A Dissection Concept for DAEs

Structural Decoupling, Unique Solvability, Convergence Theory and Half-Explicit Methods

DISSERTATION

zur Erlangung des akademischen Grades

Dr. rer. nat.
im Fach Mathematik

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

von
Dipl.-Math. Lennart Jansen

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Mathematisch-Naturwissenschaftlichen Fakultät:
Prof. Dr. Elmar Kulke

Gutachter/Gutachterin:
1. Prof. Dr. Caren Tischendorf
2. Prof. Dr. Volker Mehrmann
3. Prof. Dr. Claus Führer

eingereicht am: 16.06.14
Tag der Verteidigung: 13.10.14

Kurzzusammenfassung

Diese Arbeit befasst sich mit Differential-algebraischen Gleichungen (DAEs). DAEs spielen eine wichtige Rolle in der Modellierung, der Simulation und der Optimierung von Netzwerken und gekoppelten Problemen in vielen Anwendungsgebieten. In dieser Arbeit sind die gekoppelten Probleme aus der elektrischen Schaltungssimulation die zentrale Anwendung. Es werden in Bezug auf die Modellierung und die numerische Simulation von DAEs bereits bestehende Ergebnisse diskutiert und erweitert. Des Weiteren wird die globale eindeutige Lösbarkeit und die Sensitivität der Lösungen mit Hinsicht auf Störungen der DAEs untersucht.
Häufig wird die Modellierung von multiphysikalischen Anwendungen durch die Kopplung mehrerer einzelner DAE Systeme realisiert. Diese Herangehensweise kann hochdimensionale DAEs erzeugen, welche aufgrund von Instabilitäten nicht von klassischen numerischen Methoden, wie den BDF-Methoden, simuliert werden können. Angesichts dieser Herausforderungen werden drei Ziele formuliert: Erstens wird ein globales Lösungs-theorem formuliert und bewiesen, welches auf gekoppelte Systeme angewandt werden kann, um deren Kopplungsansatz mathematisch zu rechtfertigen. Zweitens werden numerische Methoden vorgestellt, welche unter wesentlich schwächeren Strukturannahmen stabil sind und sich daher für die Simulation von gekoppelten Systemen eignen. Drittens wird eine Strategie präsentiert, die es ermöglicht, explizite Methoden auf gekoppelte Systeme aus der Schaltungssimulation anzuwenden.
Eines der wichtigsten Werkzeuge, um diese Ziele zu erreichen sind die Indexkonzepte für DAEs. Zwei der bekanntesten Indexkonzepte sind der Tractability Index und der Strangeness Index. Beide können als Entkopplungsverfahren verstanden werden. Hier wird ein neues Indexkonzept vorgestellt, welches im Folgenden als der Dissection Index bezeichnet wird. Die Definition eines neuen Indexkonzepts wirft unweigerlich die Frage auf: Warum braucht man ein weiteres Indexkonzept? Um die oben gestellten Ziele zu erreichen, braucht man ein Entkopplungsverfahren, welches die folgenden drei Eigenschaften erfüllt: Die Komplexität des Entkopplungsverfahrens sollte nicht die Komplexität der DAE überschreiten. Das Entkopplungsverfahren sollte Eigenschaften wie Symmetrie, Monotonie und positive Definitheit erhalten. Das Entkopplungsverfahren sollte durch einen Schritt-für-Schritt Ansatz mit unabhängigen Schritten realisiert werden. Sowohl das Konzept des Tractability Index als auch das des Strangeness Index liefert kein solches Entkopplungsverfahren. Der Dissection Index hingegen kann ein solches erzeugen, wie in dieser Arbeit zu sehen sein wird. Alle theoretischen Ergebnisse dieser Arbeit werden auf gekoppelte Probleme aus der Schaltungssimulation angewandt.

Abstract

This thesis addresses differential-algebraic equations (DAEs). They play an important role in the modeling, simulation and optimization of networks and coupled problems in various applications. The main application in this thesis are coupled problems in electric circuit simulation.

We discuss and extend existing results regarding the modeling and numerical simulation of DAEs. Furthermore, we investigate the global unique solvability and the sensitivity of solutions with respect to perturbations of DAEs.

Nowadays the modeling of multi-physical applications is often realized by coupling systems of DAEs together with the help of additional coupling terms. This approach may yield high dimensional DAEs which cannot be simulated, due to instabilities, by standard numerical methods. Regarding these challenges we formulate three objectives: First we provide a global solvability theorem which can be applied to coupled systems to mathematically justify their coupling approach. Second we introduce numerical methods which are stable without needing any structural assumptions. Third we provide a way to apply explicit methods to coupled systems to be able to handle the size of the coupled systems by parallelizing the algorithms.

One of the most important tools to achieve these objectives are the index concepts for differential-algebraic equations. Two of the most popular index concepts are the Tractability Index and the Strangeness Index. They both provide a decoupling procedure. Here we introduce a new index concept which we will call the Dissection Index. The definition of a new index concept inevitably invokes the following question: Why do we need another index concept?

To achieve the objectives stated above, we need a decoupling procedure which fulfills the following three properties: The complexity of the decoupling procedure has to reflect the complexity of the DAE, i.e. the decoupling procedure should be state-independent if possible. The decoupling procedure should preserve properties like symmetry, monotonicity and positive definiteness. The decoupling procedure should be realized by a step-by-step approach with independent stages.

Both the Tractability Index concept and the Strangeness Index concept do not provide such a decoupling procedure. Whereas the Dissection Index does, as we will see in this thesis.

The theoretical results in this thesis will be applied to coupled problems in the electric circuit simulation.

Acknowledgement

There are many people who truly deserve to be mentioned in this acknowledgement. But in comparison to the support my mother has given me throughout my entire life, the support of these beloved and helpful people seems to fade.

And therefore everything left to say is:

Thank you, mother. You are awesome.

Lennart Jansen

Berlin, June 16th, 2014.

Contents

1 Introduction

This thesis addresses differential-algebraic equations (DAEs). They play an important role in the modeling, simulation and optimization of networks and coupled problems in various applications, e.g. integrated circuit design, hydraulic engineering, mechanical engineering and medicine. Often, the model equations lead to a partial differential-algebraic equation system (PDAE) meaning a mix of ordinary differential equations, partial differential equations and algebraic constraints. We focus our investigations on general differential-algebraic equations resulting from a spatial discretization of such PDAEs. We discuss and extend existing results regarding the modeling and numerical simulation of DAEs. Furthermore, we investigate the global unique solvability and the sensitivity of solutions with respect to perturbations of DAEs.

Nowadays the modeling of multi-physical applications is often realized by coupling systems of DAEs together with the help of additional coupling terms. This approach may yield high dimensional DAEs which cannot be simulated, due to instabilities, by standard numerical methods. In particular we will

1. provide a global solvability theorem which can be applied to coupled systems to mathematically justify their coupling approach.
2. provide numerical methods which are stable under almost no structural assumptions.
3. provide a way to apply explicit methods to coupled systems to be able to handle the size of the coupled systems by parallelizing the algorithms.

One of the most important tools to achieve these objectives are the index concepts for differential-algebraic equations. There are already many different index concepts available. The Differentiation Index is probably the best known index, since its concept is relatively demonstrative. It was introduced by Petzold and Campbell, see [Cam87, BCP96]. The Perturbation Index measures the degree of the influence of the derivatives of perturbation to the solution of a DAE. It was initially defined in [HLR89]. Two of the most popular index concepts are the Tractability Index [GM86, Mär02, LMT13] and the Strangeness Index [KM06]. They both provide a decoupling procedure. All of these index concepts have their own advantages and disadvantages. Here we introduce a new index concept which we will call the Dissection Index. The definition of a new index concept inevitably invokes the following question: Why do we need another index concept?

To achieve the objectives stated above, we need a decoupling procedure which fulfills the following properties:

1. The complexity of the decoupling procedure has to reflect the complexity of the DAE, i.e. the decoupling procedure should be state-independent if possible.
2. The decoupling procedure should preserve properties like symmetry, monotonicity and positive definiteness.
3. The decoupling procedure should be realized by a step-by-step analysis with independent stages.

Both the Tractability Index concept and the Strangeness Index concept do not provide such a decoupling procedure. Whereas the Dissection Index does, as we will see in this thesis.
The Dissection Index can be interpreted as a mix of the Tractability Index and the Strangeness Index. The index arises as we use the linearization concept of the Tractability Index and the decoupling procedure of the Strangeness Index. The Strangeness Index uses basis functions for its decoupling procedure while the Tractability Index uses projectors for this purpose. The advantage of projector functions is that they need less assumptions regarding the domain to be differentiable. Nevertheless, we favor basis functions since they preserve the original size of the equations while splitting them.

This thesis is structured as follows. After presenting well-known results and facts of differential-algebraic equations, we introduce the concepts of the Strangeness Index and the Tractability Index. Before we define our new index concept, we present and model the application classes which will be discussed in this thesis. These classes are electrical circuits including semiconductor devices, memristors and electromagnetic devices and mechanical multibody systems.

After introducing our new index concept and proving that it is well defined, we will analyze the sensitivity to perturbations of differential-algebraic equations. In contrast to ordinary differential equations, which can be interpreted as integral problems, differential-algebraic equations may contain differentiation problems. The appearance of these differentiation problems leads to an ill-posed problem, in the sense of Hadamard, if we consider perturbed input data, see [LRS86]. Even very small perturbations can have arbitrarily large derivatives and therefore small perturbations may have a huge influence on the solution of a differential-algebraic equation. Hence it is necessary to analyze the sensitivity to perturbations of DAEs.

In case of the perturbation analysis and also for the convergence theory it is necessary to assume that the unperturbed DAE has a global unique solution. Furthermore, we need to prove the global unique solvability of our considered coupled systems to mathematically justify their coupling approach. We will provide sufficient criteria for the global unique solvability of differential-algebraic equations with an arbitrary index. To do so we need insight of the structure of the differential-algebraic equation to apply the established solvability theories of ordinary differential equations and algebraic equations. To obtain this needed insight we make use of the Dissection Index concept.

The remaining two chapters deal with challenges of the applicability, the stability and the convergence of numerical methods. It is known that standard ODE methods like the implicit Euler methods, the BDF methods or the Radau IIA methods may loose their convergence if applied to DAEs, cf. [GP83, LMT13]. These standard ODE methods have a basic flaw: They do not reflect the product rule properly. This is not a problem as long as these methods are applied to ODEs. When we consider a DAE it may happen that the kernels, which describe the inner structure of a DAE, are not constant. If these kernels are also involved in a differentiation problem then hidden differentiations of products of functions might occur which leads to the instability of these standard ODE methods. We will introduce a class of methods which reflects the product rule properly and thereby overcomes these instability problems. In particular this will make the reformulation of the DAE superfluous.

In the last chapter we investigate half-explicit methods applied to DAEs. Since it is no longer possible to accelerate CPUs like it has been in the past, parallelizing algorithms becomes more and more important. Because they can be paralellized very efficiently, explicit methods are being focused on even more so nowadays. Hence explicit methods are focus even more nowadays because they can be paralellized very efficiently. Half-explicit methods for DAEs are often defined for semi-explicit DAEs and therefore they are rarely used in circuit simulation in contrast to mechanical applications. We introduce a new class of half-explicit multistep methods and prove their convergence.

The main application in this thesis will be the electric circuit simulation. Electrical circuits are of great importance for industrial research and therefore a mathematical understanding is needed. There already exist many works about DAE related questions regarding circuits, in particular about index analysis [Tis99, RT11] and local uniqueness and solvability [HM04, Bau12]. Besides standard elements like inductors, resistors, capacitors and source elements, electrical circuits can also contain more complex elements like semiconductor devices, memristors and electromagnetic devices. For instance, semiconductor devices and electromagnetic devices are described by a set of partial differential equations, hence they involve new questions and challenges to the analysis of electrical circuits. Previous research about semiconductor devices can be found in [SBST14, Gaj93, Gaj94, Tis03, ABG04, ST05, Sot06, Bod07, BST10], these devices are widely used in circuits because of their application as transistors. In [Chu11, Ria11, RT11, Bau12] memristors are investigated while we find previous research about electromagnetic devices in [HM76, KMST93, Wei77, Bau12, BBS11, Sch11].

In this thesis we will

1. apply our perturbation analysis to the circuit applications.
2. prove the global unique solvability of the coupled circuit model.
3. provide a topological decoupling into a semi-explicit DAE for the circuit applications which has low computational costs and preserves the symmetry and positive definiteness of the circuit model.
4. apply our class of half-explicit methods to circuit applications.

2 An Introduction to Differential-Algebraic Equations

This chapter presents well-known results and facts for Differential-Algebraic Equations (DAEs) and lays the foundations for the following chapters.
The first part of the chapter is a general introduction to differential-algebraic equations. It presents challenges and problems of this field with the help of small examples. This includes classical problems like the appearance of differentiation problems or the drift-off phenomenon.
Furthermore the concepts of the Differentiation Index [Cam87, BCP96], the Strangeness Index [KM06] and the Tractability Index [GM86, Mär02, LMT13] are introduced and discussed. These three concepts are well established analysis tools for DAEs. Their respective advantages and disadvantages will be pointed out.

2.1 Explicit ODEs vs. DAEs

Differential-algebraic equations as well as explicit Ordinary Differential Equations(ODEs) can be understood as implicit ODEs. The following definitions follow the understanding of the relations between explicit ODEs, DAEs and implicit ODEs of [LMT13].

Definition 2.1. (Implicit ODE)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D}_x, \mathcal{D}_{x'} \subset \mathbb{R}^n$ be open subsets. We consider the following equation

$$F(x'(t), x(t), t) = 0 \tag{2.1}$$

with a continuous function $F \in C(\mathcal{D}_{x'} \times \mathcal{D}_x \times \mathcal{I}, \mathbb{R}^n)$. Furthermore let F have continuous partial derivatives $\frac{\partial}{\partial x^1} F(x^1, x, t)$ and $\frac{\partial}{\partial x} F(x^1, x, t)$. We call (2.1) an implicit ODE. If $\frac{\partial}{\partial x^1} F(x^1, x, t)$ is non-singular for all triples $(x^1, x, t) \in \mathcal{D}_{x'} \times \mathcal{D}_x \times \mathcal{I}$, we call (2.1) a regular implicit ODE.

In particular explicit ODEs are regular implicit ODEs.

Definition 2.2. (Explicit ODE)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D} \subset \mathbb{R}^n$ be open subsets with $t_0 \in \mathcal{I}$ and let $f \in C(\mathcal{D} \times \mathcal{I}, \mathbb{R}^n)$ be continuous. We call

$$x'(t) = f(x(t), t), \quad x(t_0) = x^0 \tag{2.2}$$

an explicit ODE with an initial condition. Let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$. We call $x_\star \in C^1(\mathcal{I}_\star, \mathbb{R}^n)$ a solution of (2.2) on $\mathcal{I}_\star$ if the initial conditions are fulfilled, i.e. $x_\star(t_0) = x^0$, and

$$x'_\star(t) = f(x_\star(t), t) \quad \forall t \in \mathcal{I}_\star.$$

In contrast to explicit ODEs, DAEs are implicit ODEs, which are not regular.

Definition 2.3. (DAE in standard form)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D}_x, \mathcal{D}_{x'} \subset \mathbb{R}^n$ be open with $t_0 \in \mathcal{I}$. Let $F \in C(\mathcal{D}_{x'} \times \mathcal{D}_x \times \mathcal{I}, \mathbb{R}^n)$ be continuous such that the partial derivatives $\frac{\partial}{\partial x^1} F(x^1, x, t)$ and $\frac{\partial}{\partial x} F(x^1, x, t)$ are continuous with $\frac{\partial}{\partial x^1} F(x^1, x, t)$ being singular for all triples $(x^1, x, t) \in \mathcal{D}_{x'} \times \mathcal{D}_x \times \mathcal{I}$. We call

$$F(x'(t), x(t), t) = 0, \quad x(t_0) = x^0 \tag{2.3}$$

a DAE in standard form with initial conditions.We call $x_\star \in C^1(\mathcal{I}_\star, \mathbb{R}^n)$ a solution of (2.3) on $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ if the initial conditions are fulfilled, i.e. $x_\star(t_0) = x^0$, and

$$F(x'_\star(t), x_\star(t), t) = 0 \quad \forall t \in \mathcal{I}_\star.$$

We also introduce the following subclass of DAEs.

Definition 2.4. (Semi-explicit DAE)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D}_x \subset \mathbb{R}^{n_x}$ and $\mathcal{D}_y \subset \mathbb{R}^{n_y}$ be open subsets. We consider the following set of equations

$$x' = f(x, y, t) \tag{2.4a}$$
$$0 = g(x, y, t) \tag{2.4b}$$

with $f \in C(\mathcal{D}_x \times \mathcal{D}_y \times \mathcal{I}, \mathbb{R}^{n_x})$ and $g \in C(\mathcal{D}_x \times \mathcal{D}_y \times \mathcal{I}, \mathbb{R}^{n_y})$. Further, let the partial derivatives of f and g, with respect to x and y, be continuous. We call (2.4) a semi-explicit DAE.

In the case of a semi-explicit DAE only the derivatives of x appear in the equations. Therefore we call x the dynamical variables, while we call y the algebraic variables. Analogously we call the equations (2.4a) the dynamical equations and (2.4b) the algebraic equations of the semi-explicit DAE (2.4). Hence (2.4) is a special case of a DAE in the sense that it is an implicit ODE which is not regular.

In the following we present differences between explicit ODEs and DAEs with the help of a sequence of small examples.

Solvability for Arbitrary Initial Values

There are well-known solvability results for explicit ODEs which provide criteria for the solvability with respect to an arbitrary initial value x^0 at a time point t_0. Peano's theorem states that there is a $T > t_0$ such that there is at least one solution for every initial condition of the ODE on $[t_0, T]$ if f is continuous, cf. [Aul04]. We say the ODE is locally solvable, since the solution interval $[t_0, T]$ can be arbitrarily small. If the function f is locally Lipschitz continuous in x, this local solution becomes unique by the Picard-Lindelöf theorem, cf. [Aul04]. If f is even globally Lipschitz continuous in x, then for every initial condition and for every time interval $[t_0, T]$ with $T > t_0$ there is a unique solution of the ODE on the whole time interval $[t_0, T]$, cf. [GJ09].
It is not possible to formulate such results for arbitrary initial conditions in the DAE case. We consider the following DAE as a counter example.

Example 2.5. Let $\mathcal{I} := [t_0, T] \subset \mathbb{R}$ be a compact time interval and let $t \in \mathcal{I}$. Assume $f : \mathcal{I} \to \mathbb{R}$ to be continuous.

$$x' = y \tag{2.5a}$$

$$y = f(t) \tag{2.5b}$$

This DAE consists of one differential equation (2.5a) and one algebraic equation (2.5b). In contrast to the solvability of explicit ODEs for arbitrary initial conditions, see Peano's theorem and the Picard-Lindelöf theorem, the DAE (2.5) is only solvable for initial values satisfying $y(t_0) = f(t_0)$.

Differentiation Problem

We can write an explicit ODE (2.2) in integral notation

$$x(t) = x^0 + \int_{t_0}^{t} f(x(s), s)\mathrm{d}s$$

such that we deal with an integration problem. Hence, it is possible to notate an explicit ODE (2.2) without derivatives. This is no longer the case for DAEs, in general. If we change the algebraic equation in Example 2.5 by setting x equal to f instead of y, we get a very similar looking DAE, which has a totally different solution structure.

Example 2.6. Let $\mathcal{I} := [t_0, T] \subset \mathbb{R}$ be a compact time interval and let $t \in \mathcal{I}$. Let $f : \mathcal{I} \to \mathbb{R}$ be continuously differentiable.

$$\begin{aligned} x' &= y \\ x &= f(t) \end{aligned}$$

This time the dynamical variable x is fixed algebraically. In fact all solution components are algebraically fixed, since $x(t) = f(t)$ and $y(t) = x'(t) = f'(t)$. We cannot choose any initial value freely. Additionally, y depends on the derivative of the right hand side $f(t)$. So we are dealing with a differentiation problem instead of an integration problem. As an obvious analytic consequence, the right hand side $f(t)$ must be sufficiently smooth, as already mentioned in the example.
It is possible to create a differentiation problem of second order if we add one more differentiation to the equations of Example 2.6.

Example 2.7. Let $\mathcal{I} := [t_0, T] \subset \mathbb{R}$ be a compact time interval and let $t \in \mathcal{I}$. Let $f : \mathcal{I} \to \mathbb{R}$ be twice continuously differentiable.

$$\begin{aligned} x_2' &= y \\ x_1' &= x_2 \\ x_1 &= f(t) \end{aligned}$$

This set of equations is solved by $x_1 = f(t)$, $x_2 = f'(t)$ and $y = f''(t)$. Hence, the solution of y is the second derivative of f.
The appearance of differentiation problems within DAEs motivates a classification for DAEs which counts the order of the involved differentiation problem. This kind of classification is called the index of a DAE. There are several different index concepts, put simply they are all intended for counting the order of the involved differentiation problem. These index concepts are the most important tools of DAE analysis. We give an introduction of some of the most popular index concepts in the Sections 2.2, 2.3 and 2.4.

Numerical Approximation of the Difference Quotient

The differentiation problems in DAEs induce smoothness assumptions regarding the right hand side f. Additionally, there are numerical problems created by the differentiation problems. If we discretize Example 2.6 by the implicit Euler method and a time step size h, we obtain

$$\begin{aligned} \frac{x_n - x_{n-1}}{h} &= y_n \\ x_n &= f(t_n), \end{aligned}$$

with x_n and y_n being the numerical approximations of $x(t_n)$ and $y(t_n)$, respectively. It follows directly that y_n is given by the difference quotient of $f(t)$ at t_n,

$$y_n = \frac{f(t_n) - f(t_n - h)}{h}.$$

The difference quotient of a differentiable function $f(t)$ at a time point t_n converges to its derivative at the same time point $f'(t_n)$, i.e.

$$\lim_{h \to 0} \frac{f(t_n) - f(t_n - h)}{h} = f'(t_n).$$

But this convergence may fail if the difference quotient is computed on a machine with finite precision arithmetic. Numerically calculated values are only as accurate as the rounding accuracy of the used computer system. We call the rounding error δ. The rounding error can be seen as a random number in $[-10^{-16}, 10^{-16}]$ if one uses the double precision floating-point format. This phenomenon is called the loss of significance.
To visualize this problem let $f(t) = sin(t)$ and let δ_n be the rounding error of $sin(t_n)$. Then we actually calculate

$$\begin{aligned} y_n &= \frac{sin(t_n) + \delta_n - sin(t_{n-1}) - \delta_{n-1}}{h} \\ &= \frac{sin(t_n) - sin(t_{n-1})}{h} + \frac{\delta_n - \delta_{n-1}}{h} \end{aligned}$$

Therefore the numerical solution y_n will not converge against $cos(t_n)$ for $h \to 0$, since the rounding error δ_{n-1} does not converge against δ_n. In Figure 2.1 we see the numerical error $e_n := |y_n - cos(t_n)|$ at $t_n = 1$ for different time step sizes h.

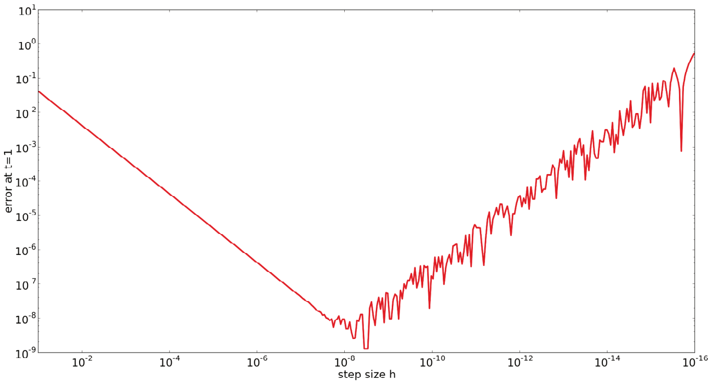


Figure 2.1: Numerical error of the difference quotient.

This basic flaw of the difference quotient is one of the main problems in the numerical treatment of DAEs. This fact is well-known since differentiation is an ill-posed problem, in the sense of Hadamard, if it is connected with perturbed input data, see [LRS86]. Hence, small time step sizes no longer yield small errors, in general. This problem can get even worse if we consider Example 2.7. Applying the implicit Euler method again we achieve

$$y_n = \frac{f(t_n) - 2f(t_{n-1}) + f(t_{n-2})}{h^2},$$

with a constant time step size h. This confronts us with a harder version of the problem of Example 2.6, since the rounding error will be multiplied by $\frac{1}{h^2}$.

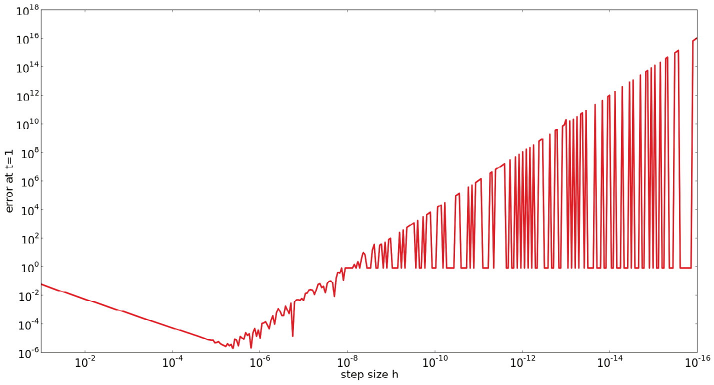


Figure 2.2: Numerical error of the difference quotient of the second derivative.

We choose again $f(t) = sin(t)$ and $t_n = 1$. Figure 2.2 describes the relation between the time step size and the accuracy of y_n reflecting the second derivative of f. As soon as the step size h drops below 10^{-8} it may happen that $f(t_n) - 2f(t_{n-1}) + f(t_{n-2})$ becomes smaller than 10^{-16} which then will be presented by a subnormal number. This explains the behavior of the error for time step sizes smaller than 10^{-8}. Summarizing, this means that a high order of the differentiation problem may lead to difficulties during the solving of the DAE.

Mixed Variables

In all the previous examples it is obvious which of the variables have to be differentiable and which are directly algebraically fixed. This does not have to be the case in general and also not in most applications. We consider the following example.

Example 2.8. Let be $\mathcal{I} := [-1, 1]$ and $t \in \mathcal{I}$.

$$(z_0 - z_1)' = z_0 + z_1 \tag{2.6a}$$

$$z_0 + z_1 = 4|t|. \tag{2.6b}$$

The general solution of this problem is given by $z_0(t) = (2+t)|t|+c$ and $z_1(t) = (2-t)|t|-c$ for some $c \in \mathbb{R}$. The initial value of either z_0 or z_1 can be chosen, but not both at the same time. So which of the solution components is algebraically fixed? They are both not fixed but the combination $(z_0+z_1)(t)$ is fixed, as we can see in the second equation (2.6b). This example shows also an important smoothness property of DAEs. In the examples 2.6 and 2.7 it was shown that the right hand side can underlie smoothness requirements. Now we see that smoothness properties of the variables are not trivial either. In Example 2.8 non of the solution components for themselves are even differentiable once. But there appears a derivative of a combination of the variables in the equations and this combination $(z_0 - z_1)(t) = 2t|t| + 2c$ is in fact continuously differentiable. Motivated by the class of semi-explicit DAEs (2.4) we call the parts of the variables, whose derivatives appear in the equations, dynamic. The remaining parts of the variables are called algebraic. Notice that these parts must not necessarily be a set of notated variables, but they can also be combinations of the variables.

As an application example for mixed variables consider the mathematical pendulum, cf. [Ste06]. This is one of the most basic DAE examples in mechanical applications.

Example 2.9. Let $\mathcal{I} := [0, 2 \cdot 10^{-6}]$ be the time interval.

$$\begin{aligned} p_1' &= v_1 \\ p_2' &= v_2 \\ mv_1' &= -2p_1\lambda \\ mv_2' &= -2p_2\lambda - mg \\ 0 &= p_1^2 + p_2^2 - L^2 \end{aligned}$$

with $m > 0$ being the mass of the object, g being the gravity of earth and L being the length of the pendulum.

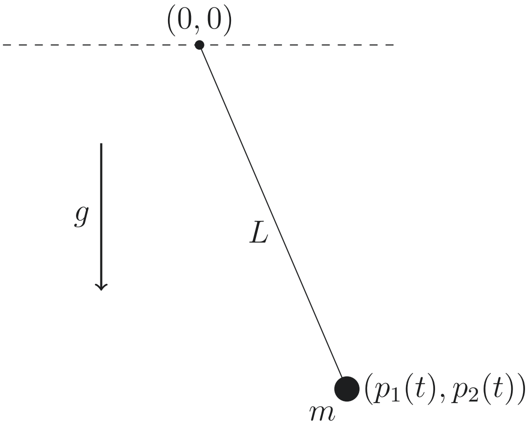


Figure 2.3: Mechanical example: mathematical pendulum

This DAE is interesting, in a mathematical point of view, because of its last equation. Obviously it is the only algebraic equation, since $m > 0$. Now the question is: Which parts of p_1 and p_2 are algebraically fixed? In fact the fixed combination of p_1 and p_2 depends on the current states of p_1 and p_2 themselves. Analyzing DAEs theoretically becomes much harder, if such state depending combinations show up.

Consistent Initial Values

The next two examples show what may happen during a simulation if the initial values are chosen randomly.

Example 2.10. Let $\mathcal{I} := [0, 1]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} x_2' &= y \\ x_1' &= e^y - 1 \\ x_1 &= -1 \end{aligned}$$

The exact solution of this example is given by $x_1(t) = -1$, $x_2(t) = x_2^0 \in \mathbb{R}$, $y(t) = 0$. We assume that we do not know the exact solution but we have to choose initial values. We observe what happens if we choose $x_1^0 = 0$, $x_2^0 = 0$, $y^0 = 0$ and discretize the example by the implicit Euler scheme. First we get the discretized system for the first time step

$$\begin{aligned} \frac{x_{2,1} - x_{2,0}}{h} &= y_1 \\ \frac{x_{1,1} - x_{1,0}}{h} &= e^{y_1} - 1 \\ x_{1,1} &= -1. \end{aligned}$$

We obtain, after reorganizing the equations,

$$x_{2,1} = hy_1 \tag{2.7a}$$

$$y_1 = ln(1 - \frac{1}{h}) \tag{2.7b}$$

$$x_{1,1} = -1 \tag{2.7c}$$

if we then insert the chosen initial values. Since the time step size h is larger than zero it has to be larger than one because equation (2.7b) is not solvable for $0 < h \leqslant 1$. But if we have to choose $h > 1$ we cannot solve the example since the solution interval is given by $\mathcal{I} := [0, 1]$.

When at least one solution passes through an initial value, we call the initial value consistent, cf. [LMT13]. The initial values in Example 2.10 are inconsistent, since no solution passes through a point with x_1 being zero. As in Example 2.8, the selection of the initial values becomes a non trivial topic in practice. Dealing with an explicit ODE we can just choose initial values for all solution components, but when we deal with DAEs this is no longer the case. It can be very difficult to find any initial value due to the size of the DAEs which appear in practice.

Not only the size of a system can be a problem when we want to calculate consistent initial values. It can also be unclear which are the conditions we have to fulfill to obtain consistent initial values. In Example 2.5 the only condition is the algebraic equation (2.5b). If we choose initial values which fulfill this algebraic equation, we already get consistent initial values for this example. Conditions arising from algebraic equations are called obvious constraints. In Example 2.6 it is not enough to fulfill the obvious constraints. Here the first equation imposes an additional condition on the initial values, because of the differentiation problem. Conditions arising from dynamical equations are called hidden constraints.

