems of this kind, see Remark 1.3.8, 4.

[9] cf. Definition 4.1.4, Appendix.

(1.14) may be rewritten as

$$A(x(t),t)((Px)'(t) - P'(t)x(t)) + b(x(t),t) = 0, \tag{1.16}$$

and hence, the function space which the solution of (1.14) should belong to again appears to be C^1_N (see (1.11)).

Because of (1.14) $f'_y = A(x,t)$ holds, and for $B := f'_x$ we have

$$B(y,x,t) = [A(x,t)y]'_x + b'_x(x,t).$$

Notice now that all solutions of (1.14) lie in

$$M_0(t) := \{x \in \mathcal{D}_f : W_0(t)b(x,t) = 0\}. \tag{1.17}$$

Moreover, the space S, which is closely related to the tangent space of $M_0(t)$, is given by

$$S(x,t) := \{z \in I\!R^n : W_0(t)B(y,x,t)z = 0\} \underset{(\mathbf{A1})}{=} \{z \in I\!R^n : W_0(t)b'_x(x,t)z = 0\}.$$

Definition 1.3.6 *If $A(x,t)$ is singular and has constant rank, then (1.14) is said to be index-1 tractable on open $\mathcal{G} \subseteq \mathcal{G}_f$ if*

$$\begin{aligned}
&\Longleftrightarrow N(t) \cap S(x,t) = \{0\},\\
&\Longleftrightarrow N(t) \oplus S(x,t) = I\!R^n,\\
&\Longleftrightarrow G_1(y,x,t) := A(x,t) + B(y,x,t)Q(t) \ \textit{is nonsingular },\\
&\Longleftrightarrow A_1(y,x,t) := G_1(y,x,t)(I - P(t)P'(t)Q(t)) \ \textit{is nonsingular },
\end{aligned}$$

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

Let us focus on the index-2 case. Suppose that $A_1(y,x,t)$ is singular and let $W_1(y,x,t)$ be a projector along im $G_1(y,x,t) =$ im $A_1(y,x,t)$. The relevant spaces on this level are

$$S_1(y,x,t) := \{z \in I\!R^n : W_1(y,x,t)B(y,x,t)P(t)z = 0\}$$

and

$$N_1(y,x,t) := \ker A_1(y,x,t).$$

Denote, analogously as in the linear case, by $Q_1(y,x,t)$ a projector onto $N_1(y,x,t)$ and $P_1(y,x,t) := I - Q_1(y,x,t)$.

Definition 1.3.7 *(1.14) is said to be index-2 tractable on open $\mathcal{G} \subseteq \mathcal{G}_f$ if $G_1(y,x,t)$ is singular on $\mathcal{G}$, $dim N(t) \cap S(x,t)$ is constant on $\mathcal{G}$, and*

$$\begin{array}{ll} \Longleftrightarrow & N_1(y,x,t) \cap S_1(y,x,t) = \{0\}, \\ \Longleftrightarrow & N_1(y,x,t) \oplus S_1(y,x,t) = I\!R^n, \\ \Longleftrightarrow & G_2(y,x,t) := A_1(y,x,t) + B(y,x,t)P(t)Q_1(y,x,t) \text{ is nonsingular} \end{array}$$

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

Remark 1.3.8 *1. Note again that the index definitions introduced above do not depend on the special choice of the different projectors and that the equivalences hold due to Lemma 4.1.3 from the Appendix.*

2. Observe that the smoothness assumptions from Definition 1.3.2 and from Definition 1.3.7 for linear DAEs do not coincide, since for the latter we did not make assumptions on the smoothness of N_1. In fact, the proper smoothness requirements are still a current matter of research [45].

3. In practice, since for nonlinear DAEs P_1 may depend on the solution, we do not consider the corresponding expression for A_2, because it would involve the term $\frac{d}{dt}(P(t)P_1(y,x,t))$, which is difficult to handle. This fact also leads to difficulties for a definition of a tractability index higher than 2. Indeed, for arbitrary nonlinear DAEs, many questions remain open concerning how to take into account the different rotating subspaces appropriately.

*4. If $\ker A(x,t)$ depends on x (i.e., **A1** is not fulfilled), then the tractability-index should be defined considering the enlarged system (cf. [42],[44]), which contains $2n$ equations:*

$$x'(t) - y(t) = 0, \tag{1.18}$$

$$f(y(t), x(t), t) = 0. \tag{1.19}$$

This system has semi-explicit form and a constant leading nullspace. Observe that the enlarged system corresponding to Example 1.2.2 is index-2 tractable.

Analogously as for linear DAEs, consider the canonical projector $Q_{1S}(y,x,t)$ fulfilling $\operatorname{im} Q_{1S}(y,x,t) = N_1(y,x,t)$ and $\ker Q_{1S}(y,x,t) = S_1(y,x,t)$. Recall that this projector is given by

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

if $Q_1(y,x,t)$ is an arbitrary projector onto $N_1(y,x,t)$. In the following, we will always assume that $Q_1(y,x,t)$ is the canonical projector and that again due to $N(t) \subseteq S_1(y,x,t)$ it always holds that

$$Q_1(y,x,t)Q(t) = 0. \tag{1.20}$$

Let us emphasize now the importance of the assumption of $(y,x,t) \in \mathcal{G}$, $\mathcal{G}$ open, in the above definitions. This assumption is supposed to be given, since we aim at numerical computations. On the one hand, for the 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. With the help of the tractability-concept, a detailed analysis of perturbed IVP leads to results concerning the numerical solvability of DAEs. In fact, for $\mu = 1,2$ it turned out that a DAE satisfying certain structural conditions has the perturbation index μ if it is index-μ tractable. For a detailed discussion see e.g. [41],[57],[44]. On the other hand, algorithms for computing a consistent initialization that involve expressions of projectors (e.g. the one presented in [14] and the one from Chapter 2) are based on the assumption that these expressions hold in a neighbourhood of the values we are interested in. If this is not the case, the equations should better be reformulated by means of analytical transformations before starting numerical computations. We illustrate the problem by means of the following two examples.

Example 1.3.9 *Consider*

$$\begin{aligned} x_1' + x_2' + x_3 &= 1, \\ x_2 x_3 &= 1, \\ x_1 + (x_2^2 + x_2)x_3 &= 0, \end{aligned}$$

$x_i : \mathcal{I}_f \to I\!R$. The relevant elements of the matrix chain read

$$\begin{aligned} A &= \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad Q = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \\ B(x,t) &= \begin{pmatrix} 0 & 0 & 1 \\ 0 & x_3 & x_2 \\ 1 & (2x_2+1)x_3 & x_2^2+x_2 \end{pmatrix}, \\ A_1(x,t) &= \begin{pmatrix} 1 & 1 & 1 \\ -x_3 & 0 & x_2 \\ 1-(2x_2+1)x_3 & 0 & x_2^2+x_2 \end{pmatrix}. \end{aligned}$$

Observe that the matrix A_1 is singular for (y,x,t) fulfilling $x_2x_3 = 1$, but nonsingular at points from an arbitrary small neighbourhood of a solution. Note that in such a case the tractability-index is not defined, since we cannot find an open $\mathcal{G}$. Observe further, that this example has differential index 2, but that if we slightly perturb the second equation, then the differential index becomes 1.

Indeed, a numerical approach to compute consistent initial values may fail when considering such an example. For this example, the algorithm described in [14] fails in practice, since the index switches in the neighbourhood of the solution.

Moreover, if we transform the equations analytically into

$$\begin{aligned} x_1' + x_2' + x_3 &= 1, \\ x_2x_3 &= 1, \\ x_1 + x_2 + 1 &= 0, \end{aligned}$$

then an open $\mathcal{G}$ can be found, the tractability-index is defined and is 2, and also the algorithm from [14] works.

Remark 1.3.10 *Consider again example 1.1.2. In this case, we would obtain*

$$A_1(x,t) = \begin{pmatrix} 1 & 0 & -1 \\ 0 & 1-2x_2 & 0 \\ 0 & x_1 - x_3 & 1-x_2 \end{pmatrix}.$$

Hence, A_1 is singular for $x_2 = 1$ and nonsingular for $x_2 = 0$. The tractability-index is not defined in the first case, since we cannot find an open $\mathcal{G}$.

1.3.3 Solvability Results

Under the exposed assumptions, the following existence and uniqueness theorem for index-1 tractable DAEs holds.

Theorem 1.3.11 *If (1.14) is index-1 tractable on $\mathcal{G} \subseteq \mathcal{G}_f$, then for $x_0 \in M_0(t_0)$,*

$$M_0(t) := \{x \in \mathcal{D} : W_0(t)b(x,t) = 0\},$$

$t_0 \in \mathcal{I}$, there exists a locally unique solution $x(\cdot) : \mathcal{I} \to I\!R^n$ in C^1_N with $x(t_0) = x_0$.

Proof: (cf.[25],[41]). Note that

$$M_0(t) = \{x \in \mathcal{D} : \exists y \quad A(x,t)y + b(x,t) = 0\}.$$

Consider $(y,x,t) \in \mathcal{G}$. Introducing the new variables

$$\begin{aligned} w &:= P(t)y + Q(t)x, \\ u &:= P(t)x, \end{aligned}$$

we rewrite

$$A(x,t)y + b(x,t) = A(u + Q(t)w, t)w + b(u + Q(t)w, t)$$

and define $\tilde{f}$ by

$$\tilde{f}(w,u,t) := A(u + Q(t)w, t)w + b(u + Q(t)w, t).$$

Observe that for $\tilde{f}$ at $w_0 = P(t_0)y_0 + Q(t_0)x_0$, $u_0 = P(t_0)x_0$, while y_0 is chosen according to $x_0 \in M_0(t_0)$, it holds

$$\tilde{f}(w_0,u_0,t_0) = 0 \quad \text{and} \quad \tilde{f}'_w(w_0,u_0,t_0) = G_1(y_0,x_0,t_0),$$

i.e., $\tilde{f}'_w(w_0,u_0,t_0)$ is nonsingular. Due to the Implicit Function Theorem in a small neighbourhood there exists a uniquely defined continuous function

$w(u,t)$ with continuous partial Jacobian $w'_u(u,t)$ satisfying $0 = \tilde{f}(w,u,t) = A(u+Q(t)w,t)w + b(u+Q(t)w,t)$.
Let now u be the locally unique solution of the regular initial value problem

$$\begin{aligned} u'(t) - P'(t)u(t) &= P(t)(I+P'(t))w(u(t),t), && (1.21)\\ u(t_0) &= P(t_0)x_0. && (1.22) \end{aligned}$$

Realize that for the solution of (1.21)-(1.22) it holds that $Q(t)u(t) = 0$. This can be verified considering the following regular initial value problem, which results when multiplying (1.21) and (1.22) by Q:

$$\begin{aligned} Q(t)u'(t) + Q'(t)u(t) - Q'(t)Q(t)u(t) &= 0,\\ Q(t_0)u(t_0) &= 0. \end{aligned}$$

Since for $\alpha(t) = Q(t)u(t)$ this implies $\alpha'(t) - Q'(t)\alpha(t) = 0$ and $\alpha(t_0) = 0$, the function α vanishes identically, which implies $u(t) = P(t)u(t)$.

Then, $x(t) = u(t) + Q(t)w(u(t),t)$ belongs to C^1_N because $P(t)x(t) = u(t)$ is continuously differentiable, whereas the part $Q(t)w(u(t),t)$ depends only continuously on t in general.
Hence, such an $x(t)$ is the C^1_N solution passing through (x_0,t_0) since

$$\begin{aligned} 0 &= \tilde{f}(w(u(t),t),u(t),t)\\ &= A(u(t)+Q(t)w(u(t),t),t)w(u(t),t) + b(u(t)+Q(t)w(u(t),t),t)\\ &\underset{(1.21)}{=} A(u(t)+Q(t)w(u(t),t),t)[u'(t) - P'(t)u(t) - P(t)P'(t)w(u(t),t)]\\ &\quad +b(u(t)+Q(t)w(u(t),t),t)\\ &\underset{PP'P=0}{=} A(u(t)+Q(t)w(u(t),t),t)[u'(t) - P'(t)[u(t)+Q(t)w(u(t),t)]]\\ &\quad +b(u(t)+Q(t)w(u(t),t),t)\\ &= A(x(t),t)((Px)'(t) - P'(t)x(t)) + b(x(t),t). \end{aligned}$$

q.e.d.

In Section 2.4 we will see how to obtain solvability results for index-2 tractable DAEs applying Theorem 1.3.11 to a corresponding (reduced) index-1 tractable DAE.

Chapter 2

Consistent Initial Values for DAEs

2.1 Introduction

Roughly speaking, the problem of determining consistent initial values for differential-algebraic equations (DAEs) can be described as follows. For ordinary differential equations, initial values have to be prescribed for all variables to determine a unique solution. However, differential-algebraic equations consist of differential equations coupled with derivative-free equations, commonly referred to as constraints[1]. Hence, not all components appear in dynamic form. Indeed, some of them are precisely determined by the constraints. Thus, no initial values can be prescribed for them.

According to ODE theory, we define for DAEs:

Definition 2.1.1 *A vector $x_0 \in I\!R^n$ is a consistent initial value of (1.14) if there exists a solution of (1.14) that fulfils $x(t_0) = x_0$.*

In practice, we are also interested in the corresponding values of the derivatives appearing in the DAE. Due to (1.15), the following definition will characterize these values properly.

Definition 2.1.2 *A vector $(x_0, P(t_0)y_0)$ is a consistent initialization of (1.14) fulfilling* **A1** *if x_0 is a consistent initial value and $(x_0, P(t_0)y_0)$ fulfils the equation $A(x_0,t_0)P(t_0)y_0 + b(x_0,t_0) = 0$.*

[1] In the literature, they are often referred to as algebraic equations.

Note that the singularity of $A(x,t)$ implies that (1.14) contains some derivative-free equations, which we will denote by explicit constraints. A consistent initialization has to fulfil precisely those equations. Moreover, the differentiation of these explicit constraints may lead to further derivative-free equations, called hidden constraints, which a consistent initialization has to fulfil, too. This occurs for higher index (≥ 2) DAEs. Hence, in general, the computation of consistent initial values may become a really hard task. We briefly resume some approaches from the literature in Section 2.2. In this thesis, we will restrict ourselves to index-2 DAEs.

Remark 2.1.3 *In the index-1 case, Theorem 1.3.11 implies that the set of consistent initial values is given by $M_0(t)$.*

For the index-2 case the consistent initial values have to lie in a subset

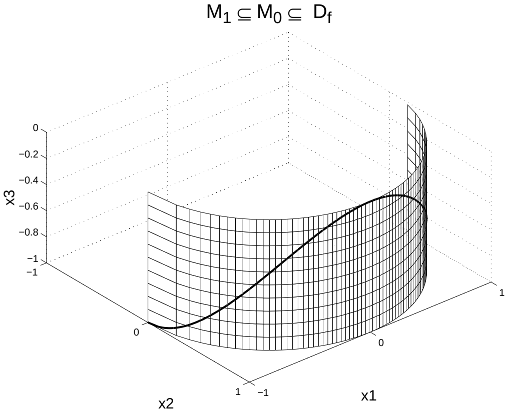
$$M_1(t) \subset M_0(t),$$

which is defined by the so-called hidden constraints and $M_0(t)$.

Let us consider the following index-2 example to get an idea of what explicit and hidden constraints may look like.

Example 2.1.4 *Consider*

$$\begin{aligned} x_1' - x_1 &= 0, \\ x_2' - \frac{x_3}{x_2} &= 0, \\ x_1^2 + x_2^2 - 1 &= 0, \end{aligned}$$



$\mathcal{D}_f = \mathcal{D} = (-1,1) \times (0,1) \times (-1,0)$. *It is easy to realize that the explicit constraint is given by $x_1^2+x_2^2-1=0$, while the hidden constraint arises from $x_3 = -x_1^2$. Consequently, consistent initial values have to fulfil both equations and lie in $\mathcal{D}_f$ due to $M_1 \subseteq M_0 \subseteq \mathcal{D}_f$:*

$$\begin{aligned} M_0 &= \{x \in \mathcal{D} : x_1^2 + x_2^2 - 1 = 0\}, \\ M_1 &= \{x \in \mathcal{D} : x_1^2 + x_2^2 - 1 = 0, \quad x_3 = -x_1^2\}. \end{aligned}$$

Since we will focus on index-2 DAEs, we are interested in a proper characterization of the appearing hidden constraints. For the systems arising from circuit simulation by MNA, their structural properties simplify the problems related to consistent initialization considerably [11]. Here, we are aiming at characterizing these structural properties when considering more general index-2 DAEs. The spaces and projectors related to the tractability index will precisely provide the description of the required structural conditions. In Section 2.3 we will introduce some structural assumptions and properties that make it possible to give the requested general description. This characterization for a wide class of nonlinear DAEs will be presented in Section 2.4, by considering an index reduction. By using this characterization, in Section 2.5 we propose an approach that permits, in many applications, a relatively easy computation of a consistent initial value. In Section 2.6 we illustrate how this approach applies to DAEs in Hessenberg form. Finally, in Section 2.7 we consider a special structure, and analyse the consequences of starting up with the implicit Euler or the trapezoidal rule from an inconsistent initial value that satisfies the original DAE equations.

2.2 Previous Work on the Initialization Problem

For index-1 DAEs, the problem of determining consistent initial values is quite well-understood, since no hidden constraints have to be taken into account.

In particular, the system

$$A(x_0,t_0)P(t_0)y_0 + b(x_0,t_0) = 0, \tag{2.1}$$

$$P(t_0)(x_0-\alpha) + Q(t_0)y_0 = 0, \tag{2.2}$$

$\alpha \in \mathbb{R}^n$ is helpful, if it is solvable [41]. The Jacobian of this system is nonsingular because of the index-1 requirement and its solution is consistent (cf. Theorem 1.3.11). Notice that $Q(t_0)y_0 = 0$ is introduced to guarantee $y_0 = P(t_0)y_0$, obtaining a quadratic system. Some remarks concerning the implementation of this approach can be found in [35].

In practice, for index-1 DAEs a consistent initial value can be computed by means of different approaches. For instance, Brown et al.[7] describe how the computation is performed in the software package DASSL (cf. [49],[5]) and an extension of it, DASPK (cf. [6]). Two possible approaches are considered:

1. For semi-explicit problems it is assumed that a value for the dynamic component is given, i.e., for Px_0, and that we have to compute the corresponding values for Qx_0 and Py_0.

2. For DAEs with a nonsingular matrix f'_x it is discussed how to compute x_0 if y_0 is given.

Moreover, in [7] an extension of the method for higher index Hessenberg[2] DAEs is announced. Hessenberg systems are also considered by Amodio and Mazzia [1], where consistent initial values are computed realizing the differentiation by special finite differences.

For arbitrary unstructured higher index cases, the problem becomes much more complicated. According to the definition of the differential index, we can define the derivative array equations

$$f^{\nu}(x^{\nu+1}, \ldots, x', x, t) = 0 \tag{2.3}$$

as the set of equations derived by differentiating the original DAE (1.1) ν-times, where ν is the differential index. Consequently, most of the approaches based on this consideration aim at computing a complete vector (x_0, y_0) fulfilling (2.3).

Leimkuhler et al. [38] considered numerical differentiation to approximate the derivatives of the derivative array equations (2.3) together with a set of user-specified information on initial conditions. The resulting overdetermined system was solved in a least square sense. This is complicated due to the rank deficiency of the Jacobian. Gopal and Biegler [23] considered (2.3), supposed that a set of initial values was given, and minimized the deviation of the consistent values from the specified ones by a successive linear programming approach. Their algorithm provides good results also for small examples of differential index 3.

Other authors consider the fact that, since (2.3) contains more equations than really necessary for computing consistent initial values, a more detailed

[2] For a definition see Section 2.6.

analysis of the equations is worthwhile.

Pantelides [48] constructed an algorithm using graph theory methods to differentiate subsets of the system. By considering a bipartite graph, this algorithm determines the so-called structural index[3], a number of differentiations to obtain consistent initial values and a selection of variables for which we may prescribe suitable initial values. The approach bases on assignments between equations and variables, locating subsets of equations for which the number of new equations generated upon differentiation of the subset exceeds the number of new variables appearing in them. However, if for instance, for $Q = \begin{pmatrix} 0 & \\ & I \end{pmatrix}$, the number of equations that should be differentiated is less than or equal to the cardinality of the variables of (Px', Qx) appearing in these equations, then equations that ought to be differentiated may escape detection. Example 2.4.2 is given to illustrate this limitation. Another systematically different structural algorithm was developed by Unger et al. [59] (see also Kröner et al.[33],[34]) by using a structural version of the symbolic algorithm for the index reduction proposed by Gear [19]. For linear systems an algorithm based on this idea was already presented in Bachmann et al. [3].

Here, we are aiming at computing consistent initial values for index-2 DAEs by considering a characterization of necessary differentiations by means of the projectors related to the tractability index. Referring to this, we also build upon previous work. Assuming that the relevant projectors are only time-dependent, Hansen [32] proposed an approach that applies index reduction and formula manipulation methods. Taking this idea up, Lamour [36] used a similar description of the part of the solution we have to differentiate, while the differentiated part was replaced by its finite differences.

Due to the fact that, in practice, some necessary assumptions on the projectors were not given, Lamour then considered the possibility to obtain the consistent initialization by differentiating (1.14) once. In this context, März [43],[46] introduced a characterization of those equations of (1.14) that really have to be differentiated by means of a suitable projector. In [14], a modification of this approach was incorporated to a method for computing

[3] This index is determined considering the zero pattern of the matrices and not their actual values.

a consistent initialization. To this end, an index-reduction technique analogous to the one that will be presented in Section 2.4 was already carried out under slightly different assumptions. The difference between the algorithm developed there and the one we present here will be extensively discussed in Section 2.5.

2.3 Some Properties of the Spaces and Projectors

In contrast to the index-1 case, where $M_0(t)$ is filled by solutions (see Theorem 1.3.11), for the index-2 case the so-called hidden constraints define a subset

$$M_1(t) \subset M_0(t),$$

which fulfils the requirement that for each point $x_0 \in M_1(t)$ there exists a solution through x_0.

Another difficulty for index-2 DAEs consists in describing the so-called index-2 components, which belong to the space $N(t) \cap S(x,t)$. These components are determined neither by a differential equation nor by a derivative-free equation, but by inherent differentiation.

Later on, we will see that the hidden constraints can be described properly using the projector W_1 introduced in Section 1.3 if we make some assumptions on the space $N(t) \cap S(x,t)$ and suppose that sufficient smoothness is given[4]. Let us first consider some structural properties that are well-known from the literature dealing with the tractability index. Thereupon, we will deduce some new structural properties that result if we suppose that $N(t) \cap S(x,t)$ depends only on t (cf. Assumption **A2**, pp. 29).

Lemma 2.3.1 *[57],[43] If* **A1** *is given, then for all* $(y,x,t) \in \mathcal{G}_f$ *it holds:*

1. $W_1(y,x,t)A(x,t) = 0$, $W_1(y,x,t)B(y,x,t)Q(t) = 0$,

2. $W_1(y,x,t) = W_1(y,x,t)W_0(t)$,

[4] In Section 2.4 the exact smoothness requirement will be introduced as the need arises.

3. $\ker W_1(y,x,t)B(y,x,t) = \ker W_1(y,x,t)b'_x(x,t) = S_1(y,x,t)$.

4. *If (1.14) is index-2 tractable on* $\mathcal{G} \subseteq \mathcal{G}_f$*, then for all* $(y,x,t) \in \mathcal{G}$ *the following equations are valid:*

   (a) $N(t) \cap S(x,t) = \mathrm{im}\ Q(t)Q_1(y,x,t)$,

   (b) $\ker W_1(y,x,t)B(y,x,t) = \ker Q_1(y,x,t) = \ker P(t)Q_1(y,x,t)$,

   (c) For $G_2(y,x,t) := A_1(y,x,t) + B(y,x,t)P(t)Q_1(y,x,t)$ *it holds that*

$$\begin{array}{rcl} G_2^{-1}(y,x,t)A(x,t) & = & P_1(y,x,t)P(t), \\ G_2^{-1}(y,x,t)B(y,x,t) & = & G_2^{-1}(y,x,t)B(y,x,t)P(t)P_1(y,x,t) \\ & & +Q_1(y,x,t) + Q(t) \\ & & +P_1(y,x,t)P(t)P'(t)Q(t). \end{array}$$

Proof: For the sake of simplicity, we drop the arguments.
1) With $0 = W_1G_1 = W_1(A + BQ)$ we obtain

$$W_1G_1P = W_1A = 0 \ \text{ and } \ W_1G_1Q = W_1BQ = 0.$$

2) Note that im $(I - W_0) = \ker W_0 = $ im $A \subseteq \ker W_1$ and that, therefore, $W_1(I - W_0) = 0$ or $W_1 = W_1W_0$.
3), 4b) and 4c) follow by straightforward computation.
Let us consider 4a).
($\supseteq$) For every $z \in$ im $QQ_1 \subseteq N$ we have $z \in N$. Further, there exists a $w \in I\!\!R^n$ such that $z = QQ_1w$. Thus,

$$Bz = BQQ_1w = (A_1 + AP'Q - A)Q_1w = A(P' - I)Q_1w \in \mathrm{im}\ A$$

is satisfied, i.e., $z \in S$.
($\subseteq$) For every $z \in N \cap S$ it holds that $z = Qz$ and that we can find a $v \in I\!\!R^n$ such that $Bz = Av$ is valid. Then define $\tilde{v} := Qv + Pv - PP'z$, which implies $Bz - AP'z = A\tilde{v}$. For $u := z - P\tilde{v}$ we thus obtain:

$$A_1u = A_1Qz - A\tilde{v} = Bz - AP'z - A\tilde{v} = 0,$$

i.e., $u \in N_1 = $ im Q_1. This finally implies $z = Qz = Qu = QQ_1u$, i.e., $z \in$ im QQ_1.

q.e.d.

Let us now develop some hitherto unexplored structural properties.

Lemma 2.3.2 *Suppose that* **A1** *is given. Then, for all* $(y, x, t) \in \mathcal{G}_f$ *it holds*

$$\begin{aligned} N(t) \cap S(x,t) &= \ker[A(x,t) + W_0(t)b'_x(x,t)] \\ &= \ker[A(x,t) + W_0(t)b'_x(x,t)Q(t)], \qquad (2.4) \\ \text{im } A_1(y,x,t) &= \text{im } G_1(y,x,t) \\ &= \text{im } [A(x,t) + W_0(t)b'_x(x,t)Q(t)] \\ &= \text{im } A(x,t) \oplus \text{im } W_0(t)b'_x(x,t)Q(t). \qquad (2.5) \end{aligned}$$

Proof: For simplicity, we drop the arguments of the matrices.
The equalities from (2.4) arise from **A1**, the definitions of $N(t)$ and $S(x,t)$, and from

$$\ker[A + W_0 b'_x] = \ker A \cap \ker W_0 b'_x = \ker A \cap \ker W_0 b'_x Q = \ker[A + W_0 b'_x Q].$$

Consider the equalities from (2.5). Since $W_0 b'_x = W_0 B$ is given by **A1**, im $W_0 BQ \subseteq$ im W_0, and im $W_0 \cap$im $A = \{0\}$, we only have to show im $(A+BQ) =$ im $A +$ im $W_0 BQ$.
($\subseteq$) For any $z \in$ im $(A + BQ)$ we find a v_1 such that

$$z = (A + BQ)v_1 = Av_1 + (I - W_0)BQv_1 + W_0 BQv_1.$$

Note that we have im $(I - W_0) =$ im A and, therefore, im $(I - W_0)BQ \subseteq imA$. Thus we find a v_2 fulfilling

$$(I - W_0)BQv_1 = Av_2.$$

Therefore, $z = A(v_1 + v_2) + W_0 BQv_1$.
($\supseteq$) For any $z \in$ (im $A +$ im $W_0 BQ$) we find $v_1 \in$ im A and $v_2 \in$ im $W_0 BQ$ such that

$$z = Av_1 + W_0 BQv_2.$$

Moreover, since im $(I-W_0)BQ \subseteq imA$, we find a v_3 such that $(I-W_0)BQv_2 = Av_3$. Hence, we obtain

$$z = (A + BQ)(P(v_1 - v_3) + Qv_2).$$

q.e.d.

We have already noted that $N(t) \cap S(x,t)$ describes the so-called index-2 components, which are determined by inherent differentiation. Hence, it seems to be reasonable that assumptions on this space may imply useful structural properties of the DAE. A reasonably claimed assumption should be given for linear DAEs and in the applications we are interested in. Thus, we assume that

A2 : $N(t) \cap S(x,t)$ depends smoothly on t and does not depend on x for $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$.

Define the smooth projectors[5] $T(t)$ onto $N(t) \cap S(x,t)$ and $U(t) := I - T(t)$ correspondingly. Note further that this assumption can easily be checked considering $\ker[A(x,t) + W_0(t)b'_x(x,t)]$ (cf. Lemma 2.3.2). This lemma also allows a relatively easy computation of a projector $T(t)$.

Remark 2.3.3 *Note that by Lemma 2.3.2 it follows that*

$$\operatorname{rank} W_1(x,t) = \operatorname{rank} T(t) \quad \textit{and} \quad \operatorname{rank} G_1(y,x,t) = \operatorname{rank} U(t)$$

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

Lemma 2.3.4 *The Assumptions* **A1** *and* **A2** *yield the following structural properties:*

1. $W_0(t)B(y,x,t) = W_0(t)B(y,x,t)U(t)$,

2. *For $(W_0b)(x,t) := W_0(t)b(x,t)$ it holds that $(W_0b)(x,t) = (W_0b)(U(t)x,t)$,*

3. $S(x,t) = S(U(t)x,t)$,

4. $\operatorname{im} G_1(y,x,t)$ *depends only on $(U(t)x,t)$. Thus, we can choose W_1 in such a way that $W_1(x,t) = W_1(U(t)x,t)$,*

5. $S_1(y,x,t) = S_1(U(t)x,t)$,

[5] cf. Definition 4.1.4, Appendix.