In the previous example the implicit Euler failed to calculate any numerical solution of the DAE due to the choice of the initial values. In the next example inconsistent initial values are chosen again but this time the implicit Euler provides a numerical solution.

Example 2.11. Let $\mathcal{I} := [0,1]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} x_2' &= y \\ x_1' &= y \\ x_1 &= 1 \end{aligned}$$

The exact solution of this example is given by $x_1(t) = 1$, $x_2(t) = x_2^0 \in \mathbb{R}$, $y(t) = 0$. We observe again what happens if we choose $x_1^0 = 0$, $x_2^0 = c \in \mathbb{R}$, $y^0 = 0$ and discretize the example with the implicit Euler scheme. With the first Euler step we obtain

$$\begin{aligned} x_{2,1} &= c + hy_1 \\ y_1 &= \frac{1}{h} \\ x_{1,1} &= 1. \end{aligned}$$

While for a general $n \geqslant 1$ we get

$$\begin{aligned} \frac{x_{2,n} - x_{2,n-1}}{h} &= y_n \\ \frac{x_{1,n} - x_{1,n-1}}{h} &= y_n \\ x_{1,n} &= 1 \end{aligned}$$

which leads to

$$x_{2,n} = x_{2,n-1} = c + 1$$

$$\begin{aligned} y_n &= 0 \\ x_{1,n} &= 1 \end{aligned}$$

for $n \geqslant 2$. In this case the Euler indeed calculates a numerical solution but this solution does not converge against the exact solution in the x_2 component regarding the initial value. The inconsistent choice of the initial values altered the trajectory of the x_2 component independently of h. Therefore Example 2.11 is even more vicious than Example 2.10 because this time the problem is not obvious.

Explicit Methods

Another challenge in the field of DAE numerics is the usage of explicit methods, cf. [ASW93, BH93, Arn98, Mur97, Ost93]. Even in Example 2.5 an explicit method cannot be used without extra efforts. Using the explicit Euler scheme we obtain

$$\begin{aligned} \frac{x_n - x_{n-1}}{h} &= y_{n-1} \\ y_{n-1} &= f(t_{n-1}). \end{aligned}$$

Since y_n does not appear in these equations, obviously it is not possible to solve these equations with respect to x_n and y_n. Of course, in this case we could just discretize the differential equation explicitly while we treat the algebraic one implicitly, i.e. we evaluate it in t_n, and obtain

$$\begin{aligned} \frac{x_n - x_{n-1}}{h} &= y_{n-1} \\ y_n &= f(t_n). \end{aligned}$$

But this approach does not work for Example 2.6. Even if we evaluate the algebraic equation in t_n, the variable y_n still does not appear in the equations.

$$\begin{aligned} \frac{x_n - x_{n-1}}{h} &= y_{n-1} \\ x_n &= f(t_n). \end{aligned}$$

Variable Time Step Size

Not only explicit methods are more difficult to use, also non constant time step size becomes harder to apply, cf. [LMT13]. If we consider Example 2.7 we are dealing with a second order differentiation problem. The problem is that even analytically the difference quotient for the second derivative converges only for a constant time step size regardless of the rounding error. So if there is at least a second order differentiation problem involved in a DAE the usage of numerical solvers with a variable time step size is not trivial. We apply the implicit Euler method on the equation of Example 2.7

$$\begin{aligned} x_2' &= y \\ x_1' &= x_2 \\ x_1 &= f(t) \end{aligned}$$

and obtain the time discretized version

$$\begin{aligned} \frac{x_{2,n} - x_{2,n-1}}{h_n} &= y_n \\ \frac{x_{1,n} - x_{1,n-1}}{h_n} &= x_{2,n} \\ x_{1,n} &= f(t_n). \end{aligned}$$

After reorganization we achieve an explicit description for the numerical solution

$$\begin{aligned} x_{1,n} &= f(t_n) \\ x_{2,n} &= \frac{f(t_n) - f(t_{n-1})}{h_n} \\ y_n &= \frac{\frac{f(t_n)-f(t_{n-1})}{h_n} - \frac{f(t_{n-1})-f(t_{n-2})}{h_{n-1}}}{h_n}. \end{aligned}$$

We define $h = \max(h_{n-1}, h_n)$, then there is a $c > 0$ such that we can write

$$y_n = \frac{h_n + h_{n-1}}{2h_n} f''(t_n) + O(h)$$

with the help of a Taylor series as long as $\frac{h_n}{h_{n-1}} \leqslant c$ and $\frac{h_{n-1}}{h_n} \leqslant c$. That directly tells us that the difference

$$y(t_n) - y_n = \frac{h_n - h_{n-1}}{2h_n} f''(t_n) + O(h)$$

will converge against zero for any function $f \in C^2(\mathbb{R}, \mathbb{R})$ if and only if $h_n = h_{n-1}$.

Convergence Problems for Classical ODE Methods

Next we consider an example from [GP83] in standard formulation.

Example 2.12. Let $\mathcal{I} := [0, 3]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} x_1' + \eta t x_2' &= -(1+\eta) x_2 \\ x_1 + \eta t x_2 &= e^{-t} \end{aligned}$$

The main difference to the previous example is the time dependency of the coefficients. This example is of huge numerical significance because the implicit Euler fails to solve it if we choose $\eta < -\frac{1}{2}$. This holds even if we use a constant step size and choose consistent initial values.

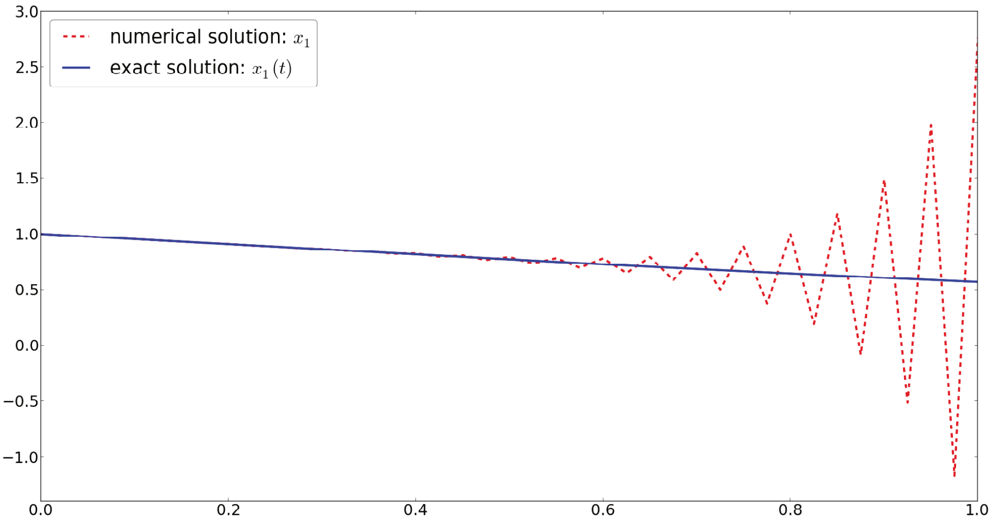


Figure 2.4: Numerical and exact solutions of the η-DAE with $\eta = -0.55$ using the implicit Euler.

Hence we can no longer rely on classical methods like the implicit Euler in general. An other alarming behavior of the η-DAE is, that for $\eta \geqslant -\frac{1}{2}$ the example is very easy to solve with the implicit Euler. This shows that the numerical complexity may depend on parameters like η.

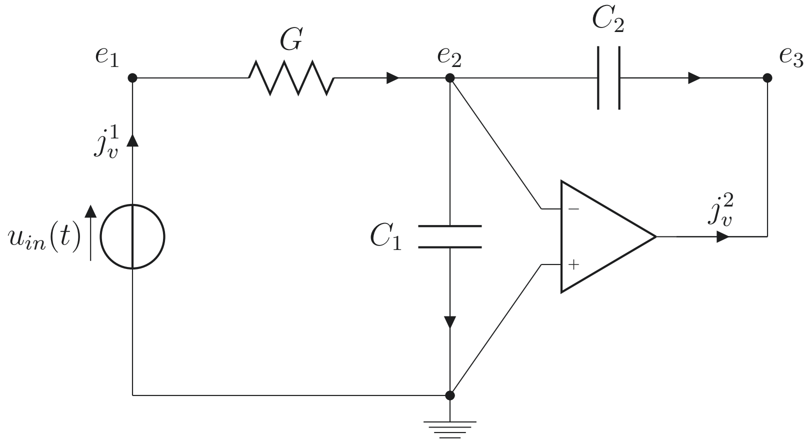
Asymptotic Instability

The next example is an electric circuit, which is called the Miller Integrator, cf. [MG05, Pul12]. The corresponding equations are given by:

Example 2.13. Let $\mathcal{I} := [0, 2 \cdot 10^{-6}]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} G(e_1 - e_2) + j_v^1 &= 0 \\ (C_1 + C_2)e_2' - C_2 e_3' - G(e_1 - e_2) &= 0 \\ C_2(e_3 - e_2)' + j_v^2 &= 0 \\ e_1 &= u_{in}(t) \\ e_3 - 2e_2 &= 0, \end{aligned}$$

where the e_i are the electrical potentials at the nodes and the j_v^i are the currents over the voltage source and the operational amplifier. Further $u_{in}(t)$ is the voltage of the voltage source which we set to $u_{in}(t) = sin(2\pi 10^6 t)$.



We transform and factorize the equations of Example 2.13 to determine the order of the differentiation problem in the example.

$$j_v^1 = G(e_2 - u_{in}(t)) \tag{2.8a}$$

$$e_1 = u_{in}(t) \tag{2.8b}$$

$$e_3 - 2e_2 = 0 \tag{2.8c}$$

$$-\frac{C_1}{G}(e_3 - 2e_2)' + \frac{C_2 - C_1}{GC_2} j_v^2 + e_2 = u_{in}(t) \tag{2.8d}$$

$$C_2(e_3 - 2e_2)' + C_2 e_2' + j_v^2 = 0, \tag{2.8e}$$

Equation (2.8d) yields a description of j_v^2 depending on the first derivative of $e_3 - 2e_2$

$$j_v^2 = \frac{GC_2}{C_2 - C_1}(u_{in}(t) - e_2 + \frac{C_1}{G}(e_3 - 2e_2)')$$

if $C_2 \neq C_1$. But for $C_2 = C_1$ we obtain

$$\begin{aligned} e_2 &= u_{in}(t) + \frac{C_1}{G}(e_3 - 2e_2)' \\ C_2(e_3 - 2e_2)' + C_2 e_2' + j_v^2 &= 0, \end{aligned}$$

which yields

$$j_v^2 = -C_2(e_3 - 2e_2)' - C_2(u_{in}'(t) + \frac{C_1}{G}(e_3 - 2e_2)'').$$

Therefore this problem contains a second order differentiation problem for $C_2 = C_1$. But for every other case, i.e. $C_2 \neq C_1$, it contains a first order differentiation problem. We choose $G = \frac{1}{1k\Omega}$ and $C_1 = 0.01\mu F$, then only for $C_2 = 0.01\mu F$ there is an order two differentiation problem within these equations. If we change C_2 slightly to $C_2 = 0.0105\mu F$, there is an order one differentiation problem in the equations. We see that the lower order differentiation problem gives us much more trouble if we solve this circuit for the two different sets of parameters, see Figure 2.5.

While the second order differentiation problem gives us a stable solution, even for relatively large time step sizes, the first order differentiation problem always drifts off regardless of how small we choose the time step size.

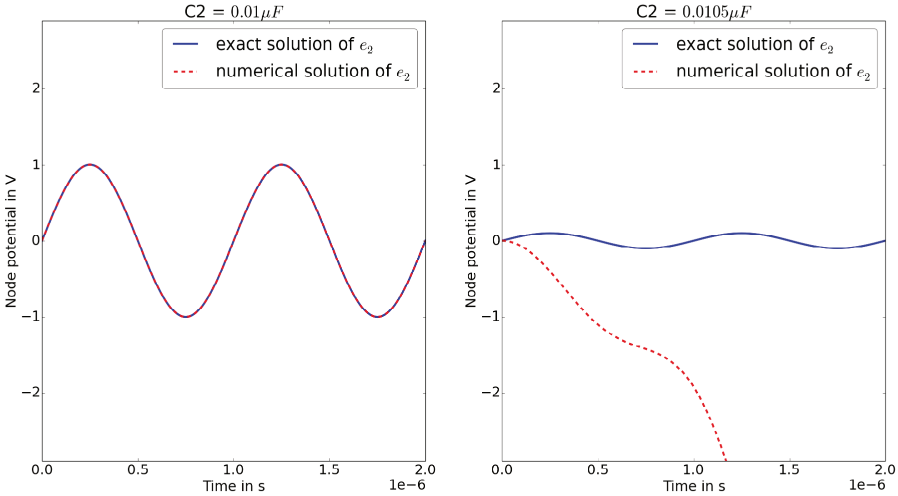


Figure 2.5: Numerical stability issues of the the Miller Integrator

To understand this behavior we examine the equations, which describe the potential e_2 for the different parameter sets. For $C_2 = C_1$ it holds

$$e_2 = u_{in}(t)$$

while we obtain

$$e_2' = \frac{G}{C_2 - C_1} e_2 - \frac{G}{C_2 - C_1} u_{in}(t) \tag{2.9}$$

for $C_2 \neq C_1$. The solutions of the homogeneous version of (2.9) are

$$e_{2,h}(t) = ce^{\frac{G}{C_2 - C_1} t}$$

with c depending on the initial condition. For C_2 slightly larger than C_1 the function $e_{2,h}(t)$ grows extremely fast. This instability in the homogeneous solutions is the reason for the drift off in Example 2.13.

The same problem can occur if we switch an algebraic equation with its derivative. We consider the following equation

$$e^{\lambda t}x = e^{(\lambda-1)t} \tag{2.10}$$

Its solution is given by

$$x(t) = e^{-t}$$

and if we solve this numerically with $\lambda = -15$ in $\mathcal{I} = [0, 2]$ we obtain a numerical solution which is similar to the exact solution.

If we want to express the same problem with a dynamical equation instead of an algebraic one, we differentiate (2.10) and obtain:

$$\lambda e^{\lambda t}x + e^{\lambda t}x' = (\lambda - 1)e^{(\lambda-1)t} \tag{2.11}$$

or even more simple $x' = -\lambda x + (\lambda - 1)e^{-t}$. This ODE is solved by $x(t) = e^{-t} + ce^{-\lambda t}$ depending on the initial value x^0 at $t_0 = 0$.

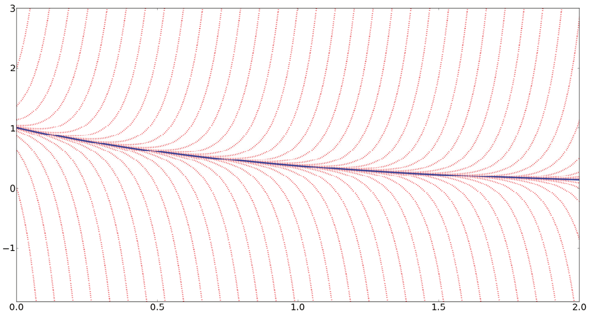


Figure 2.6: Solution trajectories of the dynamical equation

If we choose $x^0 = 1$, then the ODE (2.11) is solved by $x_\star(t) = e^{-t}$ just like the algebraic equation. For $\lambda = -15$ and $\mathcal{I} = [0, 2]$ we obtain an unstable solution again.

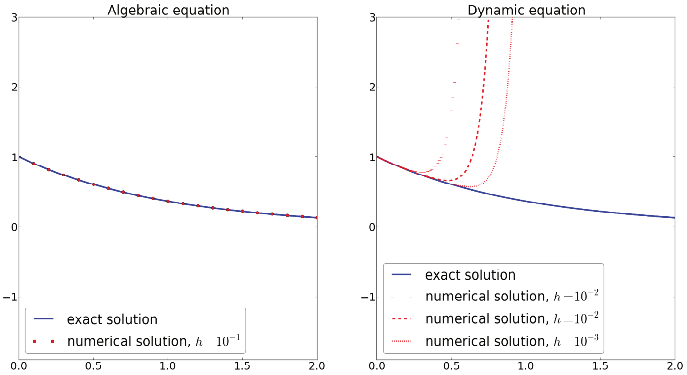


Figure 2.7: Stability behavior of the algebraic and the dynamical equation.

The reasons for this drift off phenomenon are again the unstable solution trajectories of the ODE (2.11). If we solve (2.11) by the implicit Euler or any other numerical method, we make a small error in every time step. In this particular case this small error is the reason why the numerical solution leaves the stable solution trajectory. Once on this unstable solution trajectory, the numerical solution grows unbounded.
As already mentioned, the differentiation problems involved in the DAEs motivate a classification. In the following we introduce three of the most popular classification concepts: The Differentiation Index, the Strangeness Index and the Tractability Index.

2.2 Differentiation Index

In this section we introduce the Differentiation Index for nonlinear DAEs in standard form. The Differentiation Index is probably the best known index, since its concept is relatively demonstrative. It was introduced by Petzold and Campbell, see [Cam87, BCP96]. To define the Differentiation Index we need the DAE (2.3) itself but also its derivatives. For a compact notation we define the inflated system.

Definition 2.14. (Inflated system - [KM06], p.153)
Considering the DAE (2.3) we gather the original equation and its derivatives up to order $\nu \in \mathbb{N}_0$ into an inflated system

$$F_\nu(x^{(\nu+1)}(t), ..., x'(t), x(t), t) = 0, \tag{2.12}$$

where F_ν has the form

$$F_\nu(x^{\nu+1}, ..., x^1, x, t) = \begin{bmatrix} F(x^1, x, t) \\ \frac{\partial}{\partial x^1} F(x^1, x, t)x^2 + \frac{\partial}{\partial x} F(x^1, x, t)x^1 + \frac{\partial}{\partial t} F(x^1, x, t) \\ \vdots \end{bmatrix}$$

and define the Jacobians

$$\begin{aligned}
G(x^1, x, t) &= \frac{\partial}{\partial x^1} F(x^1, x, t) \\
B(x^1, x, t) &= \frac{\partial}{\partial x} F(x^1, x, t) \\
G_\nu(x^{\nu+1}, ..., x^1, x, t) &= \frac{\partial}{\partial(x^1, \dots, x^{\nu+1})} F_\nu(x^{(\nu+1)}, ..., x^1, x, t) \\
B_\nu(x^{\nu+1}, ..., x^1, x, t) &= \frac{\partial}{\partial x} F_\nu(x^{\nu+1}, ..., x^1, x, t).
\end{aligned}$$

Over the years the definition of the Differentiation Index has been slightly modified to adjust from the linear to the nonlinear case [Cam87, CG95b, CG95a] and to deal with slightly different smoothness assumptions. We concentrate only on the nonlinear case and define the Differentiation Index with the help of the inflated system.

Definition 2.15. (Differentiation Index)
The DAE (2.3) has Differentiation Index μ, if and only if $F \in C^\mu(\mathcal{D}_{x'} \times \mathcal{D}_x \times \mathcal{I}, \mathbb{R}^n)$ and μ is the minimal number such that an explicit ODE $x' = f(x, t)$ can be extracted from $F_\mu(x^{(\mu+1)}, ..., x', x, t) = 0$ by algebraic manipulations only with f being continuous.

In a way the Differentiation Index measures the difference of a DAE to an explicit ODE by counting the number of differentiations needed for the transformation to an explicit ODE. Previously we talked about differentiation problems within a DAE. One could also say that the Differentiation Index tries to count the order of these differentiation problems. Lets take a look at a small example to get to know the inflated system and the Differentiation Index.

Example 2.16.
For $t \in [0, 10]$ consider the DAE

$$\begin{aligned}
(x_0 + x_1)' &= x_1 \\
x_0 + x_1 &= sin(t).
\end{aligned}$$

If we differentiate the system two times we get the inflated system

$$\begin{aligned}
(x_0 + x_1)' &= x_1 \\
x_0 + x_1 &= sin(t) \\
(x_0 + x_1)'' &= x_1' \\
(x_0 + x_1)' &= cos(t) \\
(x_0 + x_1)''' &= x_1'' \\
(x_0 + x_1)'' &= -sin(t).
\end{aligned}$$

With these equations we achieve by algebraic manipulations

$$\begin{aligned} x_0' &= x_1 + sin(t) \\ x_1' &= -sin(t). \end{aligned}$$

and therefore the DAE has at most Differentiation Index index 2. To guarantee that the index is not smaller than 2, we would have to check that it is not possible to create an explicit ODE with just one or none differentiation.

The major advantage of the Differentiation Index is its simple and demonstrative concept. In return the Differentiation Index concept has two major drawbacks.
First it is hard to determine whether or not the used number of differentiations is the minimal one to obtain an explicit ODE. Although it may be easy to calculate an upper bound of the Differentiation Index, to assure that the used number of differentiations is the minimal one we would have to check all smaller cases. This may become difficult and time-consuming.
And secondly the Differentiation Index needs more smoothness than necessary as we see in the next example.

Example 2.17.
For $t \in [-1, 1]$ consider the DAE

$$\begin{aligned} x_0' - x_1 &= 0 \\ x_0 + x_2 &= sin(t) + |t| \\ x_2 &= |t|. \end{aligned}$$

The Differentiation Index is at least one, since the derivatives of x_1 and x_2 do not appear in the equations. So we need to differentiate the equations at least once, which is not possible since $|t|$ is only continuous.

2.3 Strangeness Index

In this section we introduce the concept of the Strangeness Index for nonlinear DAEs in standard formulation. The Strangeness Index was first established by Kunkel and Mehrmann, see [KM06]. The Strangeness Index can be considered as a generalization of the Differentiation Index. It also uses the inflated system to analyze the structure of a DAE. The following definition of the Strangeness Index seems to be more technical than the definition of the Differentiation Index, but if we look closely they are strongly related.

Definition 2.18. (Strangeness Index - [KM06], Hypothesis 4.2.)
Given a DAE as in (2.3) with F being a μ-times continuously differentiable function, the smallest value of μ such that the following requirements are met, is called the Strangeness Index of the DAE (2.3). There are integers a and d such that the set

$$L_\mu = \{(x^{\mu+1}, \ldots, x^1, x, t) \in \mathbb{R}^{(\mu+2)n+1} | F_\mu(x^{\mu+1}, \ldots, x^1, x, t) = 0\}$$

associated with F is non-empty and such that for every $(x_0^{\mu+1}, \ldots, x_0^1, x_0, t_0) \in L_\mu$, there exists a neighborhood in which the following properties hold:

1. We have $\operatorname{rk} G_\mu(x^{\mu+1}, \ldots, x^1, x, t) = (\mu + 1)n - a$ on L_μ, such that there exists a smooth matrix function W of size $(\mu + 1)n \times a$ with pointwise maximal rank, satisfying $W^\top G_\mu = 0$ on L_μ.

2. We have $\operatorname{rk} W^\top(x^{\mu+1}, \ldots, x^1, x, t)B_\mu(x^{\mu+1}, \ldots, x^1, x, t) = a$, such that there exists a smooth matrix function Q of size $n \times d$ with $d = n - a$ and pointwise maximal rank, satisfying $W^\top B_\mu Q = 0$ on L_μ.

3. We have $\operatorname{rk} G(x^1, x, t)Q(x^{\mu+1}, \ldots, x^1, x, t) = d$, such that there exists a smooth matrix function V of size $n \times d$ and pointwise maximal rank, satisfying $\operatorname{rk} V^\top GQ = d$ on L_μ.

One important application of the Strangeness Index is the field of multibody-systems or mechanical systems. Because of the physical knowledge it is often known in advance that only a part of the equations will be used in the inflated system. And these parts are indeed as smooth as we need them to be. So for big application fields we could say that the high smoothness requirements, i.e. F being μ-times continuously differentiable, are only a technical problem, but a problem nevertheless.
The Strangeness Index does not aim for a transformation into an explicit ODE. It is rather based on an approach to provide an implicit ODE for a part of the variables and a set of algebraic equations for the rest of the variables. Therefore the Strangeness Index of a DAE can be expected to be one lower than the Differentiation Index, if they are both well defined. We want to make use of the concept of the Strangeness Index later and so we need to amplify it a little more. For this purpose we consider a linear DAE in standard form with constant coefficients and Strangeness Index μ. Such a DAE can be described by

$$F(x'(t), x(t), t) = Gx'(t) + Bx(t) - q(t)$$

with matrices G, B and a time depending function q. The inflated system of such a linear DAE inherits its linear form and can be written as

$$G_\mu \left(x'(t), ..., x^{(\mu+1)}(t)\right)^\top + B_\mu x(t) = \left(q(t), ..., q^{(\mu)}(t)\right)^\top := q_\mu(t).$$

Multiplying the transposed matrix function $W^\top$ of the first point of the definition of the Strangeness Index from the left provides an algebraic equation

$$W^\top B_\mu x(t) = W^\top q_\mu(t). \tag{2.13}$$

Since $W^\top B_\mu$ has a kernel of the dimension d, which is described by the matrix function Q, the equation (2.13) does not take care of all parts of the solution function x. Especially the x parts that lie in the image of Q are not governed by (2.13). But in the third part of the definition we get the last matrix function V such that the derivatives of the missing parts of x can be found in

$$V^\top G x'(t) + V^\top B x(t) = V^\top q(t),$$

since $\operatorname{rk} V^\top G Q = d$.

Consider again Example 2.16 with $t \in [0, 10]$ and

$$\begin{aligned}(x_0 + x_1)' &= x_1\\ x_0 + x_1 &= sin(t).\end{aligned}$$

Example 2.16 has Differentiation Index 2 hence we expect it to have Strangeness Index 1. For $\nu = 0$ the Jacobians of the inflated system are given by

$$G_0 = \begin{pmatrix}1 & 1\\ 0 & 0\end{pmatrix}, \quad B_0 = \begin{pmatrix}0 & -1\\ 1 & 1\end{pmatrix}.$$

This leads to the matrix functions

$$W = \begin{pmatrix}0\\ 1\end{pmatrix}, \; V = \begin{pmatrix}1\\ 0\end{pmatrix} \text{ and } W^\top B_0 = \begin{pmatrix}1 & 1\end{pmatrix}$$

and therefore we get

$$Q = \begin{pmatrix}1\\ -1\end{pmatrix} \text{ and } V^\top G_0 Q = \begin{pmatrix}0\end{pmatrix}.$$

With $\operatorname{rk} V^\top G_0 Q = 0 \neq 1 = n - \operatorname{rk} W^\top B_1$ the DAE does not have Strangeness Index 0. For $\nu = 1$ the Jacobians of the inflated system are given by

$$G_1 = \begin{pmatrix}1 & 1 & 0 & 0\\ 0 & 0 & 0 & 0\\ 0 & -1 & 1 & 1\\ 1 & 1 & 0 & 0\end{pmatrix}, \quad B_1 = \begin{pmatrix}0 & -1\\ 1 & 1\\ 0 & 0\\ 0 & 0\end{pmatrix}.$$

This leads to a matrix function

$$W = \begin{pmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix} \text{ and } W^\top B = \begin{pmatrix} 1 & 1 \\ 0 & -1 \end{pmatrix}$$

With $W^\top B$ non-singular it follows directly that the DAE has Strangeness Index 1. The Strangeness Index of Example 2.16 turns out as expected.
Again we need to prove that there is no smaller integer which fulfills the index conditions. Furthermore the Strangeness Index requires that F is μ-times continuously differentiable. Example 2.17 provides a problem again, because we cannot differentiate the equations. If we try to calculate the matrix functions of the Strangeness Index for the equations

$$\begin{aligned} x_0' - x_1 &= 0 \\ x_0 + x_2 &= sin(t) + |t| \\ x_2 &= |t|, \end{aligned}$$

we obviously start with $W^\top = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}^T$, which leaves us with the last two equations of Example 2.17

$$\begin{aligned} x_0 + x_2 &= sin(t) + |t| \\ x_2 &= |t|. \end{aligned}$$

We can choose $Q^\top = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^T$. By depicting Example 2.17 in standard form we get

$$B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } BQ = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and therefore $\operatorname{rk} Q > \operatorname{rk} BQ$. The third condition of the Strangeness Index is breached. So Example 2.17 does not have Strangeness Index 1 and can not be differentiated. Hence there is no uniform Strangeness Index on the whole time interval $[-1, 1]$.
The systematic decomposition of a DAE by the Strangeness Index is only implied here by the properties in Definition 2.18. Later on we will take a closer look at it on the basis of linear DAEs. The idea of the Strangeness Index is extremely powerful but its definition via the inflated system makes it hard to tap its full potential.

2.4 Tractability Index

Next we introduce the concept of the Tractability Index. It has been mainly developed by März, cf. [GM86, Mär02, LMT13]. The main features of the Tractability Index are its minimal smoothness requirements and its step by step decoupling strategy for DAEs. First the Tractability Index concept was formulated for DAEs in standard form, cf. [GM86]. But now it uses a more general class of DAEs.

Definition 2.19. (DAEs with nonlinear derivative term, [LMT13])
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D} \subset \mathbb{R}^n$ be open subsets. Let $f \in C(\mathbb{R}^m \times \mathcal{D} \times \mathcal{I}, \mathbb{R}^n)$ be continuous such that the partial derivatives $\frac{\partial}{\partial y} f(y,x,t)$ and $\frac{\partial}{\partial x} f(y,x,t)$ are also continuous with $\frac{\partial}{\partial y} f(y,x,t)$ being singular for all triples $(y,x,t) \in \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}$. We call

$$f(d'(x(t),t), x(t), t) = 0, \quad x(t_0) = x^0 \tag{2.14}$$

with $d \in C^1(\mathcal{D} \times \mathcal{I}, \mathbb{R}^m)$ a DAE with nonlinear derivative term. Let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$. We call $x_\star \in C(\mathcal{I}_\star, \mathcal{D})$ with $d(x_\star(.),.) \in C^1(\mathcal{I}_\star, \mathbb{R}^m)$ a solution of (2.14) on $\mathcal{I}_\star$ if the initial conditions are fulfilled, i.e. $x_\star(t_0) = x^0$, and

$$f(d'(x_\star(t),t), x_\star(t), t) = 0 \quad \forall t \in \mathcal{I}_\star.$$

To avoid unnecessary gaps and overlaps between im $\frac{\partial}{\partial x} d$ and ker $\frac{\partial}{\partial y} f$ the nonlinear derivative term needs to be chosen reasonably.

Definition 2.20. (Properly stated derivative term)
The DAE (2.14) has a properly stated derivative term on $\mathcal{D} \times \mathcal{I}$, if im $\frac{\partial}{\partial x} d$ and ker $\frac{\partial}{\partial y} f$ are C^1-subspaces in $\mathbb{R}^m$, and the transversality condition

$$\ker \frac{\partial}{\partial y} f(y,x,t) \oplus \operatorname{im} \frac{\partial}{\partial x} d(x,t) = \mathbb{R}^m, \quad \forall (y,x,t) \in \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}, \tag{2.15}$$

holds.

The concept of the Tractability Index does not use the inflated system, which is also called derivative array. The independence of the Tractability Index of the derivative array is discussed in [Mär98]. Hence we need another strategy to handle nonlinear DAEs (2.14). This is done by composing the linear Taylor polynomial of the DAE in the x-argument around a reference function x_*. Before we define the linearization of a nonlinear DAE we need to define a suitable set of reference functions.

Definition 2.21. (Reference function set)
Let $\mathcal{G} \subseteq \mathcal{D} \times \mathcal{I}$ be open and let $\nu \in \mathbb{N}$. We denote by $C_*^\nu(\mathcal{G})$ the set of all $C^{\max(2,\nu)}$ functions with a graph in $\mathcal{G}$. That means, that $x_* \in C_*^\nu(\mathcal{G})$ if and only if $x_* \in C^{\max(2,\nu)}(\mathcal{I}_*, \mathbb{R}^m)$ with $(x_*(t), t) \in \mathcal{G}$ for all $t \in \mathcal{I}_* \subset \mathcal{I}$.

The ν has to be sufficiently large since we do not know in advance which part of the reference functions belongs to a differentiation problem. Next we define the linearization of a nonlinear DAE associated to a reference function.

Definition 2.22. (Linearization)
Consider a nonlinear DAE (2.14) and an integer $\nu \in \mathbb{N}$. Let $x_* \in C_*^\nu(\mathcal{G})$ be a reference function with $\mathcal{G} \subseteq \mathcal{D} \times \mathcal{I}$. We call the linear DAE

$$A_*(t)(D_*(t)x(t))' + B_*(t)x(t) = q_*(t), \quad t \in \mathcal{I}_*, \tag{2.16}$$

with the coefficients

$$\begin{aligned} D_*(t) &:= d_x(x_*(t), t), \\ A_*(t) &:= \frac{\partial}{\partial y} f(d'(x_*(t), t), x_*(t), t), \\ B_*(t) &:= \frac{\partial}{\partial x} f(d'(x_*(t), t), x_*(t), t), \\ q_*(t) &:= -f(d'(x_*(t), t), x_*(t), t), \quad t \in \mathcal{I}_* \end{aligned}$$

the linearization of the nonlinear DAE (2.14) along the reference function x_* with a right hand side $q_*(t)$. Here $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial x}$ denote the partial derivatives with respect to the first and second argument of f.

Since the Tractability Index shall be independent of the choice of the reference function we can not simply pick one special reference function for the linearization. We need a more general approach and define the following matrix functions.

Definition 2.23. (Placeholder matrix functions)
Consider a nonlinear DAE (2.14) and define the continuous matrix functions A, D and B by

$$D(x, t) := \frac{\partial}{\partial x} d(x, t) \tag{2.17a}$$

$$A(x^1, x, t) := \frac{\partial}{\partial y} f(D(x, t)x^1 + d_t(x, t), x, t), \tag{2.17b}$$

$$B(x^1, x, t) := \frac{\partial}{\partial x} f(D(x, t)x^1 + d_t(x, t), x, t), \tag{2.17c}$$

for $x^1 \in \mathbb{R}^n, x \in \mathcal{D}, t \in \mathcal{I}$. Here $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial x}$ again denote the partial derivatives with respect to the first and second argument of f.

Similar to the properly stated derivative term we need these matrix functions to match with each other in the following sense.

Definition 2.24. (Regular matrix pencil)
Let be $A, B \in \mathbb{R}^{n \times n}$. The ordered matrix pair $\{A, B\}$ and the matrix pencil $\lambda A + B$ respectively are called non-singular or regular if there is a constant $\lambda \in \mathbb{R}$ so that $\det\big(\lambda A + B\big) \not\equiv 0$. Otherwise they are called singular.

We say a linear DAE (2.16) has a regular matrix pencil if $\{A_*(t)D_*(t), B_*(t)\}$ is a regular matrix pencil for all $t \in \mathcal{I}_*$ and a nonlinear DAE (2.14) has a regular matrix pencil if $\{A(x^1, x, t)D(x, t), B(x^1, x, t)\}$ is a regular matrix pencil for all $x^1 \in \mathbb{R}^m, x \in \mathcal{D}, t \in \mathcal{I}$. Now we can sum up some basic assumptions for our DAEs.

Assumption 2.25. (Basic assumptions)

1. The DAE (2.14) has a properly stated derivative term.
2. If $\ker \frac{\partial}{\partial y} f(y, x, t)$ depends on y, then d is supposed to be in $C^2(\mathcal{D} \times \mathcal{I}, \mathbb{R}^m)$.
3. The DAE has a regular matrix pencil.

With these assumptions the proper formulation of the derivative term passes down to the placeholder matrix functions.

Lemma 2.26. (Placeholder matrix functions are proper formulated)
Consider a DAE (2.14) under the Assumptions 2.25, then the decomposition

$$\ker A(x^1, x, t) \oplus \operatorname{im} D(x, t) = \mathbb{R}^n, \quad \forall (x^1, x, t) \in \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}, \tag{2.18}$$

is true, and the subspaces $\ker A$ and $\operatorname{im} D$ are C^1-subspaces in $\mathbb{R}^n$. Therefore all linearizations of (2.14) have a properly stated derivative term.

The proof can be found in [LMT13, pp. 210–212].
As a last preparation define Q_0 as the projector function onto $\ker D(x, t)$ on $\mathcal{D} \times \mathcal{I}$ and set $P_0 := I - Q_0$. We call P_0 and Q_0 admissible if and only if they are continuous. The projector valued function R defined by

$$\begin{aligned} \operatorname{im} R(x^1, x, t) &= \operatorname{im} D(x, t), \\ \ker R(x^1, x, t) &= \ker A(x^1, x, t), \end{aligned}$$

for $x^1 \in \mathbb{R}^m, x \in \mathcal{D}, t \in \mathcal{I}$, is named border projector of the DAE. Furthermore we define the generalized inverse $D(x^1, x, t)^- \in (\mathbb{R}^n, \mathbb{R}^m)$ by the means of the four conditions

$$D(x,t)D(x^1,x,t)^-D(x,t) = D(x,t) \tag{2.19a}$$
$$D(x^1,x,t)^-D(x,t)D(x^1,x,t)^- = D(x^1,x,t)^- \tag{2.19b}$$
$$D(x,t)D(x^1,x,t)^- = R(x^1,x,t) \tag{2.19c}$$
$$D(x^1,x,t)^-D(x,t) = P_0(x,t). \tag{2.19d}$$

With these preparations we can define a chain of projectors, which will allow us to define the Tractability Index.

Definition 2.27. (Projector chain - Tractability Index)
Let the DAE (2.14) satisfy the basic assumptions 2.25. Let $\mathcal{G} \subset \mathcal{D} \times \mathcal{I}$ be open and connected. Let the projector function Q_0 onto $\ker D$ be continuous on $\mathcal{G}$, $P_0 = I - Q_0$. Let D^- be the generalized inverse of D defined by (2.19). For the given level $\nu \in \mathbb{N}$, we call the sequence $G_0, ..., G_\nu$ an admissible matrix function sequence associated to the DAE on the set $\mathcal{G}$, if it is built by the rule

$$\begin{aligned}
G_0 &:= AD, B_0 := B, N_0 := \ker G_0 \\
G_i &:= G_{i-1} + B_{i-1}Q_{i-1} \\
B_i &:= B_{i-1}P_{i-1} - G_i D^-(D\Pi_i D^-)' D\Pi_{i-1} \\
N_i &:= \ker G_i
\end{aligned}$$

Set $\widehat{N}_i := (N_0 + ... + N_{i-1}) \cap N_i$ and choose X_i such that $\widehat{N}_i \oplus X_i := N_0 + ... + N_{i-1}$. Then Q_i is a projector with $im\ Q_i := N_i$ and $X_i \subset \ker Q_i$ and P_i is the complementary projector again, i.e. $P_i := I - Q_i$. At last set $\Pi_i := \Pi_{i-1} P_i$.

Additionally we assume that,

1. the matrix function G_i has constant rank r_i on $\mathbb{R}^{i \cdot m} \times \mathcal{G}, i = 0, ..., \nu$,
2. the intersection $\widehat{N}_i$ has constant dimension $u_i := dim \widehat{N}_i$,
3. the product function Π_i is continuous and $D\Pi_i D^-$ is continuously differentiable on $\mathbb{R}^{i \cdot m} \times \mathcal{G}$ for $i = 0, ..., \nu$.

With the help of the rank values r_i we can define the Tractability Index.

Definition 2.28. (Tractability Index)
Let the DAE (2.14) satisfy the basic assumptions (2.25) and let $\mathcal{G} \subseteq \mathcal{D} \times \mathcal{I}$ be an open connected set. Then the DAE (2.14) is said to be

1. regular with Tractability Index-μ on $\mathcal{G}$, if on $\mathcal{G}$ an admissible matrix function sequence exists such that $r_{\mu-1} < r_\mu = m$,
2. regular on $\mathcal{G}$, if it is regular on $\mathcal{G}$ with any index.

The constants $0 \leqslant r_0 \leqslant ... \leqslant r_{\mu-1} < r_\mu$ are named characteristic values of the regular DAE. The open connected subset $\mathcal{G}$ is called a regularity region or regularity domain.

Consider again Example 2.17:

$$\begin{aligned}
x_0' - x_1 &= 0 \\
x_0 + x_2 &= sin(t) + |t| \\
x_2 &= |t|,
\end{aligned}$$

With

$$A = D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } B = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

we can write Example 2.17 as a DAE with a properly stated derivative term

$$A(Dx)' + Bx = q(t)$$

with $q(t) := (0, sin(t) + |t|, |t|)^T$. To determine the Tractability Index we have to calculate the matrix chain starting with

$$G_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } B_0 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Next we choose the projectors

$$Q_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

with $im\ Q_0 = N_0 = \text{span}\left(\begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^\top, \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}^\top\right)$. Following the construction rules, we obtain

$$G_1 = \begin{pmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \text{ and } B_1 = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Now we have to choose the next pair of projectors, but this time we need to pay attention to the admissibility conditions. We get $N_1 = \text{span}\left(\begin{pmatrix} 1 & 1 & 0 \end{pmatrix}^\top\right)$ and therefore we may choose

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

Q_1 is an admissible choice, since $N_0 = ker\ Q_1$. Finally we obtain a non-singular matrix

$$G_2 = \begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

The Tractability Index of Example 2.17 is 2. In contrast to the Differentiation Index and the Strangeness Index, the Tractability Index is well defined for Example 2.17.
Hence, the Tractability Index needs less smoothness assumptions than the Differentiation Index and the Strangeness Index. But the projector chain of the Tractability Index tends to become complex.

2.5 Summary and Outlook

In this chapter we presented some of the challenges of DAEs. Especially we pointed out that there are problems which do only occur if we are dealing with DAEs and not with explicit ODEs.

We introduced some of the most popular index concepts as tools to analyze these DAE related effects. The analysis of DAEs becomes most difficult if the projectors or the matrix functions of the respective index concept are state dependent. For example the matrix chain of the Tractability Index is state dependent for electrical circuits, cf. [ET00]. But at the same time there are topological results indicating that such complicated projectors are not necessary, see [Tis99]. We close this chapter with the following question:

Can we create an index concept based on the established concepts which provides a state independent decoupling if possible?

3 Fields of Application

After introducing parts of the known theory of Differential-Algebraic Equations, two major application fields will be discussed in this chapter.

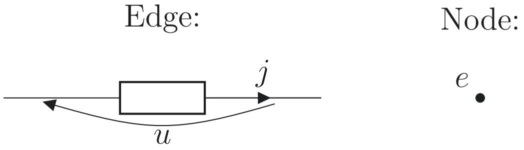
The first field of application is the investigation of electrical circuits. Electrical circuits are of great importance for industrial research and therefore a mathematical understanding is needed. There are already many works about DAE related questions regarding circuits, in particular about index analysis [Tis99, RT11] and local uniqueness and solvability [HM04, Bau12]. Besides standard elements like inductors, resistors, capacitors and source elements, electrical circuits can also contain more complex elements like semiconductor devices, memristors and electromagnetic devices. For instance semiconductor devices and electromagnetic devices are described by a set of partial differential equations, hence they involve new questions and challenges to the analysis of electrical circuits. Previous research about semiconductor devices can be found in [SBST14, Gaj93, Gaj94], these devices are widely used in circuits because of their application as transistors. In [Chu11, Ria11, RT11, Bau12] memristors are investigated while we find previous research about electromagnetic devices in [HM76, KMST93, Wei77, Bau12, BBS11, Sch11].

The field of electrical circuits simulation can be embedded in a more general network approach, cf. [JT14]. This general network approach also includes other kinds of flow networks like water, gas and blood flow networks. In Chapter 5 we will present global existence and uniqueness results and in Chapter 7 we will provide a topological decoupling for DAEs arising from electrical circuits. Similar results for gas or water networks can be found in [GJH$^+$14, JP13].

The second application field is that of mechanical multibody systems. The mechanical systems are divided into several model levels such that the models become more complex as the level of the model increases. On the highest level dynamical force elements, friction, spatial motion, contact laws and force laws as well as holonomic and nonholonomic constraints are considered. Such systems are investigated in [Ste06, ESF98, Hau89, Sim95].

3.1 Circuit Applications

We start introducing an electric circuit by understanding it as a directed graph $G := (N, E)$, with the nodes N and the arbitrarily orientated edges E. The quantities of an electric circuit are the currents j and voltages u over the edges and the electric potentials e at the nodes.



In order to obtain a well defined model in terms of uniqueness we need to choose one node as a reference node, cf. [DK84]. The potential of this reference node will be fixed, in general it can be chosen to be zero. We call this reference node the mass node. When we refer to an arbitrary node in the following we do not include the mass node. The network topology for elements with two contacts is retained by the incidence matrix $A \in \{-1, 0, 1\}^{(|N|-1)\times|E|}$. The matrix A describes the relation between all edges and all nodes except the mass node. The incidence matrix is defined by:

$$(A)_{ij} := \begin{cases} 1 & \text{, if the edge j leaves node i,} \\ -1 & \text{, if the edge j enters node i,} \\ 0 & \text{, else.} \end{cases}$$

To model the circuit with the help of the incidence matrix we use the Kirchhoff's laws, which deal with the conservation of charge and energy in electrical circuits. The two Kirchhoff's laws read:

- Kirchhoff's voltage law: At every instant of time the algebraic sum of voltages along each loop of the network is equal to zero.

- Kirchhoff's current law: At every instant of time the algebraic sum of currents entering or leaving one node of the network is equal to zero.

Let a connected electric network be given and j, u be the vectors of all edge currents and voltages and let e be the vector of all the node potentials. Then Kirchhoff's laws imply

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

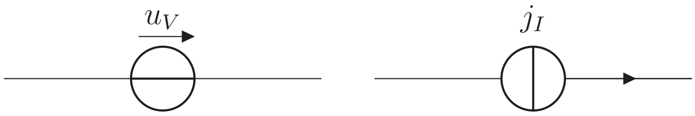
and

$$A^T e = u, \tag{3.2}$$

cf. [DK84]. It is useful to substitute node potentials for the edge voltages. This is due to the fact that the network graph usually contains more edges than nodes, hence we will obtain a smaller system size.

3.1.1 Basic Elements

The Kirchhoff's laws provide two model equations for an electrical circuit, hence we lack one last equation since we are dealing with the three electric quantities j, u and e. The last equation is provided by the different branch elements. The branch elements are described by a relation between their currents and their voltages. This holds for all branch elements with the exception of source elements, they give a direct description of their voltages or their currents. There is a number of basic branch elements which present these relation in an explicit way. We divide all these basic branch elements into four classes. We already mentioned one of these classes: the source elements. There are two basic source elements, the voltage and the current sources. Their electric symbols are:



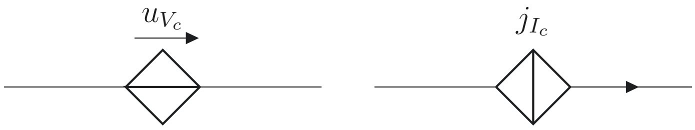
The quantities of independent sources are described by a time dependent function. Let $n_V \in \mathbb{N}$ and $n_I \in \mathbb{N}$ be the numbers of the voltage and current sources, respectively. Let $\mathcal{I} \subset \mathbb{R}$ be a compact time interval, then there are two characteristic functions $v_s : \mathcal{I} \to \mathbb{R}^{n_V}$ and $i_s : \mathcal{I} \to \mathbb{R}^{n_I}$ such that

$$u_V = v_s(t) \text{ and } j_I = i_s(t),$$

with $u_V \in \mathbb{R}^{n_V}$ the voltages over the voltage sources and $j_I \in R^{n_I}$ the currents along the current sources. If source elements are not independent they are called controlled sources. In that case their characteristic functions can depend on other quantities of the circuit, i.e. the voltages and current of the controlled sources can be described as:

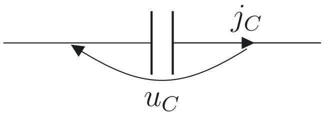
$$u_{V_c} = v_s(u, j, t) \text{ and } j_{I_c} = i_s(u, j, t).$$

The electric symbols of controlled sources are:



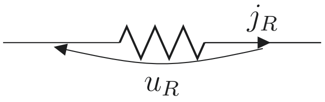
The next class of branch elements are the capacitors. Capacitors store energy in their electric field. Let $n_C \in \mathbb{N}$ be the number of the capacitors then we call $q_C : \mathbb{R}^{n_C} \times \mathcal{I} \to \mathbb{R}^{n_C}$ the characteristic function of the capacitors. The function q_C describes the electric charges of the capacitors. With the characteristic function q_C we can formulate a relation between the currents j_C and the derivative of the voltages u_C of the capacitors. Further we present the electric symbol of a capacitor:

$$j_C = \tfrac{\mathrm{d}}{\mathrm{d}t} q_C(u_C, t)$$



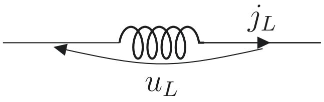
with $j_C, u_C \in \mathbb{R}^{n_C}$ the currents and voltages of the capacitors. We call all branch elements which provide a relation between their currents and the derivative of their voltages capacitor-like elements. Another class of basic branch elements are the resistors. Resistors limit the flow of their current by generating voltage drops. Let $n_R \in \mathbb{N}$ be the number of the resistors, then we call $g_R : \mathbb{R}^{n_R} \times \mathcal{I} \to \mathbb{R}^{n_R}$ the characteristic function of the resistors. The function g_R describes the conductance of the resistors. With the characteristic function g_R we can formulate a relation between the currents j_R and the voltages u_R of the resistors. Further we present the electric symbol of a resistor:

$$j_R = g_R(u_R, t)$$

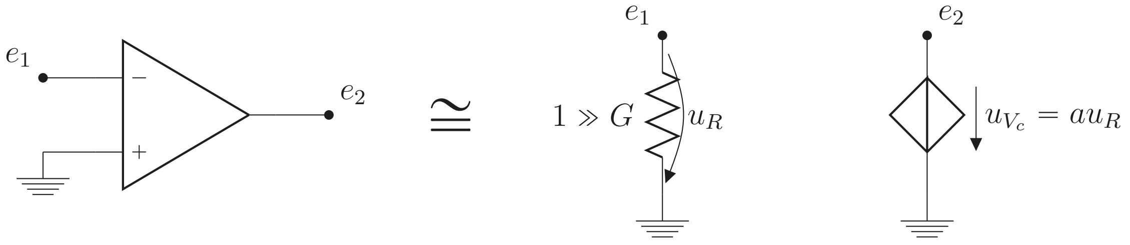


with $j_R, u_R \in \mathbb{R}^{n_R}$ the currents and voltages of the resistors. We call all branch elements which provide a relation between their currents and their voltages resistor-like elements. The last of the four basic branch elements are the inductors. Inductors store energy in their magnetic field. Let $n_L \in \mathbb{N}$ be the number of the inductors then we call $\phi_L : \mathbb{R}^{n_L} \times \mathcal{I} \to \mathbb{R}^{n_L}$ the characteristic function of the inductors. The function ϕ_L describes the magnetic flux of the inductors. With the characteristic function ϕ_L we can formulate a relation between the derivative of the currents j_L and the voltages u_L of the inductors. Further we present the electric symbol of an inductor:

$$\tfrac{\mathrm{d}}{\mathrm{d}t} \phi_L(j_L, t) = u_L$$



with $j_L, u_L \in \mathbb{R}^{n_L}$ the currents and voltages of the inductors. We call all branch elements which provide a relation between the derivative of their currents and their voltages inductor-like elements. In the next sections of this chapter we present other kinds of branch elements and classify them as capacitor-like, resistor-like or inductor-like elements. As an example for a more complex electric element we present the operational amplifier without feedback, which was already used in Example 2.13. An operational amplifier can be described with the help of one resistor element and one controlled voltage source. The purpose of an operational amplifier is to control the potential at a node such that it is equivalent to the potential at another node multiplied by a factor a.



The conductance G of the resistor has to be very small. In the ideal case it would be zero such that no current flows through the resistor, then we are able to measure the potential e_1 at the first node as the voltage drop over the resistor. With this voltage we control the voltage source such that the potential e_2 equals ae_1.

3.1.2 Semiconductor Device Model

In this section we develop a model for semiconductor devices. Many parts of the results of this section were developed together with Sascha Baumanns, Monica Selva Soto and Caren Tischendorf, cf. [SBST14]. In circuit simulation programs semiconductor devices are often described by compact models depending on hundreds of parameters, most of them without a direct physical interpretation. With the rapid development of chip technology these models became more and more complex, in particular the calibration of the parameters of these models became complicated. In order to overcome this difficulty several coupled models, i.e. models that consist of coupled differential-algebraic and partial differential equations, have been proposed for the simulation of electrical circuits over the last years, see e.g. [GS00, ABGT03]. Here we concentrate on the model originally proposed in [ABGT03] and studied further in [Tis03, ABG04, ST05, Sot06, Bod07, BST10]. It couples the DAE describing the behavior of the basic circuit elements and the circuit's topology to partial differential equations modeling the semiconductor devices in it. While in [ABGT03, Tis03, ABG04, Bod07] the properties of this model, as Partial Differential-Algebraic Equation (PDAE), are studied, in [ST05, Sot06, BST10] the DAEs that result after spatial discretization of the PDEs in the system are taken into account. In [GS00] the model considered here and other coupled models for the simulation of electrical circuits are described and some simulation results are presented.

In contrast to [ST05, Sot06, BST10], here we consider the DAEs that result if higher dimensional Drift-Diffusion (DD) equations are used for modeling the semiconductor devices in the system. With the help of auxiliary functions, originally introduced by Gajewski [Gaj93], we present a description for the current at the semiconductor contacts in such a way that current conservation is given for the continuous as well as for the discretized model. Considering the capacitive and conductive behavior of semiconductor devices, we also change the way the currents at the semiconductor contacts enter the Kirchhoff's current law (KCL) equations for the circuit.

Keeping in mind that this semiconductor device model will be coupled to an electric circuit model, voltages and currents of the circuit model are used as input and output variables in the semiconductor model. The new features regarding the modeling are the inclusion of different materials for the semiconductor device and the modeling of the current coupling equation.

The geometrical model of a semiconductor device consists of two subsets $\Omega_S, \Omega_O \subset \mathbb{R}^d$ with $d = 1, 2, 3$. We call Ω_S the semiconductor part and Ω_O the oxide part. The semiconductor part as well as the oxide part are assumed to be open, nonempty and bounded. The semiconductor part and the oxide part are disjoint but adjacent to each other, i.e.

$$\Omega_O \cap \Omega_S = \varnothing \text{ but } \Gamma_I := \bar{\Omega}_O \cap \bar{\Omega}_S \neq \varnothing$$

Let Ω_S and Ω_O have Lipschitz boundaries Γ_S and Γ_O. Split these boundaries into

$$\Gamma_S = \Gamma_{D,S} \dot{\cup} \Gamma_{N,S} \dot{\cup} \Gamma_I \text{ and } \Gamma_O = \Gamma_{D,O} \dot{\cup} \Gamma_{N,O} \dot{\cup} \Gamma_I$$

with the interface boundary Γ_I. Here $\dot{\cup}$ denotes the disjoint union. We call $\Gamma_{D,S}, \Gamma_{D,O} \neq \varnothing$ the Dirichlet boundaries and $\Gamma_{N,S}$ and $\Gamma_{N,O}$ the Neumann boundaries of the semiconductor and the oxide part, respectively. Suppose the semiconductor device to have n_{SC} pairwise disjoint metal contacts $\Gamma_{C_i} \subset \mathbb{R}^d$ such that $\Gamma_{D,S} \cup \Gamma_{D,O} = \cup_{1 \leqslant i \leqslant n_{SC}} \Gamma_{C_i}$ with

$$\Gamma_{C_i} \cap \Gamma_{D,S} = \varnothing \text{ or } \Gamma_{C_i} \cap \Gamma_{D,O} = \varnothing, \quad \forall 1 \leqslant i \leqslant n_{SC}.$$

Each contact of the semiconductor device is joined to a node of the circuit. Let the device be connected to $n_T + 1$ nodes, then we define $n_T + 1$ terminals T_i as the union of the

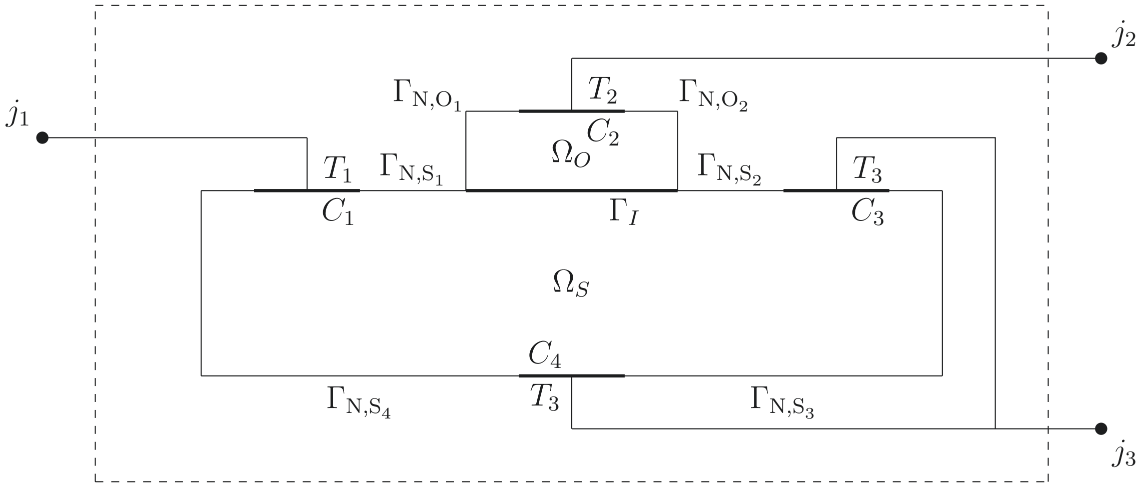


Figure 3.1: Coupling a semiconductor device ($\Omega_S \cup \Omega_O$)

contacts which are connected to the i-th node. We assume that terminals belong either to the Dirichlet boundary of the semiconductor part or to the Dirichlet boundary of the oxide part. Furthermore let the $(n_T + 1)$-th terminal belong to the Dirichlet boundary of the semiconductor part. In the equations that follow $t \in [t_0, T]$ and $x \in \Omega$ are the independent variables, t represents the time while x represents the space. Let us denote by $\mathcal{I} = [t_0, T]$ the considered time interval.

Let $n(x,t)$ be the electron density, $p(x,t)$ the hole density and $\varphi_S(x,t)$ the electrostatic potential on the semiconductor part with $n, p, \varphi_S : \Omega_S \times \mathcal{I} \to \mathbb{R}$. Further let $C : \Omega_S \to \mathbb{R}$ be the doping profile of the semiconductor while $R : \mathbb{R}^2 \to \mathbb{R}$ describes the balance of generation and recombination of electrons and holes.

With these variables and functions we can formulate the drift-diffusion equations, which describe the dynamical behavior of the electrons and the holes in the semiconductor part, i.e. $x \in \Omega_S$, cf. [Moc83, Sel84, Mar86].

$$-\frac{\partial}{\partial t}n + \frac{1}{q}\nabla \cdot J_n(n, \varphi_S, x) = R(n,p) \qquad \text{on } \Omega_S \times \mathcal{I}, \tag{3.3a}$$

$$\frac{\partial}{\partial t}p + \frac{1}{q}\nabla \cdot J_p(p, \varphi_S, x) = -R(n,p) \qquad \text{on } \Omega_S \times \mathcal{I}, \tag{3.3b}$$

$$\nabla \cdot (-\varepsilon_S \nabla \varphi_S) = q(p - n + C(x)) \qquad \text{on } \Omega_S \times \mathcal{I}. \tag{3.3c}$$

where q is the elementary charge, $\varepsilon_S > 0$ is the semiconductor dielectric constant. Notice that the doping profile C does not depend on t while n, p and φ_S depend on x and t but their arguments are being dropped for a clearer view. The functions

$$J_n(n, \varphi_S, x) := q\mu_n(x)(U_T \nabla n - n \nabla \varphi_S),$$
$$J_p(p, \varphi_S, x) := -q\mu_p(x)(U_T \nabla p + p \nabla \varphi_S)$$

describe the current densities

$$j_n(x,t) := J_n(n(x,t), \varphi_S(x,t), x) \text{ and } j_p(x,t) := J_p(p(x,t), \varphi_S(x,t), x)$$

caused by electrons and holes for $x \in \Omega_S$ and $t \in \mathcal{I}$. Furthermore, μ_n and μ_p form the mobilities of electrons and holes, we assume them to be non-negative and bounded functions of x while U_T is a constant which represents the thermal voltage.