6.

$$\begin{aligned}\ker A(x,t) &= \ker\Big[A(x,t)+W_1(U(t)x,t)B(y,x,t)\Big]\\ &= \ker\Big[A(x,t)+W_1(U(t)x,t)b'_x(U(t)x,t)\Big].\end{aligned}$$

Proof:
(1) For every $z \in N(t)\cap S(x,t) \subseteq S(x,t) := \{z : W_0(t)B(y,x,t)z = 0\}$ it trivially holds that $W_0(t)B(y,x,t)z = 0$. Therefore, $W_0(t)B(y,x,t)T(t) = 0$. Notice that, in fact, we could write this for a projector onto $S(x,t)$. Nevertheless, since $S(x,t)$ often depends on the solution, we will see that for forthcoming considerations it is advantageous to consider $T(t)$.

(2) From point (1) and **A1** it follows that

$$0 = W_0(t)B(y,x,t)T(t) = W_0(t)b'_x(x,t)T(t)$$

and, therefore,

$$(W_0b)(x,t) - (W_0b)(U(t)x,t) = \int_0^1 (W_0b)'_x(sx+(1-s)U(t)x,t)T(t)ds = 0.$$

(3) The equality (3) follows directly from (2):

$$\begin{aligned}S(x,t) &= \{z \in I\!R^n : W_0(t)B(y,x,t)z = 0\}\\ &= \{z \in I\!R^n : (W_0b)'_x(U(t)x,t)z = 0\} = S(U(t)x,t).\end{aligned}$$

(4) Taking into account the splitting from 2.3.2

$$\text{im } G_1(y,x,t) = \text{im } A(x,t) \oplus \text{im } W_0(t)b'_x(x,t)Q(t)$$

and the fact that imA depends only on t and that $W_0b'_xQ$ depends only on $(U(t)x,t)$, it follows that im $G_1(y,x,t)$ depends only on $(U(t)x,t)$.

(5) The relation (5) follows from

$$\begin{aligned}S_1(y,x,t) &= \{z \in I\!R^n : W_1(U(t)x,t)B(y,x,t)P(t)z = 0\}\\ &= \{z \in I\!R^n : W_1(U(t)x,t)b'_x(U(t)x,t)P(t)z = 0\} = S_1(U(t)x,t).\end{aligned}$$

(6) Finally, we obtain (6) since from

$$\begin{aligned} A(x,t) + W_1(U(t)x,t)B(y,x,t) &= (I - W_1(U(t)x,t))A(x,t) \\ &\quad + W_1(U(t)x,t)B(y,x,t) \end{aligned}$$

we can conclude

$$\begin{aligned} &\ker\left[A(x,t) + W_1(U(t)x,t)B(y,x,t)\right] \\ &= \ker A(x,t) \cap \ker W_1(U(t)x,t)B(y,x,t)P(t) \\ &= \ker A(x,t) \cap \ker W_1(U(t)x,t)b'_x(U(t)x,t)P(t) = \ker A(x,t). \end{aligned}$$

q.e.d.

Remark 2.3.5

- *Observe that, since (Tx) represents the index-2 components, $(W_0b)(x,t) = (W_0b)(U(t)x,t)$ means that these components can not appear in the explicit constraints. This is obviously given, since they precisely are determined by inherent differentiation.*
- *Note that if we have $(W_0b)(x,t) = (W_0b)(U(t)x,t)$ it holds*

$$(W_0b)'_x(x,t) = (W_0b)'_x(U(t)x,t)$$

but that, in general[6]*,*

$$(W_0b)'_t(x,t) \neq (W_0b)'_t(U(t)x,t).$$

Let us further suppose that there exists a time-depending, smooth space $L(t)$ such that

$$\text{im } G_1(y,x,t) \oplus L(t) = I\!R^n,$$

and that thus it is possible to choose a projector $W_1(U(t)x,t)$ with a only time-depending, smooth im $W_1(U(t)x,t)$. Indeed this assumption is given for Hessenberg systems, because W_1 is constant itself (see Section 2.6), and for the equations arising from Modified Nodal Analysis (cf. Chapter 3), where a constant space L can be found. Moreover, for linear systems, the existence of such a space is given if we assume that im $G_1(t)$ is smooth (cf. [43],[46]). Note further that, for general nonlinear systems, an even constant space L

[6] Observe that, if U is constant, then it also holds $(W_0b)'_t(x,t) = (W_0b)'_t(Ux,t)$.

can always be found locally. Therefore, since the computation of consistent initial values only requires local considerations, we do not state this as an explicit assumption.

Since im $A \subseteq$ im G_1 and thus $L \cap \text{im } A = \{0\}$, we can define a smooth projector $\hat{W}_1(t)$ fulfilling:

$$\text{im } \hat{W}_1(t) = \text{im } W_1(U(t)x,t) \text{ and } \ker \hat{W}_1(t) \supseteq \text{im } A(x,t), \tag{2.6}$$

which will become important later on. For this projector it holds that[7]

$$\hat{W}_1(t)(I - W_0(t)) = 0, \tag{2.7}$$

$$W_1(U(t)x,t)\hat{W}_1(t) = \hat{W}_1(t), \quad \text{and} \quad \hat{W}_1(t)W_1(U(t)x,t) = W_1(U(t)x,t). \tag{2.8}$$

Note that by the same argumentation as in Lemma 2.3.4,2, for $(\hat{W}_1 b)(x,t) := \hat{W}_1(t)b(x,t)$ we have

$$(\hat{W}_1 b)(x,t) = (\hat{W}_1 b)(U(t)x,t). \tag{2.9}$$

Let us finally consider the relations between $T(t)$ and $Q(t)$. Since im $T(t) = N(t) \cap S(x,t) \subseteq N(t) = \text{im } Q(t) = \ker P(t)$, it holds that $P(t)T(t) = 0$. Moreover, in the following we assume that for a fixed $Q(t)$ we consider a suitable projector $T(t)$ in such a way that also $T(t)P(t) = 0$ is satisfied. Note that this can always be assumed due to

$$(\text{im } P(t)) \cap (N(t) \cap S(x,t)) = \{0\}.$$

Thus, in the following we can make use of the relations:

$$Q(t)T(t) = T(t) = T(t)Q(t) \quad \text{and} \quad P(t)U(t) = P(t) = U(t)P(t). \tag{2.10}$$

By choosing the projectors such that they suit to each other, it becomes clear that they are adequate to decouple x successively into the different kinds of components.

2.4 Index Reduction by Differentiation

It is well known that the differentiation of a DAE or of parts of it sometimes reduces its index. For instance, if the considered equations are in Hessenberg

[7] cf. Lemma 4.1.2, Appendix.

form[8], this fact is well-understood. If not, this becomes much more complicated and a suitable selection of the parts of the DAE that have to be differentiated becomes necessary. A discussion of several well-known index reduction methods from the literature can be found in [24].

Here we will follow up the technique of März (cf. [46],[43]) to reduce the index of index-2 tractable DAEs of the form (1.14). For a better understanding of the approach we will present, we will first discuss linear systems, then give a motivation for nonlinear systems, and afterwards demonstrate how an index reduction can be reached for the nonlinear DAEs that fulfil the assumptions from Section 2.3.

2.4.1 Linear DAEs

First of all, let us illustrate the announced index reduction by means of an academic example.

Example 2.4.1 *Let us consider the linear time-independent index-2 DAE*

$$Ax' + Bx - q := \begin{pmatrix} 1 & 0 & 0 & 0 \\ & 0 & 0 & 0 \\ & & 0 & 0 \\ & & & 0 \end{pmatrix} x' + \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} x - q = 0$$

or, as single equations,

$$\begin{aligned} x_1' + x_4 &= q_1, \\ x_1 + x_2 &= q_2, \\ x_2 &= q_3, \\ x_3 &= q_4. \end{aligned}$$

Obviously, we do not require the differentiation e.g. of the fourth equation to obtain an explicit expression for the solution x_1, x_2, x_3, x_4. But the general application of the differential index[9] *requires the computation of $\frac{d}{dt}(Ax'+Bx-q)$. Using the given semi-explicit structure we would only differentiate all*

[8] cf. Section 2.6.

[9] see Definition 1.1.1.

explicit constraints. With the projector $W_0 = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$ *along* im A *we could write this in the form* $\frac{d}{dt}(W_0(Ax'+Bx-q)) = \frac{d}{dt}(W_0(Bx-q))$*. However, if, for* $Q = \begin{pmatrix} 0 & & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix}$*, we use a projector* W_1 *along* im G_1 *with* $G_1 = A + BQ = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$*, which is given by* $W_1 = \begin{pmatrix} 0 & & & \\ & 1 & -1 & \\ & & 0 & \\ & & & 0 \end{pmatrix}$*, we actually differentiate only the necessary constraint by considering* $\frac{d}{dt}(W_1(Ax' + Bx - q)) = \frac{d}{dt}(W_1(Bx-q))$*.*

In general, for linear time-independent index-2 systems

$$Ax'(t) + Bx(t) = q(t), \tag{2.11}$$

it is quite easy to realize that, if we replace the part $W_1(Bx(t) - q(t))$ by its differentiated form, i.e., if we consider

$$Ax'(t) + (W_1Bx)'(t) + (I - W_1)Bx(t) = (I - W_1)q(t) + (W_1q)'(t),$$

which, in the form (1.9), reads

$$\Big(A + W_1B\Big)x'(t) + (I - W_1)Bx(t) = (I - W_1)q(t) + (W_1q)'(t), \tag{2.12}$$

then this DAE has index 1, that means, we obtain an index reduction. To this end, let us consider the nullspace of the corresponding matrix $\tilde{A} = A + W_1B$. Due to Lemma 2.3.4, 6, we have $\tilde{N} = \ker\tilde{A} = \ker A = N$, i.e., the same derivatives appear in the two DAEs, which is our objective.

Moreover, according to Definition 1.3.2, we have to show that the corresponding matrix

$$\tilde{G}_1 = A + W_1B + (I - W_1)BQ = A + BQ + W_1B$$

is nonsingular. This holds, since $\tilde{G}_1 z = 0$ yields $(A+BQ)z = G_1 z = A_1 z = 0$, i.e., $z = Q_1 z$, and $W_1 B z = 0$, i.e., $Q_1 z = 0$.
Observe that the solutions of (2.12) belong to the same class C_N^1 as (2.11). As a consequence, this suggests the following representation for $M_1(t)$:

$$M_1(t) := \{x \in M_0(t) : \exists y \quad Ay + Bx = q(t), \quad W_1 B y = (W_1 q)'(t)\}.$$

To verify this representation, we notice that every solution of (2.11) remains also a solution of (2.12). Conversely, we have to show that if we start on M_0, then the whole solution of (2.12) lies there, too. Hence, let us suppose that $x_\star(t)$ is a solution of (2.12) with $x_\star(t_0) \in M_0(t_0)$, which implies $W_1 B x_\star(t_0) = W_1 q(t_0)$ and $x_\star(t_0) \in \tilde{M}_0(t_0)$, where it holds that

$$\tilde{M}_0(t) = \{x \in \mathcal{D} : \exists y \; Ay + (I - W_1)Bx = (I - W_1)q(t),\; W_1 B y = (W_1 q)'(t)\}.$$

By (2.12) we have

$$\begin{aligned} Ax_\star'(t) + (I - W_1)Bx_\star(t) &= (I - W_1)q(t), \\ (W_1 B x_\star)'(t) &= (W_1 q)'(t). \end{aligned}$$

Consider the function $\alpha(t) := W_1(Bx_\star(t) - q(t))$. Then $\alpha'(t) = (W_1 B x_\star)'(t) - (W_1 q)'(t) = 0$, and since $x_\star(t_0) \in M_0(t_0)$ implies $\alpha(t_0) = 0$, the function α vanishes identically, which implies that

$$Ax_\star'(t) + Bx_\star(t) = q(t)$$

is satisfied. Hence, Theorem 1.3.11 implies that, if $(W_1 q)'(t)$ exists and is continuous, then, for each $x_0 \in M_1(t_0)$, there exists a C_N^1 solution passing through it.

Example 2.4.2 *[48] Let us consider the following example, which precisely does not meet the assumptions from [48], to emphasize that the obtained description is adequate, independent of the structure of (2.11):*

$$\begin{aligned} x_1' - (x_1 + 2x_2 + 3x_3) &= 0, \\ x_1 + x_2 + x_3 + 1 &= 0, \\ 2x_1 + x_2 + x_3 &= 0. \end{aligned}$$

The matrices A and B are given by

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & -2 & -3 \\ 1 & 1 & 1 \\ 2 & 1 & 1 \end{pmatrix},$$

and the relevant projectors are $Q = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ *and* $W_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & -1 & 1 \end{pmatrix}$. *Hence, it becomes clear that* $x_1 - 1 = 0$ *has to be differentiated.*

In contrast, the algorithm from [48] fails. This is due to the fact that, on the one hand, the last two equations should be differentiated. On the other hand, x_2 *and* x_3 *appear in these two equations. Thus, the number of linearly independent equations that should be differentiated is equal to the cardinality of the variables of* (Px', Qx) *appearing in these equations (cf. p. 25). As a consequence, the algorithm terminates without detecting all equation subsets that have to be differentiated.*

For linear time-dependent index-2 systems

$$A(t)x'(t) + B(t)x(t) = q(t) \tag{2.13}$$

with $(W_1B), (W_1q) \in C^1$ an index reduction can be achieved considering (cf. [46])

$$\begin{aligned} A(t)x'(t) + W_1(t)(W_1Bx)'(t) + (I - W_1(t))B(t)x(t) \\ = (I - W_1(t))q(t) + W_1(t)(W_1q)'(t), \end{aligned}$$

which, in the form (1.9), reads

$$\begin{aligned} \Big(A(t) + W_1(t)B(t)\Big)x'(t) + W_1(t)(W_1B)'(t)x(t) + (I - W_1(t))B(t)x(t) \\ = (I - W_1(t))q(t) + W_1(t)(W_1q)'(t). \end{aligned} \tag{2.14}$$

Since we have again $\tilde{N}(t) = N(t)$ (Lemma 2.3.4, 6), in this case, the corresponding matrix $\tilde{G}_1(t)$ reads

$$\begin{aligned} \tilde{G}_1(t) &= A(t) + W_1(t)B(t) + (I - W_1(t))B(t)Q(t) + W_1(t)(W_1B)'(t)Q(t) \\ &= A(t) + W_1(t)B(t) + B(t)Q(t) + W_1(t)(W_1B)'(t)Q(t). \end{aligned}$$

Multiplying $\tilde{G}_1(t)z = 0$ by $(I - W_1(t))$ yields $(A(t) + B(t)Q(t))z = 0$, i.e., for $\tilde{z} := ((I + P(t)P'(t)Q(t))z$, we obtain $\tilde{z} = Q_1(t)\tilde{z}$ because of $G_1(t) = A_1(t)(I + P(t)P'(t)Q(t))$. Hence, it holds

$$0 = W_1(t)B(t)z + W_1(t)(W_1B)'(t)Q(t)z = W_1(t)B(t)(I + P(t)P'(t)Q(t))z,$$

i.e., $Q_1(t)\tilde{z} = 0$, and thus $\tilde{z} = 0$, and $z = 0$. Hence, $\tilde{G}_1(t)$ is nonsingular, i.e., (2.14) has index 1 in fact.

For $M_1(t)$, this suggests the representation

$$\begin{aligned} M_1(t) \quad := \quad & \{x \in M_0(t) : \exists y \quad A(t)y + B(t)x = q(t), \\ & W_1(t)[B(t)y + (W_1B)'(t)x - (W_1q)'(t)] = 0\}. \end{aligned} \tag{2.15}$$

At this point it has to be emphasized that again the same derivatives appear in both DAEs, and, thus C^1_N is the appropriate solution space for both of them.

Let us now suppose that sufficient smoothness is given and consider the DAE

$$\begin{aligned} & A(t)x'(t) + W_1(t)(W_0Bx)'(t) + (I - W_1(t))B(t)x(t) \\ & \qquad = (I - W_1(t))q(t) + W_1(t)(W_0q)'(t). \end{aligned}$$

which, in the form (1.9), reads

$$\begin{aligned} & \Big(A(t) + W_1(t)B(t)\Big)x'(t) + W_1(t)(W_0B)'(t)x(t) + (I - W_1(t))B(t)x(t) \\ & \qquad = (I - W_1(t))q(t) + W_1(t)(W_0q)'(t). \end{aligned} \tag{2.16}$$

Note that $\tilde{N}(t) = N(t)$ is given again due to Lemma 2.3.4,6, and that (2.16) has index 1, too. In this case, provided that $(W_0B), (W_0q) \in C^1$ is given, the corresponding matrix $\tilde{G}_1(t)$ reads

$$\begin{aligned} \tilde{G}_1(t) \quad &= \quad A(t) + W_1(t)B(t) + (I - W_1(t))B(t)Q(t) + W_1(t)(W_0B)'(t)Q(t) \\ &= \quad A(t) + W_1(t)B(t) + B(t)Q(t) + W_1(t)(W_0B)'(t)Q(t). \end{aligned}$$

Again, $\tilde{G}_1(t)z = 0$ yields $(A(t) + B(t)Q(t))z = 0$, i.e., for $\tilde{z} := (I + P(t)P'(t)Q(t))z$, $\tilde{z} = Q_1(t)\tilde{z}$. Observe that the definition of $\tilde{z}$ implies $Q(t)z = Q(t)\tilde{z}$ and that $\tilde{z} = Q_1(t)\tilde{z}$ thus leads to $Q(t)z = T(t)Q(t)z$ by Lemma 2.3.1,4a.

Thus, with Lemma 2.3.4,1, Lemma 2.3.1, and making use of $PU = P$, we obtain

$$\begin{aligned} W_1(t)(W_0B)'(t)Q(t)z \quad &= \quad W_1(t)(W_0BU)'(t)T(t)Q(t)z \\ &= \quad W_1(t)B(t)P(t)U'(t)T(t)Q(t)z \\ &= \quad W_1(t)B(t)P(t)P'(t)Q(t)z. \end{aligned}$$

Hence, $(W_1(t)B(t) + W_1(t)(W_0B)'(t)Q(t))z = 0$ implies

$$W_1(t)B(t)(I + P(t)P'(t)Q(t))z = 0,$$

i.e., $Q_1(t)\tilde{z} = 0$ and thus $\tilde{z} = 0$, which yields $z = 0$.

Consequently, $M_1(t)$ may be represented by

$$\begin{aligned} M_1(t) \quad := \quad & \{x \in M_0(t) : \exists y \quad A(t)y + B(t)x = q(t), \\ & W_1(t)[B(t)y + (W_0B)'(t)x - (W_0q)'(t)] = 0\}. \end{aligned} \tag{2.17}$$

Observe that this definition coincides with the preceding (2.15), since for $x \in M_0(t)$ we have

$$\begin{aligned} & W_1(t)[(W_1B)'(t)x - (W_1q)'(t)] \\ & \quad = \quad W_1(t)[(W_1W_0B)'(t)x - (W_1W_0q)'(t)] \\ & \quad = \quad W_1(t)W_1'(t)[(W_0(t)B(t))x - (W_0(t)q(t))] \\ & \qquad\quad + W_1(t)[(W_0B)'(t)x - W_1(t)(W_0q)'(t)] \\ & \underset{x \in M_0(t)}{=} \quad W_1(t)[(W_0B)'(t)x - (W_0q)'(t)]. \end{aligned} \tag{2.18}$$

Note further that for both reductions the space $N(t)$ corresponding to the original index-2 DAE and the space $\tilde{N}(t)$ corresponding to both reduced index-1 DAEs coincide, i.e., that the same derivatives as in the original index-2 DAE appear in the index-1 DAEs.

Let us now focus on the smoothness we require for the solution of (2.16). In contrast to (2.12) and (2.14), where C_N^1 characterizes the required smoothness properly, we need some more smoothness for (2.16), since we derive $W_0(t)B(t)x$. Observe that, due to the fact that $W_0(t)B(t) = W_0(t)B(t)U(t)$ holds, sufficient smoothness is given if we suppose that the solution belongs to the space

$$C^1_{N\cap S} := \left\{ x \in C(\mathcal{I}_f, I\!R^n); Ux \in C^1(\mathcal{I}_f, I\!R^n) \right\}. \tag{2.19}$$

Note that $C^1_{N\cap S} \subseteq C^1_N$ due to $PU = P$.

Conversely, it can be shown that $Ux \in C^1$ if $(W_0B), (W_0q) \in C^1$, and $\ker W_0(t)B(t)Q(t)$ is supposed to have constant rank. Since $U(t) = P(t) + U(t)Q(t)$, we only have to ascertain that $UQx \in C^1$. To this end, observe that $\ker U(t)Q(t) = \ker W_0(t)B(t)Q(t)$ is given, since

$$z \in \ker U(t)Q(t) \Leftrightarrow Q(t)z = T(t)z \Leftrightarrow Q(t)z \in S \Leftrightarrow z \in \ker W_0(t)B(t)Q(t).$$

Consequently, $W_0(t)B(t)x(t) = W_0(t)q(t)$ implies

$$UQx = UQ(W_0BQ)^+W_0BQx = UQ(W_0BQ)^+\Big(W_0q - W_0BPx\Big) \in C^1.$$

Nevertheless, for the time-independent Example 2.4.1 for instance, we have $U = \begin{pmatrix} 1 &&& \\ & 1 && \\ && 1 & \\ &&& 0 \end{pmatrix}$ and thus we suppose, unnecessarily, that x_2 and x_3 are smooth. Thus, $C^1_{N\cap S}$ characterizes sufficient smoothness, but not the really necessary one.

This fact motivated the introduction of the diagonal matrix I_{W_1} defined by

$$I_{W_{1},i,i} = \begin{cases} 1 & if \quad \exists j \in [1, n] : W_{1i,j} \not\equiv 0, \\ 0 & else. \end{cases}$$

Note that I_{W_1} is a projector and that $W_1I_{W_1} = W_1$. Making use of this definition, instead of (2.16) we can consider the DAE

$$\Big(A(t) + W_1(t)(I_{W_1}W_0B)\Big)x'(t) + W_1(t)(I_{W_1}W_0B)'(t)x(t) + \\ (I - W_1(t))B(t)x(t) = (I - W_1(t))q(t) + W_1(t)(I_{W_1}W_0q)'(t). \qquad (2.20)$$

Remark 2.4.3 *Observe further that, instead of I_{W1}, we could write any constant matrix K_{W1} fulfilling $W_1(t)K_{W1} \equiv W_1(t)$.*

Unfortunately, the existence of $(K_{W_1}W_0Bx)'(t)$ is obviously not given in general if we only assume $x \in C^1_N$ for the solution. Thus, in the following we may suppose that $x \in C^1_{N\cap S}$. For the applications in Chapter 3, this will be discussed in more detail (Remark 3.2.9).

The additional smoothness requirement seems to be a reason for considering (2.20) less appropriately than (2.14). Nevertheless, we will see that the results obtained by (2.20) may be more convenient for nonlinear systems.

Observe that, for (2.13), we have only presented possible descriptions for $M_1(t)$ so far. To verify them, we should prove, analogously as for the linear DAEs with constant coefficients, that starting on $M_0(t)$, the solutions of (2.14) and (2.16), respectively, remain in $M_0(t)$. For (2.14) this was done in [46]. For (2.16), this will be a consequence of the results from Section 2.4.3.

2.4.2 Motivation for Nonlinear DAEs

Let us assume that the Assumptions **A1** and **A2** are given. Our aim is to obtain an index-reduction for nonlinear DAEs by adapting the approach (2.20) of the previous section to nonlinear systems.

At a first glance, the above discussion may suggest that, due to the required smoothness, an adequate index reduction can always be obtained by considering the system

$$\begin{aligned} &(I - W_1(x(t),t))f(x'(t),x(t),t) \\ &+ W_1(x(t),t)\frac{d}{dt}\Big\{W_1(x(t),t)f(x'(t),x(t),t)\Big\} = 0, \end{aligned} \tag{2.21}$$

which would correspond to (2.14). If the projector W_1 is constant or depends on $(P(t)x,t)$ only, then it can be shown that (2.21) certainly has index 1 [43]. Moreover, the index reduction can be carried out considering the appropriate solution space C^1_N, as expected.

In practice, we have noticed that W_1 may also depend on the other parts of the solution. For instance, the charge-oriented Modified Nodal Analysis presents this property (cf. Chapter 3). For such systems, new insights reveal that the way to obtain a reasonable index reduction consists in considering

$$\begin{aligned} &(I - \hat{W}_1(t))f(x'(t),x(t),t) \\ &+ W_1(x(t),t)\frac{d}{dt}\Big\{I_{W_1}W_0(t)f(x'(t),x(t),t)\Big\} = 0, \end{aligned} \tag{2.22}$$

where the term $(I - \hat{W}_1(t))f(x'(t),x(t),t)$ describes the equations that are not replaced by derived ones and I_{W1} is defined analogously as for the linear case. The choice of such a projector $\hat{W}_1$ becomes important in the nonlinear case and is possible due to (2.6). Moreover, the roles of the projector W_0 and of the matrix I_{W_1} are analogous as in (2.20).

One can get an idea of why the index reduction described by (2.21) is not appropriate for general nonlinear DAEs by considering the following example.

Example 2.4.4 *Consider the index-2 DAE*

$$\begin{aligned} x_1' + x_4 &= q_1, \\ x_1 + x_2 x_3 &= q_2, \\ x_2 &= q_3, \\ x_3 &= q_4, \end{aligned}$$

$x_i, q_i : \mathcal{I}_f \to I\!R$. For the projector Q chosen as for Example 2.4.1, the projectors W_1 and $\hat{W}_1$ are given by

$$W_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & -x_3 & -x_2 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \hat{W}_1 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Let us consider the expression corresponding to (2.21):

$$\begin{aligned} x_1' + x_4 &= q_1, \\ x_1' - (x_3 - q_4)x_2' - (x_2 - q_3)x_3' + q_3' x_3 + q_4' x_2 + 2x_2 x_3 - q_3 x_3 - q_4 x_2 &= q_2', \\ x_2 &= q_3, \\ x_3 &= q_4. \end{aligned}$$

This equation has the differential index 1, but:

1. *Observe that in this case,* $\ker A \neq \ker \tilde{A}(x)$*, i.e., there appear derivatives differing from those in the original index-2 DAE.*

2. *Observe that* $\ker \tilde{A}(x)$ *depends on x. Hence, according to Remark 1.3.8,4 the tractability-index should be defined considering the corresponding enlarged system (1.18)-(1.19), for which the index is 2.*

3. *The perturbation index of this system is 2, as can be easily seen when considering* $q_1 = q_2 = q_3 = q_4 = 0$ *and the following perturbation (cf. Example 1.2.2):*

$$\begin{aligned} x_1' + x_4 &= 0, \\ x_1' - x_2'x_3 - x_2x_3' + 2x_2x_3 &= 0, \\ x_2 &= \epsilon \sin t^2, \\ x_3 &= \epsilon \cos t^2. \end{aligned}$$

   Straightforward computation leads to

$$x_4 := -\epsilon^2 2t \cos(2t^2) + 2\epsilon^2(\sin t^2)(\cos t^2),$$

   which implies that x_4 *grows with the derivative of the perturbation.*

Let us now consider the expression corresponding to (2.22):

$$\begin{aligned} x_1' + x_4 &= q_1, \\ x_1' + q_3'x_3 + q_4'x_2 &= q_2', \\ x_2 &= q_3, \\ x_3 &= q_4, \end{aligned}$$

For this system, all indices are defined and coincide, they are 1.

The example illustrates that the projector W_1 itself should not be differentiated. This is due to the fact that W_1 was defined considering the partial derivatives, not the equations themselves. Indeed, W_1 provides information on how to combine the equations we have to differentiate.

2.4.3 Nonlinear DAEs

In this section we consider the approach (2.22) for quasilinear DAEs fulfilling **A1**, **A2** and some specific smoothness assumptions, which will be introduced as the need arises.