Let $n_{\Gamma_D}, p_{\Gamma_D} : \Gamma_{D,S} \to \mathbb{R}$ be the boundary conditions of the electrons and the holes at the Dirichlet boundary of the semiconductor and let $\varphi_{bi} : \Gamma_{D,S} \to \mathbb{R}$ be the built-in potential of the semiconductor at the boundary $\Gamma_{D,S}$. Notice that $n_{\Gamma_D}, p_{\Gamma_D}$ and φ_{bi} are time independent. Further let $u_S : \mathcal{I} \to \mathbb{R}^{n_T}$ be the set of voltages which are applied between the first n_T terminals and the last terminal. The last terminal serves as a reference terminal. We can use the voltages instead of the potential at the contact nodes since the model is invariant under global potential variations, see [Tis03]. Therefore define the boundary function $\varphi_{u,S} : \Gamma_{D,S} \times \mathcal{I} \to \mathbb{R}$ with

$$\varphi_{u,S}(x,t) = \begin{cases} u_{S,i}(t) & , \text{ if } x \in T_i \text{ for } 1 \leqslant i \leqslant n_T, \\ 0 & , \text{ else.} \end{cases}$$

and u_i being the i-th component of u. Using these functions, we can formulate the boundary conditions for the semiconductor part:

$$\begin{aligned}
n(x,t) &= n_{\Gamma_D}(x) && \forall x \in \Gamma_{D,S},\ \forall t \in \mathcal{I}, && (3.4a)\\
p(x,t) &= p_{\Gamma_D}(x) && \forall x \in \Gamma_{D,S},\ \forall t \in \mathcal{I}, && (3.4b)\\
\varphi_S(x,t) &= \varphi_{bi}(x) + \varphi_{u,S}(x,t) && \forall x \in \Gamma_{D,S},\ \forall t \in \mathcal{I}, && (3.4c)\\
\nabla n(x,t) \cdot \nu_S(x) &= 0 && \forall x \in \Gamma_{N,S},\ \forall t \in \mathcal{I}, && (3.4d)\\
\nabla p(x,t) \cdot \nu_S(x) &= 0 && \forall x \in \Gamma_{N,S},\ \forall t \in \mathcal{I}, && (3.4e)\\
\nabla \varphi_S(x,t) \cdot \nu_S(x) &= 0 && \forall x \in \Gamma_{N,S},\ \forall t \in \mathcal{I}, && (3.4f)\\
J_n(n(x,t), \varphi_S(x,t), x) \cdot \nu_S(x) &= 0 && \forall x \in \Gamma_I,\ \forall t \in \mathcal{I}, && (3.4g)\\
J_p(p(x,t), \varphi_S(x,t), x) \cdot \nu_S(x) &= 0 && \forall x \in \Gamma_I,\ \forall t \in \mathcal{I}. && (3.4h)
\end{aligned}$$

Here, $\nu_S(x)$ denotes the outer unit normal vector at x with respect to Ω_S. The equations (3.4a)-(3.4c) describe conditions at the Dirichlet boundary and in particular the connection to the circuit nodes. The conditions (3.4d)-(3.4f) implicate that the electrons, the holes and the electrostatic potential can neither leave nor enter the semiconductor at the Neumann boundary. Equation (3.4g) and (3.4h) state that no particle current can enter the oxide part from the semiconductor part, this means that the oxide part is a perfect isolator. The only variable on such an isolator is the electrostatic potential $\varphi_O : \Omega_O \times \mathcal{I} \to \mathbb{R}$, which can be described by

$$\nabla \cdot (-\varepsilon_O \nabla \varphi_O) = 0 \tag{3.5}$$

with $\varepsilon_O > 0$ being the oxide dielectric constant. We define the function $\varphi_{u,O} : \Gamma_{D,O} \times \mathcal{I} \to \mathbb{R}$ by

$$\varphi_{u,O}(x,t) = \begin{cases} u_i(t) & , \text{ if } x \in T_i \text{ for } 1 \leqslant i \leqslant n_T, \\ 0 & , \text{ else.} \end{cases}$$

which assigns the voltages u to the contacts of the oxide part. Let $\Phi_{ms} : \Gamma_{D,O} \to \mathbb{R}$ be the metal-semiconductor work function difference. Then the boundary conditions of the Dirichlet and Neumann boundaries of the oxide part are given by

$$\begin{aligned}
\varphi_O(x,t) &= -\Phi_{ms}(x) + \varphi_{u,O}(x,t) && \forall x \in \Gamma_{D,O},\ \forall t \in \mathcal{I} && (3.6a)\\
\nabla \varphi_O(x,t) \cdot \nu_O(x) &= 0 && \forall x \in \Gamma_{N,O},\ \forall t \in \mathcal{I} && (3.6b)
\end{aligned}$$

with $\nu_O(x)$ being the outer unit normal vector at x with respect to Ω_O. To complete the set of boundary conditions for the whole semiconductor-oxide problem we still lack conditions for the potential at the interface boundary. These boundary conditions will connect the semiconductor part with the oxide part since the electrostatic potential exists in both parts. First we set the Dirichlet conditions such that the electrostatic potential of the semiconductor part and the oxide part match each other at the interface boundary

$$\varphi_S(x,t) = \varphi_O(x,t) \qquad \text{on } \Gamma_I \times \mathcal{I}. \tag{3.7}$$

Condition (3.7) guarantees the continuity of the potential at the interface boundary. Furthermore the sum of the gradients of the potential weighted with the dielectric constants of each region is supposed to be zero at the interface boundary.

$$\varepsilon_S \nabla\varphi_S(x,t)\cdot\nu_S(x) + \varepsilon_O\nabla\varphi_O(x,t)\cdot\nu_O(x) = 0 \qquad \forall x \in \Gamma_I,\ \forall t \in \mathcal{I}. \tag{3.8}$$

This boundary condition provides the continuity of the electric field. Now we have a full set of equations which describe the semiconductor model depending on the voltages of the circuit. Next we need to couple the quantities of the semiconductor device back to the circuit. Introducing the semiconductor current and oxide current functions

$$J_S(n,p,\varphi_S,x) := J_n(n,\varphi_S,x) + J_p(p,\varphi_S,x) + \partial_t(-\varepsilon_S\nabla\varphi_S), \quad J_O(\varphi_O) := \partial_t(-\varepsilon_O\nabla\varphi_O)$$

we can formulate the current coupling equation

$$j_k := \int_{T_k\cap\Gamma_{D,S}} J_S(n,p,\varphi_S,x)\cdot\nu_S \mathrm{d}s + \int_{T_k\cap\Gamma_{D,O}} J_O(\varphi_O)\cdot\nu_O\mathrm{d}s$$

which describes the current j_k of the circuit at the terminal T_k. We will now show the conservation of energy for this choice of the current coupling equation:

Lemma 3.1. (Conservation of energy)
Supposing n, p, φ_S and φ_O fulfill (3.3)-(3.8), we find the sum of the terminal currents to be zero:

$$\sum_{k=1}^{n_T+1} j_k = 0. \tag{3.9}$$

Proof.
To verify this statement notice that the divergence of the semiconductor current and the oxide current is zero:

$$\nabla\cdot J_S(n,p,\varphi_S,x) = \nabla\cdot(J_n(n,\varphi_S,x) + J_p(p,\varphi_S,x)) + \partial_t\nabla\cdot(-\varepsilon_S\nabla\varphi_S) \overset{(3.3)}{=} 0 \tag{3.10a}$$

$$\nabla\cdot J_O(\varphi_O) = \partial_t\nabla\cdot(-\varepsilon_O\nabla\varphi_O) \overset{(3.5)}{=} 0. \tag{3.10b}$$

The normal component of the semiconductor and oxid currents is also zero at the Neumann boundaries, while the sum of both is zero at the interface boundary.

$$J_S(n,p,\varphi_S,x)\cdot\nu_S \overset{(3.4d),(3.4e),(3.4f)}{=} 0 \qquad \text{on } \Gamma_{N,S}\times\mathcal{I} \tag{3.11a}$$

$$J_O(\varphi_O)\cdot\nu_O \overset{(3.6b)}{=} 0 \qquad \text{on } \Gamma_{N,O}\times\mathcal{I} \tag{3.11b}$$

$$J_S(n,p,\varphi_S,x)\cdot\nu_S + J_O(\varphi_O)\cdot\nu_O \overset{(3.4g),(3.4h),(3.8)}{=} 0 \qquad \text{on } \Gamma_I\times\mathcal{I}. \tag{3.11c}$$

With these properties and the divergence theorem follows the desired statement

$$\begin{aligned}\sum_{k=1}^{n_T+1} j_k &= \sum_{k=1}^{n_T+1}\int_{T_k\cap\Gamma_{D,S}} J_S(n,p,\varphi_S,x)\cdot\nu_S\mathrm{d}s + \sum_{k=1}^{n_T+1}\int_{T_k\cap\Gamma_{D,O}} J_O(\varphi_O)\cdot\nu_O\mathrm{d}s\\ &= \int_{\Gamma_{D,S}} J_S(n,p,\varphi_S,x)\cdot\nu_S\mathrm{d}s + \int_{\Gamma_{D,O}} J_O(\varphi_O)\cdot\nu_O\mathrm{d}s\\ &\overset{(3.11)}{=} \int_{\Gamma_S} J_S(n,p,\varphi_S,x)\cdot\nu_S\mathrm{d}s + \int_{\Gamma_O} J_O(\varphi_O)\cdot\nu_O\mathrm{d}s\\ &= \int_{\Omega_S}\nabla\cdot J_S(n,p,\varphi_S,x)\mathrm{d}x + \int_{\Omega_O}\nabla\cdot J_O(\varphi_O)\mathrm{d}x \overset{(3.10)}{=} 0.\end{aligned}$$

□

By Lemma 3.1 we can express the current at the last terminal by

$$j_{n_T+1} = -\sum_{k=1}^{n_T} j_k.$$

This enables us to describe the semiconductor device by n_T branches corresponding to the first n_T currents. In Figure 3.2 we see a semiconductor device with four terminals on the left, which is alternatively described by three branches S_1, S_2 and S_3 on the right. To structurally classify the semiconductor device as a capacitors-like element we need to provide direct relations between the first n_T currents and the derivative of the voltages at the device. Each of the relations can then be ascribed to one of the branches. As a first step define a set of auxiliary problems as an extension to the auxiliary functions in [Gaj93]: For each $1\leqslant i\leqslant n_T$, let $w_S^i\in C^2(\Omega_S)\cap C^1(\bar{\Omega}_S)$ and $w_O^i\in C^2(\Omega_O)\cap C^1(\bar{\Omega}_O)$ be solutions of

$$\begin{aligned}\nabla\cdot(-\varepsilon_S\nabla w_S^i) &= 0 & \nabla\cdot(-\varepsilon_O\nabla w_O^i) &= 0\\ w_S^i(x)_{|\Gamma_{D,S}} &= \begin{cases}1 & ,\ \text{if } x\in T_i,\\ 0 & ,\ \text{else.}\end{cases} & w_O^i(x)_{|\Gamma_{D,O}} &= \begin{cases}1 & ,\ \text{if } x\in T_i,\\ 0 & ,\ \text{else.}\end{cases}\\ \nabla w_S^i\cdot\nu_{S|\Gamma_{N,S}} &= 0 & \nabla w_O^i\cdot\nu_{O|\Gamma_{N,O}} &= 0\end{aligned} \tag{3.12}$$

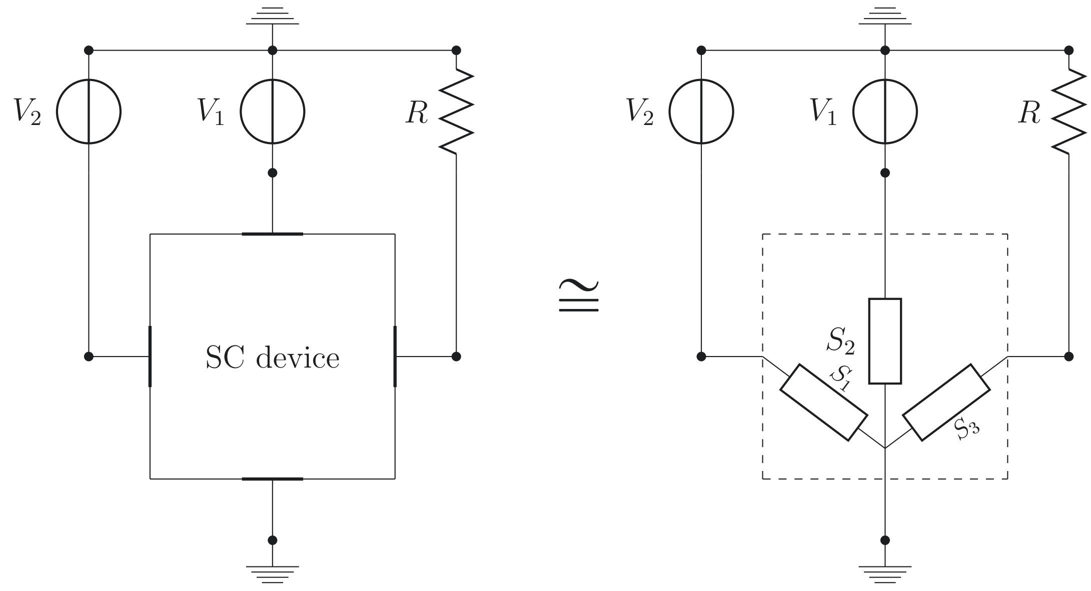


Figure 3.2: Semiconductor device model with four terminals described by three branches S_1, S_2 and S_3.

with the same interface boundary conditions as the original problem

$$w_S^i = w_O^i \quad \text{and} \quad \varepsilon_S \nabla w_S^i \cdot \nu_S + \varepsilon_O \nabla w_O^i \cdot \nu_O = 0 \quad \text{on } \Gamma_I. \tag{3.13}$$

The weak solvability of the auxiliary problem is shown in [Str14]. With the help of these auxiliary functions we define a Gramian matrix $W \in R^{n_T \times n_T}$ by

$$(W)_{ij} := \int_{\Omega_S} \varepsilon_S \nabla w_S^i \cdot \nabla w_S^j \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla w_O^i \cdot \nabla w_O^j \mathrm{d}x \quad 1 \leqslant i, j \leqslant n_T.$$

We denote the k-th row of W with W_k.

Lemma 3.2.
The matrix W is symmetric and positive definite.

Proof.
We define the spaces

$$H^1_{S,T_{n_T+1}} := \{v \in H^1(\Omega_S) |\ v_{|T_{n_T+1}} = 0\}, \quad H^1_{O,I} := \{v \in H^1(\Omega_O) |\ v_{|\Gamma_I} = 0\}$$

and

$$H^1_{SO,T_{n_T+1}} := \{v = (v_S, v_O) \in H^1_{S,T_{n_T+1}} \times H^1(\Omega_O) |\ v_{S|\Gamma_I} = v_{O|\Gamma_I}\} \tag{3.14}$$

and the bilinear form $a : H^1_{SO,T_{n_T+1}} \times H^1_{SO,T_{n_T+1}} \to \mathbb{R}$ by

$$a(v,w) := \int_{\Omega_S} \varepsilon_S \nabla v_S \cdot \nabla w_S \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla v_O \cdot \nabla w_O \mathrm{d}x.$$

The Spaces $H^1(\Omega_S)$ and $H^1(\Omega_O)$ are the well known Sobolev Spaces, see [Ada75]. Then a is a scalar product. In particular we have to show that it holds $a(v,v) = 0 \Rightarrow v = 0$, the other properties are trivial.

$$\begin{aligned} & a(v,v) = 0 \\ \Rightarrow & \int_{\Omega_S} \varepsilon_S \nabla v_S \cdot \nabla v_S \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla v_O \cdot \nabla v_O \mathrm{d}x = 0 \\ \Rightarrow & \int_{\Omega_S} \varepsilon_S \nabla v_S \cdot \nabla v_S \mathrm{d}x = 0 \text{ and } \int_{\Omega_O} \varepsilon_O \nabla v_O \cdot \nabla v_O \mathrm{d}x = 0, \end{aligned}$$

since $\nabla v_S \cdot \nabla v_S \geqslant 0$ and $\nabla v_O \cdot \nabla v_O \geqslant 0$. By $v_S \in H^1_{S,T_{n_T+1}}$ and $\int_{\Omega_S} \varepsilon_S \nabla v_S \cdot \nabla v_S \mathrm{d}x = 0$ it follows that $v_S = 0$ and therefore $v_O \in H^1_{O,I}$. This yields together with

$$\int_{\Omega_O} \varepsilon_O \nabla v_O \cdot \nabla v_O \mathrm{d}x = 0$$

that $v_O = 0$. Hence a is a scalar product and therefore W is symmetric and positive definite, since W is the Gramian matrix of a scalar product. □

Collect the currents at the terminals with $j_S := \begin{pmatrix} j_{S_1} & , \dots , & j_{S_{n_T}} \end{pmatrix}^{\mathrm{T}}$ and $j_{S_k} := -j_k$ for $1 \leqslant k \leqslant n_T$. Furthermore we denote

$$w_S(x) := \begin{pmatrix} w_S^1(x) & , \dots , & w_S^{n_T}(x) \end{pmatrix}^{\mathrm{T}} \text{ and } w_O(x) := \begin{pmatrix} w_O^1(x) & , \dots , & w_O^{n_T}(x) \end{pmatrix}^{\mathrm{T}}, \tag{3.15}$$

for all $x \in \Omega_S$ and $x \in \Omega_O$, respectively. Similar to the first set of auxiliary problems choose $\varphi_{\Gamma_S} : \Omega_S \to \mathbb{R}$ and $\varphi_{\Gamma_O} : \Omega_O \to \mathbb{R}$ such that

$$\begin{array}{rclrcl} \nabla \cdot (-\varepsilon_S \nabla \varphi_{\Gamma_S}) & = & 0 & \nabla \cdot (-\varepsilon_O \nabla \varphi_{\Gamma_O}) & = & 0 \\ \varphi_{\Gamma_S|\Gamma_{D,S}} & = & \varphi_{bi} & \varphi_{\Gamma_O|\Gamma_{D,O}} & = & -\Phi_{ms} \\ \nabla \varphi_{\Gamma_S} \cdot \nu_{|\Gamma_{N,S}} & = & 0 & \nabla \varphi_{\Gamma_O} \cdot \nu_{|\Gamma_{N,O}} & = & 0 \end{array} \tag{3.16}$$

with the same interface boundary conditions again:

$$\varphi_{\Gamma_S} = \varphi_{\Gamma_O}, \quad \varepsilon_S \nabla \varphi_{\Gamma_S} \cdot \nu_S + \varepsilon_O \nabla \varphi_{\Gamma_O} \cdot \nu_O = 0 \quad \text{on } \Gamma_I. \tag{3.17}$$

With these preparations we are able to derive a relation between the semiconductor current j_S and the derivative of the voltages u_S of the circuit applied to the semiconductor.

Lemma 3.3.
The semiconductor currents collected in j_S can be expressed as

$$\begin{aligned} j_S(t) =& \frac{\mathrm{d}}{\mathrm{d}t}(W u_S(t)) \\ & - \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S(x)\right) (J_n(n(x,t), \varphi_S(x,t), x) + J_p(p(x,t), \varphi_S(x,t), x))\, \mathrm{d}x \end{aligned}$$

Proof.
We need to use the divergence theorem, see [For96], again to obtain the desired relation. Notice that the Dirichlet boundary conditions of the auxiliary functions w_S^k and w_O^k at the Dirichlet boundary and at the interface boundary are crucial for the next steps.

$$\begin{aligned} j_{S_k} =& - j_k \\ =& - \int_{T_k \cap \Gamma_{D,S}} J_S(n,p,\varphi_S,x) \cdot \nu_S \mathrm{d}s - \int_{T_k \cap \Gamma_{D,O}} J_O(\varphi_O) \cdot \nu_O \mathrm{d}s \\ =& - \int_{\Gamma_{D,S}} w_S^k J_S(n,p,\varphi_S,x) \cdot \nu_S \mathrm{d}s - \int_{\Gamma_{D,O}} w_O^k J_O(\varphi_O) \cdot \nu_O \mathrm{d}s \\ \overset{(3.13),(3.11)}{=}& - \int_{\Gamma_S} w_S^k J_S(n,p,\varphi_S,x) \cdot \nu_S \mathrm{d}s - \int_{\Gamma_O} w_O^k J_O(\varphi_O) \cdot \nu_O \mathrm{d}s \\ =& - \int_{\Omega_S} \nabla \cdot \left(w_S^k J_S(n,p,\varphi_S,x)\right) \mathrm{d}x - \int_{\Omega_O} \nabla \cdot \left(w_O^k J_O(\varphi_O)\right) \mathrm{d}x \\ =& - \int_{\Omega_S} w_S^k \nabla \cdot J_S(n,p,\varphi_S,x) + J_S(n,p,\varphi_S,x) \cdot \nabla w_S^k \mathrm{d}x \\ & - \int_{\Omega_O} w_O^k \nabla \cdot J_O(\varphi_O) + J_O(\varphi_O) \cdot \nabla w_O^k \mathrm{d}x \\ \overset{(3.10)}{=}& - \int_{\Omega_S} J_S(n,p,\varphi_S,x) \cdot \nabla w_S^k \mathrm{d}x - \int_{\Omega_O} J_O(\varphi_O) \cdot \nabla w_O^k \mathrm{d}x \\ =& \frac{\partial}{\partial t}\left[\int_{\Omega_S} \varepsilon_S \nabla \varphi_S \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \varphi_O \cdot \nabla w_O^k \mathrm{d}x\right] \\ & - \int_{\Omega_S} \nabla w_S^k \cdot (J_n(n,\varphi_S,x) + J_p(p,\varphi_S,x)) \mathrm{d}x \end{aligned}$$

We split $\varphi_S = \bar{\varphi}_S + \varphi_{\Gamma_S} + w_S \cdot u_s(t)$ and $\varphi_O = \bar{\varphi}_O + \varphi_{\Gamma_O} + w_O \cdot u_s(t)$ such that we obtain homogenized functions at the Dirichlet boundaries, i.e.

$$\bar{\varphi}_{S|\Gamma_{D,S}} = 0, \quad \bar{\varphi}_{O|\Gamma_{D,O}} = 0. \tag{3.18}$$

Notice that $\bar{\varphi}_S$ and $\bar{\varphi}_O$ still fulfill the boundary conditions at the interface boundary since φ_{Γ_S}, φ_{Γ_O}, w_S and w_O do. Furthermore notice that φ_{Γ_S} and φ_{Γ_O} are independent of t hence

$$\frac{\partial}{\partial t}\left[\int_{\Omega_S} \varepsilon_S \nabla \varphi_{\Gamma_S} \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \varphi_{\Gamma_O} \cdot \nabla w_O^k \mathrm{d}x\right] = 0.$$

Also due to the choice of the auxiliary functions it holds:

$$
\begin{aligned}
&\int_{\Omega_S} \varepsilon_S \nabla \bar{\varphi}_S \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \bar{\varphi}_O \cdot \nabla w_O^k \mathrm{d}x \\
=& \int_{\Omega_S} \varepsilon_S \nabla \cdot \left(\nabla w_S^k \bar{\varphi}_S \right) \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \cdot \left(\nabla w_O^k \bar{\varphi}_O \right) \mathrm{d}x \\
=& \int_{\Gamma_S} \varepsilon_S \nabla w_S^k \cdot \nu_S \bar{\varphi}_S \mathrm{d}s + \int_{\Gamma_O} \varepsilon_O \nabla w_O^k \cdot \nu_O \bar{\varphi}_O \mathrm{d}s \\
\overset{(3.18)}{=}& \int_{\Gamma_I} \varepsilon_S \nabla w_S^k \cdot \nu_S \bar{\varphi}_S \mathrm{d}s + \int_{\Gamma_I} \varepsilon_O \nabla w_O^k \cdot \nu_O \bar{\varphi}_O \mathrm{d}s \\
=& \bar{\varphi}_S \left(\int_{\Gamma_I} \varepsilon_S \nabla w_S^k \cdot \nu_S + \varepsilon_O \nabla w_O^k \cdot \nu_O \mathrm{d}s \right) \overset{(3.13)}{=} 0.
\end{aligned}
$$

And with all these properties we obtain:

$$
\begin{aligned}
&\frac{\partial}{\partial t} \left[\int_{\Omega_S} \varepsilon_S \nabla \varphi_S \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \varphi_O \cdot \nabla w_O^k \mathrm{d}x \right] \\
&= \frac{\partial}{\partial t} \left[\int_{\Omega_S} \varepsilon_S (\nabla w_S \cdot u_s) \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla (w_O \cdot u_s) \cdot \nabla w_O^k \mathrm{d}x \right] \\
&= \left[\int_{\Omega_S} \varepsilon_S \nabla w_S \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla w_O \cdot \nabla w_O^k \mathrm{d}x \right] \cdot \frac{\partial}{\partial t} u_s \\
&= W_k \cdot \frac{\partial}{\partial t} u_s
\end{aligned}
$$

which allows us to write each of the currents in the following form

$$
\begin{aligned}
j_{S_k} =& \frac{\partial}{\partial t} \left[\int_{\Omega_S} \varepsilon_S \nabla \varphi_S \cdot \nabla w_S^k \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \varphi_O \cdot \nabla w_O^k \mathrm{d}x \right] \\
&- \int_{\Omega_S} \nabla w_S^k \cdot (J_n(n, \varphi_S, x) + J_p(p, \varphi_S, x) \mathrm{d}x \\
=& W_k \cdot \frac{\partial}{\partial t} u_s - \int_{\Omega_S} \nabla w_S^k \cdot (J_n(n, \varphi_S, x) + J_p(p, \varphi_S, x)) \mathrm{d}x.
\end{aligned}
$$

Put together, we achieve the desired relation of the currents and the voltages

$$
j_S = \frac{\mathrm{d}}{\mathrm{d}t}(W u_s) - \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S \right) (J_n(n, \varphi_S, x) + J_p(p, \varphi_S, x)) \mathrm{d}x.
$$

□

At this point we are able to write the complete Partial Differential-Algebraic Equation (PDAE) which describes the semiconductor model. This PDAE includes the current coupling term, the semiconductor equations and the oxide equation:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(Wu_s) - \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S\right)(J_n(n,\varphi_S,x) + J_p(p,\varphi_S,x))\mathrm{d}x &= j_S \\ -\frac{\partial}{\partial t} n + \frac{1}{q}\nabla \cdot J_n(n,\varphi_S,x) &= R(n,p) \\ \frac{\partial}{\partial t} p + \frac{1}{q}\nabla \cdot J_p(p,\varphi_S,x) &= -R(n,p) \\ \nabla \cdot (-\varepsilon_S \nabla \varphi_S) &= q(p - n + C(x)) \\ \nabla \cdot (-\varepsilon_O \nabla \varphi_O) &= 0 \end{aligned} \tag{3.19}$$

with the boundary and coupling conditions (3.4), (3.6), (3.7) and (3.8) and the auxiliary problems for φ_{Γ_S}, φ_{Γ_O}, w_S and w_O. Notice that u_s and j_S depend on t only; w, μ_n, μ_p, C, φ_{Γ_S}, φ_{Γ_O}, w_S, w_O depend on x only whereas φ_S, φ_O, n, p depend on both, x and t.
In order to numerically solve a coupled system, consisting of a semiconductor device and an electric circuit, we need to discretize the coupled system in space and solve the resulting DAE using appropriate numerical methods. Discretization in space can be done independently of the electric circuit, since it only concerns the semiconductor device.

Weak Formulation

First, we derive a weak formulation of the PDAE (3.19) which we will call an Abstract Differential-Algebraic Equation (ADAE). Consider the two real Banach spaces

$$H^1_S := \{v \in H^1(\Omega_S)|\ v_{|\Gamma_{D,S}} = 0\} \text{ and } H^1_O := \{w \in H^1(\Omega_O)|\ w_{|\Gamma_{D,O}} = 0\}$$

with the restricted norms $\|v\|_{H^1_S} := \|v\|_{H^1(\Omega_S)}$ and $\|w\|_{H^1_O} := \|w\|_{H^1(\Omega_O)}$ for all $v \in H^1_S$ and all $w \in H^1_O$, respectively.
Based on these Banach spaces define the product space

$$H^1_{SO} := \{v = (v_S, v_O) \in H^1_S \times H^1_O|\ v_{S|\Gamma_I} = v_{O|\Gamma_I}\} \tag{3.20}$$

with the scalar product $\langle v,u\rangle_{H^1_{SO}} := \int_{\Omega_S} \varepsilon_S u_S \cdot v_S \mathrm{d}x + \int_{\Omega_O} \varepsilon_O u_O \cdot v_O \mathrm{d}x$ for all $v, u \in H^1_{SO}$. Notice that ∇ is now the weak derivative. We want to choose H^1_{SO} as the solution space for the electrostatic potential. Hence, we need that H^1_{SO} is a Banach space.

Lemma 3.4.
The space H^1_{SO} defined in (3.20) is a Banach space.