Let us suppose that (1.14) is index-2 tractable. If we define I_{W1} analogously as for the linear case, then due to Lemma 2.3.4,2 it holds for $(I_{W_1}W_0b)(x,t) := I_{W_1}W_0(t)b(x,t)$ that

$$(I_{W_1}W_0b)(x,t) = (I_{W_1}W_0b)(U(t)x,t). \tag{2.23}$$

Motivated by our discussion in Section 2.4.2 and making use of (2.23) we assume that

$$\mathbf{A3}: \quad \frac{d}{dt}\Big\{(I_{W_1}W_0b)(U(t)x,t)\Big\} \quad \text{exists for all} \quad (x,t)\in\mathcal{D}_f\times\mathcal{I}_f.$$

Remark 2.4.5 *Again*[10]*, instead of* I_{W1} *we could write any constant matrix* K_{W1} *fulfilling* $W_1(\cdot)K_{W1}\equiv W_1(\cdot)$. *Observe further that, if* W_1 *is constant itself, then we can set* $K_{W_1}=W_1$. *This will become important when considering the applications in Chapter 3.*

Due to Lemma 2.3.4,4, and by the approach described in (2.22), let us consider the DAE

$$\begin{aligned} &(I-\hat{W}_1(t))\Big\{A(x(t),t)x'(t)+b(x(t),t)\Big\}\\ &+W_1(U(t)x(t),t)\frac{d}{dt}\Big\{(I_{W_1}W_0b)(U(t)x(t),t)\Big\} \quad = \quad 0. \end{aligned}$$

Since we want to analyse this equation with regard to its index, let us assume that

$$\begin{aligned} \mathbf{A4}: \qquad & W_1\frac{\partial}{\partial t}\frac{\partial}{\partial x}\Big\{(I_{W_1}W_0b)\Big\} = W_1\frac{\partial}{\partial x}\frac{\partial}{\partial t}\Big\{(I_{W_1}W_0b)\Big\},\\ & (W_1(I_{W_1}W_0b)'_x)'_x, \text{ and } \quad (W_1(I_{W_1}W_0b)'_t)'_x \quad \text{exist}\\ & \text{for all} \quad (x,t)\in\mathcal{D}_f\times\mathcal{I}_f, \text{where}\\ & (W_1(I_{W_1}W_0b)'_x)'_x, \quad (W_1(I_{W_1}W_0b)'_t)'_x\in C(\mathcal{D}_f\times\mathcal{I}_f,I\!R^n). \end{aligned}$$

Due to the quasilinear structure (1.14), to $\ker\hat{W}_1(t)\supseteq \operatorname{im} A(x,t)$ (see (2.6)), and because of (2.9) we thus consider the DAE

$$\begin{aligned} &\Big(A(x(t),t)+W_1(U(t)x(t),t)(I_{W_1}W_0b)'_x(U(t)x(t),t)\Big)x'(t)+b(x(t),t)\\ &\quad -(\hat{W}_1b)(U(t)x(t),t)+W_1(U(t)x(t),t)(I_{W_1}W_0b)'_t(x(t),t)=0. \qquad (2.24) \end{aligned}$$

Moreover, analogously as it was done for linear DAEs with constant coefficients in order to guarantee the equivalence of the solutions of (2.24) and

[10] cf. Remark 2.4.3.

(1.14), we need the additional condition that the replaced equations are fulfilled at least at one point

$$(\hat{W}_1 b)(U(t_0)x(t_0), t_0) = 0. \tag{2.25}$$

This approach suggests the following definition for $M_1(t)$

$$\begin{aligned} M_1(t) := \Bigg\{ & x \in M_0(t) : \exists y \quad A(x,t)y + b(x,t) = 0, \\ & W_1(U(t)x,t)\Big[(I_{W_1}W_0(t)b)'_x(U(t)x,t)y + (I_{W_1}W_0 b)'_t(x,t)\Big] = 0\Bigg\}. \end{aligned} \tag{2.26}$$

Let us first investigate the index of (2.24). The pencil matrices of (2.24) are given by

$$\begin{aligned} \tilde{A}(x,t) &:= A(x,t) + W_1(U(t)x,t)(I_{W_1}W_0 b)'_x(U(t)x,t) \\ \tilde{B}(y,x,t) &:= \left\{\Big(A(x,t) + W_1(U(t)x,t)(I_{W_1}W_0 b)'_x(U(t)x,t)\Big)y\right\}'_x + \\ & \quad \left\{b(x,t) - (\hat{W}_1 b)(U(t)x,t) + W_1(U(t)x,t)(I_{W_1}W_0 b)'_t(x,t)\right\}'_x. \end{aligned}$$

By Lemma 2.3.1 we have

$$W_1(U(t)x,t)(I_{W_1}W_0 b)'_x(U(t)x,t) = W_1(U(t)x,t)B(y,x,t). \tag{2.27}$$

and

$$\tilde{A}(x,t) = (A(x,t) + W_1(U(t)x,t)B(y,x,t))P(t),$$

and from Lemma 2.3.4, (6) we conclude $\ker \tilde{A}(x,t) = \ker A(x,t)$, i.e., analogously as for the linear case, the space $N(t)$ corresponding to the original index-2 DAE and the space $\tilde{N}(t)$ corresponding to the reduced index-1 DAE coincide.

According to Definition 1.3.6, to prove that (2.24) has index 1, we check the non-singularity of

$$\begin{aligned}
\tilde{G}_1(y,x,t) &:= \tilde{A}(x,t) + \tilde{B}(y,x,t)Q(t)\\
&= A(x,t) + \underbrace{W_1(U(t)x,t)(I_{W_1}W_0b)'_x(U(t)x,t)}_{3}\\
&+ \left\{\left(\underbrace{A(x,t)}_{1} + W_1(U(t)x,t)(I_{W_1}W_0b)'_x(U(t)x,t)\right)y\right\}'_x Q(t)\\
&+ \left\{\underbrace{b(x,t)}_{2} - (\hat{W}_1b)(U(t)x,t)\right\}'_x Q(t)\\
&+ \left\{W_1(U(t)x,t)(I_{W_1}W_0b)'_t(x,t)\right\}'_x Q(t).
\end{aligned}$$

To this aim we consider an arbitrary z fulfilling $\tilde{G}_1(y,x,t)z = 0$, i.e.,

$$\begin{aligned}
0 \;=\; \tilde{G}_1(y,x,t)z \underset{(2.6)}{=} &\left(A(x,t) + (\underbrace{\{A(x,t)P(t)y\}'_x}_{1} + \underbrace{b'_x(x,t)}_{2})Q(t)\right)z\\
&+ \underbrace{W_1(U(t)x,t)(I_{W_1}W_0b)'_x(U(t)x,t)P(t)}_{3}z\\
&- \hat{W}_1(t)\left\{(\hat{W}_1b)(U(t)x,t)\right\}'_x Q(t)z\\
&+ \hat{W}_1(t)\left\{W_1(U(t)x,t)(I_{W_1}W_0b)'_x(U(t)x,t)y\right\}'_x Q(t)z\\
&+ \hat{W}_1(t)\left\{W_1(U(t)x,t)(I_{W_1}W_0b)'_t(x,t)\right\}'_x Q(t)z,
\end{aligned} \tag{2.28}$$

where we make use of $W_1(U(t)x,t) = \hat{W}_1(t)W_1(U(t)x,t)$ (cf. (2.8)). We split (2.28) by multiplying it by $(I - W_1(U(t)x,t))$, and obtain, due to $W_1(U(t)x,t)\hat{W}_1(t) = \hat{W}_1(t)$,

$$0 = (I - W_1(U(t)x,t))\tilde{G}_1(y,x,t)z = (A(x,t) + B(y,x,t)Q(t))z = G_1(y,x,t)z.$$

Since $G_1(y,x,t) = A_1(y,x,t)(I + P(t)P'(t)Q(t))$, we have $\tilde{z} \in \ker A_1(y,x,t)$ for $\tilde{z} := (I + P(t)P'(t)Q(t))z$, i.e., $\tilde{z} = Q_1(y,x,t)\tilde{z}$. Hence, due to $Q(t)\tilde{z} = Q(t)z$, it holds

$$Q(t)z = T(t)Q(t)z. \tag{2.29}$$

Thus, we obtain

$$\Big\{(\hat{W}_1 b)(U(t)x,t)\Big\}'_x Q(t)z \underset{(2.29)}{=} \Big\{(\hat{W}_1 b)(U(t)x,t)\Big\}'_x T(t)Q(t)z = 0$$

and

$$\begin{aligned}\Big\{W_1(U(t)x,t)(I_{W_1}W_0 b)'_x(U(t)x,t)y\Big\}'_x Q(t)z &\underset{(2.29)}{=} \\ \Big\{W_1(U(t)x,t)(I_{W_1}W_0 b)'_x(U(t)x,t)y\Big\}'_x T(t)Q(t)z &= 0.\end{aligned}$$

Let us now consider the expression

$$\begin{aligned} & \Big\{W_1(U(t)x,t)(I_{W_1}W_0 b)'_t(x,t)\Big\}'_x Q(t)z \\ \underset{(2.29)}{=}\ & W_1(U(t)x,t)\Big\{(I_{W_1}W_0 b)(x,t)\Big\}''_{tx} T(t)Q(t)z \\ =\ & W_1(U(t)x,t)\underbrace{\Big\{\Big\{(I_{W_1}W_0 b)(U(t)x,t)\Big\}'_x T(t)\Big\}'_t}_{=0} Q(t)z \\ & - W_1(U(t)x,t)\Big\{(I_{W_1}W_0 b)(U(t)x,t)\Big\}'_x T'(t)Q(t)z \\ \underset{(2.27)}{=}\ & - W_1(U(t)x,t)B(y,x,t)T'(t)Q(t)z \\ =\ & W_1(U(t)x,t)B(y,x,t)P(t)P'(t)Q(t)z. \end{aligned}$$

Consequently, (2.28) yields

$$W_1(U(t)x,t)B(y,x,t)\Big(I + P(t)P'(t)Q(t)\Big)z = 0. \qquad (2.30)$$

Due to Lemma 2.3.1,4b equation (2.30) implies $Q_1(y,x,t)\tilde{z} = 0$, i.e., $\tilde{z} = 0$. Thus we have $z = 0$. This means that the matrix $\tilde{G}_1(y,x,t)$ is nonsingular, i.e., the DAE (2.24) has index 1.

What about the equivalence of the equations (1.14) and (2.24)? It seems to be clear that, if sufficient smoothness is given, every solution of (1.14) remains also a solution of (2.24). Conversely, we have to show that, if we start on M_0, then the whole solution of (2.24) lies there, too. Let $x_\star$ be a solution of (2.24) with $x_\star(t_0) \in \tilde{M}_0$ fulfilling (2.25), where $\tilde{M}_0$ corresponds to this index-1 problem. Therefore, (2.24) is fulfilled particularly for $x_\star(t)$. Multiplying the corresponding equation (2.24) by $\hat{W}_1(t)$ provides then

$$W_1(U(t)x_\star(t),t)\frac{d}{dt}\Big\{(I_{W_1}W_0b)(U(t)x_\star(t),t)\Big\} = 0. \tag{2.31}$$

Using this result and multiplying the corresponding equation (2.24) by $W_0(t)$ we thus obtain

$$(W_0b)(U(t)x_\star(t),t) - W_0(t)(\hat{W}_1b)(U(t)x_\star(t),t) = 0. \tag{2.32}$$

Further, with (2.25) the condition (2.32) implies $x_\star(t_0) \in M_0(t_0)$. Let us now suppose that[11]

$$\mathbf{A5}: \quad \frac{d}{dt}\Big\{(\hat{W}_1b)(U(t)x,t)\Big\} \quad \text{exists for all} \quad (x,t) \in \mathcal{D}_f \times \mathcal{I}_f.$$

If x_* is sufficiently smooth[12] to guarantee the existence of the forthcoming

[11] This assumption seems to be reasonable, since if we replace some equations by derived ones, and if we want to guarantee them by supposing only that they are fulfilled at one point, their smoothness seems to be a necessary requirement. Observe that, nevertheless, this is less than the assumption that $\frac{d}{dt}\Big\{(W_0b)(U(t)x,t)\Big\}$ exists.

[12] The required smoothness of the solution is given if, we have $x_\star \in C^1_{N\cap S}$for instance.

expressions, then (2.8) implies

$$
\begin{aligned}
\frac{d}{dt}\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} &= \frac{d}{dt}\Big\{\hat{W}_1(t)(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&= \hat{W}_1{}'(t)\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&\quad + W_1(U(t)x_\star(t),t)\hat{W}_1(t)\frac{d}{dt}\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&= \hat{W}_1{}'(t)\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&\quad + W_1(U(t)x_\star(t),t)I_{W_1}W_0(t)\hat{W}_1(t)\frac{d}{dt}\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&= \hat{W}_1{}'(t)\Big\{(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&\quad + W_1(U(t)x_\star(t),t)\frac{d}{dt}\Big\{I_{W_1}W_0(t)(\hat{W}_1 b)(U(t)x_\star(t),t)\Big\} \\
&\quad - W_1(U(t)x_\star(t),t)(I_{W_1}W_0\hat{W}_1)'(t)(\hat{W}_1 b)(U(t)x_\star(t),t) \\
&\underset{(2.32)}{=} W_1(U(t)x_\star(t),t)\frac{d}{dt}\Big\{(I_{W_1}W_0 b)(U(t)x_\star(t),t))\Big\} \\
&\quad + \Big(\hat{W}_1{}'(t) - W_1(U(t)x_\star(t),t)(I_{W_1}W_0\hat{W}_1)'(t)\Big)(\hat{W}_1 b)(U(t)x_\star(t),t) \\
&\underset{(2.31)}{=} \Big(\hat{W}_1{}'(t) - W_1(U(t)x_\star(t),t)(I_{W_1}W_0\hat{W}_1)'(t)\Big)(\hat{W}_1 b)(U(t)x_\star(t),t).
\end{aligned}
$$

For $\alpha(t) = (\hat{W}_1 b)(U(t)x_\star(t),t)$ we thus obtain

$$\alpha'(t) = \Big(\hat{W}_1{}'(t) - W_1(U(t)x_\star(t),t)(I_{W_1}W_0\hat{W}_1)'(t)\Big)\alpha(t),$$

and, because of $x_\star(t_0) \in M_0$, $\alpha(t_0) = 0$. Hence, α vanishes identically, i.e., $(\hat{W}_1 b)(U(t)x_\star(t),t) = 0$.

This proves the following:

Theorem 2.4.6 *Suppose that (1.14) is index-2 tractable on $\mathcal{G} \subseteq \mathcal{G}_f$. If the assumptions* **A1-A5** *are given, then equation (2.24) has index-1 on $\mathcal{G}$ and the* *sufficiently smooth*[13] *solutions of the index-2 equation (1.14) and the index-1 equation (2.24) fulfilling (2.25) are the same.*

Remark 2.4.7

- *Let us emphasize that for these solutions the space* C^1_N *may not characterize the required smoothness properly, since*

$$\frac{d}{dt}\Big\{(I_{W_1}W_0 b)((Ux)(t),t)\Big\},$$
$$\frac{d}{dt}\Big\{(\hat{W}_1 b)((Ux)(t),t)\Big\}$$

  may involve more derivatives than $(Px)'(t)$. *Nevertheless, the above results imply that it is not necessary to assume* $(Tx)(t)$ *to be differentiable, i.e., the space* $C^1_{N\cap S}$ *characterizes sufficient smoothness.*

- *A similar index-reduction was already carried out in [14]. Observe that the assumptions made here slightly differ from those in [14], where* $\ker A(x,t)$ *and* $\operatorname{im} A(x,t)$ *were supposed to be constant,* $\operatorname{im} A_1(y,x,t)$ *and* $\ker A_1(y,x,t)$ *were supposed to depend only on* (x,t), *and a rather complicated structural condition was assumed, but no direct restrictions on* $N\cap S(\cdot)$ *were made. Consequently, to obtain a result corresponding to Theorem 2.4.6, it was necessary to consider* C^1*-solutions.*

Since the solutions are the same, the results described in Section 1.3.3 imply that we can transfer the solvability results for index-1 tractable DAEs to the considered index-2 tractable DAEs. In fact, if the DAE (1.14) is index-2 tractable on $\mathcal{G} \subseteq \mathcal{G}_f$, then for $x_0 \in M_1(t_0)$, where

$$M_1(t) := \Big\{x \in \mathcal{D} : \exists y \quad A(x,t)y + b(x,t) = 0,$$
$$W_1(U(t)x,t)\Big[(I_{W_1}W_0 b)'_x(U(t)x,t)y + (I_{W_1}W_0 b)'_t(x,t)\Big] = 0\Big\}$$

holds, there exists a locally unique solution $x(\cdot) : \mathcal{I} \to I\!\!R^n$ of the corresponding index-1 DAE ((2.24) fulfilling (2.25)) with $x(t_0) = x_0$. Hence, if it is supposed that all the possible solutions are sufficiently smooth, then $x(\cdot)$ is

[13] If no better characterization is given, we can suppose that the solutions have to lie in $C^1_{N\cap S}$. For the applications in Chapter 3, see Remark 3.2.9.

also a locally unique solution of the index-2 tractable DAE (1.14). Consequently, in this case the above representation for $M_1(t)$ is appropriate. Of course, in practice it is desirable to have assumptions that are easy to verify, even if they are more restrictive than strictly necessary. Thus, we formulate a simplified result that follows directly from the above discussion.

Corollary 2.4.8 *If the DAE*

$$A(x(t),t)x'(t) + b(x(t),t) = 0$$

fulfilling **A1**, **A2**, **A4** *is index-2 tractable on* $\mathcal{G} \subseteq \mathcal{G}_f$, *and*

$$A(x,t)y + b(x,t)$$

and

$$W_1(U(t)x,t)\Big[(I_{W_1}W_0(t)b)'_x(U(t)x,t)y + (I_{W_1}W_0b)'_t(x,t)\Big]$$

are continuously differentiable for all $(y,x,t) \in \mathcal{G}_f$, *then for* $x_0 \in M_1(t_0)$, *where*

$$M_1(t) := \Big\{x \in \mathcal{D} : \exists y \quad A(x,t)y + b(x,t) = 0,$$
$$W_1(U(t)x,t)\Big[(W_0b)'_x(U(t)x,t)y + (W_0b)'_t(x,t)\Big] = 0\Big\}$$

holds, there exists a locally unique C^1*-solution* $x(\cdot) : \mathcal{I} \to I\!R^n$ *with* $x(t_0) = x_0$.

Proof: Observe that on the one hand, the smoothness requirements are stronger than **A3** and **A5**. On the other hand, the assumptions imply that the corresponding reduced index-1 DAE is continuously differentiable. Consequently, for this index-1 DAE the Implicit Function Theorem implies that we obtain a continuously differentiable function $w(u,t)$ in the proof of Theorem 1.3.11. Hence, for the obtained solution it holds that $x \in C^1$, and thus sufficient smoothness for Theorem 2.4.6 is given.

q.e.d.

However, we want to emphasize once again that these smoothness requirements are not necessary. In Chapter 3, we will see that these assumptions are unnecessarily strong if we consider DAEs arising from circuit simulation.

2.5 The Computation of Consistent Initial Values

In this section, we will develop a step-by-step method to compute consistent initial values. For this approach, we assume sufficient smoothness to be given in order to guarantee that the expression for $M_1(t)$ presented in the above section is appropriate. For more clarity, we first motivate the approach with an example.

2.5.1 Motivation

Several approaches to compute consistent initial values (e.g. [48],[14],[12]) consist in performing the following steps:

1. Describe the hidden constraints.

2. Determine a selection of variables or a component for which we may prescribe suitable initial values.

3. Construct a full rank system that provides the values for the remaining ones.

Here, we want to show that, under certain structural properties, we can compute a consistent initial value for index-2 DAEs as follows:

1. Describe the hidden constraints.

2. Compute a value x^0 that satisfies the explicit equations of the DAE[14], $x^0 \in M_0(t)$.

3. Correct this value in order to fulfil the hidden constraints, where the correction is also computed considering a full rank system, i.e., calculate a value $x_0 \in M_1(t)$.

[14] Observe that for index-1 DAEs, all the values that fulfil the equations of the DAE are consistent. Hence, this approach can be considered a step-by-step approach.

Let us illustrate the difference between the approach from [14] and the one we are aiming at here considering again Example 1.3.4.

$$\begin{aligned} x_1' + x_1 + x_2 &= q_1, \\ x_2' + x_3 + x_4 &= q_2, \\ x_2 &= q_3, \\ x_4 &= q_4. \end{aligned}$$

Straightforward computation shows that $\operatorname{im} A_1 = \ker W_1 = \ker \begin{pmatrix} 0 & & & \\ & 0 & & \\ & & 1 & \\ & & & 0 \end{pmatrix}$.

Thus, according to Section 2.4, the hidden constraint arises from $x_2' = q_3'(t)$. The approach from [14] would identify x_1 as the only variable for which we can prescribe a value, assign $x_{10} = \alpha_1$ and compute, afterwards, the corresponding consistent values $x_{20} = q_3(t_0), x_4 = q_4(t_0), x_3 = q_2(t_0) - q_3'(t_0) - q_4(t_0)$, and the corresponding values for x_{10}', x_{20}'.

The idea of the approach pursued now is, in contrast, a step-by-step computation of the consistent initial value. To this end, we calculate first a value $(x_1^0, x_2^0, x_3^0, x_4^0, {x_1'}^0, {x_2'}^0)$ fulfilling

$$\begin{aligned} {x_1'}^0 + {x_1}^0 + {x_2}^0 &= q_1(t_0), \\ {x_2'}^0 + x_3^0 + x_4^0 &= q_2(t_0), \\ x_2^0 &= q_3(t_0), \\ x_4^0 &= q_4(t_0), \end{aligned}$$

and correct then the value of the component that is determined by inherent differentiation, i.e. x_3, as well as the value of the derivative of the component that appears in dynamic form, but is not really dynamic, i.e. x_2', in order to obtain consistent initial values. The resulting consistent values read then

$$(x_{10}, x_{20}, x_{30}, x_{40}, x_{10}', x_{20}') := (x_1^0, x_2^0, x_3^0 + {x_2'}^0 - q_3'(t_0), x_4^0, {x_1'}^0, q_3'(t_0)).$$

For the sake of clarity, we again discuss first the approach for linear systems and present after that the generalization for nonlinear systems.

2.5.2 Linear DAEs

For a better understanding of the approaches for computing consistent initial values for index-2 DAEs, let us first divide linear DAEs into the different parts:

(i) the inherent regular ODE,

(ii) the part describing the inherent differentiation problem,

(iii) the purely algebraic part, composed by

- the algebraic part that contains the component that appears in dynamic form in (ii),
- the algebraic part that does not contain the component that appears in dynamic form in (ii).

Taking into account

$$\begin{aligned} I &= P(t)(P_1(t)+Q_1(t))+Q(t)(U(t)+T(t)) \\ &= P(t)P_1(t)+T(t)+P(t)Q_1(t)+U(t)Q(t) \end{aligned}$$

if we multiply (1.9) by

$$P(t)P_1(t)G_2^{-1}(t),\;\; T(t)G_2^{-1}(t),\;\; P(t)Q_1(t)G_2^{-1}(t),\;\; \text{and}\;\; U(t)Q(t)G_2^{-1}(t),$$

we obtain, by Lemma 2.3.1, the system

$$\begin{aligned} P(t)P_1(t)x'(t)+P(t)P_1(t)P'(t)Qx(t)& \\ +P(t)P_1(t)G_2^{-1}(t)B(t)P(t)P_1(t)x(t) &= P(t)P_1(t)G_2^{-1}(t)q(t), \quad (2.33) \\ -Q(t)Q_1(t)P(t)Q_1(t)x'(t)& \\ -Q(t)Q_1(t)P'(t)Q(t)x(t)& \\ +T(t)x(t)+T(t)Q_1(t)P(t)Q_1(t)x(t)& \\ +T(t)G_2^{-1}(t)B(t)P(t)P_1(t)x(t) &= T(t)G_2^{-1}(t)q(t), \quad (2.34) \\ P(t)Q_1(t)x(t) &= P(t)Q_1(t)G_2^{-1}(t)q(t), \quad (2.35) \\ U(t)Q(t)G_2^{-1}(t)B(t)P(t)P_1(t)x(t)& \\ +U(t)Q(t)x(t) &= U(t)Q(t)G_2^{-1}(t)q(t). \quad (2.36) \end{aligned}$$

With the denotations

$$\begin{aligned} u(t) &:= P(t)P_1(t)x(t), \\ v(t) &:= P(t)Q_1(t)x(t), \\ w(t) &:= T(t)x(t), \\ y(t) &:= U(t)Q(t)x(t)=Q(t)U(t)x(t), \end{aligned}$$

this system can be rewritten as[15]

$$\begin{aligned} u'(t) - (PP_1)'(t)(u(t)+v(t)) & \\ + P(t)P_1(t)G_2^{-1}(t)B(t)u(t) &= P(t)P_1(t)G_2^{-1}(t)q(t), & (2.37)\\ - Q(t)Q_1(t)v'(t) & \\ + Q(t)Q_1(t)(PQ_1)'(t)(u(t)+v(t)) + w(t) & \\ + T(t)Q_1(t)v(t) + T(t)G_2^{-1}(t)B(t)u(t) &= T(t)G_2^{-1}(t)B(t)q(t), & (2.38)\\ v(t) &= P(t)Q_1(t)G_2^{-1}(t)q(t), & (2.39)\\ U(t)Q(t)G_2^{-1}(t)B(t)u(t) + y(t) &= U(t)Q(t)G_2^{-1}(t)q(t). & (2.40) \end{aligned}$$

Observe that (2.39) leads to an expression for the component v . Hence, making use of this expression, (2.37) can be reformulated as a regular ODE for u. From (2.40) we see that y represents the algebraic part that is not concerned with the inherent differentiation. Finally, (2.38) represents the equations that involve the inherent differentiation and determine the component w, the so-called index-2 component. Note that we have to differentiate $PQ_1G_2^{-1}q$.

Thus, an adequate formulation of initial value problems for linear index- 2 tractable DAEs reads

$$\begin{aligned} A(t)x'(t) + B(t)x(t) &= q(t),\\ P(t_0)P_1(t_0)(x_0-\alpha) &= 0, \end{aligned}$$

for a given $\alpha \in I\!R^n$. For a proof see [47].

Considering the expression (2.17) for $M_1(t)$, in order to compute a consistent initialization it is sufficient to solve the following system[16]

$$\begin{aligned} A(t_0)y_0 + B(t_0)x_0 &= q(t_0), & (2.41)\\ P(t_0)P_1(t_0)(x_0-\alpha) &= 0, & (2.42)\\ W_1(t_0)[B(t_0)y_0 + (W_0B)'(t_0)x_0] - W_1(t_0)(W_0q)'(t_0) &= 0, & (2.43) \end{aligned}$$

[15] Here we suppose that the required smoothness of the projectors is given, cf. Remark 1.3.8.

[16] This corresponds to the approach described in [14]. There it was described also for nonlinear systems with specific structural properties, but it was assumed that $W_0' = 0$, $Q' = 0$.

for an arbitrary α. Note that (2.42) fixes the dynamic components and (2.43) describes the hidden constraints.

Let us verify that the obtained system uniquely determines $(P(t_0)z_y, z_x)$. For a solution $(P(t_0)z_y, z_x)$ of the homogeneous system, it holds that $P(t_0)P_1(t_0)z_x = 0$, and

$$\begin{aligned} A(t_0)z_y + B(t_0)z_x &= 0, & (2.44)\\ W_1(t_0)B(t_0)z_y + W_1(t_0)(W_0B)'(t_0)z_x &= 0, & (2.45) \end{aligned}$$

multiplying (2.44) by $G_2^{-1}(t_0)$ leads to:

$$\begin{aligned} P_1(t_0)P(t_0)z_y + G_2^{-1}(t_0)B(t_0)P(t_0)P_1(t_0)z_x + Q_1(t_0)z_x &\\ +Q(t_0)z_x + P_1(t_0)P(t_0)P'(t_0)Q(t_0)z_x &= 0 \quad (2.46) \end{aligned}$$

making use of the relations from Lemma 2.3.1,4c. Multiplication by $Q_1(t_0)$ yields $Q_1(t_0)z_x = 0$, which then implies $P(t_0)z_x = 0$, i.e., $z_x \in N(t_0)$. Moreover, multiplying (2.44) by $W_0(t_0)$ leads to $W_0(t_0)B(t_0)z_x = 0$, i.e., $z_x \in S(t_0)$, which implies $z_x = T(t_0)z_x$.
Consequently, (2.46) provides

$$P_1(t_0)P(t_0)z_y + Q(t_0)z_x + P_1(t_0)P(t_0)P'(t_0)Q(t_0)z_x = 0. \quad (2.47)$$

Let us now consider (2.45) taking into account $z_x = T(t_0)z_x$:

$$\begin{aligned} W_1(t_0)B(t_0)z_y + W_1(t_0)(W_0B)'(t_0)T(t_0)Q(t_0)z_x &= \\ W_1(t_0)B(t_0)z_y + W_1(t_0)(\underbrace{W_0BT}_{=0})'(t_0)Q(t_0)z_x &\\ -W_1(t_0)W_0(t_0)B(t_0)T'(t_0)Q(t_0)z_x &= \\ W_1(t_0)B(t_0)z_y - W_1(t_0)B(t_0)P(t_0)T'(t_0)Q(t_0)z_x &= \\ W_1(t_0)B(t_0)z_y + W_1(t_0)B(t_0)P(t_0)P'(t_0)Q(t_0)z_x &= 0. \end{aligned}$$

Due to Lemma 2.3.1,4b we obtain

$$Q_1(t_0)z_y + Q_1(t_0)P(t_0)P'(t_0)Q(t_0)z_x = 0.$$

Together with (2.47) this implies

$$P(t_0)z_y + Q(t_0)z_x + P(t_0)P'(t_0)Q(t_0)z_x = 0, \quad (2.48)$$

which leads to $Q(t_0)z_x = 0$ by multiplication with $Q(t_0)$. Finally, this implies $P(t_0)z_y = 0$.

Let us now consider the value we obtain solving the system

$$\begin{aligned} A(t_0)y_0 + B(t_0)x_0 &= q(t_0), & (2.49)\\ U(t_0)(x_0 - x^0) &= 0, & (2.50)\\ W_1(t_0)[B(t_0)y_0 + (W_0B)'(t_0)x_0] + W_1(t_0)(W_0q)'(t_0) &= 0 & (2.51) \end{aligned}$$

if x^0 denotes a value fulfilling

$$A(t_0)y^0 + B(t_0)x^0 = q(t_0) \qquad (2.52)$$

for a suitable $P(t_0)y^0$.