Proof.
First we show that H^1_{SO} is a normed vector space. Let v and w be elements in H^1_{SO} hence $v, w \in H^1_S \times H^1_O$. Since H^1_S and H^1_O are Banach spaces, the product space $H^1_S \times H^1_O$ is a Banach space. Therefore $v + \lambda w \in H^1_S \times H^1_O$ for every $\lambda \in \mathbb{R}$. Since the extra condition of H^1_{SO} is linear it holds that $v + \lambda w \in H^1_{SO}$ for every $\lambda \in \mathbb{R}$. The norm of $H^1_S \times H^1_O$ is also a norm of H^1_{SO} since H^1_{SO} is a subset of $H^1_S \times H^1_O$.
Next, we show the completeness of H^1_{SO}. Consider $(v_n)_{n\in\mathbb{N}} \in H^1_{SO}$ to be a Cauchy sequence. It implies v_n also to be a Cauchy sequence in $H^1_S \times H^1_O$. Since this space is a Banach space, $(v_n)_{n\in\mathbb{N}}$ has a limit v in $H^1_S \times H^1_O$. The domains have Lipschitz boundaries and therefore the trace operator is continuous, see [Eva10]. Together with the linearity of the extra condition of H^1_{SO} it follows that $v \in H^1_{SO}$. Hence, the Cauchy sequence converges with a limit in H^1_{SO}. □

We already chose homogenization functions for φ, now we choose the ones for n and p. Therefore let be $n_D, p_D \in C^2(\Omega_S) \cap C^0(\bar{\Omega}_S)$ with

$$n_D = n_{\Gamma_D} \text{ and } p_D = p_{\Gamma_D} \quad \text{on } \Gamma_{D,S}.$$

Finally, we consider

$$H^1_S := \{v \in H^1(\Omega_S) |\ v_{|\Gamma_{D,S}} = 0\}$$

which will serve as a solution space for the electrons n and the holes p. Let the homogenized electrons and holes be $\bar{n}, \bar{p} \in H^1_S$ and let the homogenized electrostatic potential be $\bar{\varphi} = (\bar{\varphi}_S, \bar{\varphi}_O) \in H^1_{SO}$. Then we obtain the following relations:

$$\begin{aligned}
n(x,t) &= \bar{n}(x,t) + n_D(x) && \forall x \in \Omega_S\, \forall t \in \mathcal{I}, \\
p(x,t) &= \bar{p}(x,t) + p_D(x) && \forall x \in \Omega_S\, \forall t \in \mathcal{I}, \\
\varphi_S(x,t) &= \bar{\varphi}_S(x,t) + \varphi_{\Gamma_S}(x) + w_S(x) \cdot u_s(t) && \forall x \in \Omega_S\, \forall t \in \mathcal{I}, \\
\varphi_O(x,t) &= \bar{\varphi}_O(x,t) + \varphi_{\Gamma_O}(x) + w_O(x) \cdot u_s(t) && \forall x \in \Omega_O\, \forall t \in \mathcal{I}.
\end{aligned}$$

Let $\vartheta = (\vartheta_S, \vartheta_O) \in H^1_{SO}$ be an arbitrary test function for the electrostatic potential φ while the electron and hole densities n and p share the same arbitrary test function $\theta \in H^1_S$. In general, n and p could have their individual test functions. Further define the functional

$$\ell(\vartheta_S) := \int_{\Omega_S} q(p_D(x) - n_D(x) + C(x))\vartheta_S(x)\mathrm{d}x \tag{3.21}$$

and obtain the homogenized weak formulation of the PDAE, also called the ADAE

$$
\begin{aligned}
&\int_{\Omega_S} \theta(x)\frac{\partial}{\partial t}\bar{n}(x,t)\mathrm{d}x + \int_{\Omega_S} \frac{1}{q}\bar{J}_n(\bar{n}(x,t),\bar{\varphi}_S(x,t),x,u_s(t))\cdot\nabla\theta(x)\mathrm{d}x \\
&\qquad\qquad\qquad\qquad + \int_{\Omega_S} \bar{R}(\bar{n}(x,t),\bar{p}(x,t),x)\theta(x)\mathrm{d}x = 0, \\
&\int_{\Omega_S} \theta(x)\frac{\partial}{\partial t}\bar{p}(x,t)\mathrm{d}x - \int_{\Omega_S} \frac{1}{q}\bar{J}_p(\bar{p}(x,t),\bar{\varphi}_S(x,t),x,u_s(t))\cdot\nabla\theta(x)\mathrm{d}x \\
&\qquad\qquad\qquad\qquad + \int_{\Omega_S} \bar{R}(\bar{n}(x,t),\bar{p}(x,t),x)\theta(x)\mathrm{d}x = 0, \\
&\int_{\Omega_S} \varepsilon_S\nabla\bar{\varphi}_S(x,t)\cdot\nabla\vartheta_S(x)\mathrm{d}x + \int_{\Omega_O} \varepsilon_O\nabla\bar{\varphi}_O(x,t)\cdot\nabla\vartheta_O(x)\mathrm{d}x \\
&\qquad\qquad\qquad\qquad - \int_{\Omega_S} q(\bar{p}(x,t)-\bar{n}(x,t))\vartheta_S(x)\mathrm{d}x = \ell(\vartheta_S)
\end{aligned}
\tag{3.22}
$$

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

$$
\begin{aligned}
\bar{J}_n(\bar{n},\bar{\varphi}_S,x,u_s) &:= J_n(\bar{n}+n_D(x),\bar{\varphi}_S+\varphi_{\Gamma_S}(x)+w_S(x)\cdot u_s,x) \\
\bar{J}_p(\bar{p},\bar{\varphi}_S,x,u_s) &:= J_p(\bar{p}+p_D(x),\bar{\varphi}_S+\varphi_{\Gamma_S}(x)+w_S(x)\cdot u_s,x) \\
\bar{R}(\bar{n},\bar{p},x) &:= R(\bar{n}+n_D(x),\bar{p}+p_D(x)).
\end{aligned}
\tag{3.23}
$$

Finite Element Discretization

In the following we derive a semi-discretized version of the semiconductor device model via a finite element discretization. An extensive review on methods for the discretization of the drift-diffusion equations is given in [MSW99]. We consider here an ordinary finite element method approach [Moc83, Mar86]. After discretization the semiconductor device model can be described as a DAE, which is added to the other circuit elements in Section 3.1.5. To achieve the Galerkin equations of the finite element discretization we define the two discrete solution spaces

$$
\begin{aligned}
H^1_{S,h} &:= \operatorname{span}\{\theta_1,\ldots,\theta_M\} \subset H^1_S \\
H^1_{SO,h} &:= \operatorname{span}\{(\vartheta_{S,1},\vartheta_{O,1}),\ldots,(\vartheta_{S,N},\vartheta_{O,N})\} \subset H^1_{SO}
\end{aligned}
$$

with $\theta_i \in H^1_S$ and $(\vartheta_{S,i},\vartheta_{O,i}) \in H^1_{SO}$ being pairwise linear independent vectors. We denote their bases as well as the associated Galerkin coefficients by

$$
\begin{aligned}
\Theta(x) &= \big(\theta_1(x),...,\theta_M(x)\big)^T & N(t) &= \big(\bar{n}_{h,1}(t),...,\bar{n}_{h,M}(t)\big)^T \\
\Phi_S(x) &= \big(\vartheta_{S,1}(x),...,\vartheta_{S,N}(x)\big)^T & P(t) &= \big(\bar{p}_{h,1}(t),...,\bar{p}_{h,M}(t)\big)^T \\
\Phi_O(x) &= \big(\vartheta_{O,1}(x),...,\vartheta_{O,N}(x)\big)^T & \Psi(t) &= \big(\bar{\varphi}_{h,1}(t),...,\bar{\varphi}_{h,N}(t)\big)^T .
\end{aligned}
$$

By the Galerkin approach we obtain

$$
\begin{aligned}
\bar{n}_h(x,t) &= \sum_{i=1}^{M} \bar{n}_{h,i}(t)\theta_i(x) = \Theta(x)\cdot N(t),\\
\bar{p}_h(x,t) &= \sum_{i=1}^{M} \bar{p}_{h,i}(t)\theta_i(x) = \Theta(x)\cdot P(t),\\
\bar{\varphi}_h(x,t) &= \left(\sum_{i=1}^{N} \bar{\varphi}_{h,i}(t)\vartheta_{S,i}(x), \sum_{i=1}^{N} \bar{\varphi}_{h,i}(t)\vartheta_{O,i}(x)\right) = (\Phi_S(x)\cdot\Psi(t), \Phi_O(x)\cdot\Psi(t)).
\end{aligned}
$$

Inserting this into the ADAE formulation (3.22) we get the Galerkin equations

$$
\begin{aligned}
\int_{\Omega_S} \theta_i\Theta \mathrm{d}x \frac{\partial}{\partial t} N(t) + \int_{\Omega_S} \frac{1}{q}\bar{J}_n(\bar{n}_h, \bar{\varphi}_{h,S}, x, u_s(t))\cdot\nabla\theta_i \mathrm{d}x + \int_{\Omega_S} \bar{R}(\bar{n}_h, \bar{p}_h, x)\theta_i \mathrm{d}x &= 0\\
\int_{\Omega_S} \theta_i\Theta \mathrm{d}x \frac{\partial}{\partial t} P(t) - \int_{\Omega_S} \frac{1}{q}\bar{J}_p(\bar{p}_h, \bar{\varphi}_{h,S}, x, u_s(t))\cdot\nabla\theta_i \mathrm{d}x + \int_{\Omega_S} \bar{R}(\bar{n}_h, \bar{p}_h, x)\theta_i \mathrm{d}x &= 0\\
\left(\int_{\Omega_S} \varepsilon_S \nabla\Phi_S\cdot\nabla\vartheta_{S,j}\mathrm{d}x + \int_{\Omega_O} \varepsilon_O\nabla\Phi_O\cdot\nabla\vartheta_{O,j}\mathrm{d}x\right)\Psi(t)&\\
- \int_{\Omega_S} q\Theta\vartheta_{S,j}\mathrm{d}x(P(t)-N(t)) &= \ell(\vartheta_{S,j})
\end{aligned}
$$

for all $t \in \mathcal{I}$, $i = 1, ..., M$ and $j = 1, ..., N$. The Galerkin coefficients are only inserted into the linear parts for cleared depiction. Solely the arguments of the variables that depend on t only are shown to underline their independence of x. The semiconductor current equation of Lemma 3.3 reads

$$
\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}(Wu_s(t)) - j_S(t)\\
&= \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S(x)\right)\left(\bar{J}_n(\bar{n}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t)) + \bar{J}_p(\bar{n}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t))\right)\mathrm{d}x.
\end{aligned}
$$

Denote the coefficients in front of the time derivatives and the electrostatic potential by

$$
(Z)_{ij} = \int_{\Omega_S} \theta_i(x)\theta_j(x)\mathrm{d}x, \tag{3.24a}
$$

$$
(T)_{ij} = \int_{\Omega_S} \varepsilon_S\nabla\vartheta_{S,i}(x)\cdot\nabla\vartheta_{S,j}(x)\mathrm{d}x + \int_{\Omega_O} \varepsilon_O\nabla\vartheta_{O,i}(x)\cdot\nabla\vartheta_{O,j}(x)\mathrm{d}x, \tag{3.24b}
$$

$$
(H)_{ij} = \int_{\Omega_S} q\theta_j(x)\vartheta_{S,i}(x)\mathrm{d}x. \tag{3.24c}
$$

Notice that Z and T are symmetric and positive definite, since they are Gramian matrices of scalar products. This can be proven analogously to Lemma 3.2. Then we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(Wu_s(t)) - \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S(x)\right) (\bar{J}_n(\bar{n}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t))& \\
+\bar{J}_p(\bar{p}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t)))\mathrm{d}x &= j_S(t), \\
Z\frac{\mathrm{d}}{\mathrm{d}t}N(t) + \int_{\Omega_S} \frac{\partial}{\partial x}\Theta(x)\frac{1}{q}\bar{J}_n(\bar{n}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t))\mathrm{d}x& \\
+\int_{\Omega_S} \Theta(x)\cdot \bar{R}(\bar{n}_h(x,t), \bar{p}_h(x,t), x)\mathrm{d}x &= 0, \\
Z\frac{\mathrm{d}}{\mathrm{d}t}P(t) - \int_{\Omega_S} \frac{\partial}{\partial x}\Theta(x)\frac{1}{q}\bar{J}_p(\bar{p}_h(x,t), \bar{\varphi}_{h,S}(x,t), x, u_s(t))\mathrm{d}x& \\
+\int_{\Omega_S} \Theta(x)\cdot \bar{R}(\bar{n}_h(x,t), \bar{p}_h(x,t), x)\mathrm{d}x &= 0, \\
T\Psi(t) - H(P(t) - N(t)) &= \ell(\Phi_S).
\end{aligned}$$

Choosing an approximation of the integrals

$$\begin{aligned}
\bar{g}_S(u_s, N, P, \Psi) \approx &- \int_{\Omega_S} \left(\frac{\partial}{\partial x} w_S(x)\right) (\bar{J}_n(\bar{n}_h(x,t), \bar{\varphi}_{h,S}(x,\cdot), x, u_s(\cdot)) \\
&+ \bar{J}_p(\bar{p}_h(x,\cdot), \bar{\varphi}_{h,S}(x,\cdot), x, u_s(\cdot)))\mathrm{d}x, \\
\bar{h}_n(u_s, N, P, \Psi) \approx &\int_{\Omega_S} \frac{\partial}{\partial x}\Theta(x)\frac{1}{q}\bar{J}_n(\bar{n}_h(x,\cdot), \bar{\varphi}_{h,S}(x,\cdot), x, u_s(\cdot))\mathrm{d}x \\
&+ \int_{\Omega_S} \Theta(x)\cdot \bar{R}(\bar{n}_h(x,\cdot), \bar{p}_h(x,\cdot), x), \qquad (3.25) \\
\bar{h}_p(u_s, N, P, \Psi) \approx &- \int_{\Omega_S} \frac{\partial}{\partial x}\Theta(x)\frac{1}{q}\bar{J}_p(\bar{p}_h(x,\cdot), \bar{\varphi}_{h,S}(x,\cdot), x, u_s(\cdot))\mathrm{d}x \\
&+ \int_{\Omega_S} \Theta(x)\cdot \bar{R}(\bar{n}_h(x,\cdot), \bar{p}_h(x,\cdot), x)\mathrm{d}x, \\
h_\Psi(N, P) \approx &H(P(\cdot) - N(\cdot)) + \ell(\Phi_S)
\end{aligned}$$

we obtain the discrete system

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(Wu_s(t)) + \bar{g}_S(u_s(t), N(t), P(t), \Psi(t)) &= j_S(t), \\
Z\frac{\mathrm{d}}{\mathrm{d}t}N(t) + \bar{h}_n(u_s(t), N(t), P(t), \Psi(t)) &= 0, \\
Z\frac{\mathrm{d}}{\mathrm{d}t}P(t) + \bar{h}_p(u_s(t), N(t), P(t), \Psi(t)) &= 0,
\end{aligned}$$

$$T\Psi(t) - h_\Psi(N(t), P(t)) = 0$$

with W, Z and T being symmetric and positive definite.

Finally, we need to provide a possibility to calculate $\varphi_{\Gamma_S}, \varphi_{\Gamma_O}$ and w_S, w_O since they are solutions of PDEs themselves. First we derive the Galerkin equations for w_S^k and w_O^k from the PDEs (3.12). Let $w_{\Gamma_{D,S}}^k \in C^2(\Omega_S) \cap C^1(\bar{\Omega}_S)$ and $w_{\Gamma_{D,O}}^k \in C^2(\Omega_O) \cap C^1(\bar{\Omega}_O)$ be functions which fulfill the boundary conditions of the auxiliary problems (3.12) with respect to w_S^k and w_O^k, respectively. For the homogenized functions

$$\bar{w}_S^k := w_S^k - w_{\Gamma_{D,S}}^k, \quad \bar{w}_O^k := w_O^k - w_{\Gamma_{D,O}}^k$$

we get the system

$$\int_{\Omega_S} \varepsilon_S \nabla \bar{w}_S^k \cdot \nabla \vartheta_S \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \bar{w}_O^k \cdot \nabla \vartheta_O \mathrm{d}x = \ell_w^k(\vartheta_S, \vartheta_O)$$

with the functional $\ell_w^k(\vartheta_S, \vartheta_O) := \int_{\Omega_S} \nabla \cdot (\varepsilon_S \nabla w_{\Gamma_{D,S}}^k)\vartheta_S \mathrm{d}x + \int_{\Omega_O} \nabla \cdot (\varepsilon_O \nabla w_{\Gamma_{D,O}}^k)\vartheta_O \mathrm{d}x$ to be solved. We denote their associated Galerkin solutions by

$$\bar{w}_h^k(x) = \left(\sum_{i=1}^N \bar{w}_{h,i}^k \vartheta_{S,i}(x), \sum_{i=1}^N \bar{w}_{h,i}^k \vartheta_{O,i}(x) \right) = (\Phi_S(x) \cdot \omega^k, \Phi_O(x) \cdot \omega^k).$$

with

$$\omega^k := (\bar{w}_{h,1}^k, ..., \bar{w}_{h,N}^k)^T.$$

The associated Galerkin equations are then given by

$$\int_{\Omega_S} \varepsilon_S \nabla \Phi_S \cdot \nabla \vartheta_{S,i} \mathrm{d}x \cdot \omega^k + \int_{\Omega_O} \varepsilon_O \nabla \Phi_O \cdot \nabla \vartheta_{O,i} \mathrm{d}x \cdot \omega^k = \ell_w^k(\vartheta_{S,i}, \vartheta_{O,i}) \qquad \forall i = 1, ..., N$$

which means that

$$T\omega^k = \ell_w^k(\Phi_S, \Phi_O) \tag{3.26}$$

has to be solved for the Galerkin coefficients ω^k for $k = 1, ..., n_T$. Analogously, we compute Galerkin approximations of the auxillary functions φ_{Γ_S} and φ_{Γ_O} that are solutions of the PDEs (3.16), (3.17).

Let $\varphi_{\Gamma_{D,S}} \in C^2(\Omega_S) \cap C^1(\bar{\Omega}_S)$ and $\varphi_{\Gamma_{D,O}} \in C^2(\Omega_O) \cap C^1(\bar{\Omega}_O)$ be functions which fulfill the boundary conditions of the auxiliary problems (3.16), (3.17). For the homogenized functions

$$\bar{\varphi}_{\Gamma_S} := \varphi_{\Gamma_S} - \varphi_{\Gamma_{D,S}}, \quad \bar{\varphi}_{\Gamma_O} := \varphi_{\Gamma_O} - \varphi_{\Gamma_{D,O}}$$

we get the system

$$\int_{\Omega_S} \varepsilon_S \nabla \bar{\varphi}_{\Gamma_S} \cdot \nabla \vartheta_S \mathrm{d}x + \int_{\Omega_O} \varepsilon_O \nabla \bar{\varphi}_{\Gamma_O} \cdot \nabla \vartheta_O \mathrm{d}x = \ell_{\varphi_\Gamma}(\vartheta_S, \vartheta_O)$$

with the functional $\ell_{\varphi_\Gamma}(\vartheta_S, \vartheta_O) := \int_{\Omega_S} \nabla \cdot (\varepsilon_S \nabla \varphi_{\Gamma_{D,S}}) \vartheta_S \mathrm{d}x + \int_{\Omega_O} \nabla \cdot (\varepsilon_O \nabla \varphi_{\Gamma_{D,O}}) \vartheta_O \mathrm{d}x$ to be solved. We denote their associated Galerkin solutions by

$$\bar{\varphi}_{\Gamma,h}(x) = \left(\sum_{i=1}^{N} \bar{\varphi}_{\Gamma,h,i} \vartheta_{S,i}(x), \sum_{i=1}^{N} \bar{\varphi}_{\Gamma,h,i} \vartheta_{O,i}(x) \right) = (\Phi_S(x) \cdot \varphi_\Gamma, \Phi_O(x) \cdot \varphi_\Gamma)$$

with

$$\varphi_\Gamma := (\bar{\varphi}_{\Gamma,h,1}, ..., \bar{\varphi}_{\Gamma,h,N})^T.$$

The associated Galerkin equations are then given by

$$\int_{\Omega_S} \varepsilon_S \nabla \Phi_S \cdot \nabla \vartheta_{S,i} \mathrm{d}x \cdot \varphi_\Gamma + \int_{\Omega_O} \varepsilon_O \nabla \Phi_O \cdot \nabla \vartheta_{O,i} \mathrm{d}x \cdot \varphi_\Gamma = \ell_{\varphi_\Gamma}(\vartheta_{S,i}, \vartheta_{O,i}) \qquad \forall i = 1, ..., N$$

which means that

$$T \varphi_\Gamma = \ell_{\varphi_\Gamma}(\Phi_S, \Phi_O) \tag{3.27}$$

has to be solved for the Galerkin coefficients φ_Γ.

Finally, we obtain the discretized system described by a DAE by substituting $w_S(x)$ that appear in (3.25) by $\Phi_S(x) \cdot \omega + w_{\Gamma_{D,S}}(x)$ and by substituting $\varphi_{\Gamma_S}(x)$ that appear in (3.25) via the functionals $\bar{J}_n$ and $\bar{J}_p$ defined in (3.23) by $\Phi_S(x) \cdot \varphi_\Gamma + \varphi_{\Gamma_{D,S}}(x)$:

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(C_S u_s(t)) + g_S(u_s(t), N(t), P(t), \Psi(t)) &= j_S(t), \\
Z \frac{\mathrm{d}}{\mathrm{d}t} N(t) + h_n(u_s(t), N(t), P(t), \Psi(t)) &= 0, \\
Z \frac{\mathrm{d}}{\mathrm{d}t} P(t) + h_p(u_s(t), N(t), P(t), \Psi(t)) &= 0, \\
T \Psi(t) - h_\Psi(N(t), P(t)) &= 0.
\end{aligned}$$

Here, g_S, h_n and h_p represent the functions $\bar{g}_S$, $\bar{h}_n$ and $\bar{h}_p$ where the auxiliary functions $w_S(x)$ and φ_{Γ_S} are replaced by $\Phi_S(x) \cdot \omega + w_{\Gamma_{D,S}}(x)$ and $\Phi_S(x) \cdot \varphi_\Gamma + \varphi_{\Gamma_{D,S}}(x)$, respectively. The matrix C_S is the approximation of W given by the replacement of $w_S(x)$ and $w_O(x)$ by $\Phi_S(x) \cdot \omega + w_{\Gamma_{D,S}}(x)$ and $\Phi_O(x) \cdot \omega + w_{\Gamma_{D,O}}(x)$ in the definition of W, see page 43.
Later on we will need the following assumption for analytic purposes:

Assumption 3.5. The finite element method as well as the methods for the approximations of the involved integrals are such that g_S, h_n, h_p and h_Ψ are continuously differentiable functions of their arguments.

For a more compact notation define

$$\zeta(t) := \begin{pmatrix} N(t) \\ P(t) \end{pmatrix}, \quad M_\zeta := \begin{pmatrix} Z & 0 \\ 0 & Z \end{pmatrix}, \quad h_\zeta(u_s, \zeta, \Psi) := \begin{pmatrix} h_n(u_s, N, P, \Psi) \\ h_p(u_s, N, P, \Psi) \end{pmatrix}$$

and rewrite the discretized system into

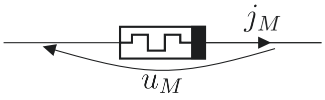
$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(C_S u_s(t)) + g_S(u_s(t), \zeta, \Psi) &= j_S(t), \\ M_\zeta \frac{\mathrm{d}}{\mathrm{d}t}\zeta + h_\zeta(u_s(t), \zeta, \Psi) &= 0, \\ T\Psi + h_\Psi(\zeta) &= 0, \end{aligned} \tag{3.28}$$

which brings us to the end of the modeling of the semiconductor device.

3.1.3 Memristor Model

Next, we add memristor elements to our system. Memristors limit the flow of their current by generating voltage drops which are affected by the history of the current. Let $n_M \in \mathbb{N}$ be the number of the memristors then we call $\phi_M : \mathbb{R}^{n_M} \times \mathcal{I} \to \mathbb{R}^{n_M}$ the characteristic function of the memristors. The function ϕ_M describes magnetic flux of the resistors. With the characteristic function ϕ_M we can formulate a relation between the charges q_M and the voltages u_M of the memristors. Further we present the electric symbol of a memristor.

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\phi_M(q_M, t) = u_M$$



with $j_M, u_M \in R^{n_R}$ being the currents and voltages of the memristors. Assume that ϕ_M is continuously differentiable with the Jacobian $M(q_M, t) := \frac{\partial}{\partial y}\phi_M(y, t)$ being non-singular. By the relation between the charge and the current $j_M = \frac{\mathrm{d}}{\mathrm{d}t}q_M$ we obtain the expression

$$j_M = M(q_M, t)^{-1}\left(u_M - \frac{\partial}{\partial t}\phi_M(q_M, t)\right) =: g_M(u_M, q_M, t) \tag{3.29}$$

for j_M in terms of the voltage u_M and the charges of the memristors q_M. Hence the memristor is a resistor-like element with memory.

As an example for a memristor we present the HP memristor [SSSW08]. Therefore define the dopant mobility $\mu_V = 10^{-13}\frac{m^2}{Vs}$, the total device length $d = 10^{-8}m$ and the limits of the memristor resistance $R_{off} = 36 \cdot 10^3\Omega$ and $R_{on} = 1 \cdot 10^2\Omega$ to write the functions

$$\phi_M(q) = R_{off}(q - \frac{\mu_V R_{on}}{2d^2}q^2)$$

and

$$g_M(u,q) = M(q)^{-1}u$$

with the Jacobian

$$M(q,t) = M(q) = R_{off}(1 - \frac{\mu_V R_{on}}{d^2}q).$$

3.1.4 Electromagnetic Device Model

This section introduces the electromagnetic device model. In contrast to the semiconductor device model we will use an already discretized model from the literature. In particular we are interested in models discretized by the Finite Integration Technique(FIT). The FIT discretization is an established tool to discretize electromagnetic devices which was developed and formulated by Thomas Weiland [Wei77, TW96, Yee66, CW01]. This discretization method yields properties which allow a structural classification of the electromagnetic device model as one of the lumped elements, analogous to the semiconductor case. Such a discretized electromagnetic model is developed in [Bau12]. The model in [Bau12] arises from the full Maxwell Equations spatially discretized with the FIT.

We call the discretized electric field $E \in \mathbb{R}^{3n}$ and the discretized magnetic flux density $B \in \mathbb{R}^{3n}$ with n depending on the refinement of the FIT discretization. Further we call $A \in \mathbb{R}^{3n}$ and $\phi \in \mathbb{R}^n$ the discretized vector and scalar potential while $M_\varepsilon, M_\sigma, M_\nu \in \mathbb{R}^{3n\times 3n}$ represent the three material properties for the permittivity, the conductivity and the reluctivity. The reluctivity of the device is set to be constant in contrast to [Bau12] to simplify the structural classification of the electromagnetic device model. Of course we exclude some materials by this simplification.

The discretized versions of the differential operators are notated with $G \in \mathbb{R}^{3n\times n}$ in the case of the gradient, $\tilde{S} \in \mathbb{R}^{n\times 3n}$ in the case of the divergence and $C \in \mathbb{R}^{3n\times 3n}$ in the case of the rotation operator. Last we define the excitation matrix $\Lambda \in \mathbb{R}^{3n\times n_\Gamma}$ which represents the boundary operator, meaning each column of Λ is the sum of the outer normal vectors at each point of the discretization grid belonging to the related contact area with n_Γ the number of contact areas. Furthermore the transposed excitation matrix $\Lambda^\top \in \mathbb{R}^{n_\Gamma \times 3n}$ represents the integral over the contact areas. The discretized operators and matrices of the FIT discretization fulfill a set of important properties, see [Wei77, TW96, Yee66, Bau12, Sch11]. In particular the discretized material relations M_ε and M_ν are positive definite diagonal matrices while M_σ is a positive semi-definite diagonal matrix. Furthermore $C\Lambda$ has full column rank and the equality $\nabla \times \nabla = 0$ is inherited by the discretized operators $CG = 0$, see [Bau12] Lemma 3.25 and Lemma 3.28. In [Bau12] the discretized model is written in the FIT potential formulation:

$$\begin{aligned} j_E - \Lambda^\top C^\top M_\nu C A &= 0, \\ \vartheta \tilde{S} M_\varepsilon G \frac{\mathrm{d}}{\mathrm{d}t}\phi + \tilde{S} M_{\bar{\nu}} A &= 0, \\ M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} G\phi + M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t}\pi + M_\sigma G\phi + C^\top M_\nu C A + M_\sigma \pi - M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t}\Lambda u_E - M_\sigma \Lambda u_E &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} A - \pi &= 0, \end{aligned} \tag{3.30}$$

with the discrete artificial material matrix $M_{\bar{\nu}}$ and $\vartheta = 0$ if we choose the Coulomb gauge or $\vartheta = 1$ if we choose the Lorenz gauge. Here j_E are the currents and u_E are the potentials at the contact areas. While the contact areas are connected to a circuit via electric wires the rest of the boundary of the electromagnetic device is grounded. In [Bau12] it was shown that the sum of the incoming and outgoing total currents over all boundary parts equals zero. Therefore the currents at the non-conductive boundary parts can be expressed as the negative sum of the currents at the contact areas.

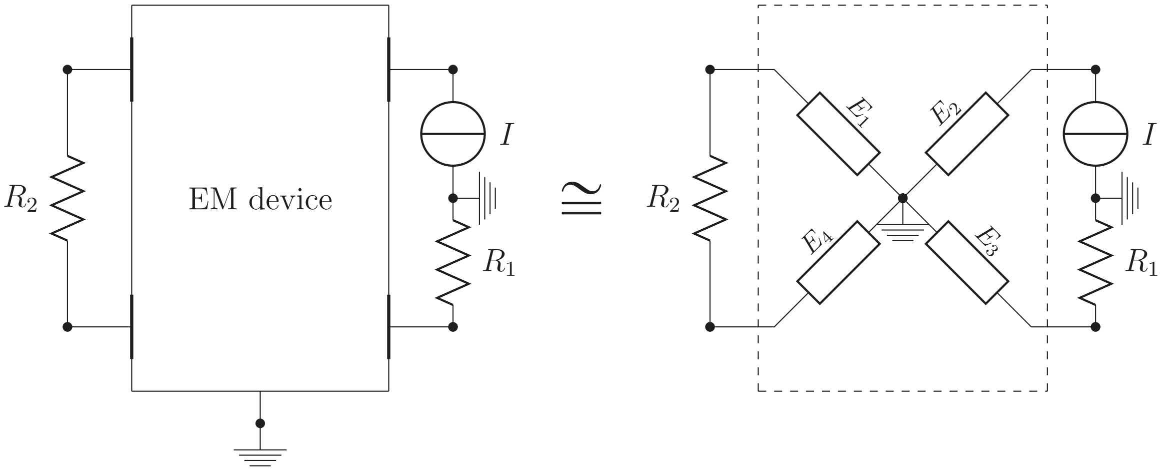


Figure 3.3: Representation of a electromagnetic device with four contact areas by bipolar circuit elements.

This behavior enables us to describe the electromagnetic device by n_Γ bipolar elements. In Figure 3.3 we see an electromagnetic device with four contact areas on the left which is alternatively represented by four branches E_1, E_2, E_3 and E_4. These four branches correspond to the four currents at the contact areas.

For our purposes it is convenient to switch the discretized electromagnetic device model back to the field formulation, i.e. a formulation in the electric field and the magnetic density. Therefore consider the discretized electric field and the discretized magnetic density in terms of the vector and scalar potential:

$$\begin{aligned} E &:= -G\phi + \Lambda u_E - \frac{\mathrm{d}}{\mathrm{d}t}A \\ B &:= CA \end{aligned}$$

Together the equations of the electromagnetic device can be written as

$$j_E = \Lambda^\top C^\top M_\nu B \tag{3.31a}$$

$$M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t}E + M_\sigma E = C^\top M_\nu B \tag{3.31b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}A = -G\phi + \Lambda u_E - E \tag{3.31c}$$

$$B = CA \tag{3.31d}$$

$$0 = \vartheta \tilde{S} M_\varepsilon G \frac{\mathrm{d}}{\mathrm{d}t}\phi + \tilde{S} M_{\bar{\nu}} A. \tag{3.31e}$$

And after applying the discretized rotation operator C to Equation (3.31c) and dropping the equations (3.31d) and (3.31e) we obtain the system in the FIT field formulation:

$$j_E = \Lambda^\top C^\top M_\nu B \tag{3.32a}$$

$$M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t}E + M_\sigma E = C^\top M_\nu B \tag{3.32b}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}B = -CE + C\Lambda u_E. \tag{3.32c}$$

FIT was also applied to a formulation in E and B in [Yee66, Wei77]. As in the semiconductor case we like to structurally classify the electromagnetic device as one of the lumped elements. Therefore we remodel the current coupling equation (3.31a) by differentiating it to work out a relation between $\frac{\mathrm{d}}{\mathrm{d}t}j_E$ and u_E:

$$\frac{\mathrm{d}}{\mathrm{d}t}j_E = \Lambda^\top C^\top M_\nu \frac{\mathrm{d}}{\mathrm{d}t}B$$

Insert the discretized Maxwell-Faraday law (3.32c) into the derived current coupling equation and get

$$\begin{aligned} &\frac{\mathrm{d}}{\mathrm{d}t}j_E = \Lambda^\top C^\top M_\nu \frac{\mathrm{d}}{\mathrm{d}t}B \\ \Leftrightarrow &\frac{\mathrm{d}}{\mathrm{d}t}j_E = \Lambda^\top C^\top M_\nu (C\Lambda u_E - CE) \\ \Leftrightarrow &\frac{\mathrm{d}}{\mathrm{d}t}j_E = \Lambda^\top C^\top M_\nu C\Lambda u_E - \Lambda^\top C^\top M_\nu CE. \end{aligned}$$

We remember that $C\Lambda$ has full column rank and M_ν is positive definite. Therefore $\Lambda^\top C^\top M_\nu C\Lambda$ is also positive definite.

Remark. The matrix $\Lambda^\top C^\top M_\nu C\Lambda$ is a diagonal matrix, if the spatial discretization is fine enough, i.e. there are at least three finite volumes between all contact areas.

We define $L_E := (\Lambda^\top C^\top M_\nu C\Lambda)^{-1}$ and $\chi_E := L_E\Lambda^\top C^\top M_\nu C$ and write

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} j_E = L_E^{-1} u_E - \Lambda^\top C^\top M_\nu C E \\
\Leftrightarrow &\frac{\mathrm{d}}{\mathrm{d}t}(L_E j_E) = u_E - L_E\Lambda^\top C^\top M_\nu C E \\
\Leftrightarrow &\frac{\mathrm{d}}{\mathrm{d}t}(L_E j_E) - u_E + \chi_E E = 0
\end{aligned}$$

Hence we end up with the following set of equations

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(L_E j_E) - u_E + \chi_E E &= 0 \\
M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} E + M_\sigma E - C^\top M_\nu B &= 0 \\
\frac{\mathrm{d}}{\mathrm{d}t} B + C E - C\Lambda u_E &= 0,
\end{aligned}$$

and we notice that the current coupling term has the structure of an inductor. The matrix L_E can be interpreted as the inductance of the EM device and for L_E holds:

Lemma 3.6.
The matrix L_E is symmetric and positive definite.