Straightforward computation shows that no contradictions arise, since (2.50) is consistent with (2.49) due to the special choice of x^0. This consistency can easily be verified if we define

$$\begin{aligned} \hat{x}_0 &= x_0 - x^0, & (2.53)\\ P(t_0)\hat{y}_0 &= P(t_0)y_0 - P(t_0)y^0, & (2.54) \end{aligned}$$

and compute $(\hat{x}_0, P(t_0)\hat{y}_0)$ from the system that results from (2.49)-(2.51) and (2.52) if $(x^0, P(t_0)y^0)$ are considered as fixed values:

$$\begin{aligned} A(t_0)\hat{y}_0 + B(t_0)\hat{x}_0 &= 0, & (2.55)\\ U(t_0)\hat{x}_0 &= 0, & (2.56)\\ W_1(t_0)[B(t_0)[y^0 + \hat{y}_0] + (W_0B)'(t_0)[x^0 + \hat{x}_0]] & & \\ -W_1(t_0)(W_0q)'(t_0) &= 0. & (2.57) \end{aligned}$$

Multiplying (2.56) by $P(t_0)P_1(t_0)$ we obtain $P(t_0)P_1(t_0)\hat{x}_0 = 0$. Thus, decoupling (2.56) analogously as in (2.33)-(2.36), we can deduce $P(t_0)Q_1(t_0)\hat{x}_0 = 0$ and $U(t_0)Q(t_0)\hat{x}_0 = 0$. Consequently, it becomes clear that $U(t_0)\hat{x}_0 = 0$ actually fixes only additionally $P(t_0)P_1(t_0)\hat{x}_0 = 0$, i.e., that (2.55)-(2.56) consist only of $n + \text{rank } PP_1$ linearly independent equations. Hence, (2.49) is consistent with (2.50) due to the special choice of x^0.

Remark 2.5.1 *This approach can be considered a step-by-step approach, since we first calculate an $x^0 \in M_0(t_0)$ (and the corresponding $P(t_0)y^0$), and afterwards compute the correction in order to obtain a consistent value $x_0 \in M_1(t_0)$ (and the corresponding $P(t_0)y_0$).*

Note that, due to

$$U(t) = U(t)(Q(t) + P(t)(P_1(t) + Q_1(t)) = U(t)Q(t) + P(t)P_1(t) + P(t)Q_1(t)$$

it becomes clear that (2.49)-(2.51) consists of more restrictions than (2.41)-(2.43). However, on the one hand we recognize from the decoupling (2.33)-(2.36) and (2.52) that, if we set $P(t_0)P_1(t_0)x_0 = P(t_0)P_1(t_0)x^0$, then we obtain

$$\begin{aligned} U(t_0)x_0 &= U(t_0)Q(t_0)x_0 + P(t_0)P_1(t_0)x_0 + P(t_0)Q_1(t_0)x_0 \\ &= U(t_0)Q(t_0)G_2^{-1}(t_0)q(t_0) - U(t_0)Q(t_0)G_2^{-1}(t_0)B(t_0)P(t_0)P_1(t_0)x^0 \\ &\quad + P(t_0)P_1(t_0)x^0 + P(t_0)Q_1(t_0)G_2^{-1}(t_0)q(t_0) = U(t_0)x^0. \end{aligned}$$

On the other hand, $U(t_0)x_0 = U(t_0)x^0$ implies

$$P(t_0)P_1(t_0)U(t_0)x_0 = P(t_0)P_1(t_0)P(t_0)U(t_0)x^0 = P(t_0)P_1(t_0)x_0.$$

Consequently, for $\alpha = x^0$ the results from (2.49)-(2.51) and (2.41)-(2.43) coincide.

Remark 2.5.2 *At first glance, the second approach seems to be neither easier to realize nor of more practical relevance. The application will show that it has some advantages:*

- *$U(t)$ (and $W_1(t)$) may be computed easier than $P(t)P_1(t)$ (and $W_1(t)$) (cf. Lemma 2.3.2).*
- *The task of determining values for $(\hat{x}_0, P(t_0)\hat{y}_0)$ by making use of (2.55)-(2.57) may look very similar to the direct computation of $(x_0, P(t_0)y_0)$ from (2.41) - (2.43). In fact, since (2.55)-(2.57) can be reformulated as a system for $(T(t_0)\hat{x}_0, P(t_0)\hat{y}_0)$, the dimension may be reduced considerably. Moreover, in this way we take advantage of the fact that sometimes the user of a simulation package uses $\alpha = x^0$ and wants to preserve this values.*

- *In Remark 2.5.3 we will discuss the advantages of considering the system corresponding to (2.55)-(2.57) for nonlinear systems.*

- *In practice, the systems (2.41) - (2.43) and (2.55)-(2.57) can be enlarged by* $Q(t_0)y_0 = 0$ *(analogously as in (2.2)) in order to obtain a nonsingular system. Moreover, it has to be noted that for a special choice of the projector* W_1*, the system (2.49)-(2.51) together with* $Q(t_0)y_0 = 0$ *can be reformulated as a quadratic system (cf. [47],[14]).*

2.5.3 Nonlinear DAEs

Analogously as in the previous section, let us suppose that we know some values $(x^0, P(t_0)y^0)$ that fulfil the equations of the DAE i.e.,

$$A(x^0, t_0)y^0 + b(x^0, t_0) = 0.$$

Then the system

$$\begin{aligned} A(x_0, t_0)y_0 + b(x_0, t_0) &= 0, && (2.58)\\ U(t_0)x_0 &= U(t_0)x^0, && (2.59)\\ W_1(U(t_0)x_0, t_0)\Big[(I_{W_1}W_0b)'_x(U(t_0)x_0, t_0)P(t_0)y_0\Big] & && \\ + W_1(U(t_0)x_0, t_0)\Big[(I_{W_1}W_0b)'_t(x_0, t_0)\Big] &= 0 && (2.60) \end{aligned}$$

will be helpful, if it is solvable, to obtain values $(x_0, P(t_0)y_0)$ fulfilling the equations of the DAE as well as the hidden constraints. Let us consider the Jacobian $J(y_0, x_0, t_0)$ and show that $(Pz_y, z_x) \in \ker J$ implies $(Pz_y, z_x) = 0$. For simplicity, we drop the arguments of the matrices:

$$J = \begin{pmatrix} A & B \\ 0 & U \\ W_1B & \{W_1(I_{W_1}W_0b)'_x y_0 + W_1(I_{W_1}W_0b)'_t\}'_x \end{pmatrix}.$$

For $(Pz_y, z_x) \in \ker J$ it holds that $Az_y + Bz_x = 0$, and multiplication by G_2^{-1} provides

$$P_1Pz_y + G_2^{-1}BPP_1z_x + Q_1z_x + Qz_x + P_1PP'Qz_x = 0. \qquad (2.61)$$

Analogously as for the linear DAEs (cf. (2.46)) we obtain $U(t_0)z_x = 0$, and from (2.61) we have:

$$P_1 P z_y + Q z_x + P_1 P P' Q z_x = 0. \tag{2.62}$$

Let us consider the expressions we obtain from the third row of $(P(t_0)z_y, z_x) \in \ker J$ in detail[17]. Firstly, observe that

$$\{W_1(I_{W1}W_0b)'_x y_0\}'_x z_x = 0$$

since $T(t_0)z_x = z_x$ and $W_1(I_{W1}W_0b)'_x(x,t) = W_1(U(t)x,t)(I_{W1}W_0b)'_x(U(t)x,t)$. Secondly, consider

$$\begin{aligned} &\{W_1(I_{W1}W_0b)'_t\}'_x(x_0,t_0)z_x \\ = \quad & W_1(U(t_0)x_0,t_0)(I_{W1}W_0b)''_{xt}(x_0,t_0)T(t_0)z_x \\ = \quad & -W_1(U(t_0)x_0,t_0)(I_{W1}W_0b)'_x(x_0,t_0)P(t_0)T'(t_0)z_x \\ = \quad & W_1(U(t_0)x_0,t_0)(I_{W1}W_0b)'_x(x_0,t_0)P(t_0)P'(t_0)Q(t_0)z_x. \end{aligned}$$

Therefore, the third row of $(Pz_y, z_x) \in \ker J$ implies

$$W_1 B z_y + W_1 B P P' Q z_x = 0,$$

which is equivalent to

$$Q_1 P z_y + Q_1 P P' Q z_x = 0, \tag{2.63}$$

and thus, analogously as in the linear case, (2.62) and (2.63) lead to

$$P z_y + Q z_x + P P' Q z_x = 0,$$

which implies $Qz_x = 0$ and, thus, $z_x = Tz_x = TQz_x = 0$ and $Pz_y = 0$.

Remark 2.5.3 • *The nonlinear system (2.58)-(2.60) can be enlarged by $Q(t_0)y_0 = 0$ (analogously as in (2.2)) in order to obtain a system with full rank Jacobian in practice. The resulting nonlinear system may be solved by the Gauss-Newton method (cf. e.g. [54]), where solutions with defect zero provide consistent initial values.*

[17] Here we consider the arguments explicitly, since they play an important role.

- *If we set first $U(t_0)x_0 = U(t_0)x^0$, instead of (2.58)-(2.60), the lower-dimensional system*

$$\begin{aligned} A(U(t_0)x_0 + T(t_0)x_0, t_0)y_0 + b(U(t_0)x_0 + T(t_0)x_0, t_0) &= 0, \\ W_1(U(t_0)x_0, t_0)\Big[(I_{W_1}W_0b)'_x(U(t_0)x_0, t_0)P(t_0)y_0\Big] & \\ + W_1(U(t_0)x_0, t_0)\Big[(I_{W_1}W_0b)'_t(U(t_0)x_0 + T(t_0)x_0, t_0)\Big] &= 0 \end{aligned}$$

  will be helpful to obtain the additionally required values $(T(t_0)x_0, P(t_0)y_0)$.

- *In addition to the aspects discussed in Remark 2.5.2, the approach (2.58)-(2.60) presents the following advantages:*

  - *For nonlinear systems from applications, Q_1 and PP_1 often depend on (x, t), while U is constant. Consequently, some of the difficulties that may appear for the generalization of (2.41) - (2.43) for nonlinear systems (cf.[14]) can be avoided considering (2.58)-(2.60). In fact, in [14] the full rank of the Jacobian of the obtained system was verified only for special cases.*
  - *For nonlinear systems, α cannot be chosen arbitrarily, and x^0 may be a reasonable guess. Moreover, for the special structure described in Section 2.7, that is precisely given in circuit simulation, the correction for x^0 is relatively easy to compute, because it results from a linear system.*

Example 2.5.4 *Let us consider again Example 2.1.4, which is in Hessenberg form. For this system, the above approach means that if we choose values for x_1 and x_2 on the cylinder, then the corresponding value for x_3 is determined by the equation describing the parabola, which is intuitively clear.*

Let us finally illustrate that if x^0 is chosen arbitrarily, then the system (2.58)-(2.60) may be unsolvable.

Example 2.5.5 *Consider*

$$\begin{aligned} x_1' - x_1 &= 0, \\ x_2' - \frac{x_3^2 - 0.5}{x_2} &= 0, \\ x_1^2 + x_2^2 - 1 &= 0, \end{aligned}$$

$x_i : \mathcal{I}_f \to I\!R$. It is easy to recognize that the explicit constraint is given by $x_1^2 + x_2^2 - 1 = 0$, while the hidden constraint arises from $x_1^2 + x_3^2 - 0.5 = 0$. Consequently, consistent initial values have to fulfil both equations. Let us now consider the following two cases:

- *$x_1^0 = 0$, $x_2^0 = 1$ fulfil the explicit constraint $x_1^2 + x_2^2 - 1 = 0$. By considering the system (2.58)-(2.60) and supposing $x_3 > 0$, we obtain the corresponding consistent values $(x_{10}, x_{20}, x_{30}) = (0, 1, \sqrt{0.5})$.*
- *$x_1^0 = \sqrt{0.9}$, $x_2^0 = \sqrt{0.1}$ also fulfil the explicit constraint $x_1^2 + x_2^2 - 1 = 0$. For these values, the system (2.58)-(2.60) is not solvable in $I\!R$.*

In Section 2.7 we will focus on a special structure of DAEs that implies that (2.58)-(2.60) can be formulated as a linear system and is, consequently, uniquely solvable for all $x^0 \in M_0$.

2.6 Application to DAEs in Hessenberg Form

Consider index-2 DAEs in Hessenberg form, i.e., systems

$$\begin{aligned} x_1'(t) &= b_1(x_1(t), x_2(t), t), && (2.64)\\ 0 &= b_2(x_1(t), t), && (2.65) \end{aligned}$$

with $B_{21}(\cdot)B_{12}(\cdot)$ nonsingular, for $B_{ij}(\cdot) := \frac{\partial b_i}{\partial x_j}(\cdot)$, $i, j = 1, 2$. This structure leads to

$$\begin{aligned} A &= \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}, \quad B(x_1, x_2, t) = \begin{pmatrix} B_{11}(x_1, x_2, t) & B_{12}(x_1, x_2, t) \\ B_{21}(x_1, t) & 0 \end{pmatrix}, \\ A_1(x_1, x_2, t) &= \begin{pmatrix} I & B_{12}(x_1, x_2, t) \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

Since $N = N \cap S(\cdot)$ is always constant, and

$$T = Q = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix}, \quad W_1 = W_0 = \begin{pmatrix} 0 & 0 \\ 0 & I \end{pmatrix},$$

the assumptions **A1**, **A2** are fulfilled, while **A3**, which coincides with **A5** due to the fact that W_1 is constant, is always supposed to be given. Moreover, we realize that the Tx and the Qx components coincide, a fact that simplifies the structure considerably.

Let us suppose that a value x_1^0 fulfilling

$$0 = b_2({x_1}^0, t_0)$$

is given. According to (2.58)-(2.60), in order to compute a consistent initialization for nonlinear index-2 Hessenberg systems, it would be adequate to consider, if it is solvable, the nonlinear system

$$\begin{aligned} y_{10} &= b_1(x_{10}, x_{20}, t_0), & (2.66)\\ x_{10} &= x_1^0, & (2.67)\\ 0 &= B_{21}(x_{10}, t_0)y_{10} + [b_2]'_t(x_{10}, t_0), & (2.68) \end{aligned}$$

where (y_{10}, x_{10}, x_{20}) are the unknowns.

Remark 2.6.1 *Note that instead of solving (2.66)-(2.68) we can fix $x_{10} = x_1^0$ and consider the system*

$$\begin{aligned} y_{10} &= b_1(x_{10}, x_{20}, t_0),\\ 0 &= B_{21}(x_{10}, t_0)y_{10} + [b_2]'_t(x_{10}, t_0), \end{aligned}$$

where (y_{10}, x_{20}) are the unknowns. This quadratic system may be solved by the Newton method (cf., in contrast, Remark 2.5.3).

In particular, for Hessenberg systems of the special structure

$$\begin{aligned} x_1'(t) &= \tilde{b}_1(x_1(t), t) + \mathcal{B}_1(x_1(t), t)x_2(t), & (2.69)\\ 0 &= \tilde{b}_2(x_1(t), t), & (2.70) \end{aligned}$$

a consistent initialization can be computed by solving only a linear system. In this case, (2.58)-(2.60) implies that, if we know values ${x_1}^0$ fulfilling (2.70), then we set $x_{10} = x_1^0$ for computing a consistent initialization and solve then the linear system that reads:

$$\begin{aligned} y_{10} &= \tilde{b}_1(x_{10}, t) + \mathcal{B}_1(x_{10}, t_0)x_{20},\\ 0 &= \tilde{B}_{21}(x_{10}, t_0)y_{10} + [\tilde{b}_2]'_t(x_{10}, t_0), \end{aligned}$$

where (y_{10}, x_{20}) are the unknowns.

Example 2.6.2 *The stabilized Euler-Lagrange equations described in [22] present the structure (2.69)-(2.70). We write the equations in a form that emphasizes their Hessenberg structure:*

$$\begin{aligned} p' &= v - G(p)^T\mu, \\ v' &= M(p)^{-1}f(p,v) - M(p)^{-1}G(p)^T\lambda, \\ 0 &= G(p)v, \\ 0 &= g(p), \end{aligned}$$

where $p, v \in I\!\!R^{n_p}$ are position and velocity variables, $\lambda \in I\!\!R^{n_\lambda}$ are Lagrange multipliers with $n_\lambda \leq n_p$, $M(p)$ is the positive definite mass matrix, $f(p,v)$ are the applied outer forces, $g(p)$ are the constraints, and $G(p) := \frac{\partial}{\partial p}g(p)$ is the constraint matrix with full rank n_λ.

The hidden constraints arise from

$$\begin{aligned} 0 &= G(p)v' + \left(\frac{d}{dt}G(p)\right)v =: G(p)v' + \tilde{G}(p,v)p', \\ 0 &= G(p)p'. \end{aligned}$$

Notice that for the above approach we suppose that we have values (v^0, p^0) that fulfil:

$$\begin{aligned} 0 &= G(p^0)v^0, \\ 0 &= g(p^0), \end{aligned}$$

where $x_{10} := x_1^0$ corresponds to

$$\begin{aligned} p_0 &= p^0, \\ v_0 &= v^0. \end{aligned}$$

Moreover, accordingly to (2.69)-(2.70), to compute $(x_{20}, x'_{10}) = (\mu_0, \lambda_0, p'_0, v'_0)$ we consider the system

$$\begin{aligned} p'_0 &= v_0 - G(p_0)^T\mu_0, \\ v'_0 &= M(p_0)^{-1}f(p_0,v_0) - M(p_0)^{-1}G(p_0)^T\lambda_0, \\ 0 &= G(p_0)v'_0 + \tilde{G}(p_0,v_0)p'_0, \\ 0 &= G(p_0)p'_0. \end{aligned}$$

Hence, a consistent initialization can be obtained successively by means of:

$$
\begin{aligned}
p_0 &= p^0, \\
v_0 &= v^0, \\
\mu_0 &= 0, \\
p_0' &= v_0, \\
\lambda_0 &= (G(p_0)M(p_0)^{-1}G(p_0)^T)^{-1}(G(p_0)M(p_0)^{-1}f(p_0,v_0) \\
&\quad + \tilde{G}(p_0,v_0)v_0), \\
v_0' &= M(p_0)^{-1}(f(p_0,v_0) - G(p_0)^T\lambda_0).
\end{aligned}
$$

Note that the correction we perform affects:

- *the values of μ that have to be* 0 *on the one hand,*
- *the values of λ that are completely fixed by p_0 and $v_0$on the other hand,*
- *and, finally, suitable values of the derivatives.*

Example 2.6.3 *For the index-2 formulation of the trajectory prescribed path control problem (TPPC) discussed in [4],[5], x_2 occurs nonlinearly. Thus, the structure (2.69)-(2.70) is not given. Consequently, the corresponding consistent value cannot be computed by solving only a linear system. Nevertheless, the solution of the corresponding nonlinear system arising if x_{10} is prescribed, is then exactly the one that can be found explicitly in [37].*

2.7 Analyzing a Special Structure

In the following we will focus on a special structure that considerably simplifies the task of solving the over-determined system (2.58)-(2.60), but does not correspond to the Hessenberg form. Indeed, we focus on a structure that is given in the applications we are interested in (cf. Chapter 3). This structure implies that (2.58)-(2.60) becomes a linear system with respect to $(T(t_0)x_0, P(t_0)y_0)$.

Let us assume that

$$
\begin{aligned}
\mathbf{A6}: \quad & \text{im}\ A(x,t), \quad \ker A(x,t) \quad \text{and}\ N(t)\cap S(x,t) \\
& \qquad\qquad \text{are constant for} \quad (x,t)\in\mathcal{D}_f\times\mathcal{I}_f,
\end{aligned}
$$

since this is given in the applications we are interested in. Observe that this assumption precisely implies that Q, P, T, U, W_0 are constant projectors.

Let us further suppose in the following that (1.14) has the structure

$$\mathbf{A7}: \quad A(Ux(t),t)x'(t) + \tilde{b}(Ux(t),t) + \mathcal{B}(Ux(t),t)Tx(t) = 0 \tag{2.71}$$

for a matrix $\mathcal{B}$, i.e., we suppose, that the $N \cap S$-component occurs only linearly.

Lemma 2.7.1 *Due to* **A7***, it holds that*

$$W_0\mathcal{B}(Ux,t)T = 0 \tag{2.72}$$

Proof: Note that because of the structure (2.71) it holds: $B(\cdot)T = \mathcal{B}(\cdot)T$. Therefore (see Lemma 2.3.4),

$$W_0\mathcal{B}(\cdot)T = W_0B(\cdot)T = 0.$$

q.e.d.

2.7.1 Calculation of Consistent Initial Values by Solving a Linear Sytem

Analogously as in Section 2.5, we start from a value x^0 that fulfils the equations of the DAE, but is probably not consistent. This means, we suppose we know values (x^0, Py^0) fulfilling

$$A(Ux^0,t_0)y^0 + \tilde{b}(Ux^0,t_0) + \mathcal{B}(Ux^0,t_0)Tx^0 = 0. \tag{2.73}$$

Actually, we are looking for a consistent value, i.e., a value x_0 that fulfils the equations of the DAE

$$A(Ux_0,t_0)y_0 + \tilde{b}(Ux_0,t_0) + \mathcal{B}(Ux_0,t_0)Tx_0 = 0, \tag{2.74}$$

as well as the hidden constraints[18]

$$W_1(Ux_0,t_0)\Big[(I_{W_1}W_0\tilde{b})'_x(Ux_0,t_0)Py_0 + (I_{W_1}W_0\tilde{b})'_t(Ux_0,t_0)\Big] = 0. \tag{2.75}$$

[18] Observe that for a constant projector U we have $(I_{W_1}W_0b)'_t(x,t) = (I_{W_1}W_0b)'_t(Ux,t)$ due to $(W_0b)(x,t) = (W_0b)(Ux,t)$. Therefore, in (2.60), this partial derivative with respect to time corresponds to $(I_{W_1}W_0\tilde{b})'_t(Ux_0,t_0)$.

Let us now recall the notation (cf. (2.53),(2.54))

$$\begin{aligned}\hat{x}_0 &= x_0 - x^0,\\ P\hat{y}_0 &= Py_0 - Py^0,\end{aligned}$$

and establish a relation between $\hat{x}_0$ and $\hat{y}_0$. If we set $Ux_0 = Ux^0$ and subtract (2.73) from (2.74) we obtain:

$$A(Ux_0,t_0)\hat{y}_0 + \mathcal{B}(Ux_0,t_0)T\hat{x}_0 = 0. \tag{2.76}$$

Moreover, due to (2.75), we know that $(\hat{x}_0, P\hat{y}_0)$ has to fulfil

$$\begin{aligned} W_1(Ux_0,t_0)\Big[(I_{W_1}W_0\tilde{b})'_x(Ux_0,t_0)P[y^0+\hat{y}_0]&\\ + W_1(Ux_0,t_0)\Big[(I_{W_1}W_0\tilde{b})'_t(Ux_0,t_0)\Big] &= 0.\end{aligned}$$

Theorem 2.7.2 *Suppose that* **A6**, **A7**, **A3 - A5** *hold, and sufficient smoothness*[19] *is given. Then we obtain consistent initial values* (x_0, Py_0) *starting from the possibly inconsistent values* (x^0, Py^0) *setting* $Ux_0 := Ux^0$, *computing the unique solution* $(\hat{x}_0, P\hat{y}_0)$ *of the linear system*

$$\begin{aligned} A(Ux_0,t_0)\hat{y}_0 + \mathcal{B}(Ux_0,t_0)T\hat{x}_0 &= 0, && (2.77)\\ U\hat{x}_0 &= 0, && (2.78)\\ W_1(Ux_0,t_0)(I_{W_1}W_0\tilde{b})'_x(Ux_0,t_0)P[y^0+\hat{y}_0] &&\\ + W_1(Ux_0,t_0)(I_{W_1}W_0\tilde{b})'_t(Ux_0,t_0) &= 0 && (2.79)\end{aligned}$$

and setting

$$\begin{aligned} x_0 &= x^0 + \hat{x}_0,\\ Py_0 &= Py^0 + P\hat{y}_0,\end{aligned}$$

for which (2.74) and (2.75) are fulfilled.

Proof:
Note that on the one hand, Lemma 2.7.1 implies that (2.77) can be rewritten as

$$A(Ux_0,t_0)\hat{y}_0 + (I - W_0(t_0))\mathcal{B}(Ux_0,t_0)T\hat{x}_0 = 0.$$

[19] cf. Section 2.4

Hence, the row rank of (2.77)-(2.79) is less than or equal to

$$\text{rank } A + \text{rank } U + \text{rank } W_1 = \text{rank } P + \text{rank } U + \text{rank } T = n + \text{rank } P$$

due to Remark 2.3.3.
On the other hand, if we suppose that (Pz_y, z_x) is a solution of the homogeneous system, then it can be deduced $(Pz_y, z_x) = 0$ analogously as in Section 2.5.3. Thus, the system is uniquely solvable.

q.e.d.

Specifics related to the realization in circuit simulation can be found in Chapter 3.

In the following we analyse the differences between the numerical solutions we obtain starting from values (x^0, Py^0) and from the corresponding consistent values (x_0, Py_0).

2.7.2 Consequences for the Implicit Euler Method

Recall that when solving (1.14) numerically by means of an implicit Euler method in the first step we solve the system:

$$A(x_1, t_1)\frac{x_1 - x_0}{h} + b(x_1, t_1) = 0.$$

Making use of the above results, we note that the same systems have to be solved starting at x_0 or at x^0, because $Ux_0 = Ux^0$ implies $Px_0 = Px^0$.

Remark 2.7.3 • *In practice, since the Jacobian required for the Newton method may depend on Tx, the results we obtain starting with the initial guess x^0 may differ from those achieved starting with the initial guess x_0. This applies, for instance, to the systems described in the Examples 2.6.2 and 2.6.3.*

- *If a system has the structure*

$$\mathbf{A8}: \quad A(Ux(t), t)x'(t) + b(Ux(t), t) + \mathcal{B}Tx(t) = 0 \qquad (2.80)$$

for a constant matrix $\mathcal{B}$, then the same initial guess is used in both cases to start the Newton iteration due to the fact that $Ux_0 = Ux^0$

and because $J = J(Ux, t)$ holds for the Jacobian. In this case exactly the same results are obtained starting from x_0 and x^0. This applies, for instance, to those systems that arise from circuit simulation (cf. Chapter 3).

2.7.3 Consequences for the Trapezoidal Rule

We now focus on systems of the form :

$$\mathbf{A9}: \quad Ax'(t) + \tilde{b}(Ux(t), t) + \mathcal{B}Tx(t) = 0,$$

where $A, \mathcal{B}, U, T$ are constant.

For ODEs

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

the trapezoidal rule reads:

$$\frac{x_1 - x_0}{h} = \frac{f(x_1, t_1) + f(x_0, t_0)}{2}.$$

Remark 2.7.4 *Recall that the convergence and stability properties of the trapezoidal rule are not desirable (cf. e.g. [31]). Thus, the trapezoidal rule should be used only in combination with the Backward Difference Formulae (BDF) [62], for instance.*

There are several possibilities to adapt this method to DAEs. We will consider the method presented in [62], which introduces the approximation[20]

$$A\frac{dx}{dt}(t_1) = 2A\frac{x_1 - x_0}{h} - A\frac{dx}{dt}(t_0)$$

in equations of the structure **A9**.

Lemma 2.7.5 *The structure* **A9** *implies the following properties for the matrix chain of the tractability index we obtain:*

- *A is constant*

[20] See also [25], where, more generally, Runge-Kutta methods are considered for DAEs, in particular. These methods are denoted by IRK(DAE).

- $B = B(Ux, t)$.
- *From the above we obtain:*
  $G_1 = G_1(Ux,t)$, $A_1 = A_1(Ux,t)$, $Q_1 = Q_1(Ux,t)$,
  $G_2 = G_2(Ux,t) = G_1(Ux,t) + B(Ux,t)PQ_1(Ux,t)$ *and* $W_1 = W_1(Ux,t)$.

Proof: The assertions follow by straightforward computation.

We will see that if we apply the trapezoidal rule for DAEs of this shape, the systems we have to solve starting from (x_0, Py_0) or by (x^0, Py^0) are not the same. Nevertheless, we will show that the obtained results are only different for Tx, i.e., the error that is introduced because we do not fulfil the hidden constraint affects only the value of the $N \cap S$-component of the next step. The values for the remaining components are the same starting from (x_0, Py_0) or (x^0, Py^0).

To prove this phenomenon we consider the system we obtain starting from the value (x_0, Py_0):

$$2 \cdot A\frac{x_1 - x_0}{h} - Ay_0 + \tilde{b}(Ux_1, t_1) + \mathcal{B}Tx_1 = 0.$$

From the relation $A\hat{y}_0 + \mathcal{B}T\hat{x}_0 = 0$ (cf. (2.76)) we obtain

$$2 \cdot A\frac{x_1 - x_0}{h} - Ay^0 + \mathcal{B}T\hat{x}_0 + \tilde{b}(Ux_1, t_1) + \mathcal{B}Tx_1 = 0. \tag{2.81}$$

With the aid of the projectors of the tractability index it is possible to recognize that the term $\mathcal{B}T\hat{x}_0$, which is the only discrepancy with respect to the corresponding system we obtain starting from (x^0, Py^0), affects precisely Tx_1. To this end, we split (cf. (2.33)-(2.36)) the equation (2.81) multiplying it by

$$(PP_1(\cdot) + PQ_1(\cdot) + UQ)G_2^{-1}(\cdot) \quad \text{and} \quad TG_2^{-1}(\cdot),$$

evaluated at (Ux_0, t_0). Since we only use these terms to split the system, this can be done considering them a constant expression, because we already know $Ux_0 = Ux^0$. Further, this implies that we can use the same projectors for the splitting of the system we obtain starting from the value (x^0, Py^0). With Lemma 2.3.1,4c we obtain

$$\begin{aligned} 2 \cdot PP_1\frac{x_1 - x_0}{h} - PP_1y^0 + (PP_1 + PQ_1 + UQ)G_2^{-1}\tilde{b}(Ux_1, t_1) &= 0, &(2.82)\\ -2 \cdot QQ_1\frac{x_1 - x_0}{h} - QQ_1y^0 + T\hat{x}_0 + Tx_1 + TG_2^{-1}\tilde{b}(Ux_1, t_1) &= 0. &(2.83)\end{aligned}$$

Observe that for calculating Ux_1 we only have to consider (2.82). This can easily be seen considering the range of the equations. Realize that (2.83) contains as many linearly independent equations as rank T = rank QQ_1. Taking into account that Tx_1 only appears in (2.83), these equations have to fix Tx_1, while Ux_1 is fixed by (2.82). Furthermore, in the above system we can notice that $T\hat{x}_0$ only appears in (2.83). Therefore, the same value for Ux_1 is obtained starting from (x_0, Py_0) or (x^0, Py^0).

Remark 2.7.6 *Notice further that we obtain*

$$Tx_1 = -T\hat{x}_0 + 2 \cdot QQ_1 \frac{x_1 - x_0}{h} + QQ_1 y^0 - TP_1 G_2^{-1} \tilde{b}(Ux_1, t_1).$$

Thus, if further steps are undertaken, the error induced by the inconsistency alternates the sign. For instance, if we suppose that the value calculated for Ux_1 *is accurate, and denote by* x^1 *the corresponding value obtained starting from* (x^0, Py^0)*, then the above discussion would imply*

$$Ux_1 = Ux^1, \quad Tx_1 = -T\hat{x}_0 + Tx^1.$$

Consequently, if x_2 *and* x^2 *denote the values obtained starting from* x_1 *and* x^1*, respectively, analogously as above it results*

$$Ux_2 = Ux^2, \quad Tx_2 = -(Tx_1 - Tx^1) + Tx^2 = T\hat{x}_0 + Tx^2.$$

Remark 2.7.7 *Let us finally remark that if we consider systems of the structure*

$$Ax'(t) + b(Ux(t), t) + \mathcal{B}(t)Tx(t) = 0, \tag{2.84}$$

for instance, the splitting (2.82) - (2.83) does not work any more, since the matrices are evaluated at different times. Consequently, it does not hold that the error in the $N \cap S$*-component cannot be transferred to other components. In Section 3.6 we will give an example that illustrates this effect. Notice also that the observation we made with respect to the implicit Euler method holds analogously for the structure (2.84).*

2.8 Some Concluding Remarks

Let us consider in detail on the differences and advantages of the presented approach with respect to some of those discussed in Section 2.2.

Since differentiations are numerically difficult, the approaches (e.g. [38],[23]) based on the consideration of the derivative array (2.3) have to cope with the disadvantage of considering unnecessarily high derivatives. Indeed, for index-2 DAEs, the derivative array involves second derivatives of the original DAE. In contrast, the approach presented here differentiates a part of the original DAE only once.

Recall that the algorithm from [48] also derives only suitable parts of the original DAE, but, as mentioned before, some equations that have to be differentiated may escape detection. However, for index-2 DAEs with structural index 2 some parts of the original DAE are derived twice. Recently [52], it was realized that the structural index may also exceed the differential index, even for DAEs with constant coefficients. Indeed, this applies to simple examples from circuit simulation. Consequently, the structural determination of the index and of the consistent initial values is not reliable.

With respect to the other approaches based on the tractability-index we require relatively weak assumptions on the projectors. Concretely, the approaches from [32],[36] require that $PQ_1 = (PQ_1)(t)$ is given, while in [43] $W_1 = W_1(P(t)x, t)$ was requested to hold. In this context it has to be mentioned that, in contrast to **A2**, these assumptions are not given in the applications of Chapter 3[21].

As mentioned before, the algorithms from [48] and [14] determine a selection of variables and a component, respectively, for which we may prescribe suitable initial values. Both approaches base on the assumption that for nonlinear DAEs the obtained systems are solvable. Thus, even if the algorithm from [48] works, the resulting system has to be solvable with respect to the variables left unspecified. Analogously, in [14] the system corresponding to (2.41)-(2.43) has to be solvable for nonlinear DAEs. Thus, a further advan-

[21] These assumptions are not given for the equations arising from MNA, since for the charge-oriented MNA, for example, the projectors PQ_1 and W_1 even depend on (Ux) (cf. [15]).

tage of (2.58)-(2.60) is that we always can enlarge it by $Q(t_0)y_0 = 0$ in order to obtain a full column rank Jacobian and that, moreover, this system is often linear in applications.

In contrast to the approach [23], which computes initial values minimizing the deviation of the variables from a specified guess, by (2.58)-(2.60) we compute initial values for which the deviation affects only the so-called index-2 component (corresponding to $N \cap S$). Consequently, even if we start the different approaches with an x^0 as the initial guess, they may lead to different consistent initial values. Since in practice, the user of a simulation package knows Px^0 sometimes and wants to preserve this values for Px_0, this becomes another advantage of (2.58)-(2.60).

Finally, it has to be noticed that (2.58)-(2.60) can be considered to be added-on certain algorithms that compute consistent initial values for index-1 variables. For instance, the computation of a value $(x^0, P(t_0)y^0)$ may be carried out analogously as consistent initial values for index-1 DAEs are computed by the approach described in point 2 on p. 24.

An important and interesting matter of research resulting from the problems related to the computation of consistent initial values are the numerical consequences of starting an integration process with inconsistent values. For instance, for index-1 DAEs of the form $Ax'(t) + b(x(t), t) = 0$ such considerations were presented in [55]. There it was realized that a variety of implicit Runge-Kutta methods converge at the same rate whether or not the initial conditions are consistent. For systems arising in circuit simulation, in [61] some considerations related to this were made for the implicit Euler method. For investigations of this kind, the results presented in the Sections 2.7.2 and 2.7.3 become of special interest, since a better understanding of the properties of the index-2 components becomes possible.

Let us emphasize at last the following aspects:

- Since differentiation problems are ill-posed in the sense of Hadamard, i.e., small perturbations in the input data can provide arbitrarily large perturbations in the output data, it is advisable to perform as few differentiations as possible in general. Thus, the approaches based on the consideration of $M_1(t)$ (see Section 2.4) benefit from the fact that they involve considerably less differentiations than those based on the derivative array (2.3).

- The requested assumptions are more general and easier to verify than those of other approaches based on projectors related to the tractability-index.
- The approach (2.58)-(2.60) provides new insights for the understanding of the effect that inconsistent initial values may have on numerical solutions (cf. Section 2.7.2 and 2.7.3).

Consequently, it is intended to continue this work focusing on the following:

- (2.58)-(2.60) will be tested [60], while it has only been implemented for systems arising from circuit simulation (see Chapter 3) so far.
- It would be interesting to investigate the effect that inconsistent initial values may have on numerical solutions when considering further integration methods. Here we focus on the implicit Euler method and the trapezoidal rule only, since these are commonly used in circuit simulation.

At last, we want illustrate that if the structural assumptions from Section 2.3 are not met, there may not exist such a direct relation between a value fulfilling the equations of the DAE and a consistent initial value.

Example 2.8.1 *For $t \geq 1$, $x_i : \mathcal{I}_f \to I\!R$, $x_2, x_3 > 0$ consider:*

$$\begin{aligned} x_1' + x_2' + x_3 &= 0, \\ x_2 x_3 &= 1 + \ln t, \\ x_1 + x_2 + x_2 x_3 &= 0. \end{aligned}$$

For this index-2 example $N \cap S(\cdot)$ does not depend only on t:

$$N \cap S(x,t) = \{z \in I\!R^3 : \quad z_1 + z_2 = 0, \quad x_3 z_2 + x_2 z_3 = 0\}.$$

Note that a consistent initial value is given by $x_3 = \frac{1}{t_0}$, $x_2 = t_0(1 + \ln t_0)$, $x_1 = -(1 + t_0)(1 + \ln t_0)$, but that no linear relation exists between these values and other values that fulfil the equations.

2.9 An Example: The NAND-Gate

The NAND-gate is a logical gate that computes the elementary logical operation NAND (Not AND). It consists of two n-channel enhancement MOSFETs (MEs), one n-channel depletion MOSFET (MD) and one load capacitance C (cf. [28],[57]).

The drain voltage of MD is constant at $V_{DD} = 5$V. The bulk voltages are not at ground: $V_{BB} =$ -2.5 V. The source voltages of both MEs are at ground. The gate voltages of both enhancement MOSFETs are controlled by two voltage sources V_1 and V_2.

Roughly speaking, the MOSFETs act as a switch between drain and source: they will close if the voltage between gate and source drops below a certain threshold value. This means that as soon as V_1 or V_2 are low, the corresponding MEs will lock. If V_1 and V_2 exceed a given threshold, then a drain current will flow through both MEs and the voltage at node 1 will break down. Hence, depending on the input voltages, a response is generated at node 1, representing the Not AND-operation. The response at node 1 will only be LOW (FALSE) if both V1 and V2 exceed a given threshold voltage U_T, i.e. both are HIGH (TRUE).

We consider the MOSFET- model[22] from [16], that implies that the NAND-gate equations are index-2 tractable [57]. The MOSFETs MD and ME differ only in parameter values.

The equations can be found in the Appendix. The vector of unknowns reads

$$\begin{aligned}&(q, q_{1gd}, q_{1gs}, q_{1db}, q_{1sb}, q_{2gd}, q_{2gs}, q_{2db}, q_{2sb}, q_{3gd}, q_{3gs}, q_{3db}, q_{3sb},\\ &\quad e_1, e_2, e_3, e_4, e_5, e_6, e_7, e_8, e_9, e_{10}, e_{11}, e_{12}, j_1, j_2, j_{BB}, j_{DD})\,.\end{aligned}$$

Straightforward computation shows that a projector onto the space $N \cap S(\cdot)$

[22] Note tat if we consider a different model (e.g. from [56]), then the index of the NAND-Gate becomes 1 (cf.[29]).

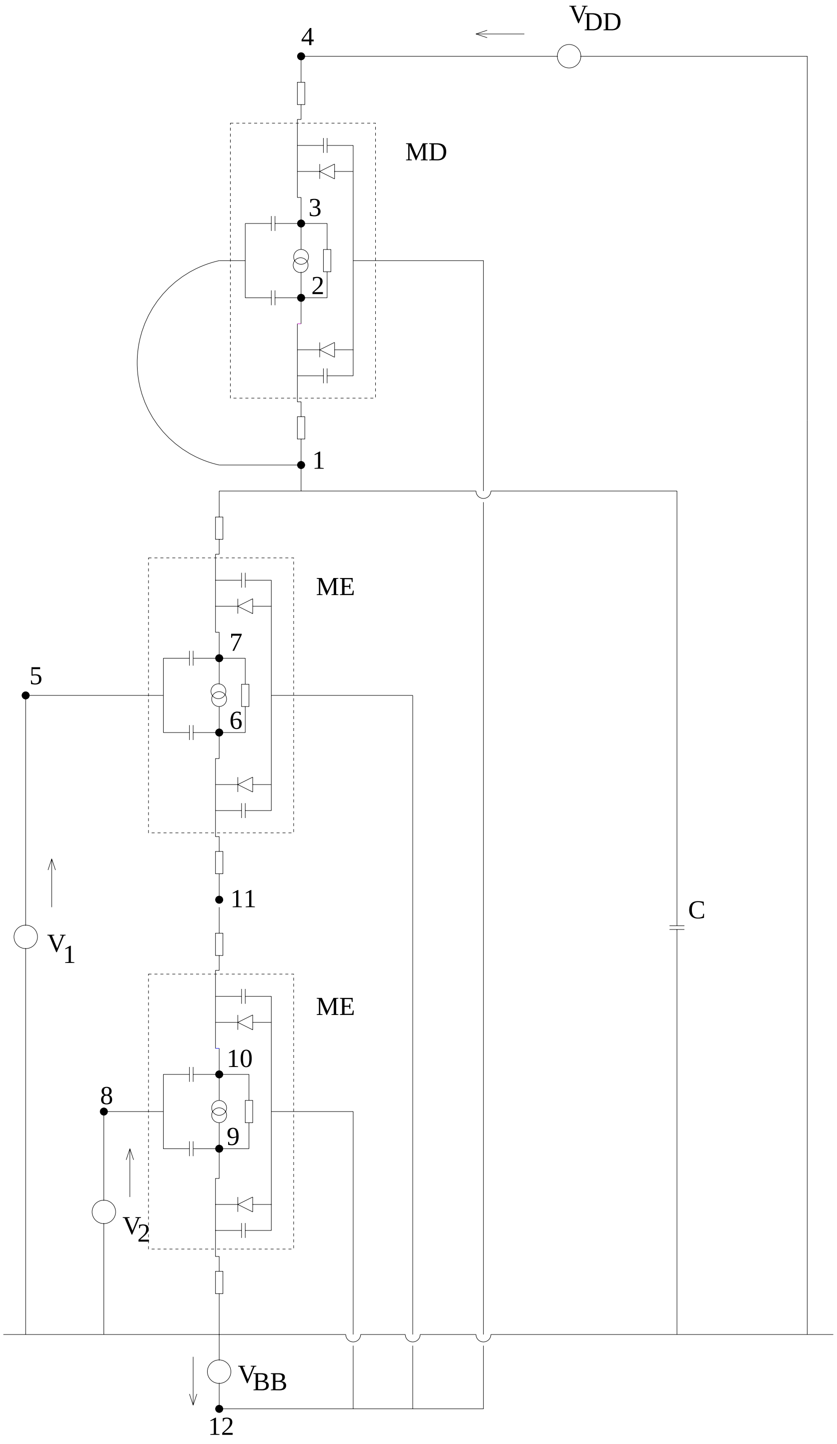


Figure 2.1: NAND-Gate

is given by

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

This means that in order to obtain a consistent value it may be necessary to correct the currents through V_1, V_2, V_{BB}.

We focus on $V_1(t_0) = V_2(t_0) = 0$. Let us suppose that $(x^0, P(t_0)y^0) = (x^0, 0)$, i.e., we consider the so-called DC-operating point. For $V_1'(t_0) = 10^9$, $V_2'(t_0) = V_{BB}' = 0$ the corrections $(\hat{x}^0, P(t_0)\hat{y}^0)$ computed by the algorithm described in Section 3.5 read:

$$\hat{x}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ -7.40744E-05 \\ 0 \\ 7.40744E-05 \\ 0 \end{pmatrix}, \; y_0 = \hat{y}_0 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 4.07947E-05 \\ 3.32797E-05 \\ 4.07947E-05 \\ 3.32797E-05 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The result shows that the difference between x^0 and x_0 consists in a current that flows through V_1, through the enhancement MOSFET that is incident with node 5, and through V_{BB}. Note that inside the MOSFET the current is divided. Since we have $V_2'(t_0) = V_{BB}' = 0$, it results that no additional current flows through V_2.

The values for x^0 and x_0 can be found in the Appendix. The values for x_0 correspond to those obtained in [14] for $\alpha = x^0$. The values obtained when considering only linear capacitances (cf. [57]) can be found in [11].

Chapter 3

Application to Circuit Simulation

The index and the structure of the equations we obtain in electric circuit simulation depend, among other things, on the scheme for setting up the equations. We will restrict ourselves to one of the most frequently used modelling techniques, the modified nodal analysis (MNA). In Section 3.1 we introduce the equations arising from two different formulations, the conventional MNA and the charge-oriented MNA, in order to analyse their special structure afterwards. These equations consist of the nodal equations (given by Kirchhoff's current law) and the characteristic equations of the voltage-defining elements, i.e., inductances and voltage sources. In the case of the charge-oriented MNA, the voltage-charge and current-flux equations are also added to the system. Because of the large dimension of many circuits (often 10^5 circuit elements), it is difficult, in general, to determine their structural properties. Nevertheless, if the positive definiteness of the Jacobians of the element-characterizing functions is assumed, the problem simplifies considerably. Beyond this, since Kirchhoff's laws describe linear relations, some of the variables occur only linearly, even if the capacitances, inductances and resistances are highly nonlinear. Indeed, in order to guarantee the structure requested in Chapter 2, only the voltage-controlled voltage sources (VCVS), current-controlled voltage sources (CCVS), voltage-controlled current sources (VCCS), and current-controlled current sources (CCCS) have to be analyzed. The class of controlled sources we will consider is exactly the one for which the structure of the spaces associated to the DAE can be described analogously as for networks without controlled sources. In particular, Section

3.2 is devoted to ascertaining that the MNA equations have the structural properties presented in Section 2.7 if a network contains only the class of controlled sources described in Section 3.1.3.

Insight in how to compute practically consistent initial values by applying the technique from Section 2.7.1 is given in Section 3.3. Moreover, in Section 3.4 it will be briefly outlined how this computation can be considerably simplified. Concretely, we reduce the computational costs determining the hidden constraints by means of a graph-theoretical approach that makes it unnecessary to figure out the corresponding projectors explicitly. Finally, in Section 3.5 we will give some details about the realization and discuss, in Section 3.6, an example that illustrates the effects described in the Sections 2.7.2 and 2.7.3.

The results presented in this chapter were partly developed in [15],[12],[11]. In particular, Section 3.1 and the first part of Section 3.2 have been taken from [15], where the assumptions for the controlled current sources have been slightly modified in order to guarantee **A2**. Here, we aim at summarizing the results from [15],[12],[11] in connection with the general theory developed in the previous chapter. Since these articles contain more examples and details concerning the realization, the interested reader is referred to them.

3.1 The Modified Nodal Analysis (MNA)

In the following we discuss lumped electric circuits containing nonlinear and possibly time-variant resistances, capacitances, inductances, voltage sources and current sources. Usually, circuit simulation tools are based on these kinds of network elements. For two-terminal (one-port) lumped elements, the current through the element and the voltage across it are well-defined quantities. For lumped elements with more than two terminals, the current entering any terminal and the voltage across any pair of terminals are well defined at all times (cf. [10]). Hence, general n-terminal elements are completely described by $(n-1)$ currents entering the $(n-1)$ terminals and the $(n-1)$ branch voltages across each of these $(n-1)$ terminals and the reference terminal n.

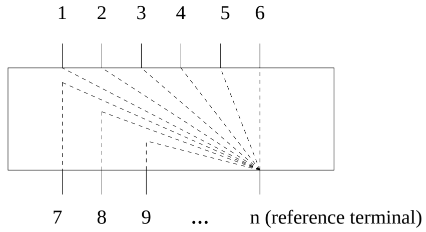


Figure 3.1: n-terminal circuit element

In particular, n-terminal resistances can be modelled by an equation system of the form

$$j_k = r_k^e(u_1, ..., u_{n-1}, t) \quad \text{for} \quad k = 1, ..., n-1$$

if j_k represents the current entering the terminal k and u_l describes the voltage across the pair of terminals $\{l, n\}$ (for $k, l = 1, ..., n-1$). Kirchhoff's Current Law implies the current entering the terminal n to be given by $j_n = -\sum_{k=1}^{n-1} j_k$. The conductance matrix $G^e(u_1, ..., u_{n-1}, t)$ is then defined by the Jacobian

$$G^e(u_1, ..., u_{n-1}, t) := \begin{pmatrix} \frac{\partial r_1^e}{\partial u_1} & \cdots & \frac{\partial r_1^e}{\partial u_{n-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial r_{n-1}^e}{\partial u_1} & \cdots & \frac{\partial r_{n-1}^e}{\partial u_{n-1}} \end{pmatrix}.$$

The index e shall specify the correlation to a special element of a circuit. Later on we will introduce the conductance matrix $G(u, t)$ describing all resistances of a circuit. Correspondingly, the capacitance matrix $C^e(u_1, ..., u_{n-1}, t)$ of a general n-terminal capacitance is given by

$$C^e(u_1, ..., u_{n-1}, t) := \begin{pmatrix} \frac{\partial q_1^e}{\partial u_1} & \cdots & \frac{\partial q_1^e}{\partial u_{n-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial q_{n-1}^e}{\partial u_1} & \cdots & \frac{\partial q_{n-1}^e}{\partial u_{n-1}} \end{pmatrix}$$

if the voltage-current relation is defined by means of charges by

$$j_k = \frac{d}{dt} q_k^e(u_1, ..., u_{n-1}, t) \quad \text{for} \quad k = 1, ..., n-1.$$

In order to illustrate what the matrices C^e may look like, let us consider a MOSFET-model as an example of a common n-terminal element.

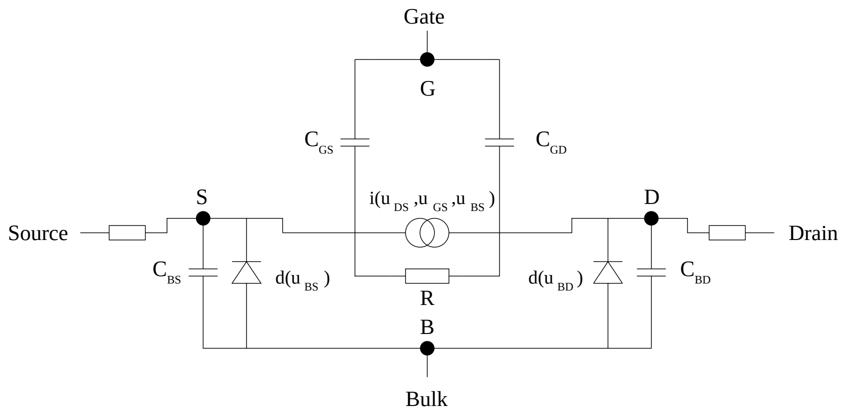


Figure 3.2: MOSFET-model

Choosing the source node S as the reference node, we have the reference voltages u_{GS}, u_{DS}, and u_{BS}. For the currents we obtain

$$\begin{aligned}
j_G &= C_{GS}\dot{u}_{GS} + C_{GD}(\dot{u}_{GS} - \dot{u}_{DS}),\\
j_D &= -C_{GD}(\dot{u}_{GS} - \dot{u}_{DS}) - C_{BD}(\dot{u}_{BS} - \dot{u}_{DS})\\
&\quad + d(u_{BS} - u_{DS}) + i(u_{GS}, u_{DS}, u_{BS}) + \frac{1}{R}u_{DS},\\
j_B &= C_{BS}\dot{u}_{BS} + C_{BD}(\dot{u}_{BS} - \dot{u}_{DS}) - d(u_{BS}) - d(u_{BS} - u_{DS}).
\end{aligned}$$

Note that j_S is given by the formula $j_S = -j_G - j_D - j_B$ due to Kirchoff's Current Law. Now it is easy to verify that

$$C^e(u_{GS}, u_{DS}, u_{BS}) = \begin{pmatrix} C_{GS} + C_{GD} & -C_{GD} & 0 \\ -C_{GD} & C_{GD} + C_{BD} & -C_{BD} \\ 0 & -C_{BD} & C_{BS} + C_{BD} \end{pmatrix}$$

for the MOSFET-model from [16].

Inductances can be modelled by means of fluxes by

$$u_k = \frac{d}{dt}\phi_k^e(j_1, ..., j_{n-1}, t) \quad \text{for} \quad k = 1, ..., n-1.$$

Then, the inductance matrix $L^e(j_1, ..., j_{n-1}, t)$ of a general n-terminal inductance is given by the Jacobian

$$L^e(j_1, ..., j_{n-1}, t) := \begin{pmatrix} \frac{\partial \phi_1^e}{\partial j_1} & \cdots & \frac{\partial \phi_1^e}{\partial j_{n-1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi_{n-1}^e}{\partial j_1} & \cdots & \frac{\partial \phi_{n-1}^e}{\partial j_{n-1}} \end{pmatrix}.$$

A commonly used method for network analysis in circuit simulation packages like TITAN[1] and SPICE[2] is the Modified Nodal Analysis (MNA). It represents a systematic treatment of general circuits and is important when computers perform the analysis of networks automatically. The scheme to set up the MNA equations is:

1. Write node equations by applying Kirchhoff's Current Law (KCL) to each node except for the datum node:

   $$Aj = 0. \tag{3.1}$$

   The vector j represents the branch current vector. The matrix A is called the (reduced) incidence matrix, which is defined by

   $$a_{ik} := \begin{cases} +1 & \text{if branch k leaves node } i \\ -1 & \text{if branch k enters node } i \\ 0 & \text{if branch } k \text{ is not incident with node } i \end{cases}$$

   for all the nodes i but the datum node (cf. [10]).
   Observe that the incidence matrix describes the network graph, the branch-node relations.

2. Replace the currents j_k of voltage-controlled elements by the voltage-current relations of these elements in equation (3.1).

3. Add the current-voltage relations for all current-controlled elements.

Note that, in case of multi-terminal elements with n terminals, we speak of branches if they represent a pair of terminals $\{l, n\}$ with $1 \leq l \leq n-1$.

In general, the MNA leads to quasilinear DAEs. In order to obtain more detailed information about the structure of these DAEs, we split the (reduced) incidence matrix A into the element-related incidence matrices

$$A = (A_C, A_L, A_R, A_V, A_I),$$

where A_C, A_L, A_R, A_V, and A_I describe the branch-current relations for capacitive branches, inductive branches, resistive branches, branches of voltage

[1] Infineon Technologies (formerly SIEMENS AG).

[2] Developed in the 70s by the University of California, Berkeley.

sources and branches of current sources, respectively. Denote by e the node potentials (excepting the datum node) and by j_L and j_V the current vectors of inductances and voltage sources. Defining the vector of functions for current and voltage sources by i and v, respectively, we obtain the following quasilinear DAE-system from the MNA:

$$\begin{aligned} A_C \frac{dq(A_C^T e,t)}{dt} + A_R r(A_R^T e,t) + A_L j_L + A_V j_V & \\ + A_I i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) &= 0, \qquad (3.2) \\ \frac{d\phi(j_L,t)}{dt} - A_L^T e &= 0, \qquad (3.3) \\ A_V^T e - v(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) &= 0. \qquad (3.4) \end{aligned}$$

Note that the vectors $A_C^T e$, $A_L^T e$, $A_R^T e$ and $A_V^T e$ describe the branch voltages for the capacitive, inductive, resistive and voltage source branches, respectively.

Remark 3.1.1 *Due to the fact that the currents through resistances are functions of the branch potentials, we do not include them separately as controlling functions. Of course, if the network does not contain controlled sources, then the source functions reduce to functions $i(t)$ and $v(t)$ that depend on time only.*

Nowadays, circuit simulation packages use two different approaches for solving (3.2)-(3.4): the conventional and the charge-oriented one.

3.1.1 The Conventional MNA

For the conventional MNA the vector of unknowns consists of all node voltages and all branch currents of current-controlled elements.

Defining

$$C(u,t) := \frac{\partial q(u,t)}{\partial u},\ q_t'(u,t) := \frac{\partial q(u,t)}{\partial t},\ L(j,t) := \frac{\partial \phi(j,t)}{\partial j},\ \phi_t'(j,t) := \frac{\partial \phi(j,t)}{\partial t}$$

we obtain[3]

$$\begin{aligned}
A_C C(A_C^T e,t) A_C^T \frac{de}{dt} + A_C q_t'(A_C^T e,t) + A_R r(A_R^T e,t) & & \\
+ A_L j_L + A_V j_V + A_I i(A^T e, A_C^T \frac{de}{dt}, j_L, j_V, t) &= 0, & (3.5)\\
L(j_L,t)\frac{dj_L}{dt} + \phi_t'(j_L,t) - A_L^T e &= 0, & (3.6)\\
A_V^T e - v(A^T e, A_C^T \frac{de}{dt}, j_L, j_V, t) &= 0. & (3.7)
\end{aligned}$$

Later on we will also need

$$G(u,t) := \frac{\partial r(u,t)}{\partial u},\ r_t'(u,t) := \frac{\partial r(u,t)}{\partial t}.$$

3.1.2 The Charge-oriented MNA

In comparison with the conventional MNA, the vector of unknowns consists additionally of the charge of capacitances and the flux of inductances. Moreover, the original voltage-charge and current-flux equations are added to the system. The resulting system is then of the form (cf. [26])

$$\begin{aligned}
A_C \frac{dq}{dt} + A_R r(A_R^T e,t) + A_L j_L + A_V j_V & & \\
+ A_I i(A^T e, \frac{dq}{dt}, j_L, j_V, t) &= 0, & (3.8)\\
\frac{d\phi}{dt} - A_L^T e &= 0, & (3.9)\\
A_V^T e - v(A^T e, \frac{dq}{dt}, j_L, j_V, t) &= 0, & (3.10)\\
q - q_C(A_C^T e,t) &= 0, & (3.11)\\
\phi - \phi_L(j_L,t) &= 0. & (3.12)
\end{aligned}$$

[3] Note that we have

$$\frac{dq(A_C^T e,t)}{dt} = C(A_C^T e,t) A_C^T \frac{de}{dt} + q_t'(A_C^T e,t).$$

Therefore, $i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) = i_*(A^T e, A_C^T \frac{de}{dt}, j_L, j_V, t)$ for a suitable function i_*. An analogous relation is valid for the controlled voltage-sources. For simplicity, we drop the index $*$.

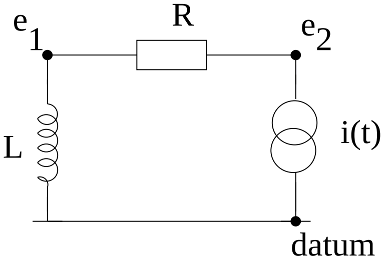
This formulation seems to be more convenient in practice. A detailed discussion of advantages of the charge-oriented MNA with respect to the conventional MNA can be found in [26].