Proof*.*
The matrix M_ν is a positive definite diagonal matrix and $C\Lambda$ has full column rank. □

Due to its inductor-like structure we anticipate that parts of the potentials of the nodes connected to an electromagnetic device might be involved in a differentiation problem, since the topological index conditions of a circuit state that a cutset of inductors and current sources leads to an differentiation problem of order one, see [Tis99]. To avoid a coupling by components involved in a differentiation problem change the coupling term by multiplying Maxwell-Faraday's law (3.32c) by $C^\top M_\nu$

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(L_E j_E) - u_E + \chi_E E &= 0 \\
M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} E + M_\sigma E - C^\top M_\nu B &= 0 \\
\frac{\mathrm{d}}{\mathrm{d}t}(C^\top M_\nu B) + C^\top M_\nu C E - C^\top M_\nu C\Lambda u_E &= 0
\end{aligned}$$

and define the auxiliary current density

$$J := C^\top M_\nu B - \chi_E^T j_E.$$

Furthermore we define the curl-curl matrix

$$M_{CC} := C^\top M_\nu C - \chi_E^T \Lambda^\top C^\top M_\nu C \Lambda \chi_E = C^\top M_\nu C - \chi_E^T L_E^{-1} \chi_E$$

for a more compact notation and obtain the FIT inductor-like formulation with a coupling from the circuit to the electromagnetic device via the currents at the contact areas:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}(L_E j_E) - u_E + \chi_E E &= 0, \\ M_\varepsilon \frac{\mathrm{d}}{\mathrm{d}t} E + M_\sigma E - J - \chi_E^T j_E &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}t} J + M_{CC} E &= 0. \end{aligned} \tag{3.33}$$

With this formulation we complete the modeling of the electromagnetic device.

3.1.5 Modified Nodal Analysis

In this section we join all the elements of the previous sections together into a network framework. Therefore we use the Modified Nodal Analysis(MNA). The classical MNA deals with capacitors, resistors, inductors, voltage and current sources as electric elements, see [CL75, CDK87, DK84]. The equations of the MNA arises as we rearrange the incidence Matrix to

$$A = \begin{pmatrix} A_C & A_R & A_L & A_V & A_I \end{pmatrix},$$

with A_C, A_R, A_L, A_V and A_I the incidence matrices of the capacitors, resistors, inductors, voltage sources and current sources. We also split the current with respect to these elements and obtain with Kirchhoff's first law (3.1)

$$A_C j_C + A_R j_R + A_L j_L + A_V j_V + A_I i_I = 0.$$

Next we insert the characteristic functions of the capacitors, resistors and current sources, add the inductor and voltage source equations and also replace the voltages by the electric node potentials with Kirchhoff's second law (3.2) for each element. Then the well known MNA can be formulated based on Kirchhoff's current law, Kirchhoff's voltage law and the physical element relations, see [Tis99].

$$A_C \frac{\mathrm{d}}{\mathrm{d}t} Q_C(A_C^\top e, t) + A_R g_R(A_R^\top e, t) + A_L j_L + A_V j_V + A_I i_s(t) = 0,$$

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\phi_L(j_L,t) - A_L^\top e &= 0,\\ A_V^\top e - v_s(t) &= 0.\end{aligned}$$

As a basic physical assumption the branch elements should not produce any energy on their own. This can be mathematically covered by the next assumption.

Assumption 3.7. (Passive Elements)
We assume the characteristic functions $q_C(u,t), g_R(u,t)$ and $\phi_L(j,t)$ to be continuously differentiable with the Jacobians

$$C(u,t) := \frac{\partial}{\partial u} q_C(u,t), \quad G(u,t) := \frac{\partial}{\partial u} g_R(u,t) \text{ and } L(j,t) := \frac{\partial}{\partial j}\phi_L(j,t)$$

being positive definite.

Furthermore we assume that the circuit is connected and not shorted.

Assumption 3.8.
Let A_V have full column rank and let $\begin{pmatrix} A_C & A_R & A_L & A_V \end{pmatrix}$ have full row rank.

To extend the MNA to semiconductor devices, memristors and electromagnetic devices sort the network edges like before in such a way that the incidence matrix A forms a block matrix with blocks describing the different types of network elements, that is,

$$A = \begin{pmatrix} A_C & A_S & A_R & A_M & A_L & A_E & A_V & A_I \end{pmatrix}.$$

For a more compact notation it is convenient to group the capacitor-like elements, the resistor-like elements and the inductor-like elements together by defining

$$A_{\mathcal{C}} := \begin{pmatrix} A_C & A_S \end{pmatrix}, \quad q_{\mathcal{C}}(A_{\mathcal{C}}^\top e, t) := \begin{pmatrix} q_C(A_C^\top e, t) \\ C_S A_S^\top e \end{pmatrix}, \quad g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) := \begin{pmatrix} 0 \\ g_S(A_S^\top e, \zeta, \Psi) \end{pmatrix}$$

and

$$A_{\mathcal{R}} := \begin{pmatrix} A_R & A_M \end{pmatrix}, \quad g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t) := \begin{pmatrix} g_R(A_R^\top e, t) \\ g_M(A_M^T e, q_M, t) \end{pmatrix}$$

and

$$A_{\mathcal{L}} := \begin{pmatrix} A_L & A_E \end{pmatrix}, \quad j_{\mathcal{L}} := \begin{pmatrix} j_L \\ j_E \end{pmatrix}, \quad \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) := \begin{pmatrix} \phi_L(j_L, t) \\ L_E j_E \end{pmatrix}, \quad \chi_{\mathcal{L}} := \begin{pmatrix} 0 \\ \chi_E \end{pmatrix}.$$

As before we use the two Kirchhoff laws (3.1) and (3.2). Additionally we use (3.28), (3.29) and (3.33) to obtain:

$$\begin{aligned}
A_{\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^\top e,t)+g_{\mathcal{C}}(A_{\mathcal{C}}^\top e,\zeta,\Psi)\right)+A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^\top e,q_M,t)+A_{\mathcal{L}}j_{\mathcal{L}}+A_Vj_V+A_Ii_s(t)&=0,\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\mathcal{L}}(j_{\mathcal{L}},t)-A_{\mathcal{L}}^\top e+\chi_{\mathcal{L}}E&=0,\\
A_V^\top e-v_s(t)&=0,\\
M_\zeta\frac{\mathrm{d}}{\mathrm{d}t}\zeta+h_\zeta(A_S^\top e,\zeta,\Psi)&=0,\\
T\Psi(t)-h_\Psi(\zeta)&=0,\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_M(q_M,t)-A_M^Te&=0,\\
M_\varepsilon\frac{\mathrm{d}}{\mathrm{d}t}E+M_\sigma E-J-\chi_{\mathcal{L}}^Tj_{\mathcal{L}}&=0,\\
\frac{\mathrm{d}}{\mathrm{d}t}J+M_{CC}E&=0
\end{aligned}\tag{3.34}$$

with $t \in \mathcal{I}$ and $\mathcal{I}$ a compact time interval. We call (3.34) the extended MNA.
The matrices C_S and L_E are positive definite for the models of the semiconductor and electromagnetic devices investigated in the last sections, see Lemma 3.2 and Lemma 3.6. Therefore Assumption 3.7 must only be extended to:

Assumption 3.9.
We assume the characteristic functions $q_C(u,t), g_R(u,t), \phi_M(q,t)$ and $\phi_L(j,t)$ to be continuously differentiable with the Jacobians

$$\begin{aligned}
C(u,t) &:= \frac{\partial}{\partial u}q_C(u,t), \quad G(u,t) := \frac{\partial}{\partial u}g_R(u,t)\\
M(q,t) &:= \frac{\partial}{\partial q}\phi_M(q,t) \text{ and } L(j,t) := \frac{\partial}{\partial j}\phi_L(j,t)
\end{aligned}$$

being positive definite.

Analogously Assumption 3.8 now reads:

Assumption 3.10.
Let A_V have full column rank and let $\begin{pmatrix}A_{\mathcal{C}} & A_{\mathcal{R}} & A_{\mathcal{L}} & A_V\end{pmatrix}$ have full row rank.

Thereby we finish the circuit modeling sections.

3.2 Mechanical Applications

In this section we present the mechanical applications, in particular multibody applications. In [Sim95, ESF98, Ste06] a classification of several forms of the equations of motion is given. In this section we will shortly recall the modeling levels 0 and 1 [Sim95].
The modeling level 0 is based on the standard formulation of equations of motion. For modeling level 0 consider the position variables $p(t) \in \mathbb{R}^{n_p}$, the velocity variables $v(t) \in \mathbb{R}^{n_v}$ and the Lagrange multipliers $\lambda(t) \in \mathbb{R}^{n_\lambda}$. Here for the dimensions of the positions and the velocities it holds $n_p = n_v$. Then the equations of motion of modeling level 0 are given by

$$p' = v \tag{3.35a}$$

$$M(p)v' = f(p,v,t) - G^\top(p)\lambda \tag{3.35b}$$

$$0 = g(p) \tag{3.35c}$$

with the initial values

$$p(t_0) = p_0,\ v(t_0) = v_0,\ \lambda(t_0) = \lambda_0,$$

on the domain $\mathcal{I} = [t_0, T]$. The n_p equations (3.35a) are called kinematic equations of motion. Furthermore, the equations of motion are affected by the so called holonomic constraints (3.35c). Holonomic constraints never influence the velocities. From the constraints $g(p) = 0$ one obtains the constraint matrix $G(p) = \frac{\partial}{\partial p} g(p)$ which column-wise contains the inaccessible directions of motion. The n_v equations (3.35b) are called dynamical equations of motion. They can be derived from the equilibrium of forces and momenta and include the mass matrix $M(p)$, the vector $f(p,v,t)$ of the applied and gyroscopic forces, the constraint matrix $G(p)$ of the holonomic constraints, the associated constraint forces $G^T(p)\lambda$, and the Lagrange multipliers λ. The mass matrix $M(p)$ is positive semi-definite, since the kinetic energy is a positive semi-definite quadratic form, and it includes the inertia properties of the multibody system.

In the modeling level 1 case we deal with spatial multibody systems with dynamical force elements which are influenced by friction effects. The influence of the friction is modeled as an applied force such that f additionally depends on the Lagrange multipliers λ. The dynamical force elements, like multibody systems with additional control devices, hydraulic or electromagnetic components, are modeled by new variables r, which specify the state of such dynamical force elements by an ordinary differential equation

$$r' = b(p, v, r, t),$$

c.f. [ESF98]. In the case of spatial multibody systems, which are discussed in [ESF98], it is possible that $n_p < n_v$ and therefore we need a transformation matrix $Z(p) \in \mathbb{R}^{n_p \times n_v}$, which relates the position variables to the velocities. The transformation matrix $Z(p)$ is not the identity I_{n_p} if there are rotations in three dimensional space. In the two dimensional case we have $Z(p) = I_{n_p}$, i.e. $p' = v$. Note that the transformation matrix $Z(p)$ mainly depends on the choice of the velocity vector. Different choices for the velocity vector are presented in [Ami92, Hau89, RR88, Whi59]. Summing up, the equations of motion of modeling level 1 have the form

$$\begin{aligned} p' &= Z(p)v, && (3.36a)\\ M(p,t)v' &= f(p,v,r,\lambda,t) - Z^\top(p)G^\top(p,t)\lambda, && (3.36b)\\ r' &= b(p,v,r,\lambda,t), && (3.36c)\\ 0 &= g(p,t) && (3.36d) \end{aligned}$$

with the initial values

$$p(t_0) = p_0,\ v(t_0) = v_0,\ r(t_0) = r_0,\ \lambda(t_0) = \lambda_0.$$

We have the initial value problem for the equations of motion of modeling level 1 on the domain $\mathcal{I} = [t_0, T]$. Note in addition, that in contrast to [Sim95] for reasons of symmetry, the dynamical equations of motion are multiplied by the transformation matrix $Z(p)$, implicitly contained in $M(p,t)$ and $f(p,v,r,t)$.

We close this section by a set of additional assumptions, which bound the index of mechanical applications by 3, see Section 4.5.

Assumption 3.11.

The mass matrix $M(p,t)$ and the block matrix

$$\begin{pmatrix} M(p,t) & Z(p)^\top G(p,t)^\top - \frac{\partial}{\partial\lambda} f(p,v,r,\lambda,t) \\ G(p,t)Z(p) & 0 \end{pmatrix}$$

are non-singular. This yields that $G(p,t)Z(p)$ has full row rank and that the Schur-complement

$$G(p,t)Z(p)M^{-1}(p,t)G_\lambda(p,v,r,\lambda,t)$$

is non-singular with $G_\lambda(p,v,r,\lambda,t) := Z(p)^\top G(p,t)^\top - \frac{\partial}{\partial\lambda} f(p,v,r,\lambda,t)$.

3.3 Summary and Outlook

We introduced two application fields for DAEs in this chapter. The main focus regarding the applications in this work will be the analysis of electrical circuits. The classical modified nodal analysis deals with capacitors, resistors, inductors, voltage and current sources as network elements. We extended the list of elements by semiconductor devices, memristors and electromagnetic devices. In particular we were able to structurally identify the new elements with the basic ones.

This identification will be used in Chapter 5 to prove global existence and uniqueness results for circuits including the semi-discretized semiconductor devices, the memristors and the semi-discretized electromagnetic devices. Furthermore a topological decoupling for DAEs arising from electrical circuits will be derived in Chapter 7. The most important properties of the topological decoupling will be its cheap calculation and the applicability of half-explicit methods to the decoupled DAE.

To obtain the existence results and the topological decoupling we introduce a new index concept in Chapter 4. The presented mechanical applications will be analyzed with this index concept such that we can described the influence of perturbation onto these applications.

4 The Concept of the Dissection Index

Both the Tractability Index and the Strangeness Index are important tools for the analysis of DAEs. The Tractability Index and the Strangeness Index are decoupling procedures which analyze the structure of a DAE. The Strangeness Index concept includes over- and under determined systems and it is well suited for the analysis of DAEs in Hessenberg-form, cf. [KM06]. The main assets of the Tractability Index are its low smoothness assumptions and its step-by-step approach, cf. [LMT13].
Here we introduce a new index concept, which can be interpreted as a mix of the Tractability Index and the Strangeness Index. The index arises as we use the linearization concept of the Tractability Index and the decoupling procedure of the Strangeness Index.
But before we introduce this mixed index concept we need a good reason to do so, since the definition of an index concept entails much technical work. We already mentioned in Section 2.3 that the Strangeness Index requires too much differentiability and that its calculation may be laboriously since it has no step-by-step approach. While the Tractability Index does not share these weaknesses the projector chain of the Tractability Index tends to become unnecessarily complex, which we demonstrate in the following examples:

Example 4.1.
Let $\mathcal{I} := [t_0, T] \subset \mathbb{R}$ be a compact time interval and let $t \in \mathcal{I}$. Let $f : \mathcal{I} \to \mathbb{R}$ be continuously differentiable.

$$(1+t^2)x' = y \tag{4.1a}$$
$$x = f(t) \tag{4.1b}$$

The solution of Example 4.1 can be explicitly given by:

$$\begin{aligned} x(t) &= f(t) \\ y(t) &= (1+t^2)f'(t). \end{aligned}$$

Here x is algebraically defined by equation (4.1b) and y is defined by the differentiation problem in equation (4.1a). An index concept should reflect the simple structure of Example 4.1 by using only constant projectors or basis functions to analyze it. Especially for a decoupling procedure simple projectors or basis functions are essential.
In contrast to our wish for constant operators both the matrix chain of the Tractability Index and the basis functions of the Strangeness Index are time dependent. First we

consider the matrix chain of the Tractability Index and start by denoting

$$G_0 = \begin{pmatrix} 1+t^2 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

This yields the first two projectors

$$Q_0 = \begin{pmatrix} 0 & 0 \\ c & 1 \end{pmatrix} \quad \text{and} \quad P_0 = \begin{pmatrix} 1 & 0 \\ -c & 0 \end{pmatrix}$$

with an arbitrary continuous function c. We obtain

$$G_1 = G_0 + B_0 Q_0 = \begin{pmatrix} 1+t^2-c & -1 \\ 0 & 0 \end{pmatrix}, \quad B_1 = \begin{pmatrix} c & 0 \\ 1 & 0 \end{pmatrix} \text{ and } Q_1 = \begin{pmatrix} 1 & 0 \\ 1+t^2-c & 0 \end{pmatrix},$$

which yields the non-singular matrix

$$G_2 = G_1 + B_1 Q_1 = \begin{pmatrix} 1+t^2 & -1 \\ 1 & 0 \end{pmatrix}.$$

Hence Example 4.1 is of Tractability Index 2, but there is no admissible matrix chain with only constant projectors since for every continuous function c at least one of the projectors Q_0 or Q_1 is time dependent.

The Strangeness Index also cannot go without time dependent basis functions. It is sufficient to consider the Jacobian G_1 of the inflated system, which is given by

$$G_1 = \begin{pmatrix} 1+t^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 2t & -1 & 1+t^2 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}.$$

The image of this Jacobian is time dependent, hence the matrix W is also time dependent. Aside from practical reasons like the efficient implementation of a decoupling procedure, these time dependencies may even invoke extra smoothness conditions. This is in particular hurtful for the tractability concept, since one of its greatest strengths are its minimal smoothness conditions. Indeed the Tractability Index does not need any differentiability of the right hand side for its definition, but it does need the differentiability of the $D\Pi_i D^-$ terms. While the differentiability of the right hand side may be necessary for the solvability of the DAE in case of a differentiation problem, the differentiability of the $D\Pi_i D^-$ terms may be completely needless.

We demonstrate this problem with an example from the circuit simulation.

Example 4.2.
We consider the electric circuit in Figure 4.1 with two inductors, which both have an inductance of $L_1 = L_2 = 1$, an independent current source with a continuously differentiable function i_s and a controlled voltage source defined by $v_s(e_1) = e_1 - a(e_1)$ with $a \in C^1(\mathbb{R}, \mathbb{R})$, $\frac{\partial}{\partial e_1} a(e_1) > 1$. We consider the time interval $\mathcal{I} = [0, 1]$. The node potential e_1 and the currents j_1 and j_2 through the inductors are the solution of the equations:

$$j_1' + e_1 = 0 \tag{4.2a}$$
$$j_2' + a(e_1) = 0 \tag{4.2b}$$
$$j_1 + j_2 + i_s(t) = 0. \tag{4.2c}$$

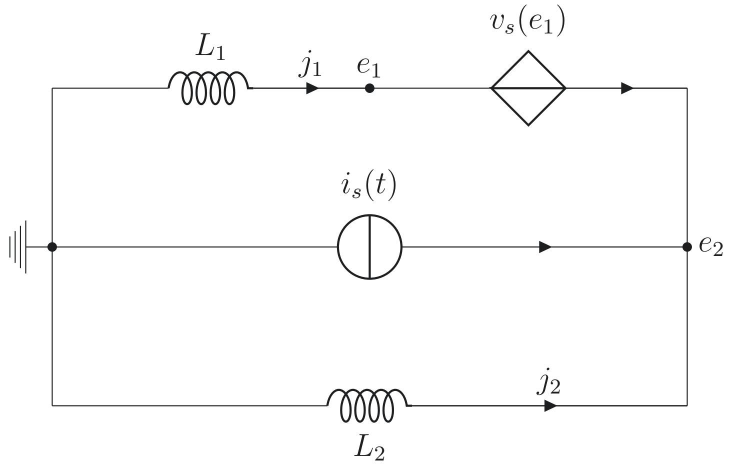


Figure 4.1: Electric circuit including a controlled voltage source.

By multiplying the matrices $\begin{pmatrix} 1 & 1 \end{pmatrix}$ and $\begin{pmatrix} 1 & -1 \end{pmatrix}$ to the left of the first two equations we obtain:

$$(j_1 - j_2)' + e_1 - a(e_1) = 0 \tag{4.3}$$
$$e_1 + a(e_1) = -(j_1 + j_2)' \tag{4.4}$$
$$j_1 + j_2 + i_s(t) = 0, \tag{4.5}$$

which yields the necessity of the differentiability of i_s, which was already assumed above. By inserting (4.5) in (4.4) and denoting

$$\begin{pmatrix} j_1 \\ j_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{pmatrix} \bar{j} + \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \tilde{j}$$

we get $\bar{j} := j_1 - j_2$ and $\tilde{j} = i_s(t)$, hence we obtain a system in $\bar{j}$ and e_1:

$$\begin{aligned} \bar{j}' + e_1 - a(e_1) &= 0 \\ e_1 + a(e_1) &= i_s'(t). \end{aligned}$$

The function $f(e_1) := e_1 + a(e_1)$ is strongly monotone due to $\frac{\partial}{\partial e_1}(e_1 + a(e_1)) = 1 + a'(e_1) > 2$ and therefore bijective and there is an inverse function Ψ such that

$$\begin{aligned} \bar{j} &= \int_0^t a(\Psi(i_s'(s))) - \Psi(i_s'(s))\mathrm{d}s + \bar{j}(0) \\ e_1 &= \Psi(i_s'(t)). \end{aligned}$$

In particular we notice that it is not necessary that a is two times differentiable for (4.2) to be solvable. In contrast we need a to be two times differentiable for the Tractability Index to be well defined. Define

$$A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & a'(e_1) \\ 1 & 1 & 0 \end{pmatrix}$$

and thereby obtain

$$G_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad N_0 = \text{span}(\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}).$$

This yields the first two projectors

$$Q_0 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ c_1 & c_2 & 1 \end{pmatrix} \quad \text{and} \quad P_0 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -c_1 & -c_2 & 0 \end{pmatrix}$$

depending on two continuous functions c_1 and c_2. With the help of Q_0 we obtain

$$G_1 = \begin{pmatrix} 1 + c_1 & c_2 & 1 \\ c_1 a'(e_1) & 1 + c_2 a'(e_1) & a'(e_1) \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad N_1 = \text{span}(\begin{pmatrix} 1 \\ a'(e_1) \\ -1 - c_1 - c_2 a'(e_1) \end{pmatrix}).$$

Then N_1 and N_0 yield the projectors

$$Q_1 = \begin{pmatrix} 1 - a'(e_1)c_3 & c_3 & 0 \\ a'(e_1)(1 - a'(e_1)c_3) & a'(e_1)c_3 & 0 \\ (-1 - c_1 - c_2 a'(e_1))(1 - a'(e_1)c_3) & (-1 - c_1 - c_2 a'(e_1))c_3 & 0 \end{pmatrix}$$

and

$$P_1 = \begin{pmatrix} a'(e_1)c_3 & -c_3 & 0 \\ a'(e_1)(a'(e_1)c_3 - 1) & 1 - a'(e_1)c_3 & 0 \\ (-1 - c_1 - c_2a'(e_1))(a'(e_1)c_3 - 1) & (1 + c_1 + c_2a'(e_1))c_3 & 1 \end{pmatrix}$$

with a continuous function c_3. In order to continue the matrix chain we need to calculate the derivative of $D\Pi_1 D^- = DP_1D^-$ with the generalized inverse D^- of D. In this case D^- is given by:

$$D^- = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ -c_1 & -c_2 \end{pmatrix}.$$

Hence we can calculate

$$DP_1D^- = \begin{pmatrix} a'(e_1)c_3 & -c_3 \\ a'(e_1)(a'(e_1)c_3 - 1) & 1 - a'(e_1)c_3 \end{pmatrix}.$$

The Tractability Index requires DP_1D^- to be continuously differentiable, hence it is only defined, if the functions c_3, $a'(e_1)c_3$, $(a'(e_1))^2c_3 - a'(e_1)$ and $1 - a'(e_1)c_3$ are continuously differentiable. In general the function c_3 could depend on all state variables and the time, but for the differentiability of DP_1D^- it is sufficient to consider a function c_3 which depends on e_1. We split $\mathbb{R} = \mathbb{R}_0 \bigcup \mathbb{R}_C$ with $\mathbb{R}_0 := \{e_1 \in \mathbb{R} |\ c_3(e_1) = 0\}$ and $\mathbb{R}_C := \{e_1 \in \mathbb{R} |\ c_3(e_1) \neq 0\}$. The set $\mathbb{R}_C$ is open, since c_3 is continuous. Hence c_3 and $a'(e_1)c_3$ being continuously differentiable yields that $a'(e_1)$ is continuously differentiable on $\mathbb{R}_C$. Let $\bar{x} \in \mathbb{R}_0$, then we obtain the existence of the following limit by $(a'(e_1))^2c_3 - a'(e_1)$ being continuously differentiable:

$$\begin{aligned} &\lim_{h\to 0} \frac{(a'(\bar{x}+h))^2c_3(\bar{x}+h) - a'(\bar{x}+h) - ((a'(\bar{x}))^2c_3(\bar{x}) - a'(\bar{x}))}{h} \\ &= \lim_{h\to 0} (\frac{(a'(\bar{x}+h))^2c_3(\bar{x}+h)}{h} - \frac{a'(\bar{x}+h) - a'(\bar{x})}{h}) \\ &= \lim_{h\to 0} (\frac{(a'(\bar{x}+h))^2(c_3(\bar{x}) + c_3'(\xi)h)}{h} - \frac{a'(\bar{x}+h) - a'(\bar{x})}{h}) \\ &= \lim_{h\to 0} ((a'(\bar{x}+h))^2c_3'(\xi) - \frac{a'(\bar{x}+h) - a'(\bar{x})}{h}) \end{aligned}$$

with $\xi \in (\bar{x}, \bar{x}+h)$. With a' being continuous this yields the existence of

$$\lim_{h\to 0} \frac{a'(\bar{x}+h) - a'(\bar{x})}{h},$$

hence a' is differentiable at all points $\bar{x} \in \mathbb{R}_0$. Therefore the Tractability Index is only defined for (4.2), if a is two times differentiable.
The same behavior can also be observed for circuits without controlled elements.

Example 4.3.

We consider the electric circuit in Figure 4.2 with two capacitors, two resistors, one inductor and one current source. Let the conductance of both resistors be $G_1 = G_2 = 1$ and also the inductance of the inductor and the capacitance of the second capacitor be $L = C_2 = 1$. Let i_s be an arbitrary continuously differentiable function and define the characteristic curve of the first capacitor and the function q_C by:

$$q_{C_1}(x) := \begin{cases} x, & x < 1 \\ 0.5x^2 + 0.5, & 1 \leqslant x < 2 \\ 2x - 1.5, & 2 \leqslant x \end{cases} \qquad \text{and} \qquad q_C(x) = \begin{pmatrix} q_{C_1}(x_1) \\ x_2 \end{pmatrix}.$$

Then the equations of the MNA are given by:

$$\begin{aligned} A_C \frac{\mathrm{d}}{\mathrm{d}t} q_C(A_C^T e) + A_R A_R^T e + A_L j_L + A_I i_s(t) &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t} j_L - A_L^T e &= 0 \end{aligned}$$

with

$$A_C = \begin{pmatrix} 0 & 0 \\ 1 & 1 \\ -1 & -1 \\ 0 & 0 \end{pmatrix}, \quad A_R = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{pmatrix}, \quad A_L = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad A_I = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

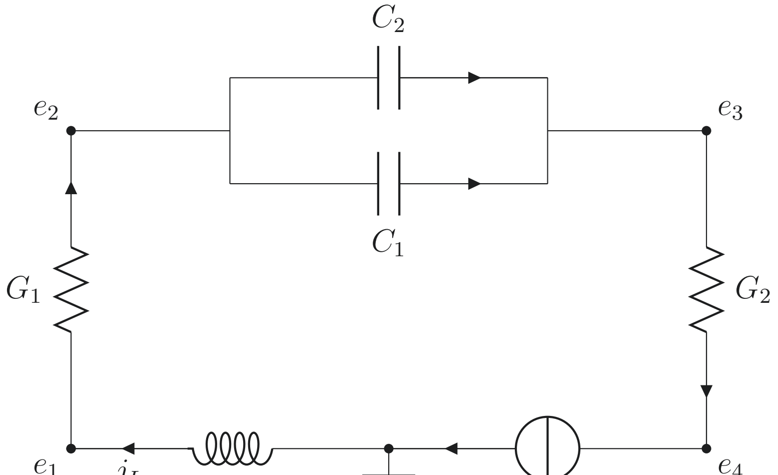


Figure 4.2: Electric circuit without controlled elements.

The Jacobian of q_{C_1} is given by

$$C_1(x) := \begin{cases} 1, & x < 1 \\ x, & 1 \leqslant x < 2 \\ 2, & 2 \leqslant x. \end{cases}$$

Furthermore we denote the matrices

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad D = \begin{pmatrix} 0 & C_1(e_2 - e_3) & -C_1(e_2 - e_3) & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and

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

Therefore we get the first G of the matrix chain

$$G_0 = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & C & -C & 0 & 0 \\ 0 & -C & C & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad N_0 = \operatorname{span}(\begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix})$$

with $C := C_1(e_2 - e_3) + 1$. We choose Q_0 and obtain G_1 and N_1 as follows

$$Q_0 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, G_1 = \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 1 + C & -C & 0 & 0 \\ 0 & 1 - C & C & -1 & 0 \\ 0 & -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } N_1 = \operatorname{span}(\begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ 1 \end{pmatrix}).$$

This yields the projectors

$$Q_1 = \begin{pmatrix} 0 & -c_1 & c_1 & 0 & -1 \\ 0 & -c_1 & c_1 & 0 & -1 \\ 0 & -c_1 & c_1 & 0 & -1 \\ 0 & -c_1 & c_1 & 0 & -1 \\ 0 & c_1 & -c_1 & 0 & 1 \end{pmatrix} \quad \text{and} \quad P_1 = \begin{pmatrix} 1 & c_1 & -c_1 & 0 & 1 \\ 0 & 1 + c_1 & -c_1 & 0 & 1 \\ 0 & c_1 & 1 + c_1 & 0 & 1 \\ 0 & c_1 & -c_1 & 1 & 1 \\ 0 & -c_1 & c_1 & 0 & 0 \end{pmatrix}$$

with a continuous function c_1. With the help of the borderline projector

$$R = \begin{pmatrix} \frac{C_1(e_2-e_3)}{C_1(e_2-e_3)+1} & \frac{C_1(e_2-e_3)}{C_1(e_2-e_3)+1} & 0 \\ \frac{1}{C_1(e_2-e_3)+1} & \frac{1}{C_1(e_2-e_3)+1} & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$DP_1 = D - DQ_1 = D + \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -c_1 & c_1 & 0 & -1 \end{pmatrix}$$

we obtain

$$DP_1D^- = DD^- + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ * & * & * \end{pmatrix} = R + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ * & * & * \end{pmatrix} = \begin{pmatrix} \frac{C_1(e_2-e_3)}{C_1(e_2-e_3)+1} & \frac{C_1(e_2-e_3)}{C_1(e_2-e_3)+1} & 0 \\ \frac{1}{C_1(e_2-e_3)+1} & \frac{1}{C_1(e_2-e_3)+1} & 0 \\ * & * & * \end{pmatrix}$$

with arbitrary entries $*$. Due to

$$\frac{1}{C_1(x)+1} := \begin{cases} \frac{1}{2}, & x < 1 \\ \frac{1}{x+1}, & 1 \leqslant x < 2 \\ \frac{1}{3}, & 2 \leqslant x \end{cases}$$

DP_1D^- is not differentiable. It does not mean that we cannot find projectors Q_0, P_0, Q_1 and P_1 such that DP_1D^- is differentiable. It may be possible to choose a different Q_0, such that the new N_1 enables us to choose a Q_1, which leads to a differentiable term DP_1D^-. However, this observation leads us to our last desired property of an index concept: It should be possible to describe the projectors or basis functions of a stage of the matrix chain without additional constraints from the other stages. This dependency in the tractability concept becomes a burden especially if the index of the DAE becomes larger than 2. In this case of admissible projectors, the kernel N_1 invokes a condition on the choice of Q_2 but at the same time depends on the choice of Q_0. This weakens the step-by-step concept of the Tractability Index.