3.1.3 The Index of the MNA Equations

In the following we discuss the index and the structure of the equations introduces above. For this purpose, some special cutsets[4] and loops[5] will be important. Therefore we define:

Definition 3.1.2 *[15]*

1. *An **L-I cutset** is a cutset consisting of inductances and/or current sources only.*



Conventional MNA:

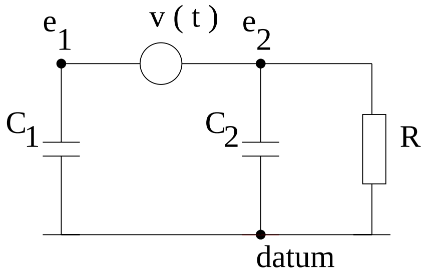
$$\begin{aligned} j_L + \frac{1}{R}(e_1 - e_2) &= 0, \\ -\frac{1}{R}(e_1 - e_2) + i(t) &= 0, \\ Lj_L' - e_1 &= 0. \end{aligned}$$

Figure 3.3: Example of an L-I cutset

2. *A **C-V loop** is a loop consisting of capacitances and voltage sources only.*

[4] A set of branches of a connected graph is called a cutset if the removal af all the branches of the set causes the remaining graph to have two separate parts and the removal of all but any one of the branches of the set leaves the remaining graph connected (cf. [10]).

[5] A subgraph of a graph is called a loop if it is connected and precisely two branches of it are incident with each node (cf. [10]).



Conventional MNA:

$$\begin{aligned} C_1 e_1' + j_V &= 0, \\ -j_V + C_2 e_2' + \frac{1}{R} e_2 &= 0, \\ e_1 - e_2 &= v(t). \end{aligned}$$

Figure 3.4: Example of a C-V loop

Theorem 3.1.3 *[15] Consider lumped electric circuits containing resistances, capacitances, inductances, and voltage and current sources. Let the capacitance, inductance and conductance matrices of all capacitances, inductances, and resistances, respectively, be positive definite.*[6] *Furthermore, let the following conditions for the controlled sources*[7] *be satisfied:*

1. *The controlled voltage sources do not form a part of any C-V loop and their controlling elements fulfil the conditions exposed in the Tables 3.1 and 3.2.*

2. *Each controlled current source fulfils at least one of the following conditions:*

   (a) *It does not form a part of any L-I cutset and the controlling elements fulfil the conditions exposed in the Tables 3.3 and 3.4.*

   (b) *There exists a path formed by capacitances that connects its incidence nodes. The controlling elements fulfil the conditions exposed in Table 3.6 for CCCS, and the VCCS are controlled by voltages fulfilling the conditions from Table 3.5.*

   (c) *There exists a path formed by capacitances and voltage sources that connects its incidence nodes. The controlling elements fulfil*

[6] For capacitances and inductances with affine characteristics the positive definiteness implies that they are strictly locally passive (cf. [17]).

[7] More precisely, we can permit controlling voltages and currents that can be expressed in terms of the listed voltages and currents. For instance, the current through a resistance that forms a loop with capacitances only can be expressed as a function of the voltage across those capacitances and can thus be allowed. The interested reader is referred to [15].

the conditions exposed in Table 3.7 for CCCS, and the VCCS are controlled by voltages fulfilling the conditions from Table 3.5.

Then, the conventional MNA leads to a DAE with index[8] < 2 *if and only if the network contains neither L-I cutsets nor C-V loops. Otherwise, the conventional MNA leads to an index-2 DAE.*

The controlling voltages of a VCVS can be voltages of:
1. capacitances,
2. independent voltage sources.

Table 3.1: VCVS - condition (1)

The controlling currents of a CCVS can be currents of:
1. inductances,
2. independent current sources.

Table 3.2: CCVS - condition (1)

The controlling voltages of a VCCS can be voltages of:
1. capacitances,
2. voltage sources.

Table 3.3: VCCS - condition (2a)

Theorem 3.1.4 *[15] The same conclusions as in Theorem 3.1.3 are valid under the same assumptions if we consider the charge-oriented MNA instead of the conventional MNA.*

[8] For reasons of simplicity, we do not consider the index-0 cases, which correspond to regular ODEs, and the index-1 cases, separately.

The controlling currents of a CCCS can be currents of:

1. inductances,
2. independent current sources,

Table 3.4: CCCS - condition (2a)

The controlling voltages of a VCCS can be voltages of:

1. capacitances,
2. voltage sources,
3. resistances.

Table 3.5: VCCS - conditions (2b), (2c)

The reader who is not interested in details concerning the controlled sources may suppose that the considered network contains only independent sources. By doing so, the upcoming discussion can be considerably simplified.

Remark 3.1.5 *1. The presented criteria can be checked locally. It is neither necessary to find special trees nor to make additional assumptions on the functions and parameters that define the controlled sources. Usually, it is not difficult to check whether a model of a network element including controlled sources satisfies these conditions or not.*

2. If no assumptions on the controlled sources are made, different problems arise. On the one hand, if arbitrary controlling elements for the controlling sources are considered, then the index of the network equations may depend on the parameters defining them (cf. [51]). On the other hand, if controlled sources are allowed to form part of L-I-cutsets of C-V-loops, it is possible to be confronted with higher index (>2) problems (cf. [27]).

Example 3.1.6 *Consider again the MOSFET-model given in Figure 4.9. The VCCS from source to drain is controlled by the branch voltages u_{GS}, u_{DS}, and u_{BS}. For these, the conditions (2a)-(2c) are satisfied since there are capacitive ways from gate to source, from drain to source as well as from bulk to source, and there exists a capacitive way from source to drain.*

The controlling current of a CCCS can be the current of:

1. inductances,
2. independent current sources,
3. resistances,
4. voltage sources that do not form a part of a C-V loop.

Table 3.6: CCCS - condition (2b)

The controlling current of a CCCS can be the current of:

1. inductances,
2. resistances,
3. independent current sources.

Table 3.7: CCCS - condition (2c)

Corollary 3.1.7 *[15] The assumption of Theorem 3.1.3 on the resistances can be slightly reduced. In fact, only the positive definiteness of the conductance matrix corresponding to those resistances that do not form a loop with capacitances and/or voltage sources is required.*

This statement follows immediately from Theorem 3.1.3 if we consider the resistances as VCCS.

The rather extensive proofs of the Theorems 3.1.3 and 3.1.4 are given in [15] for the differential and the tractability index. For nonlinear time-independent circuits without controlled sources, a proof can be found in [58]. The proofs are based, among others, on the structural properties discussed below. In [15] there can also be found a detailed discussion of these results in comparison with other results from the literature devoted to circuit theory.

Here, we will focus only on the structural properties that are relevant with respect to the assumptions of the preceding chapters[9].

[9] Note that for stability, for instance, some other structural properties become relevant [46].

3.2 Some Structural Properties of the MNA Equations

In this section we introduce some projectors onto the spaces defined by the element-related incidence-matrices. These projectors will permit a proper description of the conditions we impose on the controlled sources. Moreover, they will precisely enable us to reveal the structural properties of the MNA equations.

Theorem 3.2.1 *[58] In practice, the following relations are satisfied for the (reduced) incidence matrix $A = (A_C A_L A_R A_V A_I)$.*

1. *The matrix $(A_C A_L A_R A_V)$ has full row rank, because cutsets of current sources are forbidden.*

2. *The matrix A_V has full column rank, because loops of voltage sources are forbidden.*

3. *The matrix $(A_C A_R A_V)$ has full row rank if and only if the circuit does not contain a cutset consisting of inductances $\underline{\textit{and/or}}$ current sources only.*

4. *Let Q_C be any projector onto $\ker A_C^T$. Then, the matrix $Q_C^T A_V$ has full column rank if and only if the circuit does not contain a loop consisting of capacitances $\underline{\textit{and}}$ voltage sources only.*

Note that loops containing only capacitances are excluded in point 4, whereas cutsets containing only inductances are included in point 3 of Theorem 3.2.1. For a complete proof of Theorem 3.2.1 we refer to [58].

We denote by Q_C, Q_{V-C}, Q_{R-CV}, $\bar{Q}_C$, and $\bar{Q}_{V-C}$ a projector onto $\ker A_C^T$, $\ker A_V^T Q_C$, $\ker A_R^T Q_C Q_{V-C}$, $ker A_C$, and $\ker Q_C^T A_V$, respectively. The complementary projectors will be denoted by $P := I - Q$, with the corresponding subindex. We observe that

$$\operatorname{im} P_C \subset \ker P_{V-C}, \quad \operatorname{im} P_{V-C} \subset \ker P_{R-CV} \quad \text{and} \quad \operatorname{im} P_C \subset \ker P_{R-CV},$$

and that thus $Q_C Q_{V-C}$ is a projector onto $\ker(A_C\, A_V)^T$, and $Q_C Q_{V-C} Q_{R-VC}$ is a projector onto $\ker(A_C\, A_R\, A_V)^T$. To shorten denotations, we use the abbreviation $Q_{CRV} := Q_C Q_{V-C} Q_{R-CV}$. Moreover, without loss of generality, these projectors are supposed to fulfil $Q_{CRV}Q_C = Q_{CRV}$. Remark that the projector P_{CRV} does not coincide with the projector P_{R-CV} in general.

Using the introduced projections we obtain the following corollary from Theorem 3.2.1.

Corollary 3.2.2 *[15] Theorem 3.2.1 implies that*

1. $Q_{CRV} = 0$ *if and only if the network does not contain L-I cutsets,*
2. $\bar{Q}_{V-C} = 0$ *if and only if the network does not contain C-V loops.*

In order to obtain a description of assumption (1) of Theorem 3.1.3 by means of projectors, we split the incidence matrix A_V into $(A_{Vt}A_{Vco})$ for independent and controlled sources, respectively.

Lemma 3.2.3 *[15] The condition that controlled voltage sources do not form a part of a C-V loop is equivalent to* $\bar{Q}_{V-C} = \begin{pmatrix} (\bar{Q}_{V-C})_t \\ 0 \end{pmatrix}$. *Here,* $(\bar{Q}_{V-C})_t$ *denotes the upper part of* $\bar{Q}_{V-C}$ *corresponding to* A_{Vt}.

For a proof see [15].

Hence, assumption (1) of Theorem 3.1.3 implies that

$$\begin{aligned} \bar{Q}^T_{V-C} v(A^T e, \frac{dq(A_C^T e, t)}{dt}, j_L, j_V, t) &= \bar{Q}^T_{V-C} v_t(t), && (3.13) \\ v(A^T e, \frac{dq(A_C^T e, t)}{dt}, j_L, j_V, t) &= v_*(A_C^T e, j_L, t) && (3.14) \end{aligned}$$

is given for a suitable function v_* and for a vector $v_t(t)$ that contains the functions of independent voltage sources and zeros instead of the functions of controlled voltage sources. In the following we will drop the index *.

In order to transcribe the assumptions made for controlled current sources, we split the incidence matrix A_I into $(A_{It}, A_{Ia}, A_{Ib}, A_{Ic})$ and the current vector i correspondingly for the independent current sources and the controlled current sources that fulfil (2a), (2b) and (2c), respectively. If a controlled current source fulfils more than one of the conditions (2a), (2b) and (2c), the corresponding column of A_I should be assigned to only one of the matrices A_{Ia}, A_{Ib}, and A_{Ic}.

Lemma 3.2.4 *[15] The condition that controlled current sources do not form a part of an L-I cutset is equivalent to the relation* $Q_{CRV}^T A_I = (Q_{CRV}^T A_{It}\ 0)$.

For a proof see [15].

Thus, assumption (2a) of Theorem 3.1.3 implies that

$$\begin{aligned} Q_{CRV}^T A_I i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) &= Q_{CRV}^T A_{It} i_t(t), \qquad (3.15) \\ i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) &= i_a((A_C A_V)^T e, j_L, t) \qquad (3.16) \end{aligned}$$

for a suitable function i_a.

Furthermore, assumption (2b) of Theorem 3.1.3 implies by definition that

$$\begin{aligned} Q_C^T A_{Ib} &= 0, \qquad (3.17) \\ i(A^T e, \frac{dq(A^T e,t)}{dt}, j_L, j_V, t) &= i_b((A_C A_R A_V)^T e, j_L, \bar{P}_{V-C} j_V, t) \qquad (3.18) \end{aligned}$$

for a suitable function i_b.

Finally, assumption (2c) of Theorem 3.1.3 implies that

$$\begin{aligned} Q_{V-C}^T Q_C^T A_{Ic} &= 0, \qquad (3.19) \\ i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) &= i_c((A_C A_R A_V)^T e, j_L, t) \qquad (3.20) \end{aligned}$$

holds for a suitable function i_c.

Regarding (3.15), (3.17), and (3.19), the assumptions imply that

$$Q_{CRV}^T A_I i(A^T e, \frac{dq(A_C^T e,t)}{dt}, j_L, j_V, t) = Q_{CRV}^T A_{It} i_t \qquad (3.21)$$

is always fulfilled.

In the forthcoming sections we will show that the conventional and the charge oriented MNA lead to DAEs that fulfil the structural assumptions from the Chapters 1 and 2 if the premises from Theorem 3.1.3 are given.

3.2.1 The Conventional MNA

For shorter expressions, we drop the arguments of the matrices in the following if they are clear from the context. In order to distinguish between constant and non-constant terms, we will use a dot as an argument for non-constant terms.

Writing the system (3.5)-(3.7) as a nonlinear DAE (1.13) with $A(y,x,t) := f'_y(y,x,t)$ and $B(y,x,t) := f'_x(y,x,t)$, we obtain that

$$A(\cdot) = \begin{pmatrix} A_C C(\cdot) A_C^T & 0 & 0 \\ 0 & L(\cdot) & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and, using (3.14), (3.16), (3.18) and (3.20),

$$B(\cdot) = \begin{pmatrix} A_C \bar{C}(\cdot) A_C^T + A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T & A_L + A_I \frac{di(\cdot)}{dj_L} & A_V + A_{Ib} \frac{di_b(\cdot)}{dj_V} \bar{P}_{V-C} \\ -A_L^T & \bar{L}(\cdot) & 0 \\ A_V^T - \frac{dv(\cdot)}{de} A_C^T & -\frac{dv(\cdot)}{dj_L} & 0 \end{pmatrix}$$

with

$$\bar{C}(u', u, t) = \frac{d}{du} C(u,t) u' + \frac{d}{du} q'_t(u,t)$$

and

$$\bar{L}(j'_L, j_L, t) = \frac{d}{dj_L} L(j_L, t) j'_L + \frac{d}{dj_L} \phi'_t(j_L, t).$$

With regard to the positive definiteness assumption we may choose the following constant projectors onto $N = \ker A(\cdot)$ and along $\operatorname{im} A(\cdot)$:

$$Q = \begin{pmatrix} Q_C & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix}, \quad W_0 = \begin{pmatrix} Q_C^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \end{pmatrix}.$$

Observe that **A1** is fulfilled, since $\ker A(\cdot)$ and $\operatorname{im} A(\cdot)$ are constant.

Moreover, using (3.17), we obtain

$$\begin{aligned} S(\cdot) = \{z:\ Q_C^T (A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T z_e & \\ + Q_C^T (A_L + A_I \frac{di(\cdot)}{dj_L}) z_L + Q_C^T A_V z_V &= 0, \\ (A_V^T - \frac{dv(\cdot)}{de} A_C^T) z_e - \frac{dv(\cdot)}{dj_L} z_L &= 0\}. \end{aligned}$$

Lemma 3.2.5 *[15] For this Q we may choose*

$$T = \begin{pmatrix} Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C} \end{pmatrix}, \quad U = \begin{pmatrix} P_{CRV} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & \bar{P}_{V-C} \end{pmatrix},$$

and

$$W_1 = \hat{W}_1 = \begin{pmatrix} Q_{CRV}^T & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}^T \end{pmatrix}.$$

Proof: To prove the expression given for T, we ascertain that

$$N \cap S(\cdot) = \text{im } Q_{CRV} \times \{0\} \times \text{im } \bar{Q}_{V-C}.$$

Firstly, we show that the relation "$\subseteq$" is true. Assuming $z \in N \cap S(\cdot)$ we know that $z_e = Q_C z_e$, $z_L = 0$ and $z \in S(\cdot)$. Hence, we have

$$Q_C^T A_R G(\cdot) A_R^T Q_C z_e + Q_C^T A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T Q_C z_e + Q_C^T A_V z_V = 0, \quad (3.22)$$

$$A_V^T Q_C z_e = 0. \quad (3.23)$$

Then, equation (3.23) provides additionally $z_e = Q_{V-C} z_e$. Thus, due to (3.19) and (3.15) -(3.16), multiplying (3.22) by Q_{V-C}^T we obtain

$$Q_{V-C}^T Q_C^T A_R G(\cdot) A_R^T Q_C Q_{V-C} z_e = 0.$$

Since $G(\cdot)$ was assumed to be positive definite, this implies $A_R^T Q_C Q_{V-C} z_e = 0$, i.e., $A_R^T z_e = 0$ and so $z_e \in \text{im } Q_{CRV}$. Now the relation (3.22) implies that $Q_C^T A_V z_V = 0$, i.e., $z_V = \bar{Q}_{V-C} z_V$.
Secondly, we show that the relation "$\supseteq$" is satisfied. Assume that $z_e = Q_{CRV} z_e$, $z_L = 0$, and $z_V = \bar{Q}_{V-C} z_V$. Then $z \in N$ holds trivially and

$$(A_V^T - \frac{dv(\cdot)}{de} A_C^T) z_e - \frac{dv(\cdot)}{dj_L} z_L = 0 \quad (3.24)$$

is fulfilled. Additionally, we obtain that

$$Q_C^T [(A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T) z_e + (A_L + A_I \frac{di(\cdot)}{dj_L}) z_L + A_V z_V] = 0.$$

To prove the expression given for $W_1 = \hat{W}_1$ we note that straightforward computation leads to

$$A_1(\cdot) = \begin{pmatrix} A_C C(\cdot) A_C^T + A_R G(\cdot) A_R^T Q_C + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T Q_C & 0 & A_V + A_{Ib} \frac{di_b(\cdot)}{dj_V} \bar{P}_{V-C} \\ -A_L^T Q_C & L(\cdot) & 0 \\ A_V^T Q_C & 0 & 0 \end{pmatrix}$$

and verify

$$\text{im } A_1(\cdot) = \ker Q_{CRV}^T \times I\!R^{n_L} \times \ker \bar{Q}_{V-C}^T.$$

Firstly, note that $\text{im } A_1(\cdot) \subseteq \ker Q_{CRV}^T \times I\!R^{n_L} \times \ker \bar{Q}_{V-C}^T$ holds trivially because of Lemma 3.2.4.
Secondly, to show $\text{im } A_1(\cdot) \supseteq \ker Q_{CRV}^T \times I\!R^{n_L} \times \ker \bar{Q}_{V-C}^T$, we assume that $z \in \ker Q_{CRV}^T \times I\!R^{n_L} \times \ker \bar{Q}_{V-C}^T$, i.e., $Q_{CRV}^T z_1 = 0$ and $\bar{Q}_{V-C}^T z_3 = 0$. Then, there is an α_0 such that

$$z_3 = A_V^T Q_C \alpha_0. \tag{3.25}$$

Due to $Q_{CRV}^T A_I = (Q_{CRV}^T A_{It}\ 0)$ (cf. Lemma 3.2.4) the relation

$$z_1 - A_R G(\cdot) A_R^T Q_C P_{V-C} \alpha_0 - A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T Q_C P_{V-C} \alpha_0 \in \ker Q_{CRV}^T$$

holds, i.e., there are α_1, α_2 and γ_1 such that

$$\begin{aligned} & z_1 - A_R G(\cdot) A_R^T Q_C P_{V-C} \alpha_0 - A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T Q_C P_{V-C} \alpha_0 \\ = \ & A_C C(\cdot) A_C^T \alpha_1 + A_R G(\cdot) A_R^T Q_C Q_{V-C} \alpha_2 + A_V \gamma_1. \end{aligned} \tag{3.26}$$

This is a simple conclusion of the fact that

$$\ker Q_{CRV}^T = \text{im } (A_C C(\cdot) A_C^T, A_R G(\cdot) A_R^T Q_C Q_{V-C}, A_V A_V^T),$$

since $C(\cdot)$ and $G(\cdot)$ are positive definite.

Let us now focus on the different cases that may occur for the controlled current sources. Considering (3.16) we see that

$$\begin{aligned} \frac{di_a(\cdot)}{de} (A_C A_R A_V)^T Q_C &= \frac{di_a((A_C A_V)^T e, j_L, t)}{de} (A_C A_V)^T Q_C \\ &= \frac{di_a(\cdot)}{de} (A_C A_V)^T Q_C P_{V-C}. \end{aligned} \tag{3.27}$$

Regarding (3.19) we find α_3 and γ_2 such that

$$A_{Ic} \frac{di_c(\cdot)}{de} (A_C A_R A_V)^T Q_C Q_{V-C} \alpha_2 = A_C C(\cdot) A_C^T \alpha_3 + A_V \gamma_2. \tag{3.28}$$

Using (3.17) we find α_4 and α_5 such that

$$A_{Ib} \frac{di_b(\cdot)}{de} (A_C A_R A_V)^T Q_C Q_{V-C} \alpha_2 = A_C C(\cdot) A_C^T \alpha_4, \tag{3.29}$$

$$A_{Ib} \frac{di_b(\cdot)}{dj_V} \bar{P}_{V-C} (\gamma_1 - \gamma_2) = A_C C(\cdot) A_C^T \alpha_5. \tag{3.30}$$

Choosing $\alpha := P_C(\alpha_1 - \alpha_3 - \alpha_4 - \alpha_5) + Q_C P_{V-C}\alpha_0 + Q_C Q_{V-C}\alpha_2$, $\beta := L^{-1}(\cdot)(z_2 + A_L^T Q_C \alpha)$, $\gamma := \gamma_1 - \gamma_2$ and regarding (3.25)-(3.29), we obtain that

$$z = A_1(\cdot)\begin{pmatrix}\alpha\\ \beta\\ \gamma\end{pmatrix} \in \text{im } A_1(\cdot).$$

q.e.d.

Observe that **A2** is fulfilled, since $N \cap S(\cdot)$ is constant.

Recall further that **A3** should be assumed. Nevertheless, taking into account that W_1 is constant, for the conventional MNA we can consider $K_{W_1} = W_1$ (cf. Remark 2.4.5), and thus it suffices to assume that the left hand sides of

$$\frac{d}{dt}\Big(\bar{Q}_{V-C}^T A_V^T P_C e - \bar{Q}_{V-C}^T v_t(t)\Big) \;=\; 0, \tag{3.31}$$

$$\frac{d}{dt}\Big(Q_{CRV}^T A_L j_L + Q_{CRV}^T A_{I_t} i_t(t)\Big) \;=\; 0, \tag{3.32}$$

exist, where (3.31)-(3.32) are precisely the equations that lead to the hidden constraints.

Observe that for $\hat{W}_1 = W_1$ this would also imply **A5**. Moreover, from (3.31)-(3.32) we deduce that only $P_C e, j_L \in C^1$ is required. Finally, note that **A6** is also fulfilled.

Corollary 3.2.6 *The equations of the conventional MNA*

- *fulfil assumptions* **A1**, **A2**, **A6**,
- *admit a slightly weaker version of* **A3 -A5**, *accordingly to Remark 2.4.5,*
- *require only $P_C e, j_L \in C^1$ (instead of $x \in C^1_{N\cap S}$) in Theorem 2.4.6, and*
- *have the structure $A(Px,t)x' + \tilde{b}(Ux,t) + \mathcal{B}Tx = 0$, i.e.,* **A7** *is given particularly.*

Proof: The first statements have been deduced above and the last one can easily be verified considering

$$\mathcal{B}T := \begin{pmatrix} 0 & 0 & A_V \bar{Q}_{V-C} \\ -A_L^T Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and $A_C^T e = A_C^T P_{CRV} e$, $A_R^T e = A_R^T P_{CRV} e$,$A_V^T e = A_V^T P_{CRV} e$.

q.e.d.

3.2.2 The Charge-oriented MNA

Analogously to the conventional MNA, straightforward computation leads to

$$A = \begin{pmatrix} A_C & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$B(\cdot) = \begin{pmatrix} 0 & 0 & A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T & A_L + A_I \frac{di(\cdot)}{dj_L} & A_V + A_{Ib} \frac{di_b(\cdot)}{dj_V} \bar{P}_{V-C} \\ 0 & 0 & -A_L^T & 0 & 0 \\ 0 & 0 & A_V^T - \frac{dv(\cdot)}{de} A_C^T & -\frac{dv(\cdot)}{dj_L} & 0 \\ I & 0 & -C(\cdot) A_C^T & 0 & 0 \\ 0 & I & 0 & -L(\cdot) & 0 \end{pmatrix}.$$

Hence, we may choose the following constant projectors onto ker A and along im A:

$$Q = \begin{pmatrix} \bar{Q}_C & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}, \quad W_0 = \begin{pmatrix} Q_C^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix},$$

and, using again (3.17), represent $S(\cdot)$ by

$$\begin{aligned} S(\cdot) = \{z : Q_C^T (A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T) z_e \\ + Q_C^T (A_L + A_I \frac{di(\cdot)}{dj_L}) z_L + Q_C^T A_V z_V = 0, \\ (A_V^T - \frac{dv(\cdot)}{de} A_C^T) z_e - \frac{dv(\cdot)}{dj_L} z_L = 0, \\ z_q - C(\cdot) A_C^T z_e = 0, \\ z_\phi - L(\cdot) z_L = 0\}. \end{aligned}$$

Lemma 3.2.7 *For this Q we may choose*[10]

$$T = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \bar{Q}_{V-C} \end{pmatrix}, \quad U = \begin{pmatrix} I & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 \\ 0 & 0 & P_{CRV} & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & \bar{P}_{V-C} \end{pmatrix},$$

and

$$W_1(\cdot) = \begin{pmatrix} Q_{CRV}^T & 0 & 0 & 0 & Q_{CRV}^T A_L L^{-1}(\cdot) \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}^T & \bar{Q}_{V-C}^T A_V^T H_1^{-1}(\cdot) A_C & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where $H_1(A_C^T e, t) := A_C C(A_C^T e, t) A_C^T + Q_C^T Q_C$ is a nonsingular matrix due to the positive definiteness of $C(A_C^T e, t)$. Observe further that we may set:

$$\hat{W}_1 = \begin{pmatrix} Q_{CRV}^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}^T & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Proof: To prove the expression given for T, we show that

$$N \cap S(\cdot) = \{0\} \times \{0\} \times \operatorname{im} Q_{CRV} \times \{0\} \times \operatorname{im} \bar{Q}_{V-C}.$$

Firstly, we verify the relation "$\subseteq$". Assuming $z \in N \cap S(\cdot)$ we know that $z_q = \bar{Q}_C z_q$, $z_\phi = 0$ and $z \in S(\cdot)$. Thus, we have

$$\begin{aligned} Q_C^T A_R G(\cdot) A_R^T z_e + Q_C^T A_I \frac{di(\cdot)}{de}(A_C A_R A_V)^T z_e & \\ + Q_C^T A_L z_L + Q_C^T A_I \frac{di(\cdot)}{dj_L} z_L + Q_C^T A_V z_V &= 0, \qquad (3.33) \\ (A_V^T - \frac{dv(\cdot)}{de} A_C^T) z_e - \frac{dv(\cdot)}{dj_L} z_L &= 0, \qquad (3.34) \\ z_q - C(\cdot) A_C^T z_e &= 0, \qquad (3.35) \\ z_\phi - L(\cdot) z_L &= 0. \qquad (3.36) \end{aligned}$$

The equations (3.35) and (3.36) imply $z_e = Q_C z_e$ (and thus $z_q = 0$), and $z_L = 0$, respectively. Consequently, equation (3.34) provides $z_e = \bar{Q}_{V-C} z_e$. Multiplying (3.33) by Q_{V-C}^T we obtain, again by (3.19) and (3.15) -(3.16),

$$Q_{V-C}^T Q_C^T A_R G(\cdot) A_R^T Q_C Q_{V-C} z_e = 0.$$

[10] These expressions were stated in [15] without proof.

Since $G(\cdot)$ was assumed to be positive definite, this implies $A_R^T Q_C Q_{V-C} z_e = 0$, i.e., $A_R^T z_e = 0$ and so $z_e \in \operatorname{im} Q_{CRV}$. Now the relation (3.33) implies that $Q_C^T A_V z_V = 0$, i.e., $z_V = \bar{Q}_{V-C} z_V$.
Secondly, we show that the relation "$\supseteq$" is satisfied. Assume that $z_q = 0$, $z_\phi = 0$, $z_e = Q_{CRV} z_e$, $z_L = 0$, and $z_V = \bar{Q}_{V-C} z_V$. Then $z \in N = \ker A(\cdot)$ holds trivially and $z \in S(\cdot)$ can easily be verified.

Let us now focus on the expression given for $W_1(\cdot)$. Straightforward computation yields

$$A_1(\cdot) = \begin{pmatrix} A_C & 0 & A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T & A_L + A_I \frac{di(\cdot)}{dj_L} & A_V + A_{Ib} \frac{di_b(\cdot)}{dj_V} \bar{P}_{V-C} \\ 0 & I & -A_L^T & 0 & 0 \\ 0 & 0 & A_V^T - \frac{dv(\cdot)}{de} A_C^T & -\frac{dv(\cdot)}{dj_L} & 0 \\ \bar{Q}_C & 0 & -C(\cdot) A_C^T & 0 & 0 \\ 0 & 0 & 0 & -L(\cdot) & 0 \end{pmatrix}.$$

Firstly, note that $\operatorname{im} A_1(\cdot) \subseteq \ker W_1(\cdot)$ holds, since due to Lemma 3.2.4 and (3.13) the multiplication $W_1(\cdot) A_1(\cdot)$ leads to

$$\begin{aligned} Q_{CRV}^T A_L - Q_{CRV}^T A_L L(\cdot) L^{-1}(\cdot) &= 0, \\ \bar{Q}_{V-C}^T A_V^T - \bar{Q}_{V-C}^T A_V^T H^{-1}(\cdot) A_C C(\cdot) A_C^T &= 0. \end{aligned}$$

Secondly, to show that $\operatorname{im} A_1(\cdot) \supseteq \ker W_1(\cdot)$, we consider $z \in \ker W_1(\cdot)$. Hence, it holds

$$Q_{CRV}^T z_1 + Q_{CRV}^T A_L L^{-1}(\cdot) z_5 = 0, \qquad (3.37)$$

$$\bar{Q}_{V-C}^T z_3 + \bar{Q}_{V-C}^T A_V^T H_1^{-1}(\cdot) A_C z_4 = 0. \qquad (3.38)$$

From (3.38) and (3.13) it follows that there exists an $\tilde{\alpha}_0$ such that

$$z_3 + A_V^T H_1^{-1}(\cdot) A_C z_4 - \frac{dv(\cdot)}{de} A_C^T H_1^{-1}(\cdot) A_C z_4 - \frac{dv(\cdot)}{dj_L} L^{-1}(\cdot) z_5 = A_V^T Q_C \tilde{\alpha}_0.$$

Furthermore, from (3.37) and Lemma 3.2.4 it follows that the expression

$$\begin{aligned} z_1 + A_L L^{-1}(\cdot) z_5 + [A_R G(\cdot) A_R^T + A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T] H_1^{-1}(\cdot) A_C z_4 \\ + A_I \frac{di(\cdot)}{dj_L} L^{-1}(\cdot) z_5 - A_R G(\cdot) A_R^T Q_C P_{V-C} \tilde{\alpha}_0 \\ - A_I \frac{di(\cdot)}{de} (A_C A_R A_V)^T Q_C P_{V-C} \tilde{\alpha}_0 \qquad (3.39) \end{aligned}$$

lies in $\ker Q_{CRV}^T$. Since

$$\ker Q_{CRV}^T = \mathrm{im}\ (A_C C(\cdot) A_C^T, A_R G(\cdot) A_R^T Q_C Q_{V-C}, A_V A_V^T),$$

there exist $\tilde{\alpha}_1$, $\tilde{\alpha}_2$, and $\tilde{\gamma}_1$, such that (3.39) is equal to

$$A_C C(\cdot) A_C^T \tilde{\alpha}_1 + A_R G(\cdot) A_R^T Q_C Q_{V-C} \tilde{\alpha}_2 + A_V \tilde{\gamma}_1.$$

Let us now focus, analogously as we did for the conventional MNA, on the three different cases that may occur for controlled current sources. Considering (3.16) we see that

$$\begin{aligned} \frac{di_a(\cdot)}{de}(A_C A_R A_V)^T Q_C &= \frac{di_a((A_C A_V)^T e, j_L, t)}{de}(A_C A_V)^T Q_C \\ &= \frac{di_a(\cdot)}{de}(A_C A_V)^T Q_C P_{V-C}. \end{aligned} \tag{3.40}$$

Regarding (3.19) we find $\tilde{\alpha}_3$ and $\tilde{\gamma}_2$ such that

$$A_{Ic}\frac{di_c(\cdot)}{de}(A_C A_R A_V)^T Q_C Q_{V-C} \tilde{\alpha}_2 = A_C C(\cdot) A_C^T \tilde{\alpha}_3 + A_V \tilde{\gamma}_2. \tag{3.41}$$

Using (3.17) we find $\tilde{\alpha}_4$ and $\tilde{\alpha}_5$ such that

$$A_{Ib}\frac{di_b(\cdot)}{de}(A_C A_R A_V)^T Q_C Q_{V-C} \tilde{\alpha}_2 = A_C C(\cdot) A_C^T \tilde{\alpha}_4, \tag{3.42}$$

$$A_{Ib}\frac{di_b(\cdot)}{dj_V}\bar{P}_{V-C}(\tilde{\gamma}_1 - \tilde{\gamma}_2) = A_C C(\cdot) A_C^T \tilde{\alpha}_5. \tag{3.43}$$

By the above considerations, for

$$\alpha = \begin{pmatrix} \bar{P}_C C(\cdot) A_C^T(\tilde{\alpha}_1 - \tilde{\alpha}_3 - \tilde{\alpha}_4 - \tilde{\alpha}_5) + \bar{Q}_C z_4 \\ z_2 + A_L^T(Q_C P_{V-C}\tilde{\alpha}_0 + Q_C Q_{V-C}\tilde{\alpha}_2 - H_1^{-1}(\cdot) A_C z_4) \\ Q_C P_{V-C}\tilde{\alpha}_0 + Q_C Q_{V-C}\tilde{\alpha}_2 - H_1^{-1}(\cdot) A_C z_4 \\ -L^{-1}(\cdot) z_5 \\ \tilde{\gamma}_1 - \tilde{\gamma}_2 \end{pmatrix}$$

we thus obtain $z = A_1(\cdot)\alpha$, i.e., $z \in \mathrm{im}\ A_1(\cdot)$.

q.e.d.

Observe that **A2** is fulfilled, since $N \cap S(\cdot)$ is constant.

Note again that **A3** should be assumed. Nevertheless, considering the specific W_1, for the charge-oriented MNA we can consider

$$K_{W_1} := \begin{pmatrix} Q_{CRV}^T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \bar{Q}_{V-C}^T & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}, \tag{3.44}$$

(cf. Remark 2.4.5). Thus it suffices to suppose that the left hand sides of the equations

$$\begin{aligned} \frac{d}{dt}\Big(\bar{Q}_{V-C}^T A_V^T P_C e - \bar{Q}_{V-C}^T v_t(t)\Big) &= 0, &(3.45)\\ \frac{d}{dt}\Big(Q_{CRV}^T A_L j_L + Q_{CRV}^T A_{I_t} i_t(t)\Big) &= 0, &(3.46)\\ \frac{d}{dt}(q - q_C(A_C^T e, t)) &= 0, &(3.47)\\ \frac{d}{dt}(\phi - \phi_L(j_L, t)) &= 0, &(3.48) \end{aligned}$$

exist, where (3.45)-(3.48) are the equations involved in the expressions that lead to the hidden constraints.

Observe also that **A5** corresponds to the existence of the left hand of (3.45) and (3.46) and, therefore, is given, too. Moreover, from (3.45)- (3.48) we deduce that only $q, \phi, P_C e, j_L \in C^1$ is required. Finally, note that **A6** is also fulfilled.

Corollary 3.2.8 *The equations of the charge-oriented MNA*

- *fulfil the assumptions* **A1**, **A2**, **A6**,
- *admit a slightly weaker version of* **A3**-**A5**, *according to Remark 2.4.5,*
- *require only $q, \phi, P_C e, j_L \in C^1$ (instead of $x \in C^1_{N\cap S}$) in Theorem 2.4.6, and*
- *have the structure $Ax' + \tilde{b}(Ux, t) + \mathcal{B}Tx = 0$, which corresponds to* **A9**.

Proof: The first assertions have been discussed above and the last one can easily be ascertained considering

$$\mathcal{B}T := \begin{pmatrix} 0 & 0 & 0 & 0 & A_V \bar{Q}_{V-C} \\ 0 & 0 & -A_L^T Q_{CRV} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and $A_C^T e = A_C^T P_{CRV} e$, $A_R^T e = A_R^T P_{CRV} e$, $A_V^T e = A_V^T P_{CRV} e$.

q.e.d.

Remark 3.2.9 • *Observe that the smoothness assumptions required for the conventional MNA and the charge-oriented MNA correspond to each other. In particular, for the solutions, it results that*

- *for the conventional MNA it suffices to suppose that $P_C e, j_L \in C^1$,*
- *for the charge-oriented MNA it suffices to suppose that*

$$q, \phi, P_C e, j_L \in C^1.$$

- *With respect to the smoothness that has to be given for the input signals, the above results imply that only the characteristic equations of the current and voltage sources that form part of L-I cutsets and C-V loops, respectively, have to be smooth.*

3.3 Consistent Initial Values for the MNA Equations

In this section we apply the approach from Section 2.7 for computing a consistent initialization to the MNA-equations, where the required smoothness is assumed to be given. In order to avoid the introduction of new notations, we will denote the values (x_0, Py_0) for the systems arising from the MNA by $(e_0, j_{L_0}, j_{V_0}, P_C e_0', j_{L_0}')$ and $(q_0, \phi_0, e_0, j_{L_0}, j_{V_0}, \bar{P}_C q_0', \phi_0')$, respectively. The same will be done for (x^0, Py^0) and $(\hat{x}_0, P\hat{y}_0)$.

Corollary 3.3.1 *For the MNA equations Theorem 2.7.2 implies:*

- *For the conventional MNA the solution* $(\bar{Q}_{V-C}\hat{j}_{V_0},\ P_C\hat{e}'_0)$ *of the system*

  $$\begin{array}{rcl} A_C C(A_C^T e_0, t_0) A_C^T \hat{e}'_0 + A_V \bar{Q}_{V-C}\hat{j}_{V_0} &=& 0, \\ \bar{Q}^T_{V-C} A_V^T \hat{e}'_0 + \bar{Q}^T_{V-C} A_V^T e'^0 - \bar{Q}^T_{V-C} v'_t(t_0) &=& 0 \end{array}$$

  and the solution ($Q_{CRV}\hat{e}_0$ *,* $\hat{j}'_{L_0}$*) of the system*

  $$\begin{array}{rcl} L(j_L, t_0)\hat{j}'_{L_0} - A_L^T Q_{CRV}\hat{e}_0 &=& 0, \\ Q^T_{CRV} A_L \hat{j}'_{L_0} + Q^T_{CRV} A_L {j_L}'^0 + Q^T_{CRV} A_I i'_t(t_0) &=& 0 \end{array}$$

  provide the values we require to compute consistent initial values, where $P_C e'^0$*,* ${j_L}'^0$ *,*$P_C e^0$ *and* ${j_L}^0$ *are fixed values.*

- *For the charge-oriented MNA the solution* $(\bar{Q}_{V-C}\hat{j}_{V_0},\ \bar{P}_C\hat{q}'_0)$ *of the system*

  $$\begin{array}{rcl} A_C \hat{q}'_0 + A_V \bar{Q}_{V-C}\hat{j}_{V_0} &=& 0, \\ \bar{Q}^T_{V-C} A_V^T C^{-1}(A_C^T e_0, t_0)\hat{q}'_0 + \bar{Q}^T_{V-C} A_V^T C^{-1}(A_C^T e_0, t_0) q'^0 && \\ - \bar{Q}^T_{V-C} v'_t(t_0) &=& 0 \end{array}$$

  and the solution $(Q_{CRV}\hat{e}_0$ *,* $\hat{\phi}'_0)$ *of the system*

  $$\begin{array}{rcl} \hat{\phi}'_0 - A_L^T Q_{CRV}\hat{e}_0 &=& 0, \\ Q^T_{CRV} A_L L^{-1}(j_L, t_0)\hat{\phi}'_0 + Q^T_{CRV} A_L L^{-1}(j_L, t_0)\phi'^0 + Q^T_{CRV} A_I i'_t(t_0) &=& 0 \end{array}$$

  provide the values we require to compute consistent initial values, where $\bar{P}_C q'^0$*,* ϕ'^0*,*$P_C e^0$ *and* ${j_L}^0$ *are fixed values.*

Proof: Observe that, making use of the projectors from Lemma 3.2.5, for the conventional MNA the system from Theorem 2.7.2 reads:

$$\begin{array}{rcl} A_C C(A_C^T e_0, t_0) A_C^T \hat{e}'_0 + A_V \bar{Q}_{V-C}\hat{j}_{V_0} &=& 0, \\ L(j_L, t_0)\hat{j}'_{L_0} - A_L^T Q_{CRV}\hat{e}_0 &=& 0, \\ P_{CRV}\hat{e}_0 &=& 0, \\ \hat{j}_{L_0} &=& 0, \\ \bar{P}_{V-C}\hat{j}_{V_0} &=& 0, \\ Q^T_{CRV} A_L \hat{j}'_{L_0} + Q^T_{CRV} A_L {j_L}'^0 + Q^T_{CRV} A_I i'_t(t_0) &=& 0, \\ \bar{Q}^T_{V-C} A_V^T \hat{e}'_0 + \bar{Q}^T_{V-C} A_V^T e'^0 - \bar{Q}^T_{V-C} v'_t(t_0) &=& 0. \end{array}$$

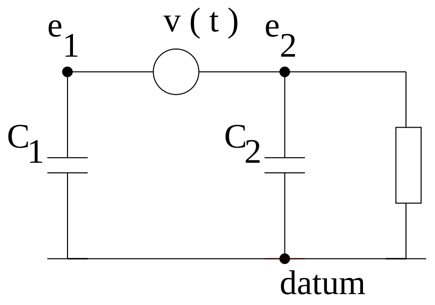


Figure 3.5: Circuit containing a C-V loop

This leads to the stated system. For the charge oriented MNA, the projectors presented in Lemma 3.2.7 yield the corresponding results.

q.e.d.

Note that the values obtained by Corollary 3.3.1 coincide with those obtained in [11] directly.

Example 3.3.2 *Let us consider the academic example from Figure 3.5 to illustrate tha approach described in Corollary 3.3.1. The equations resulting by the conventional MNA read:*

$$\begin{aligned} C_1 e_1' + j_V &= 0, \\ -j_V + C_2 e_2' + \frac{1}{R} e_2 &= 0, \\ e_1 - e_2 &= v(t_0). \end{aligned}$$

The values we obtain for the DC operation point are

$$e_1 = v(t_0), \quad e_2 = 0 \quad , j_V = 0,$$

and the corresponding consistent initialization is given by

$$\begin{aligned} e_1 &= v(t_0), \quad e_2 = 0, \quad j_V = -\frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} v'(t_0), \\ e_1' &= \frac{1}{C_1} \frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} v'(t_0), \quad e_2' = -\frac{1}{C_2} \frac{1}{\frac{1}{C_1} + \frac{1}{C_2}} v'(t_0). \end{aligned}$$

Considering the graph of the network (cf. [12]), Corollary 3.3.1 implies that the correction affects exactly those elements that form a part of the L-I cutsets and C-V loops. Indeed, the above theorem can be interpreted as follows.

Corollary 3.3.3 *[11] For networks that contain only the specified controlled sources we obtain consistent initial values starting from possibly inconsistent ones that fulfil the equations of the DAE in the following way:*

1. *Add convenient values to the values of the currents flowing through the branches of voltage sources that form a part of the C-V loops.*
2. *Add convenient values to the values of the node potentials to obtain additional branch voltages across the branches that form a part of the L-I cutsets.*

Moreover, the values obtained in Corollary 3.3.1 imply that, at time t_0, the sum of the additional power delivered by the C-V loops and L-I cutsets is equal to the sum of the additional power absorbed by the branches of the C-V loops and L-I cutsets.

A proof can be found in [11].

3.4 Graph-theoretical Determination of the Hidden Constraints

According to (3.31),(3.32) for the conventional MNA the equations that lead to the hidden constraints are

$$\begin{aligned} \frac{d}{dt}\Big(\bar{Q}_{V-C}^T A_V^T P_C e - \bar{Q}_{V-C}^T v_t(t)\Big) &= 0, && (3.49)\\ \frac{d}{dt}\Big(Q_{CRV}^T A_L j_L + Q_{CRV}^T A_{I_t} i_t(t)\Big) &= 0. && (3.50)\end{aligned}$$

On the other hand, for the charge-oriented MNA, the hidden constraints result (cf. (3.45)-(3.48)) from (3.49)-(3.50) and

$$\begin{aligned} \frac{d}{dt}(q - q_C(A_C^T e, t)) &= 0,\\ \frac{d}{dt}(\phi - \phi_L(j_L, t)) &= 0. \end{aligned}$$

Thus, we are interested in expressions for (3.49)-(3.50) without requiring the computation of the corresponding projectors. In [12] it was shown that these equations can be obtained directly from the network by making use of the following two procedures that analyse its graph[11]. In fact, these procedures exactly determine the linearly independent equations that describe the hidden constraints arising from C-V loops and L-I cutsets, respectively.

Let us first focus on the constraints (3.49) arising from the C-V loops. Recall that

$$A_V^T e - v(\cdot) = 0$$

are the characteristic equations of the voltage sources. Since $\bar{Q}_{V-C}$ describes the C-V loops, it results that

$$\bar{Q}_{V-C}^T A_V^T e - \bar{Q}_{V-C}^T v_t(t) = 0$$

corresponds to the sums of the characteristic equations of the voltage sources that form a part of the C-V loops (cf. [12]). More exactly, these equations can be determined by means of the following procedure.

PROCEDURE 1

1. Search a C-V loop in the given network graph. If no C-V loop is found, then end.

2. Write the equation resulting from the sum of the derivatives of the characteristic equations of the voltage sources contained in the C-V loop, taking into account the orientation of the loop and the reference direction of the considered branches.
   For instance, if the voltage sources $v_1, ..., v_k$ form a part of the C-V loop and we define

$$\alpha_i := \begin{cases} +1 & \text{if the orientation of the loop coincides with that of } v_i \\ -1 & \text{else,} \end{cases}$$

[11] A similar graph-theoretical analysis of the network can be found in [8] for linear passive RLC networks in order to determine state-variables. The approach is based on the construction of a normal tree, i.e., a tree that contains all independent voltage sources, no independent current sources, a maximal number of capacitive branches and a minimal number of inductive branches.

then the equation we write in this step is

$$\sum_{i=1}^{k} \alpha_i ((A_V^T e)_i' - v_i') = 0.$$

3. Form a new network graph by deleting the branch of one voltage source that forms a part of the loop, leaving the nodes unchanged.

4. Return to 1, considering the new network graph.

Let us now focus on the constraints (3.50) arising from L-I cutsets. To get an idea of where they arise from, recall that

$$A_C \frac{dq}{dt} + A_R r(A_R^T e, t) + A_L j_L + A_V j_V + A_I i(\cdot) = 0$$

are the nodal equations. Since Q_{CRV} describes the L-I cutset, it can be shown that

$$Q_{CRV}^T A_L j_L + Q_{CRV}^T A_I i_t(t) = 0$$

corresponds to the equations that arise from KCL for the L-I cutsets. Consequently, the equations corresponding to (3.50) can be determined by means of the following procedure that starts by considering the original graph (cf. [12]).

PROCEDURE 2

1. Search an L-I cutset. If one is found, then select an arbitrary inductance of this cutset. Realize that we can always find such an inductance because cutsets of current sources only are forbidden. If no L-I cutset is found, then end.

2. Write a new equation resulting by derivation of the cutset equation arising from 1.
   For instance, if the current sources $i_1, ..., i_k$ and the inductances $j_{L1}, ..., j_{L\tilde{k}}$ form a part of the L-I cutset and we define

$$\alpha_j \;:=\; \begin{cases} +1 & \text{if the orientation of the cutset coincides with that of } i_j \\ -1 & \text{else,} \end{cases}$$

$$\tilde{\alpha}_j \;:=\; \begin{cases} +1 & \text{if the orientation of the cutset coincides with that of } j_{Lj} \\ -1 & \text{else,} \end{cases}$$

then the equation obtained in this step reads

$$\sum_{j=1}^{k} \alpha_j {i'}_j + \sum_{i=1}^{\tilde{k}} \tilde{\alpha}_j j'_{Li} = 0.$$

3. Delete the chosen inductance from the network contracting its incident nodes.

4. Return to 1, considering the new network graph.

For the proofs we refer to [12]

3.5 Realization Specifics

The results from Theorem 3.1.3 and Corollary 3.3.1 have been implemented by S. Sturtzel within the project "Structural analysis of DAEs in circuit simulation"[12] in the simulation package TITAN[13]. There, a graph theoretical algorithm has been developed that provides important information for several aspects:

1. Structural analysis (the assumptions on the controlled sources from Theorem 3.1.3 have to be given).

2. Index determination (cf. Theorem 3.1.3).

3. Identification of critical variables: the variables that are involved in $N \cap S(\cdot)$ are the currents through voltage sources that form a part of C-V loops and the branch voltages of branches that form a part of L-I cutsets.

4. Description of the linear system that provides the values required for the computation of a consistent initialization (Corollary 3.3.1 and Procedures 1 and 2).

[12] The exact German title is "Untersuchung der speziellen differential-algebraischen Struktur der Netzwerkgleichungen für die Schaltkreissimulation zur Entwicklung zuverlässiger und effizienter Simulationsverfahren". This project was sponsored by the Bundesministerium für Bildung, Wissenschaft, Forschung und Technologie (BMBF) (German Federal Ministry of Education, Science, Research and Technology) within the program "Mathematical methods for solving problems in industry and economy".

[13] Infineon Technologies (formerly SIEMENS AG).

With regard to the computation of a consistent initialization, it is important to note that:

1. The derivatives of the functions $i_t(t)$, $v_t(t)$, which we require for the expressions of the hidden constraints, were available.

2. The algorithm is implemented as an add-on in the simulation package, since values (x^0, Py^0) were given (e.g., the DC-operating point).

3. Since the structure of the equations from Corollary 3.3.1 is similar to the structure of the original system, a part of them can be solved as a structural subset of the original system, taking advantage of the sparse handling.

In practice, the computation of a consistent initialization is carried out with regard to the following aspects:

1. To start the integration, i.e., in general, the DC operating point is corrected in order to obtain a consistent starting point.

2. To obtain consistent values after discontinuities of the derivatives of the input functions, i.e., at the breakpoints.

3. For a clean handling of user given initial conditions by calculating an appropriate x^0 (cf. the approach presented in [11]).

A more detailed discussion and some examples can be found in [13].

3.6 Examples

Let us consider the academic example from Figure 3.6 to illustrate the effects described in the Sections 2.7.2 and 2.7.3. The equations resulting by the charge-oriented MNA read:

$$\begin{aligned}
q_1' + \frac{1}{R}e_1 + j_V + i(j_V, t) &= 0, \\
-j_V + q_2' &= 0, \\
e_1 - e_2 &= v(t), \\
q_1 &= q_1(e_1), \\
q_2 &= C_2 e_2.
\end{aligned}$$

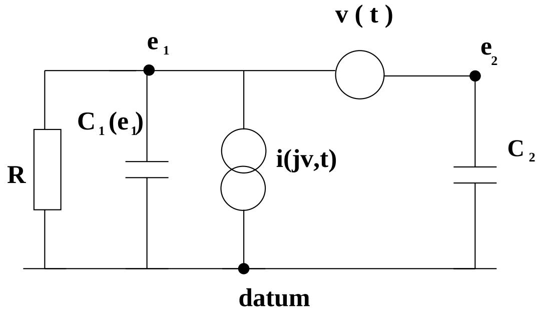


Figure 3.6: Example with a controlled source not fulfilling the assumptions from Section 3.1.3

Suppose $R = 1$, $C_2 = 1$, $v(t) = 2sin(t)$, $q_1(e_1) = e_1^2$. Depending on the definition of the controlled source $i(j_V, t)$ we may obtain different structural properties:

- Structure $Ax' + b(Ux, t) + \mathcal{B}Tx = 0$ for[14]

$$i(j_V, t) = i(t) = -sin(t) - 2 - (2sin(t) + 3)(cos(t)).$$

- Structure $Ax' + b(Ux, t) + \mathcal{B}(t)Tx = 0$ for the current controlled current-source[15]

$$i(j_V, t) = (2sin(t) + 3)j_V - sin(t) - 2.$$

Note that the two possibilities are chosen in such a way that in both cases we obtain the same solutions for the consistent value $(4, 2, 2, 2, -1)$. The different numerical effects that these structures may yield are illustrated in the Figures 3.7 and 3.8 for the implicit Euler method and the trapezoidal rule for the constant step-size $h = 0.1$.

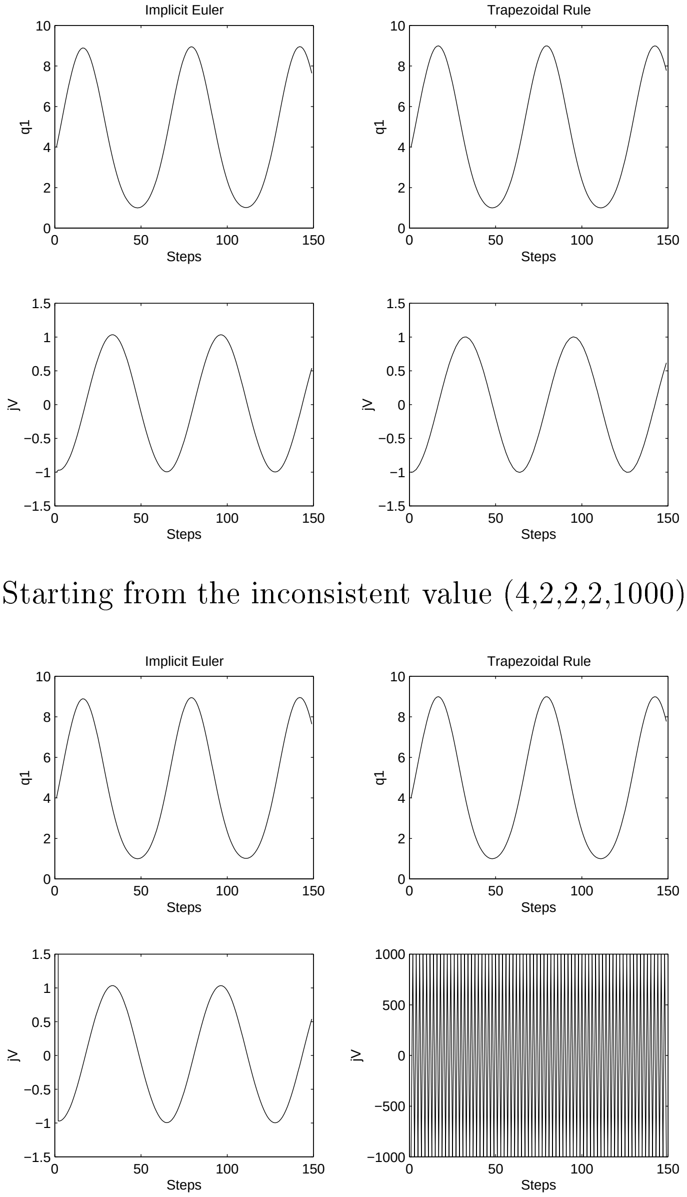
Notice on the one hand that, for the structure $Ax' + b(Ux, t) + \mathcal{B}Tx = 0$ (Figure 3.7), the implicit Euler method leads to the same results after the first

[14] Note that we assume that the current source is independent. Hence, the assumptions from Section 3.1.3 are met.

[15] Note that this kind of controlled sources was forbidden in the class described in Section 3.1.3, because the controlling current corresponds to a voltage source that forms a part of a C-V loop.

step, starting up from the consistent and from inconsistent initial value, as expected from Remark 2.7.3. Moreover, the trapezoidal rule precisely shows the effect described in Remark 2.7.6.

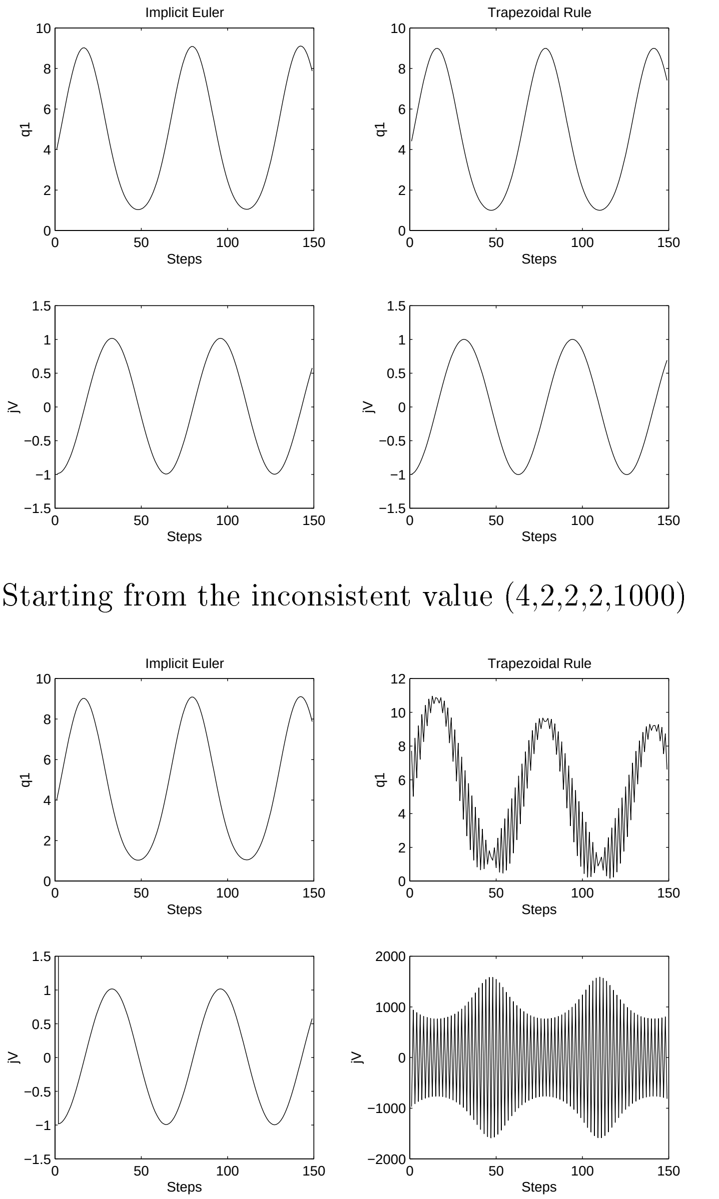
On the other hand, if the structure $Ax' + b(Ux, t) + \mathcal{B}(t)Tx = 0$ (Figure 3.8) is given, then, for the trapezoidal rule, the error made in the $N \cap S(\cdot)$ component may affect the other components, too. In this example, all components have the oscillating behaviour that is introduced by the trapezoidal rule. However, after the first step, the implicit Euler method leads to the same results, starting up from the consistent and from the inconsistent initial value, as expected from Remark 2.7.7.