4.1 Dissection Index

In the following we will introduce an index concept, which improves the following properties of the Tractability Index concept and the Strangeness Index concept:

(i) The non-linearity of the projectors and matrices.

(ii) The differentiability assumptions regarding the involved functions.

(iii) The independence between the stages of the step-by-step analysis.

While we will base our mixed index concept on a splitting strategy using basis functions, the Tractability Index uses projectors for this purpose. The advantage of projector functions is that they need less assumptions regarding the domain to be differentiable. Nevertheless we favor basis functions since they preserve the original size of the equations while splitting them.
The formulation of the Strangeness Index concept with the help of projector can be found in [Lam07].
Before we can define this mixed index concept we have to set some preparations regarding the basis functions. We start with the definition of the complementary kernel, the transposed kernel and the transposed complementary kernel.

Definition 4.4. (Complementary Functions)
Let V be a vector space and W be a subspace of V. Let U be a subspace of V such that $U \oplus W = V$. We say U is a direct difference between V and W and write $U = V \ominus W$. Notice that U is not unique.
Let $\mathcal{D} \subset \mathbb{R}^n$ be open and connected, let $\mathcal{I} \subset \mathbb{R}$ be a compact interval and let $M \in C(\mathcal{D} \times \mathcal{I}, \mathbb{R}^{m \times n})$ be a matrix function. We call $\mathbb{R}^n \ominus \ker M(x,t)$ a complementary kernel, $\ker M^T(x,t)$ the transposed kernel and $\mathbb{R}^m \ominus \ker M^T(x,t)$ a transposed complementary kernel for all $(x,t) \in \mathcal{D} \times \mathcal{I}$.

Thereby a canonical splitting of $\mathbb{R}^n$ and $\mathbb{R}^m$ is induced by the matrix function M. With the next definition we fix this splitting into matrix valued functions.

Definition 4.5. (Basis functions)
Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval and $\mathcal{D} \subset \mathbb{R}^n$ be open and connected. Let $M \in C(\mathcal{D} \times \mathcal{I}, \mathbb{R}^{m \times n})$ be a matrix function. Assume there are integers $n_y \in \mathbb{N}$ and $m_w \in \mathbb{N}$ such that

$$n_y = \dim(\ker M(x,t)) \quad \text{and} \quad m_w = \dim(\ker M^T(x,t)), \qquad \forall (x,t) \in \mathcal{D} \times \mathcal{I}$$

and define $n_x = n - n_y$ and $m_v = m - m_w$. Choose four matrix functions

$$\begin{array}{rclrcl} P & : & \mathcal{D} \times \mathcal{I} \to \mathbb{R}^{n \times n_x}, & Q & : & \mathcal{D} \times \mathcal{I} \to \mathbb{R}^{n \times n_y}, \\ V & : & \mathcal{D} \times \mathcal{I} \to \mathbb{R}^{m \times m_v}, & W & : & \mathcal{D} \times \mathcal{I} \to \mathbb{R}^{m \times m_w} \end{array}$$

such that the set of the columns of the matrix functions form basises of a complementary kernel, the kernel, a complementary transposed kernel and the transposed kernel, respectively. We call P, Q, V and W the associated basis functions of M.

The idea of using basis functions for a decoupling of a DAE goes back to Kunkel and Mehrmann, cf. [KM06].
For the associated basis functions of a matrix function M it holds

$$\begin{array}{rclrcl} \operatorname{im} P(x,t) & = & \mathbb{R}^n \ominus \ker M(x,t), & \operatorname{im} Q(x,t) & = & \ker M(x,t), \\ \operatorname{im} V(x,t) & = & \mathbb{R}^m \ominus \ker M^T(x,t), & \operatorname{im} W(x,t) & = & \ker M^T(x,t) \end{array}$$

point wise. Additionally it holds $n_x + n_y = n$, $m_v + m_w = m$ and $n_x = m_v = \operatorname{rk} M$.

Remark 4.6.
Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval and $\mathcal{D} \subset \mathbb{R}^n$ be open and connected. Let $M \in C(\mathcal{D} \times \mathcal{I}, \mathbb{R}^{m\times n})$ be a matrix function. Let P and V be basis functions of a complementary kernel and a transposed complementary kernel. Then the matrix $(V^\top MP)(x,t)$ is non-singular for all $(x,t) \in \mathcal{D} \times \mathcal{I}$. The matrix $(V^\top MP)(x,t)$ is quadratic due to $n_x = m_v = \operatorname{rk} M$. Let $z \neq 0$ then it follows $P(x,t)z \neq 0$ since P has full column rank. By $\operatorname{im} P(x,t) = \mathbb{R}^n \ominus \ker M(x,t)$ and $P(x,t)z \neq 0$ it follows that $M(x,t)(P(x,t)z) \neq 0$ which finally leads to $(V^\top MP)(x,t)z = V^\top(x,t)(M(x,t)P(x,t)z) \neq 0$ by $\operatorname{im} V(x,t) = \mathbb{R}^m \ominus \ker M^T(x,t)$.

We notice that the integers n_y, n_x, n_w and n_v may be zero. In this case the associated matrix would have zero columns, hence the matrix has no entries. Now we are able to understand these four sub-spaces as matrix valued functions, but up until now these definitions are only point wise. The next Lemma will provide us with criteria which lead to continuous or even differentiable basis functions.

Lemma 4.7. (Global differentiable basis functions, Lemma 2.1.10. in [Ste06])
Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval and $\mathcal{D} \subset \mathbb{R}^n$ be C^l-diffeomorphic to a parallelepiped in $\mathbb{R}^n$. Let be $M \in C^l(\mathcal{D} \times \mathcal{I}, \mathbb{R}^{m\times n})$. Furthermore, suppose there is an $r \in \mathbb{N}$ such that $dim(\operatorname{im} M(z,t)) = r$ for all $(z,t) \in \mathcal{D} \times \mathcal{I}$. Then there exists a matrix function $Q \in C^l(\mathcal{D} \times \mathcal{I}, \mathbb{R}^{n,n-r})$, with $\operatorname{im} Q(z,t) = \ker M(z,t)$ for all $(z,t) \in \mathcal{D} \times \mathcal{I}$.

So the differentiability of the matrix $M(x,t)$ passes down to the associated basis functions as long as we operate on domains described in Lemma 4.7.
In order to prepare the mentioned splitting strategy involving basis functions choose a fixed but arbitrary point $(x_*, t_*) \in \mathcal{D} \times \mathcal{I}$ and consider the basis functions $P_* := P(x_*, t_*)$ and $Q_* := Q(x_*, t_*)$ at this point. Combine the basis function P_* of a complementary kernel and the basis function Q_* of the kernel into one matrix $T := \begin{pmatrix} P_* & Q_* \end{pmatrix}$. Notice that this matrix is quadratic and non-singular since $n_x + n_y = n$ and together the set of the columns of P_* and Q_* form a basis of $\mathbb{R}^n$. Hence T is suited as a coordinate transformation matrix and we can split any $z \in \mathbb{R}^n$ uniquely into an $x \in \mathbb{R}^{n_x}$ and a $y \in \mathbb{R}^{n_y}$ by

$$z = \begin{pmatrix} P_* & Q_* \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = P_* x + Q_* y$$

We take the analogous considerations for a complementary transposed kernel and the transposed kernel. These considerations lead to two basis functions $V_* := V(x_*, t_*)$ and $W_* := W(x_*, t_*)$ and the non-singularity of $F := \begin{pmatrix} V_* & W_* \end{pmatrix}^\top$. With these basis functions we can formulate a factorization of an equation

$$f = 0 \quad \Leftrightarrow \quad \begin{pmatrix} V_*^\top \\ W_*^\top \end{pmatrix} f = 0 \quad \Leftrightarrow \quad \begin{pmatrix} V_*^\top f \\ W_*^\top f \end{pmatrix} = 0$$

with $f \in \mathbb{R}^m$. With this splitting we obtain two properties for nonlinear functions.

Lemma 4.8. (Basis functions)
Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval, let $\mathcal{D} \subset \mathbb{R}^n$ be open and convex and let $f, g \in C^1(\mathcal{D} \times \mathcal{I}, \mathbb{R}^n)$. Assume the existence of basis functions $P(t)$ and $Q(t)$ of a complementary kernel of $\frac{\partial}{\partial z} f(z,t)$ and the kernel of $\frac{\partial}{\partial z} f(z,t)$ being independent of z. Further let exist a basis function $W(t)$ of the transposed kernel of $\frac{\partial}{\partial z} g(z,t)Q(t)$ being also independent of z. Then for each $(z,t) \in \mathcal{D} \times \mathcal{I}$ there is a unique $x \in \mathbb{R}^{n_x}$ such that

$$f(z,t) = f(P(t)x, t) \text{ and } W^\top(t)g(z,t) = W^\top(t)g(P(t)x,t).$$

Proof*.*
Let z and t be fixed. Then there is a unique $x \in \mathbb{R}^{n_x}$ and a unique $y \in \mathbb{R}^{n_y}$ such that $z = P(t)x + Q(t)y$. Applying the mean value theorem to f we get

$$\begin{aligned} f(z,t) - f(P(t)x,t) &= \int_0^1 f_z(P(t)x + sQ(t)y, t)\mathrm{d}s(z - P(t)x) \\ &= \int_0^1 \underbrace{f_z(P(t)x + sQ(t)y, t)Q(t)}_{=0} \,\mathrm{d}s\; y = 0, \end{aligned}$$

since $\operatorname{im} Q(t) = \ker f_z(z,t)$ for all $(z,t) \in \mathcal{D} \times \mathcal{I}$. Analogously we obtain

$$\begin{aligned} &W^\top(t)g(z,t) - W^\top(t)g(P(t)x,t) \\ =&W^\top(t)(g(z,t) - g(P(t)x,t)) \\ =&W^\top(t)\int_0^1 g_z(P(t)x + sQ(t)y,t)\mathrm{d}s(z - P(t)x) \\ =&\int_0^1 \underbrace{W^\top(t)g_z(P(t)x + sQ(t)y,t)Q(t)}_{=0} \,\mathrm{d}s\; y = 0. \end{aligned}$$

□

For a better understanding of the matrix chain we demonstrate it for the case of a linear DAE with constant coefficients. But first we need the following lemma as a preparation:

Lemma 4.9. (The differentiable variable part)
Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval and $\mathcal{D} \subset \mathbb{R}^n$ be open and connected. Let $d \in C^1(\mathcal{D} \times \mathcal{I}, \mathbb{R}^m)$ and let $P(z,t)$, $Q(z,t)$, $V(z,t)$ and $W(z,t)$ be the associated basis functions of $\frac{\partial}{\partial z}d(z,t)$. Further let P and Q be independent of z, i.e. $P(z,t) = P(t)$ and $Q(z,t) = Q(t)$, and let them be continuously differentiable. We define x and y by $z = \begin{pmatrix} P(t) & Q(t) \end{pmatrix} \begin{pmatrix} x & y \end{pmatrix}^\top$. Then it holds

$$x(t) \in C^1(\mathcal{I}, \mathbb{R}^n).$$

Proof.
Due to the differentiability of d and Lemma 4.8 we get

$$\begin{aligned}
&d'(z(t),t)\\
=&d'(P(t)x(t),t)\\
=&\lim_{h\to 0} \frac{d(P(t)x(t),t) - d(P(t-h)x(t-h),t-h)}{h}\\
=&\lim_{h\to 0} \frac{d(P(t)x(t),t) - d(P(t)x(t-h),t)}{h}\\
&+\lim_{h\to 0} \frac{d(P(t)x(t-h),t) - d(P(t)x(t-h),t-h)}{h}\\
&+\lim_{h\to 0} \frac{d(P(t)x(t-h),t-h) - d(P(t-h)x(t-h),t-h)}{h}\\
=&\lim_{h\to 0} \frac{d(P(t)x(t),t) - d(P(t)x(t-h),t)}{h} + d_z(P(t)x,t)P'(t)x + d_t(P(t)x,t).
\end{aligned}$$

Apply the Mean Value Theorem and obtain

$$\begin{aligned}
&d'(z(t),t) - d_z(P(t)x,t)P'(t)x - d_t(P(t)x,t)\\
=&\lim_{h\to 0} \frac{d(P(t)x(t),t) - d(P(t)x(t-h),t)}{h}\\
=&\lim_{h\to 0} \int_0^1 d_z(sP(t)x(t) + (1-s)P(t)x(t-h),t)\mathrm{d}s\ P(t)\frac{x(t)-x(t-h)}{h}\\
=&d_z(P(t)x(t),t)P(t)\lim_{h\to 0} \frac{x(t)-x(t-h)}{h}.
\end{aligned}$$

It holds that $d(z(t),t) = d(P(t)x(t),t)$ by Lemma 4.8 hence V does not depend on $y(t)$. Therefore we get

$$\begin{aligned}
&\lim_{h\to 0} \frac{x(t)-x(t-h)}{h}\\
=&M(x(t),t)(d'(P(t)x(t),t) - d_z(P(t)x(t),t)P'(t)x(t) - d_t(P(t)x(t),t))
\end{aligned}$$

with $M(x(t),t) := (V(x(t),t)d_z(P(t)x(t),t)P(t))^{-1}V(x(t),t)$. □

Let $\mathcal{I} \subset \mathbb{R}$ be a compact interval. For $(x,t) \in \mathbb{R}^n \times \mathcal{I}$ observe the following equation

$$A(Dx)' + Bx = r(t)$$

with $A \in \mathbb{R}^{n\times m}$, $D \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times n}$, $Dx \in C^1(\mathbb{R}^n,\mathbb{R}^m)$ and $r : \mathcal{I} \to \mathbb{R}^n$ sufficiently smooth. Assume that the DAE fulfills the basic Properties 2.25. Let P, Q, V and W be associated basis functions of AD.
With $Dx(t) \in C^1(\mathcal{I},\mathbb{R}^m)$ and Lemma 4.9 it follows that $x_0(t) \in C^1(\mathcal{I},\mathbb{R}^n)$. Begin the decomposition of the linear DAE by inserting this variable splitting.

$$\begin{aligned} &A(Dx)' + Bx = r(t) \\ \Leftrightarrow &ADPx_0' + BPx_0 + BQy_0 = r(t) \end{aligned}$$

Next split the equations by multiplying $V^\top$ and $W^\top$ from the left.

$$\begin{aligned} &ADPx_0' + BPx_0 + BQy_0 = r(t) \\ \Leftrightarrow &\left\{ \begin{array}{ccccccc} V^\top ADPx_0' &+& V^\top BPx_0 &+& V^\top BQy_0 &=& V^\top r(t) \\ W^\top ADPx_0' &+& W^\top BPx_0 &+& W^\top BQy_0 &=& W^\top r(t) \end{array} \right. \\ \Leftrightarrow &\left\{ \begin{array}{ccccccc} V^\top ADPx_0' &+& V^\top BPx_0 &+& V^\top BQy_0 &=& V^\top r(t) \\ && W^\top BPx_0 &+& W^\top BQy_0 &=& W^\top r(t). \end{array} \right. \end{aligned}$$

Then the first step of the matrix chain provides

$$\begin{array}{ccccccccc} G_1 &:=& V^\top ADP, & B^{\mathrm{v}}_{x_1} &:=& V^\top BP, & B^{\mathrm{v}}_{y_1} &:=& V^\top BQ \\ &&& B^{\mathrm{w}}_{x_1} &:=& W^\top BP, & B^{\mathrm{w}}_{y_1} &:=& W^\top BQ. \end{array}$$

Inserting this notation in the equations we get

$$G_1 x_0' + B^{\mathrm{v}}_{x_1} x_0 + B^{\mathrm{v}}_{y_1} y_0 = V^\top r(t) \tag{4.6a}$$

$$B^{\mathrm{w}}_{x_1} x_0 + B^{\mathrm{w}}_{y_1} y_0 = W^\top r(t). \tag{4.6b}$$

If $B^{\mathrm{w}}_{y_1} = W^\top BQ$ would be a quadratic non-singular matrix then Equation (4.6b) would give us an explicit expression for the whole variable y_0 by

$$\tilde{y}_1 := y_0 = (B^{\mathrm{w}}_{y_1})^{-1}(W^\top r(t) - B^{\mathrm{w}}_{x_1} x_0). \tag{4.7}$$

If we then insert this into Equation (4.6a) and multiply this equation with the inverse of G_1 which is non-singular by the definition of P and Q we would obtain an explicit ODE for the whole variable $x_0 =: x_1$ as we can see as follows:

$$\begin{aligned} &G_1 x_1' + B^{\mathrm{v}}_{x_1} x_1 + B^{\mathrm{v}}_{y_1} \tilde{y}_1 = V^\top r(t) \\ \Rightarrow\ &G_1 x_1' + B^{\mathrm{v}}_{x_1} x_1 + B^{\mathrm{v}}_{y_1} (B^{\mathrm{w}}_{y_1})^{-1}(W^\top r(t) - B^{\mathrm{w}}_{x_1} x_1) = V^\top r(t) \end{aligned}$$

$$\Rightarrow G_1 x_1' + (B^{\mathrm{v}}_{x_1} - B^{\mathrm{v}}_{y_1}(B^{\mathrm{w}}_{y_1})^{-1} B^{\mathrm{w}}_{x_1}) x_1 = (V^\top - B^{\mathrm{v}}_{y_1}(B^{\mathrm{w}}_{y_1})^{-1} W^\top) r(t).$$

Together with Equation (4.7) we obtain the decoupled equations:

$$\begin{aligned} x_1' &= -G^{-1}(B^{\mathrm{v}}_{x_1} - B^{\mathrm{v}}_{y_1}(B^{\mathrm{w}}_{y_1})^{-1} B^{\mathrm{w}}_{x_1}) x_1 + G^{-1}(V^\top - B^{\mathrm{v}}_{y_1}(B^{\mathrm{w}}_{y_1})^{-1} W^\top) r(t) \\ \tilde{y}_1 &= (B^{\mathrm{w}}_{y_1})^{-1}(W^\top r(t) - B^{\mathrm{w}}_{x_1} x_1). \end{aligned}$$

If $B^{\mathrm{w}}_{y_1}$ is singular then we need the next sequence of basis functions. Therefore let $P_{y_1}, Q_{y_1}, V_{y_1}, W_{y_1}$ be the four associated basis functions of $B^{\mathrm{w}}_{y_1}$ and let P_{x_1}, Q_{x_1} be the associated basis functions of $W_{y_1}^\top B^{\mathrm{w}}_{x_1}$ with respect to the kernel and a complementary kernel. Furthermore let V_{x_1}, W_{x_1} be the associated basis functions of $G_1 Q_{x_1}$ with respect to a transposed complementary kernel and the transposed kernel.
We denote $y_0 = P_{y_1}\tilde{y}_1 + Q_{y_1} y_1$ an $x_0 = P_{x_1}\tilde{x}_1 + Q_{x_1} x_1$. With these two variable splittings and an equation splitting by a multiplication with $V_{y_1}^\top$ and $W_{y_1}^\top$ continue the splitting of the linear DAE.

$$\begin{array}{rcrcrcl} G_1 x_0' & + & B^{\mathrm{v}}_{x_1} x_0 & + & B^{\mathrm{v}}_{y_1} y_0 & = & V^\top r(t) \\ & & B^{\mathrm{w}}_{x_1} x_0 & + & B^{\mathrm{w}}_{y_1} y_0 & = & W^\top r(t) \end{array}$$

$$\Leftrightarrow \left\{ \begin{array}{rcrcrcl} G_1 x_0' & + & B^{\mathrm{v}}_{x_1} x_0 & + & B^{\mathrm{v}}_{y_1} y_0 & = & V^\top r(t) \\ & & B^{\mathrm{w}}_{x_1} x_0 & + & B^{\mathrm{w}}_{y_1} P_{y_1} \tilde{y}_1 & = & W^\top r(t) \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{rcccccl} G_1 x_0' & + & B^{\mathrm{v}}_{x_1} x_0 & + & B^{\mathrm{v}}_{y_1} y_0 & = & V^\top r(t) \\ & & V_{y_1}^\top B^{\mathrm{w}}_{x_1} x_0 & + & V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1} \tilde{y}_1 & = & V_{y_1}^\top W^\top r(t) \\ & & W_{y_1}^\top B^{\mathrm{w}}_{x_1} x_0 & & & = & W_{y_1}^\top W^\top r(t) \end{array} \right.$$

$$\Leftrightarrow \left\{ \begin{array}{rcccccl} G_1 x_0' & + & B^{\mathrm{v}}_{x_1} x_0 & + & B^{\mathrm{v}}_{y_1} y_0 & = & V^\top r(t) \\ & & V_{y_1}^\top B^{\mathrm{w}}_{x_1} x_0 & + & V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1} \tilde{y}_1 & = & V_{y_1}^\top W^\top r(t) \\ & & W_{y_1}^\top B^{\mathrm{w}}_{x_1} P_{x_1} \tilde{x}_1 & & & = & W_{y_1}^\top W^\top r(t). \end{array} \right.$$

We assume the non-singularity of $W_{y_1}^\top B^{\mathrm{w}}_{x_1} P_{x_1}$, since otherwise the system would be underdetermined. Furthermore $V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1}$ is also non-singular due to the definition of $V_{y_1}^\top$ and P_{y_1}. Therefore the two algebraic equations provide us with explicit expressions for $\tilde{y}_1$ and $\tilde{x}_1$ by:

$$\begin{aligned} \tilde{x}_1 &= (W_{y_1}^\top B^{\mathrm{w}}_{x_1} P_{x_1})^{-1} W_{y_1}^\top W^\top r(t) =: r_{x_1}(t) \\ \tilde{y}_1 &= (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} (V_{y_1}^\top W^\top r(t) - V_{y_1}^\top B^{\mathrm{w}}_{x_1} x_0) \\ &= (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} (V_{y_1}^\top W^\top r(t) - V_{y_1}^\top B^{\mathrm{w}}_{x_1} P_{x_1} \tilde{x}_1 - V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1} x_1) \\ &= (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} (V_{y_1}^\top W^\top r(t) - V_{y_1}^\top B^{\mathrm{w}}_{x_1} P_{x_1} r_{x_1}(t)) - (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1} x_1 \\ &=: r_{y_1}(t) - (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1} x_1. \end{aligned}$$

Now, split the x_0 and y_0 in the dynamic part of the equations

$$\begin{array}{rcccccl} G_1 x_0' & + & B^{\mathrm{v}}_{x_1} x_0 & + & B^{\mathrm{v}}_{y_1} y_0 & = & V^\top r(t) \\ G_1 Q_{x_1} x_1' + G_1 P_{x_1} \tilde{x}_1' & + & B^{\mathrm{v}}_{x_1} Q_{x_1} x_1 + B^{\mathrm{v}}_{x_1} P_{x_1} \tilde{x}_1 & + & B^{\mathrm{v}}_{y_1} Q_{y_1} y_1 + B^{\mathrm{v}}_{y_1} P_{y_1} \tilde{y}_1 & = & V^\top r(t). \end{array}$$

We insert the expression for $\tilde{x}_1$ and obtain

$$G_1 Q_{x_1} x_1' + B^{\mathrm{v}}_{x_1} Q_{x_1} x_1 + B^{\mathrm{v}}_{y_1} Q_{y_1} y_1 + B^{\mathrm{v}}_{y_1} P_{y_1} \tilde{y}_1 = r_1^*(t)$$

with $r_1^*(t) := V^\top r(t) - G_1 P_{x_1} r_{x_1}'(t) - B^{\mathrm{v}}_{x_1} P_{x_1} r_{x_1}(t)$. And finally insert the expression for $\tilde{y}_1$ into the equation and get

$$G_1 Q_{x_1} x_1' + (B^{\mathrm{v}}_{x_1} Q_{x_1} - B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1}) x_1 + B^{\mathrm{v}}_{y_1} Q_{y_1} y_1 = r_1(t)$$

with $r_1(t) := r_1^*(t) - r_{y_1}(t)$. Include the last pair of basis functions of the current sequence V_{x_1} and W_{x_1} by multiplying $V_{x_1}^\top$ and $W_{x_1}^\top$ from the left side to split the equation with respect to $G_1 Q_{x_1} x_2'$

$$\begin{aligned} & G_1 Q_{x_1} x_1' \; + \; (B^{\mathrm{v}}_{x_1} Q_{x_1} - B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1}) x_1 \; + \; B^{\mathrm{v}}_{y_1} Q_{y_1} y_1 \; = \; r_1(t) \\ \Leftrightarrow & \left\{ \begin{array}{ccccccc} G_2 x_1' & + & B^{\mathrm{v}}_{x_2} x_1 & + & B^{\mathrm{v}}_{y_2} y_1 & = & V_{x_1}^\top r_1(t), \\ & & B^{\mathrm{w}}_{x_2} x_1 & + & B^{\mathrm{w}}_{y_2} y_1 & = & W_{x_1}^\top r_1(t). \end{array} \right. \end{aligned}$$

with the same notation of the matrix chain

$$\begin{aligned} G_2 :=& V_{x_1}^\top G_1 Q_{x_1}, \\ B^{\mathrm{v}}_{x_2} :=& V_{x_1}^\top B^{\mathrm{v}}_{x_1} Q_{x_1} - V_{x_1}^\top B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1}, \\ B^{\mathrm{w}}_{x_2} :=& W_{x_1}^\top B^{\mathrm{v}}_{x_1} Q_{x_1} - W_{x_1}^\top B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1}, \\ B^{\mathrm{v}}_{y_2} :=& V_{x_1}^\top B^{\mathrm{v}}_{y_1} Q_{y_1}, \\ B^{\mathrm{w}}_{y_2} :=& W_{x_1}^\top B^{\mathrm{v}}_{y_1} Q_{y_1}. \end{aligned}$$

If $B^{\mathrm{w}}_{y_2}$ would be quadratic and non-singular we would notate $\tilde{y}_2 := y_1$ and $x_2 := x_1$ and the process would end. Otherwise repeat this process until $B^{\mathrm{w}}_{y_i}$ is quadratic and non-singular if possible.

While the Strangeness Index operates on DAEs in standard form the Tractability Index deals with DAEs with a properly stated derivative term, in recent works. The mixed index concept will be defined on a DAE class which includes both DAEs in standard form and properly stated DAEs.

Definition 4.10. (Semi-properly stated derivative term)
The DAE (2.14) has a semi-properly stated derivative term on $\mathcal{D} \times \mathcal{I}$, if im $\frac{\partial}{\partial x} d$ and ker $\frac{\partial}{\partial y} f$ are C-subspaces in $\mathbb{R}^m$, and the condition

$$\operatorname{im} \frac{\partial}{\partial y} f(y,x,t) = \operatorname{im} \frac{\partial}{\partial y} f(y,x,t) \frac{\partial}{\partial x} d(x,t), \quad \forall (y,x,t) \in \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}, \tag{4.8}$$

holds.

For a DAE in standard form we have $d(x,t) = x$ hence we obtain $\frac{\partial}{\partial x}d(x,t) = I$ and therefore (4.8) holds. For a proper formulated DAE the identity (4.8) follows directly from the definition of the properly stated derivative term.
In the following we will mainly work with DAEs with a semi-properly stated derivative term. Therefore we define:

Definition 4.11. (DAEs with nonlinear semi-properly stated derivative term)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D} \subset \mathbb{R}^n$ be open subsets. Let $f \in C(\mathbb{R}^m \times \mathcal{D} \times \mathcal{I}, \mathbb{R}^n)$ be continuous such that the partial derivatives $\frac{\partial}{\partial y}f(y,x,t)$ and $\frac{\partial}{\partial x}f(y,x,t)$ are also continuous with $\frac{\partial}{\partial y}f(y,x,t)$ being singular for all triples $(y,x,t) \in \mathbb{R}^m \times \mathcal{D} \times \mathcal{I}$. Furthermore let $d \in C^1(\mathcal{D} \times \mathcal{I}, \mathbb{R}^m)$ and let d and f be semi-properly formulated. We call

$$f(d'(x(t),t), x(t), t) = 0, \quad x(t_0) = x_0 \tag{4.9}$$

a DAE with a nonlinear semi-properly derivative term.

Now we are starting to formulate the mixed index concept. Consider a DAE (2.14) with a semi-properly stated derivative term and define the matrix functions

$$\begin{aligned} D(x,t) &:= \frac{\partial}{\partial x}d(x,t) \\ A(x^1,x,t) &:= \frac{\partial}{\partial y}f(D(x,t)x^1 + d_t(x,t), x, t), \\ B(x^1,x,t) &:= \frac{\partial}{\partial x}f(D(x,t)x^1 + d_t(x,t), x, t). \end{aligned}$$

with $x^1 \in \mathbb{R}^n$, $x \in \mathcal{D}$ and $t \in \mathcal{I}$. Again $\frac{\partial}{\partial y}$ and $\frac{\partial}{\partial x}$ denote the partial derivatives with respect to the first and second argument of f.
With the help of the basis functions we construct a matrix chain emulating the chain of the Tractability Index.

Definition 4.12. (Matrix chain)
Let $P(x^1,x,t)$, $Q(x^1,x,t)$, $V(x^1,x,t)$ and $W(x^1,x,t)$ be associated basis functions of $A(x^1,x,t)D(x,t)$. The variable x^1 is a auxiliary variable that can be seen as a place holder for the derivative. This auxiliary variable is called jet variable, see [LMT13]. Hence the derivatives of P and Q depend on the second derivative of x and we need to introduce new jet variables $x^i \in \mathbb{R}^n$ as placeholders for the i-th derivative, respectively. Let $i,k \in \mathbb{N}$ and define $X^k := (x^k, ..., x^1, x)$, further consider a sufficiently smooth function $g(X^{i-1},t)$ then we define a jet-derivative operator $(.)'$ as

$$g'(X^i,t) := (g(X^{i-1},t))' := \frac{\partial}{\partial t}g(X^{i-1},t) + \sum_{j=0}^{i-1}\frac{\partial}{\partial x^j}g(X^{i-1},t)x^{j+1}.$$

Let P and Q be sufficiently smooth and define

$$\begin{aligned}
G_1(X^1,t) :=& V^\top(X^1,t)A(X^1,t)D(x,t)P(X^1,t),\\
B^{\mathrm{v}}_{x_1}(X^2,t) :=& V^\top(X^1,t)B(X^1,t)P(X^1,t)+V^\top(X^1,t)A(X^1,t)(D(x,t)P(X^1,t))',\\
B^{\mathrm{v}}_{y_1}(X^2,t) :=& V^\top(X^1,t)B(X^1,t)Q(X^1,t)+V^\top(X^1,t)A(X^1,t)(D(x,t)Q(X^1,t))',\\
B^{\mathrm{w}}_{x_1}(X^1,t) :=& W^\top(X^1,t)B(X^1,t)P(X^1,t),\\
B^{\mathrm{w}}_{y_1}(X^1,t) :=& W^\top(X^1,t)B(X^1,t)Q(X^1,t)
\end{aligned}$$

as the next sequence of matrices. Let

- $P_{y_1}, Q_{y_1}, V_{y_1}, W_{y_1}$ be the four associated basis functions of $B^{\mathrm{w}}_{y_1}(X^1,t)$.
- P_{x_1}, Q_{x_1} be the basis functions of $(W^\top_{y_1}B^{\mathrm{w}}_{x_1})(X^1,t))$ with respect to the kernel and a complementary kernel, respectively.
- V_{x_1}, W_{x_1} be the basis functions of $(G_1Q_{x_1})(X^1,t)$ with respect to a complementary transposed kernel and the transposed kernel, respectively.