Starting from the consistent value (4,2,2,2,-1), h=0.1

Starting from the inconsistent value (4,2,2,2,1000)

Figure 3.7: Example of the structure $Ax' + b(Ux,t) + \mathcal{B}Tx = 0$



Starting from the consistent value (4,2,2,2,-1), h=0.1

Starting from the inconsistent value (4,2,2,2,1000)

Figure 3.8: Example of the structure $Ax' + b(Ux,t) + \mathcal{B}(t)Tx = 0$

Summary

It is well-known from a large body of literature that, for solving DAEs numerically, consistent initial values have to be calculated. This thesis deals with an approach for handling this problem for index-2 DAEs by considering projectors onto the spaces related to the DAE. There are two major aspects in the work presented here.

On the one hand, new structural properties are deduced from the assumptions **A1** and **A2** (cf. Lemma 2.3.4). Subsequently, a method is proposed to choose suitable equations of an index-2 DAE, whose differentiation leads to an index reduction (Theorem 2.4.6). This index reduction yields new theoretical results for the existence and uniqueness of solutions of index-2 DAEs which apply to a wider class of applications than previous results. Based on this method, a step-by-step approach (described in (2.58)-(2.60)) to compute consistent initial values is developed. In this way, we gain new insights about how to deal with structural properties of index-2 DAEs. In particular, it turns out that, in comparison to index-1 DAEs, the additional step that has to be undertaken in practice often consists in solving a linear system (Theorem 2.7.2). The numerical consequences of this fact are exemplified for two methods commonly used in circuit simulation, the implicit Euler method and the trapezoidal rule.

On the other hand, the application of the obtained results to the equations arising in circuit simulation by means of the modified nodal analysis (MNA) is worked out (Corollary 3.3.1), where a short overview of the specifics of their realization is given.

Appendix

Some Basic Definitions and Results

Definition 4.1.1

- *A matrix $Q \in I\!R^{n\times n}$ is a projector onto R_1 if and only if $Q^2 = Q$ and* $\operatorname{im} Q = R_1$.
- *A matrix $W \in I\!R^{n\times n}$ is a projector along R_2 if and only if $W^2 = W$ and* $\ker W = R_2$.
- *For $I\!R^n = R_1 \oplus R_2$ a matrix $Q \in I\!R^{n\times n}$ is the uniquely defined projector onto R_1 along R_2 if and only if $Q^2 = Q$,* $\operatorname{im} Q = R_1$, *and* $\ker Q = R_2$.

Lemma 4.1.2

- *Assume Q and $\bar{Q}$ to be projectors onto a subspace R_1. Then, $Q = \bar{Q}Q$ holds.*
- *Assume W and $\bar{W}$ to be projectors along a subspace R_2. Then, $W = W\bar{W}$ holds.*
- *If Q is a projector onto a subspace R_1, then $P := I - Q$ is a projector along R_1.*

A fundamental relation between the spaces, the matrix chain and the choice of the projectors related to the definition of the tractability index is given by the following lemma.

Lemma 4.1.3 *Let $A_*, B_* \in L(I\!R^n)$ be given, let Q_* be a projector onto* $\ker A_*$ *and W_* be a projector along* $\operatorname{im} A_*$. *Denote*

$$S_* := \{z \in I\!R^n : W_* B_* z = 0\}$$

Then the following conditions are equivalent:

1. *The matrix* $G_* := A_* + B_* Q_*$ *is nonsingular.*
2. $S_* \oplus \ker A_* = \mathbb{R}^n$.
3. $S_* \cap \ker A_* = 0$.

Moreover, if G_* *is nonsingular, then the relation*

$$Q_{*S} = Q_* G_*^{-1} B_*$$

holds for the canonical projector onto $\ker A_*$ *along* S_*.

For a proof see [25].

Definition 4.1.4 *A space* $R(\cdot) : \mathcal{I} \to \mathbb{R}^n$ *is said to depend smoothly on* t *if it has a* C^1*-basis, i.e., there are linear, independent* C^1*-functions*

$$n_1(\cdot), n_2(\cdot), \ldots, n_k(\cdot)$$

such that

$$R(t) = L(\{n_1(t), n_2(t), \ldots, n_k(t)\})$$

for all $t \in \mathcal{I}$.

Remark 4.1.5

- $R(\cdot)$ *depends smoothly on* t *if and only if there is a continuously differentiable projector* $Q(\cdot)$ *onto* $R(\cdot)$.
- $R(\cdot)$ *depends smoothly on* t *if and only if there is a continuously differentiable projector* $W(\cdot)$ *along* $R(\cdot)$.

Description of the NAND-Gate

For the NAND-gate model we consider in Section 2.9, the equations are given by:

$$
\begin{aligned}
\frac{e_1 - e_2}{R_s} - \frac{e_7 - e_1}{R_d} + q' + q'_{1gd} + q'_{1gs} &= 0, \\
-q'_{1gs} + q'_{1sb} + \frac{e_2 - e_1}{R_s} + \frac{e_2 - e_3}{R_{sd}} + i^D_{bd}(e_{12} - e_2) & \\
+ i^D_{ds}(e_3 - e_2, e_1 - e_2, e_{12} - e_2) &= 0, \\
-q'_{1gd} + q'_{1db} + \frac{e_3 - e_4}{R_d} - \frac{e_2 - e_3}{R_{sd}} + i^D_{bd}(e_{12} - e_3) & \\
- i^D_{ds}(e_3 - e_2, e_1 - e_2, e_{12} - e_2) &= 0, \\
\frac{e_4 - e_3}{R_d} + j_{DD} &= 0, \\
q'_{2gd} + q'_{2gs} + j_1 &= 0, \\
-q'_{2gs} + q'_{2sb} + \frac{e_6 - e_{11}}{R_s} + \frac{e_6 - e_7}{R_{sd}} + i^E_{bs}(e_{12} - e_6) & \\
+ i^E_{ds}(e_7 - e_6, e_5 - e_6, e_{12} - e_6) &= 0, \\
-q'_{2gd} + q'_{2db} + \frac{e_7 - e_1}{R_d} - \frac{e_6 - e_7}{R_{sd}} + i^E_{bd}(e_{12} - e_7) & \\
- i^E_{ds}(e_7 - e_6, e_5 - e_6, e_{12} - e_6) &= 0, \\
q'_{3gd} + q'_{3gs} + j_2 &= 0, \\
-q'_{3gs} + q'_{3sb} + \frac{e_9}{R_s} + \frac{e_9 - e_{10}}{R_{sd}} + i^E_{bs}(e_{12} - e_9) & \\
+ i^E_{ds}(e_{10} - e_9, e_8 - e_9, e_{12} - e_9) &= 0, \\
-q'_{3gd} + q'_{3db} + \frac{e_{10} - e_{11}}{R_d} - \frac{e_9 - e_{10}}{R_{sd}} + i^E_{bd}(e_{12} - e_{10}) & \\
- i^E_{ds}(e_{10} - e_9, e_8 - e_9, e_{12} - e_9) &= 0, \\
\frac{e_{11} - e_6}{R_s} - \frac{e_{10} - e_{11}}{R_d} &= 0, \\
-q'_{1db} - q'_{1sb} - i^D_{bd}(e_{12} - e_2) - i^D_{bd}(e_{12} - e_3) & \\
-q'_{2db} - q'_{2sb} - i^E_{bs}(e_{12} - e_6) - i^E_{bd}(e_{12} - e_7) & \\
-q'_{3db} - q'_{3sb} - i^E_{bs}(e_{12} - e_9) - i^E_{bd}(e_{12} - e_{10}) + j_{BB} &= 0,
\end{aligned}
$$

$$\begin{aligned}
e_5 - V_1(t) &= 0,\\
e_8 - V_2(t) &= 0,\\
e_{12} - V_{BB} &= 0,\\
e_4 - V_{DD} &= 0,\\
q - Ce_1 &= 0,\\
q_{1gd} - q_{gd}(e_1 - e_3) &= 0,\\
q_{1gs} - q_{gs}(e_1 - e_2) &= 0,\\
q_{1db} - q_{db}(e_3 - e_{12}) &= 0,\\
q_{1sb} - q_{sb}(e_2 - e_{12}) &= 0,\\
q_{2gd} - q_{gd}(e_5 - e_7) &= 0,\\
q_{2gs} - q_{gs}(e_5 - e_6) &= 0,\\
q_{2db} - q_{db}(e_7 - e_{12}) &= 0,\\
q_{2sb} - q_{sb}(e_6 - e_{12}) &= 0,\\
q_{3gd} - q_{gd}(e_8 - e_{10}) &= 0,\\
q_{3gs} - q_{gs}(e_8 - e_9) &= 0,\\
q_{3db} - q_{db}(e_{10} - e_{12}) &= 0,\\
q_{3sb} - q_{sb}(e_9 - e_{12}) &= 0.
\end{aligned}$$

The elements are modelled as follows [57].

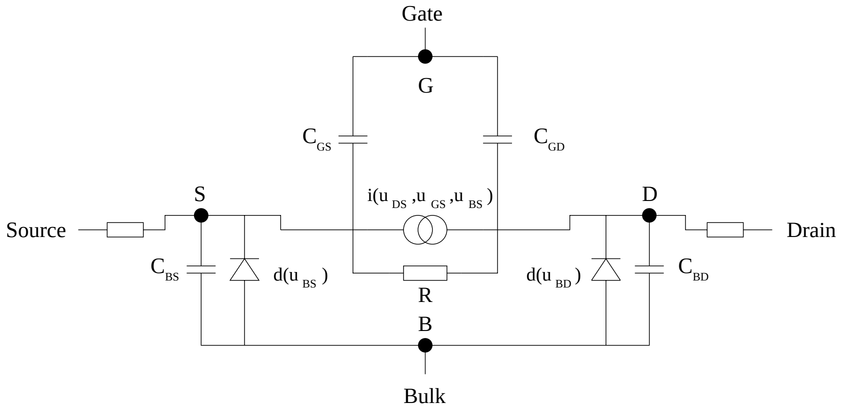


Figure 4.9: MOSFET-model

	ME	MD
i_s	10^{-14} A	10^{-14} A
U_T	25.85 V	25.85 V
U_{T0}	0.8 V	-2.43 V
β	$1.748 \cdot 10^{-3}\ A/V^2$	$5.35 \cdot 10^{-4}\ A/V^2$
γ	$0.0\ \sqrt{V}$	$0.2\ \sqrt{V}$
δ	$0.02\ V^{-1}$	$0.02\ V^{-1}$
Φ	1.01 V	1.28 V

Table 4.8: Technical parameters the MOSFETs ME and MD

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

$$i_{bs}(u) = i_{bd}(u) = \begin{cases} -i_s \cdot (exp(\frac{u}{U_T} - 1)) & \text{for} \quad u \leq 0, \\ 0 & \text{for} \quad u > 0. \end{cases}$$

The current through the controlled current source between drain and source is modelled by the function

$$i_{ds}(u_{ds}, u_{gs}, u_{bs}) =$$

$$\begin{cases} 0 & \text{for} \quad u_{gs} - U_{TE} \leq 0, \\ -\beta \cdot (1 + \delta \cdot u_{ds}) \cdot (u_{gs} - U_{TE}) & \text{for} \quad 0 < u_{gs} - U_{TE} \leq u_{ds}, \\ -\beta \cdot u_{ds} \cdot (1 + \delta \cdot u_{ds}) \cdot [2 \cdot (u_{gs} - U_{TE}) - u_{ds}] & \text{for} \quad 0 < u_{ds} < u_{gs} - U_{TE} \end{cases}$$

with $U_{TE} = U_{T_0} + \gamma \cdot (\sqrt{\Phi - u_{bs}} - \sqrt{\Phi})$.

The technical parameters for the MOSFETs ME and MD are given in Table 4.8.

The values for the resistances are chosen for all MOSFETs as

$$R_s = R_d = 4\Omega, \quad R_{sd} = 10^5\Omega.$$

The load capacitance is constant with $C = 0.5 \cdot 10^{-13}$ F. The capacitances between gate and source as well as those between gate and drain are modelled as linear capacitors, i.e.,

$$q_{gs}(u) = q_{gd} = C_1 \cdot u \quad \text{with} \quad C_1 = 0.6 \cdot 10^{-13} \quad \text{F}.$$

The capacitance between bulk and drain as well as the one between bulk and source are modelled as nonlinear capacitances (Level B in [57]):

$$q_{db}(u) = q_{sb}(u) = \begin{cases} -C_0 \cdot \Phi_B \cdot (1 - \sqrt{1 - \frac{u}{\Phi_B}}) & \text{for} \quad u \leq 0, \\ -C_0 \cdot (1 + \frac{u}{4\Phi_B}) \cdot u & \text{for} \quad u > 0, \end{cases}$$

with

$$C_0 = 0.24 \cdot 10^{-13} F \quad \text{and} \quad \Phi_B = 0.87 \quad V.$$

The voltage sources V_1 and V_2 are supposed to fulfil $V_1(t_0) = V_2(t_0) = 0$, $V_1'(t_0) = 10^9$, and $V_2'(t_0) = 0$.

The obtained DC-operating point and the corresponding consistent values read

$$x^0 = \begin{pmatrix} 2.50000E-13 \\ -8.62244E-26 \\ -1.49214E-27 \\ 3.73965E-13 \\ 3.73965E-13 \\ -3.00000E-13 \\ -7.06710E-14 \\ 3.73965E-13 \\ 1.34912E-13 \\ -7.06710E-14 \\ -6.17042E-29 \\ 1.34912E-13 \\ 8.15517E-14 \\ 5.00000E+00 \\ 5.00000E+00 \\ 5.00000E+00 \\ 5.00000E+00 \\ 0.00000E+00 \\ 1.17785E+00 \\ 5.00000E+00 \\ 0.00000E+00 \\ 1.02840E-15 \\ 1.17785E+00 \\ 1.17785E+00 \\ -2.50000E+00 \\ 0.00000E+00 \\ 0.00000E+00 \\ 1.10103E-14 \\ -1.12674E-14 \end{pmatrix}, \; x_0 = \begin{pmatrix} 2.50000E-13 \\ -8.62244E-26 \\ -1.49214E-27 \\ 3.73965E-13 \\ 3.73965E-13 \\ -3.00000E-13 \\ -7.06710E-14 \\ 3.73965E-13 \\ 1.34912E-13 \\ -7.06710E-14 \\ -6.17042E-29 \\ 1.34912E-13 \\ 8.15517E-14 \\ 5.00000E+00 \\ 5.00000E+00 \\ 5.00000E+00 \\ 5.00000E+00 \\ 0.00000E+00 \\ 1.17785E+00 \\ 5.00000E+00 \\ 0.00000E+00 \\ 1.02840E-15 \\ 1.17785E+00 \\ 1.17785E+00 \\ -2.50000E+00 \\ -7.40744E-05 \\ 0.00000E+00 \\ 7.40744E-05 \\ -1.12674E-14 \end{pmatrix},$$

where it has to be mentioned that for an easier realization the exponential function describing i_{bs} and i_{bd} was approximated by a polynomial and for i_{bs}, i_{bd} and q_{db}, q_{sb} only the cases $u < 0$ and $u > 0$, respectively, were considered.

Assumptions of Chapters 1 and 2

A1 : $N(t) := \ker A(x,t), \quad \mathrm{im}\, A(x,t)$ depend smoothly on t, and do not depend on x for $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ $(p.14)$,

A2 : $N(t) \cap S(x,t)$ depends smoothly on t, and does not depend on x for $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ $(p.29)$,

A3 : $\frac{d}{dt}\Big\{(I_{W_1}W_0b)(U(t)x,t)\Big\}$ exists for all $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ $(p.43)$,

A4 : $W_1 \frac{\partial}{\partial t}\frac{\partial}{\partial x}\Big\{(I_{W_1}W_0b)\Big\} = W_1 \frac{\partial}{\partial x}\frac{\partial}{\partial t}\Big\{(I_{W_1}W_0b)\Big\}$,
$(W_1(I_{W_1}W_0b)'_x)'_x$, and $(W_1(I_{W_1}W_0b)'_t)'_x$ exist
for all $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$, where
$(W_1(I_{W_1}W_0b)'_x)'_x, \quad (W_1(I_{W_1}W_0b)'_t)'_x \quad \in C(\mathcal{D}_f \times \mathcal{I}_f, I\!R^n)$ $(p.43)$,

A5 : $\frac{d}{dt}\Big\{(\hat{W}_1b)(U(t)x,t)\Big\}$ exists for all $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ $(p.47)$,

A6 : $\mathrm{im}\, A(x,t), \quad \ker A(x,t)$ and $N(t) \cap S(x,t)$ are constant for $(x,t) \in \mathcal{D}_f \times \mathcal{I}_f$ $(p.64)$,

A7 : $A(Ux(t),t)x'(t) + \tilde{b}(Ux(t),t) + \mathcal{B}(Ux(t),t)Tx(t) = 0$ is given $(p.65)$,

A8 : $A(Ux(t),t)x'(t) + \tilde{b}(Ux(t),t) + \mathcal{B}Tx(t) = 0$ is given $(p.67)$,

A9 : $Ax'(t) + \tilde{b}(Ux(t),t) + \mathcal{B}Tx(t) = 0$ is given $(p.68)$.

Notations of Chapter 3

MNA		Modified Nodal Analysis,
$VCVS$		voltage-controlled voltage sources,
$CCVS$		current-controlled voltage sources,
$VCCS$		voltage-controlled current sources,
$CCCS$		current-controlled current sources,
L-I cutset		cutset consisting of inductances and/or current sources only,
C-V loop		loop consisting of capacitances and voltage sources only,
A	$=$	$(A_C, A_L, A_R, A_V, A_I)$ (reduced) incidence matrix describing the branch-node relations:
		A_C capacitive branches,
		A_L inductive branches,
		A_R resistive branches,
		A_V branches of voltage sources,
		A_I branches of current sources,
Q_C		projector onto $\ker A_C^T$,
Q_{V-C}		projector onto $\ker A_V^T Q_C$,
Q_{R-CV}		projector onto $\ker A_R^T Q_C Q_{V-C}$,
$\bar{Q}_C$		projector onto $\ker A_C$,
$\bar{Q}_{V-C}$		projector onto $\ker Q_C^T A_V$,
Q_{CRV}	$:=$	$Q_C Q_{V-C} Q_{R-CV}$,
$C(u,t)$	$:=$	$\dfrac{\partial q(u,t)}{\partial u}, \quad q_t'(u,t) := \dfrac{\partial q(u,t)}{\partial t}$,
$L(j,t)$	$:=$	$\dfrac{\partial \phi(j,t)}{\partial j}, \quad \phi_t'(j,t) := \dfrac{\partial \phi(j,t)}{\partial t}$,
$G(u,t)$	$:=$	$\dfrac{\partial r(u,t)}{\partial u}, \quad r_t'(u,t) := \dfrac{\partial r(u,t)}{\partial t}$,
$H_1(A_C^T e, t)$	$:=$	$A_C C(A_C^T e, t) A_C^T + Q_C^T Q_C$,

Bibliography

[1] P. Amodio and F. Mazzia. An algorithm for the computation of consistent initial values for differential–algebraic equations. *Numerical Algorithms*, 19:13–23, 1998.

[2] U. R. Ascher and L.R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations.* SIAM, Philadelphia, 1998.

[3] R. Bachmann, L. Brüll, T. Mrziglod, and U. Pallaske. On methods for reducing the index of differential-algebraic equations. *Computers chem. Engng.*, 14:1271–1273, 1990.

[4] 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.

[5] K.E. Brenan, S.L. Campell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* North-Holland, New York, 1989.

[6] P.N. Brown, A.C. Hindemarsh, and L.R. Petzold. Using Krylov methods in the solution of large-scale differential-algebraic systems. *[J] SIAM J. Sci. Comput.*, 15(6):1467–1488, 1994.

[7] P.N. Brown, A.C. Hindemarsh, and L.R. Petzold. Consistent initial condition calculations for differential-algebraic systems. *[J] SIAM J. Sci. Comput.*, 19(5):1495–1512, 1998.

[8] P. R. Bryant. The order of complexity of electrical networks. *Proc. IEE (GB), Part C*, 106:174–188, 1959.

[9] S.L. Campbell and C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[10] C.A. Desoer and E.S. Kuh. *Basic Circuit Theory.* McGraw-Hill, Singapore, 1969.

[11] D. Estévez Schwarz. Consistent initial values for DAE systems in circuit simulation. Preprint 99–5, Humboldt-Universität, Berlin, 1999.

[12] D. Estévez Schwarz. Topological analysis for consistent initialization in circuit simulation. Preprint 99–3, Humboldt-Universität, Berlin, 1999.

[13] D. Estévez Schwarz, U. Feldmann, R. März, S. Sturtzel, and C. Tischendort. Finding beneficial DAE structures in circuit simulation. Preprint 00–7, Humboldt-Universität, Berlin, 2000.

[14] D. Estévez Schwarz and R. Lamour. The Computation of Consistent Initial Values for Nonlinear Index-2 Differential-Algebraic Equations. Preprint 99–13, Humboldt-Universität, Berlin, 1999.

[15] D. Estévez Schwarz and C. Tischendorf. Structural analysis of electric circuits and consequences for MNA. *Int. J. of Circuit Theory and Applications*, 28:131–162.

[16] 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.

[17] M. Fosséprez. *Non-linear Circuits: Qualitative Analysis of Non-linear, Non-reciprocal Circuits.* John Wiley & Sons, Chichester, 1992.

[18] F. R. Gantmacher. *Teorija matrits.* Gosudarstv. Izdat. Techn.-Teor. Lit., Moskva, 1954.

[19] C. W. Gear. Differential-Algebraic Equation Index Transformations. *SIAM J. Sci. Stat. Comput.*, 9:39–47, 1988.

[20] C. W. Gear. Differential algebraic equations, indices, and integral algebraic equations. *SIAM J. Numer. Anal.*, 27(6):1527–1534, 1990.

[21] C. W. Gear and L. R. Petzold. ODE Methods for the Solution of Differential/Algebraic Systems. *SIAM J. Numer. Anal.*, 21:716–728, 1984.

[22] C.W. Gear, G.K. Gupta, and B.J. Leimkuhler. Automatic integration of Euler-Lagrange equations with constraints. *J. Comp. and Appl. Math.*, 12/13:77–90, 1985.

[23] V. Gopal and L.T. Biegler. A successive linear programming approach for initialization and reinitialization after discontinuities of differential-algebraic equations. *SIAM J. Sci. Comput*, 20(2):447–467, 1999.

[24] E. Griepentrog. Index Reduction Methods for Differential-Algebraic Equations. In E. Griepentrog, M. Hanke, and R. März, editors, *Seminarbericht 92–1*, pages 14–29. Humboldt-Universität, Berlin, 1992.

[25] E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment.* Teubner-Texte Math. 88. Teubner, Leipzig, 1986.

[26] M. Günther and U. Feldmann. CAD-based electric-circuit modeling in industry I. Mathematical structure and index of network equations. *Surv. Math. Ind.*, 8:97–129, 1999.

[27] M. Günther and U. Feldmann. CAD based electric modeling in industry. Part II: Impact of circuit configurations and parameters. *Surv. Math. Ind.*, 8:131–157, 1999.

[28] M. Günther and P. Rentrop. Suitable One-Step Methods for Quasilinear-Implicit ODE's. Technical Report TUM-M9405, Technische Universität München, 1994.

[29] M. Günther and P. Rentrop. The NAND-gate - a benchmark for the numerical simulation of digital circuits. pages 27–33. VDE-Verlag, 1996.

[30] E. Hairer, Ch. Lubich, and M. Roche. *The numerical solution of differential-algebraic systems by Runge-Kutta methods.* Lecture Notes in Math. 1409. Springer Verlag, 1989.

[31] 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.

[32] B. Hansen. Computing Consistent Initial Values for Nonlinear Index-2 Differential-Algebraic Equations. In E. Griepentrog, M. Hanke, and R. März, editors, *Seminarbericht 92–1*, pages 142–157. Humboldt-Universität, Berlin, 1992.

[33] A. Kröner, W. Marquardt, and E. D. Gilles. Computing consistent initial conditions for differential-algebraic equations. *Comput. Chem. Eng.*, 16:131–138, 1992. (Supplement).

[34] A. Kröner, W. Marquardt, and E. D. Gilles. Getting around consistent initialization of DAE systems? *Comput. Chem. Eng.*, 21:145–158, 1996.

[35] R. Lamour. A well-posed shooting method for transferable DAE's. *Numer. Math.*, 59:815–8294, 1991.

[36] R. Lamour. A Shooting Method for Fully Implicit Index-2 Differential-Algebraic Equations. *SIAM J. Sci. Comput*, 18:94–114, 1997.

[37] B. Leimkuhler. *Approximation Methods for the Consistent Initialization of Differential-Algebraic Equations.* PhD thesis, University of Illinois, 1988.

[38] B. Leimkuhler, L.R. Petzold, and C.W. Gear. Approximation Methods for the Consistent Initialization of Differential-Algebraic Equations. *SIAM J. Numer. Anal.*, 28:205–226, 1991.

[39] R. März. A Matrix Chain for Analysing Differential-Algebraic Equations. Preprint 162, Humboldt-Universität, Berlin, 1987.

[40] R. März. Some new results concerning index-2 differential-algebraic equations. *J. Math. Anal. Appl.*, 140(1):177–199, 1989.

[41] R. März. Numerical methods for differential-algebraic equations. *Acta Numerica*, pages 141–198, 1992.

[42] R. März. On linear differential-algebraic equations and linearizations. *APNUM*, 18:267–292, 1995.

[43] R. März. Analysis und Numerik für Algebro-Differentialgleichungen. Special lecture, 1998.

[44] R. März. EXTRA-ordinary differential equations: Attempts to an analysis of differential-algebraic systems. *Progress in Mathematics*, 168:313–334, 1998.

[45] R. März. Algebro-Differentialgleichungen und ihre adjungierten Systeme in neuer einheitlicher Form. Talk given at a Seminar, 1999.

[46] R. März and A. R. Rodríguez Santiesteban. Analyzing the stability behaviour of DAE solutions and their approximations. Preprint 99–2, Humboldt-Universität, Berlin, 1999.

[47] R. März and C. Tischendorf. Solving more general index-2 differential-algebraic equations. *Comput. Math. Appl.*, 28(10-12):77–105, 1994.

[48] C. C. Pantelides. The Consistent Initialization of Differential-Algebraic Systems. *SIAM J. Sci. Statist.Comput.*, 9:213–231, 1988.

[49] L. R. Petzold. A Description of DASSL: A differential/algebraic system solver. In *Proc. 10th IMACS World Congress, August 8-13 Montreal*, 1982.

[50] 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.

[51] G. Reißig. *Beiträge zu Theorie und Anwendung impliziter Differentialgleichungen.* PhD thesis, Techn. Univ. Dresden, 1998.

[52] G. Reißig, W. S. Martinson, and P. I. Barton. Differential-algebraic equations of index 1 may have an arbitrarily high structural index. *SIAM J. Sci. Statist.Comput.*, 1999. To appear.

[53] W. C. Rheinboldt. Differential-algebraic systems as differential equations on manifolds. *Math. Comp.*, 43:473–482, 1984.

[54] H. Schwetlick. *Numerische Lösung nichlinearer Gleichungen.* VEB Deutscher Verlag der Wissenschaft, Berlin, 1979.

[55] L.F. Shampine and J. Kierzenka. Solving DAE's with inconsistent initial conditions. *Commun. Appl. Anal.*, 2(4):513–520, 1998.

[56] H. Shichman and D. A. Hodges. Insulated-gate field-effect transistor switching circuits. *IEEE J. Solid State Circuits*, SC-3:285–289, 1968.

[57] C. Tischendorf. *Solution of index-2 differential algebraic equations and its application in circuit simulation.* PhD thesis, Humboldt-Universität zu Berlin, 1996.

[58] C. Tischendorf. Topological index calculation of DAEs in circuit simulation. *Surv. Math. Ind.*, 8(3-4):187–199, 1999.

[59] J. Unger, A. Kröner, and W. Marquardt. Structural analysis of differential-algebraic equations systems-theory and applications. *Comput. Chem. Eng.*, 19(8):867–882, 1995.

[60] H. Unglaub. Master thesis. In preparation.

[61] J. Vlach, J. M. Wojciechowsi, and A. Opal. Analysis of nonlinear networks with inconsistent initial conditions. *IEEE Transactions on circuits and systems: Fundamental theory and applications*, 42(4):195–200, 1995.

[62] H. Wriedt. Transientensimulation elektrischer Netzwerke mit TRBDF. *Int. Ser. Numer. Math.*, 117:133–142, 1994.