The complementary kernel, the transposed kernel and the complementary transposed kernel are defined in Definition 4.4. We keep formulating the sequence as long as possible for $i \geqslant 2$ up to an integer $\mu \in \mathbb{N}$.
For the next sequence of matrices it is necessary to include the jet-derivative of $Q_{x_{i-1}}$ as the jet-derivatives of DP and DQ were included in the first step. In the following we assume that the occurring derivatives of the basis functions exists.
Then we construct

$$\begin{array}{rclrclrcl}
G_i &=& V^\top_{x_{i-1}}G_{i-1}Q_{x_{i-1}} & B^{\mathrm{v}}_{x_i} &=& V^\top_{x_{i-1}}B_{x_{i-1}} & B^{\mathrm{v}}_{y_i} &=& V^\top_{x_{i-1}}B_{y_{i-1}}\\
 && & B^{\mathrm{w}}_{x_i} &=& W^\top_{x_{i-1}}B_{x_{i-1}} & B^{\mathrm{w}}_{y_i} &=& W^\top_{x_{i-1}}B_{y_{i-1}}
\end{array}$$

with

$$\begin{aligned}
B_{y_{i-1}} :=& B^{\mathrm{v}}_{y_{i-1}}Q_{y_{i-1}}\\
B_{x_{i-1}} :=& B^{\mathrm{v}}_{x_{i-1}}Q_{x_{i-1}}+G_{i-1}Q'_{x_{i-1}}-B^{\mathrm{v}}_{y_{i-1}}P_{y_{i-1}}(V^\top_{y_{i-1}}B^{\mathrm{w}}_{y_{i-1}}P_{y_{i-1}})^{-1}V^\top_{y_{i-1}}B^{\mathrm{w}}_{x_{i-1}}Q_{x_{i-1}}.
\end{aligned}$$

Therefore G_i, $B^{\mathrm{v}}_{y_i}$ and $B^{\mathrm{w}}_{y_i}$ depend on (X^{i-1},t) while $B^{\mathrm{v}}_{x_i}$ and $B^{\mathrm{w}}_{x_i}$ depend on (X^i,t) due to $Q'_{x_{i-1}}$, except for $B^{\mathrm{v}}_{y_2}$ and $B^{\mathrm{w}}_{y_2}$ which may depend on (X^2,t). Let

- $P_{y_i}, Q_{y_i}, V_{y_i}, W_{y_i}$ be the four associated basis functions of $B^{\mathrm{w}}_{y_i}(X^{i-1},t)$.
- P_{x_i}, Q_{x_i} be the basis functions of $(W^\top_{y_i}B^{\mathrm{w}}_{x_i})(X^i,t)$ with respect to the kernel and a complementary kernel, respectively.
- V_{x_i}, W_{x_i} be the basis functions of $(G_iQ_{x_i})(X^i,t)$ with respect to a complementary transposed kernel and the transposed kernel, respectively.

We will use the term basis chain as a synonym for the term matrix chain.

The whole idea of a matrix chain belongs to the Tractability Index but at the same time we can already see that the decoupling strategy mimics the strategy of the Strangeness Index for linear DAEs, when compared with [KM06] on pages 56-80.
For a more consistent notation we integrate the first four basis functions to match the notation of the other ones by

$$Q_{x_0} := P, \quad Q_{y_0} := Q, \quad V_{x_0} := V, \quad W_{x_0} := W.$$

The splitting of the variables induced by the basis chain can be illustrated by the following diagram:

$$\begin{tikzcd}
 & & \tilde{x}_1 & \tilde{x}_2 & & \tilde{x}_{\mu-1} & \\
 & x_0 \arrow[ur, "P_{x_1}"] \arrow[r, "Q_{x_1}"'] & x_1 \arrow[ur, "P_{x_2}"] \arrow[rr, "\cdots"] & & x_{\mu-2} \arrow[ur, "P_{x_{\mu-1}}"] \arrow[r, "Q_{x_{\mu-1}}"'] & x_{\mu-1} \arrow[r, "Q_{x_\mu}"'] & x_\mu \\
x \arrow[ur, "P = Q_{x_0}"] \arrow[dr, "Q = Q_{y_0}"'] & & & & & & \\
 & y_0 \arrow[r, "Q_{y_1}"] \arrow[dr, "P_{y_1}"'] & y_1 \arrow[rr, "\cdots"] \arrow[dr, "P_{y_2}"'] & & y_{\mu-2} \arrow[r, "Q_{y_{\mu-1}}"] \arrow[dr, "P_{y_{\mu-1}}"'] & y_{\mu-1} \arrow[dr, "P_{y_\mu}"'] & \\
 & & \tilde{y}_1 & \tilde{y}_2 & & \tilde{y}_{\mu-1} & \tilde{y}_\mu
\end{tikzcd}$$

We notice that P is notated with Q_{x_0} even though P is a basis of a complementary kernel. But in the splitting sense bases denoted with a P indicate parts of the variables which are set algebraically. The variables are recursively defined by:

$$x = Px_0 + Qy_0, \quad x_{i-1} = P_{x_i}\tilde{x}_i + Q_{x_i}x_i, \quad y_{i-1} = P_{y_i}\tilde{y}_i + Q_{y_i}y_i. \tag{4.10}$$

In the case of a proper formulated DAE $P(x^1, x, t)$ and $Q(x^1, x, t)$ would not depend on x^1 since the kernel and a complementary kernel of $A(x^1, x, t)D(x, t)$ would only depend on the subspaces of $D(x, t)$. Additionally, the matrices $B^{\mathrm{v}}_{x_1}(X^2, t)$ and $B^{\mathrm{v}}_{y_1}(X^2, t)$ do not depend on x^2 since $P(X^1, t)$ and $Q(X^1, t)$ do not depend on x^1 and consequently $(DP)'(X^2, t)$ and $(DQ)'(X^2, t)$ do not depend on x^2. Furthermore a proper formulated derivative term yields the identity $D(x, t)Q(x, t) = 0$ hence we obtain

$$B^{\mathrm{v}}_{y_1}(X^1, t) = V^\top(X^1, t)B(X^1, t)Q(x, t).$$

With the help of the matrix chain we define the Dissection Index.

Definition 4.13. (Dissection Index)
Let the DAE (4.9) have a semi-properly stated derivative term, let f and d be sufficiently smooth, let $\mathcal{G} \subset \mathcal{D} \times \mathcal{I}$ be open and connected and let $\mu \in \mathbb{N}$. We assume all basis functions to exist and have constant rank on $\mathbb{R}^{i \cdot n} \times \mathcal{G}$ for $i = 0, ..., \mu$. We define the characteristic values r_i of the DAE as $r_0 := \operatorname{rk} AD$ and $r_i := r_{i-1} + \operatorname{rk} B^{\mathrm{w}}_{y_i}$ for $i = 1, \ldots, \mu$. Then the DAE (4.9) is said to be

1. regular with Dissection Index 0 on $\mathcal{G}$, if $r_0 = n$,
2. regular with Dissection Index μ on $\mathcal{G}$, if $r_{\mu-1} < r_\mu = n$,
3. regular on $\mathcal{G}$, if it is regular on $\mathcal{G}$ with any Dissection Index.

This definition of the Dissection Index seems very similar to the definition of the Tractability Index in [LMT13], except for the crucial fact that the characteristic values are calculated differently. One could say that we lifted the basic decoupling idea of the Strangeness Index for linear DAEs up to nonlinear DAEs by using the tools of the Tractability Index. In the following we assume that a DAE has a constant Dissection Index on its whole definition region. This is a crucial assumption for the rest of the thesis.
Notice that the matrix chain ends with $B^{\mathrm{w}}_{y_2}$ in the index 2 case, hence the calculation of $B^{\mathrm{v}}_{x_1}$ is not needed. This yields that neither $(D(x,t)P(X^1,t))'$ nor $(D(x,t)Q(X^1,t))'$ is needed in the proper formulated index 2 case. Furthermore we amplify that it is advantageous to define an alternative ending of the matrix chain. Before defining the alternative matrix ending, we are going to prove some basic property of the matrix chain.

Lemma 4.14.
Let the DAE (4.9) have a finite Dissection Index μ. Then the matrices G_i, $W^\top_{x_i} G_i P_{x_i}$, $V^\top_{y_i} B^{\mathrm{w}}_{y_i} P_{y_i}$ and $B^{\mathrm{w}}_{y_\mu}$ are non-singular for $i = 1, \ldots, \mu$.

Proof.
The columns of the basis function P are a basis of a complementary kernel of $V^\top AD$, since they are a basis of a complementary kernel of $V^\top AD$ and V is a basis function of a transposed complementary kernel of AD. Hence, $V^\top ADP$ has a trivial kernel and due to $\operatorname{rk} P = \operatorname{rk} V$ the matrix $G_1 = V^\top ADP$ is quadratic and therefore non-singular. As an induction hypothesis it now holds that G_{i-1} is non-singular. Then $G_{i-1}Q_{x_{i-1}}$ has full column rank and therefore $G_i = V^\top_{x_{i-1}} G_{i-1} Q_{x_{i-1}}$ is quadratic and non-singular, since $V_{x_{i-1}}$ is a basis function of a transposed complementary kernel of $G_{i-1}Q_{x_{i-1}}$.
Now we know that G_i is non-singular. Furthermore we know that $\begin{pmatrix} V_{x_i} & W_{x_i} \end{pmatrix}$ and $\begin{pmatrix} P_{x_i} & Q_{x_i} \end{pmatrix}$ are non-singular. Together this yields the non-singularity of

$$\begin{pmatrix} V_{x_i} & W_{x_i} \end{pmatrix}^\top G_i \begin{pmatrix} P_{x_i} & Q_{x_i} \end{pmatrix} = \begin{pmatrix} V^\top_{x_i} G_i P_{x_i} & V^\top_{x_i} G_i Q_{x_i} \\ W^\top_{x_i} G_i P_{x_i} & 0 \end{pmatrix}$$

and thereby the non-singularity of $W_{x_i}^\top G_i P_{x_i}$.
The non-singularity of $V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i}$ follows directly by Remark 4.6.
We have $\operatorname{rk} Q_{y_{\mu-1}} = \operatorname{rk} W_{y_{\mu-1}} = \operatorname{rk} P_{x_{\mu-1}} = \operatorname{rk} W_{x_{\mu-1}}$, hence $B_{y_\mu}^{\mathrm{w}} = W_{x_{\mu-1}}^\top B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}}$ is quadratic. We deal with a DAE with a Dissection Index μ. This yields

$$n = \operatorname{rk} AD + \sum_{i=1}^{\mu} \operatorname{rk} B_{y_i}^{\mathrm{w}}$$

which leads to

$$\begin{aligned}
\operatorname{rk} B_{y_\mu}^{\mathrm{w}} &= n - \operatorname{rk} AD - \sum_{i=1}^{\mu-1} \operatorname{rk} B_{y_i}^{\mathrm{w}} = \operatorname{rk} Q_{y_0} - \sum_{i=1}^{\mu-1} \operatorname{rk} B_{y_i}^{\mathrm{w}} \\
&= \operatorname{rk} Q_{y_0} - \operatorname{rk} B_{y_1}^{\mathrm{w}} - \sum_{i=2}^{\mu-1} \operatorname{rk} B_{y_i}^{\mathrm{w}} = \operatorname{rk} Q_{y_0} - \operatorname{rk} P_{y_1} - \sum_{i=2}^{\mu-1} \operatorname{rk} B_{y_i}^{\mathrm{w}} \\
&= \operatorname{rk} Q_{y_1} - \sum_{i=2}^{\mu-1} \operatorname{rk} B_{y_i}^{\mathrm{w}} \\
&\quad \vdots \\
&= \operatorname{rk} Q_{y_{\mu-1}}.
\end{aligned}$$

Hence, the matrix $B_{y_\mu}^{\mathrm{w}}$ has full rank. This yields in particular that we can choose $V_{y_i} = P_{y_i} = I$. □

Now we introduce the alternative basis chain ending by the following lemma.

Lemma 4.15.
Let the DAE (4.9) have a semi-properly stated derivative term, let f and d be sufficiently smooth and let $\mathcal{G} \subset \mathcal{D} \times \mathcal{I}$ be open and connected. Then the following two statements are equivalent to each other:

- The DAE has Dissection Index μ.
- The DAE has a Dissection Index larger than $\mu - 1$. Let W_y^* and V_y^* be the basis functions of the transposed kernel and a complementary transposed kernel of $B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}}$, respectively. Then $(W_y^*)^\top G_{\mu-1} Q_{x_{\mu-1}}$ is non-singular.

Proof.
Let the DAE have Dissection Index μ. Then $\begin{pmatrix} G_{\mu-1} Q_{x_{\mu-1}} & B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}} \end{pmatrix}$ is non-singular due to

$$\begin{pmatrix} V_{x_{\mu-1}}^\top \\ W_{x_{\mu-1}}^\top \end{pmatrix} \begin{pmatrix} G_{\mu-1} Q_{x_{\mu-1}} & B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}} \end{pmatrix} = \begin{pmatrix} V_{x_{\mu-1}}^\top G_{\mu-1} Q_{x_{\mu-1}} & V_{x_{\mu-1}}^\top B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}} \\ 0 & W_{x_{\mu-1}}^\top B_{y_{\mu-1}}^{\mathrm{v}} Q_{y_{\mu-1}} \end{pmatrix}$$

being non-singular. But thereby

$$\begin{pmatrix} (V_y^*)^\top \\ (W_y^*)^\top \end{pmatrix} \begin{pmatrix} G_{\mu-1}Q_{x_{\mu-1}} & B^{\mathrm{v}}_{y_{\mu-1}}Q_{y_{\mu-1}} \end{pmatrix} = \begin{pmatrix} (V_y^*)^\top G_{\mu-1}Q_{x_{\mu-1}} & (V_y^*)^\top B^{\mathrm{v}}_{y_{\mu-1}}Q_{y_{\mu-1}} \\ (W_y^*)^\top G_{\mu-1}Q_{x_{\mu-1}} & 0 \end{pmatrix}$$

is non-singular, hence $(W_y^*)^\top G_{\mu-1}Q_{x_{\mu-1}}$ is non-singular. The other direction of the proof can be shown analogously. □

The alternative chain ending in Lemma 4.15 is helpful since it may happen that W_y^* is easier to calculate than $W_{x_{\mu-1}}$. If we further denote $B_{x_0} := BP$ and $B_{y_0} := BQ$ we can formulate the following lemma.

Lemma 4.16.
Let a DAE (4.9) have a finite Dissection Index $\mu \in \mathbb{N}$ then it holds:

$$W_{y_i}^\top B^{\mathrm{w}}_{x_i} P_{x_i} \text{ is non-singular for } i = 1, ..., \mu.$$

Proof.
With the definition of V and W we obtain

$$n = \operatorname{rk} V + \operatorname{rk} W = \operatorname{rk} AD + \operatorname{rk} W_{x_0}$$

while the definition of the Dissection Index provides

$$n = r_0 + \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i} = \operatorname{rk} AD + \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i}$$

and therefore we achieve

$$\operatorname{rk} W_{x_0} = \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i}.$$

With the definition of P_{x_i} and P_{y_i} we obtain

$$\begin{aligned} \operatorname{rk} P_{x_i} + \operatorname{rk} P_{y_i} &= \operatorname{rk} W_{y_i}^\top B^{\mathrm{w}}_{x_i} + \operatorname{rk} B^{\mathrm{w}}_{y_i} = \operatorname{rk} \begin{pmatrix} B^{\mathrm{w}}_{x_i} & B^{\mathrm{w}}_{y_i} \end{pmatrix} \\ &= \operatorname{rk} (W_{x_{i-1}}^\top \begin{pmatrix} B_{x_{i-1}} & B_{y_{i-1}} \end{pmatrix}) \leqslant \operatorname{rk} W_{x_{i-1}}^\top = \operatorname{rk} W_{x_{i-1}} \end{aligned} \tag{4.11}$$

on $\mathbb{R}^{i \cdot m} \times \mathcal{G}$ for $i = 1, ..., \mu$, since the columns of W_{y_i} are a basis of $\ker B^{\mathrm{w}}_{y_i}$. Furthermore we get

$$\operatorname{rk} W_{x_i} = \dim(\ker Q_{x_i}^\top G_i^\top) = \dim(\ker Q_{x_i}^\top) = \dim(\operatorname{im} P_{x_i}) = \operatorname{rk} P_{x_i}$$

and

$$\operatorname{rk} B^{\mathrm{w}}_{y_i} = \dim(\operatorname{im} B^{\mathrm{w}}_{y_i}) = \dim(\operatorname{im} P_{y_i}) = \operatorname{rk} P_{y_i} \tag{4.12}$$

for $i = 1, ..., \mu$. Together we obtain

$$\operatorname{rk} P_{x_i} + \operatorname{rk} B^{\mathrm{w}}_{y_i} = \operatorname{rk} P_{x_i} + \operatorname{rk} P_{y_i} \leqslant \operatorname{rk} W_{x_{i-1}} = \operatorname{rk} P_{x_{i-1}} \text{ for } i = 2, ..., \mu \tag{4.13}$$

We assume that at least one of the inequalities in (4.13) is strict which would lead to

$$\operatorname{rk} W_{x_0} \overset{(4.11)}{=} \operatorname{rk} P_{x_1} + \operatorname{rk} P_{y_1} \overset{(4.12)}{=} \operatorname{rk} P_{x_1} + \operatorname{rk} B^{\mathrm{w}}_{y_1} \overset{(4.13)}{>} \operatorname{rk} P_{x_\mu} + \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i} \geqslant \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i}.$$

But this would be a contradiction to $\operatorname{rk} W_{x_0} = \sum_{i=1}^{\mu} \operatorname{rk} B^{\mathrm{w}}_{y_i}$ and therefore it holds

$$\operatorname{rk} P_{x_i} + \operatorname{rk} P_{y_i} = \operatorname{rk} W_{x_{i-1}}.$$

Due to $B^{\mathrm{w}}_{y_i} = W^\top_{x_{i-1}} B_{y_{i-1}}$ and V_{y_i} and W_{y_i} being basis functions with respect to the transposed complementary kernel and the transposed kernel of $B^{\mathrm{w}}_{y_i}$ we obtain

$$\operatorname{rk} P_{x_i} = \operatorname{rk} W_{x_{i-1}} - \operatorname{rk} P_{y_i} = \operatorname{rk} W_{x_{i-1}} - \operatorname{rk} V_{y_i} = \operatorname{rk} W_{y_i}$$

and by $\operatorname{rk} P_{x_i} = \operatorname{rk} W^\top_{y_i} B^{\mathrm{w}}_{x_i}$ it follows that $W^\top_{y_i} B^{\mathrm{w}}_{x_i} P_{x_i}$ is non-singular for all $i = 1, ..., \mu$. □

Before we can use the Dissection Index there are two fundamental properties which must hold to enable us to call the Dissection Index well defined. First of all the value of the index is not allowed to depend on the choice of the basis functions. Before we proof this statement we formulate two technical lemmata.

Lemma 4.17.
Let $\mathcal{G} \subset \mathbb{R}^k$ be C^l-diffeomorphic to a parallelepiped in $\mathbb{R}^k$ and let be $M \in C^l(\mathcal{G}, \mathbb{R}^{m \times n})$. Furthermore, suppose that $dim(\operatorname{im} M(z)) = r$ for all $z \in \mathcal{G}$. By Lemma 4.7 there exists a matrix valued function $Q \in C^l(\mathcal{G}, \mathbb{R}^{n \times (n-r)})$ with $\operatorname{im} Q(z) = \ker M(z)$ for all $z \in \mathcal{G}$. Let $\bar{Q} \in C^l(\mathcal{G}, \mathbb{R}^{n \times (n-r)})$ be another matrix valued function with $\operatorname{im} \bar{Q}(z) = \ker M(z)$ for all $z \in \mathcal{G}$. Then there exists a transformation function $T \in C^l(\mathcal{G}, \mathbb{R}^{(n-r) \times (n-r)})$ such that $\bar{Q}(z) = Q(z)T(z)$ and $T(z)$ being non-singular for all $z \in \mathcal{G}$.

Proof.
The matrix $Q(z)$ has full column rank, hence we can choose

$$T(z) := (Q^\top(z)Q(z))^{-1} Q^\top(z)\bar{Q}(z)$$

with $T(z) \in C^l(\mathcal{G}, \mathbb{R}^{(n-r)\times(n-r)})$ since the inverse matrix A^{-1} of a matrix A is as smooth as the matrix A itself. We assume that $T(z)$ is singular for a $z \in \mathcal{G}$, hence there would be a $x \in \mathbb{R}^{n-r}$ with $x \neq 0$ and $q^\top(z)\bar{Q}(z)x = 0$. The matrix $\bar{Q}(z)$ has full column rank, thus $\bar{Q}(z)x \neq 0$ and $\bar{Q}(z)x \in \ker M(z)$. Therefore $\bar{Q}(z)x$ is a nonzero element of the kernel of $M(z)$ which is perpendicular to a basis of the kernel of $M(z)$. This is a contradiction and therefore the assumption is wrong. □

Lemma 4.18.
Let $\mathcal{G} \subset \mathbb{R}^k$ be C^l-diffeomorphic to a parallelepiped in $\mathbb{R}^k$ and let be $M \in C^l(\mathcal{G}, \mathbb{R}^{m\times n})$. Furthermore, suppose that $dim(\operatorname{im} M(z)) = r$ for all $z \in \mathcal{G}$. By Lemma 4.7 there exists a matrix valued function $Q \in C^l(\mathcal{G}, \mathbb{R}^{n\times(n-r)})$ with $\operatorname{im} Q(z) = \ker M(z)$ for all $z \in \mathcal{G}$. Let $P \in C^l(\mathcal{G}, \mathbb{R}^{n\times r})$ and $\bar{P} \in C^l(\mathcal{G}, \mathbb{R}^{n\times r})$ be two matrix valued functions with $\operatorname{im} \bar{P}(z) \oplus \ker M(z) = \operatorname{im} \bar{P}(z) \oplus \ker M(z) = \mathbb{R}^n$ for all $z \in \mathcal{G}$. Then there exist two transformation functions $T \in C^l(\mathcal{G}, \mathbb{R}^{r\times r})$ and $M_q \in C^l(\mathcal{G}, \mathbb{R}^{(n-r)\times r})$ such that $\bar{P}(z) = P(z)T(z) + Q(z)M_q(z)$ with $T(z)$ being non-singular for all $z \in \mathcal{G}$.

Proof.
The matrix $\begin{pmatrix} P(z) & Q(z) \end{pmatrix}$ is non-singular, hence we can choose

$$\begin{pmatrix} T(z) \\ M_q(z) \end{pmatrix} := \begin{pmatrix} P(z) & Q(z) \end{pmatrix}^{-1} \bar{P}(z).$$

Assume that there is a $z \in \mathcal{G}$ such that $T(z)$ does not have full row rank. Then there would be a $x \in \operatorname{im} \bar{P}(z)$ with $x \in \operatorname{im} Q(z)$ which is a contradiction to $\operatorname{im} \bar{P}(z) \oplus \ker M(z) = \mathbb{R}^n$, $\bar{P} \in C^l(\mathcal{G}, \mathbb{R}^{n\times r})$ and $dim(\operatorname{im} M(z)) = r$. Hence $T(z)$ is quadratic and $T(z)$ has full row rank. □

With the help of these lemmata we are able to prove:

Theorem 4.19. (Rank independence)
Consider a DAE (4.9) with a semiproperly stated derivative term and let $\mathcal{G} \subset \mathcal{D} \times \mathcal{I}$ be open and connected. Let, for a given $\mu \in \mathbb{N}$, a basis functions sequence, associated to the DAE, exist. Then the characteristic values $r_0, ..., r_\mu$ and the Dissection Index itself are independent of the special choice of the involved basis functions.

Proof. To prove Theorem 4.19 we have to show that the ranks of AD and $B^{\mathrm{w}}_{y_i}$ are independent of the choice of the basis functions for all $1 \leqslant i \leqslant \mu$. Obviously $\operatorname{rk} AD$ does not depend on the basis functions.
We define two different basis chains

$$P, Q, V, W, P_{x_i}, Q_{x_i}, P_{y_i}, Q_{y_i}, V_{x_i}, W_{x_i}, V_{y_i}, W_{y_i}$$

and

$$\bar{P}, \bar{Q}, \bar{V}, \bar{W}, \bar{P}_{x_i}, \bar{Q}_{x_i}, \bar{P}_{y_i}, \bar{Q}_{y_i}, \bar{V}_{x_i}, \bar{W}_{x_i}, \bar{V}_{y_i}, \bar{W}_{y_i}$$

and show in the following that $\operatorname{rk} B^{\mathrm{w}}_{y_i}$ is equal for both basis chains for all $1 \leqslant i \leqslant \mu$.
We know that Q and $\bar{Q}$ and W and $\bar{W}$ are basis functions for the same subspace, respectively. So there are two non-singular matrices T_Q and T_W which serve as coordinate transformations such that

$$\bar{Q} = QT_Q \text{ and } \bar{W} = WT_W.$$

By Lemma 4.18 there are transformation matrices T_P, T_V, M_Q and M_W with T_P and T_V being non-singular such that

$$\bar{P} = PT_P + QM_Q \text{ and } \bar{V} = VT_V + WM_W.$$

We are able to choose a continuously differentiable transformation matrix T_Q since Q and $\bar{Q}$ are continuously differentiable, see Lemma 4.17.
For $i \geqslant 0$ we show with the help of an induction that there are suitable coordinate transformations $T_{V_{x_{i+1}}}, T_{W_{x_{i+1}}}, T_{Q_{x_{i+1}}}, T_{Q_{y_{i+1}}}, M_{W_{x_{i+1}}}$ and $M_{Q_{x_{i+1}}}$ and matrices $X_{Q_{x_{i+1}}}$ such that

$$\begin{aligned}
\bar{G}_{i+1} =& T_{V_{x_i}}^\top G_{i+1} T_{Q_{x_i}} \\
\bar{B}^{\mathrm{v}}_{x_{i+1}} =& T_{V_{x_i}}^\top B^{\mathrm{v}}_{x_{i+1}} T_{Q_{x_i}} + M_{W_{x_i}}^\top B^{\mathrm{w}}_{x_{i+1}} T_{Q_{x_i}} + T_{V_{x_i}}^\top G_{i+1} T'_{Q_{x_i}} \\
& + T_{V_{x_i}}^\top B^{\mathrm{v}}_{y_{i+1}} X_{Q_{x_i}} + M_{W_{x_i}}^\top B^{\mathrm{w}}_{y_{i+1}} X_{Q_{x_i}} \\
\bar{B}^{\mathrm{v}}_{y_{i+1}} =& T_{V_{x_i}}^\top B^{\mathrm{v}}_{y_{i+1}} T_{Q_{y_i}} + M_{W_{x_i}}^\top B^{\mathrm{w}}_{y_{i+1}} T_{Q_{y_i}} \\
\bar{B}^{\mathrm{w}}_{x_{i+1}} =& T_{W_{x_i}}^\top B^{\mathrm{w}}_{x_{i+1}} T_{Q_{x_i}} + T_{W_{x_i}}^\top B^{\mathrm{w}}_{y_{i+1}} X_{Q_{x_i}} \\
\bar{B}^{\mathrm{w}}_{y_{i+1}} =& T_{W_{x_i}}^\top B^{\mathrm{w}}_{y_{i+1}} T_{Q_{y_i}}
\end{aligned}$$

and

$$\begin{array}{rclrcl}
\bar{Q}_{y_{i+1}} & = & T_{Q_{y_i}}^{-1} Q_{y_{i+1}} T_{Q_{y_{i+1}}}, & \bar{P}_{y_{i+1}} & = & T_{Q_{y_i}}^{-1} (P_{y_{i+1}} T_{P_{y_{i+1}}} + Q_{y_{i+1}} M_{Q_{y_{i+1}}}), \\
\bar{W}_{y_{i+1}} & = & T_{W_{x_i}}^{-1} W_{y_{i+1}} T_{W_{y_{i+1}}}, & \bar{V}_{y_{i+1}} & = & T_{W_{x_i}}^{-1} (V_{y_{i+1}} T_{V_{y_{i+1}}} + W_{y_{i+1}} M_{W_{y_{i+1}}}) \\
\bar{Q}_{x_{i+1}} & = & T_{Q_{x_i}}^{-1} Q_{x_{i+1}} T_{Q_{x_{i+1}}}, & \bar{P}_{x_{i+1}} & = & T_{Q_{x_i}}^{-1} (P_{x_{i+1}} T_{P_{x_{i+1}}} + Q_{x_{i+1}} M_{Q_{x_{i+1}}}) \\
\bar{W}_{x_{i+1}} & = & T_{V_{x_i}}^{-1} W_{x_{i+1}} T_{W_{x_{i+1}}}, & \bar{V}_{x_{i+1}} & = & T_{V_{x_i}}^{-1} (V_{x_{i+1}} T_{V_{x_{i+1}}} + W_{x_{i+1}} M_{W_{x_{i+1}}})
\end{array}$$

holds with $M_{W_{x_0}} = M_W$, $M_{Q_{x_0}} = M_Q$, $X_{Q_{x_0}} = M_Q$, $T_{V_{x_0}} = T_V$, $T_{Q_{x_0}} = T_P$, $T_{W_{x_0}} = T_W$ and $T_{Q_{y_0}} = T_Q$.

Induction start $(i = 0)$
With these coordinate transformation matrices we can write

$$\begin{aligned}
\bar{G}_1 &= \bar{V}^\top AD\bar{P} = (T_V^\top V^\top + M_W^\top W^\top)AD(PT_P + QM_Q) = T_V^\top V^\top ADPT_P \\
&= T_V^\top G_1 T_P, \\
\bar{B}^{\mathrm{w}}_{y_1} &= \bar{W}^\top B\bar{Q} = T_W^\top W^\top BQT_Q \\
&= T_W^\top B^{\mathrm{w}}_{y_1} T_Q, \\
\bar{B}^{\mathrm{w}}_{x_1} &= \bar{W}^\top B\bar{P} = T_W^\top W^\top B(PT_P + QM_Q) \\
&= T_W^\top B^{\mathrm{w}}_{x_1} T_P + T_W^\top B^{\mathrm{w}}_{y_1} M_Q, \\
\bar{B}^{\mathrm{v}}_{y_1} &= \bar{V}^\top A(D\bar{Q})' + \bar{V}^\top B\bar{Q}
\end{aligned}$$

$$\begin{aligned}
&= (T_V^\top V^\top + M_W^\top W^\top)A(DQT_Q)' + (T_V^\top V^\top + M_W^\top W^\top)BQT_Q \\
&= T_V^\top V^\top A(DQT_Q)' + T_V^\top V^\top BQT_Q + M_W^\top W^\top BQT_Q \\
&= T_V^\top V^\top ADQT_Q' + T_V^\top V^\top A(DQ)'T_Q + T_V^\top V^\top BQT_Q + M_W^\top W^\top BQT_Q \\
&= T_V^\top V^\top A(DQ)'T_Q + T_V^\top V^\top BQT_Q + M_W^\top W^\top BQT_Q \\
&= T_V^\top B_{y_1}^{\mathrm{v}} T_Q + M_W^\top B_{y_1}^{\mathrm{w}} T_Q
\end{aligned}$$

and

$$\begin{aligned}
\bar{B}_{x_1}^{\mathrm{v}} =& \bar{V}^\top A(D\bar{P})' + \bar{V}^\top B\bar{P} \\
=& (T_V^\top V^\top + M_W^\top W^\top)A(D(PT_P + QM_Q))' + (T_V^\top V^\top + M_W^\top W^\top)B(PT_P + QM_Q) \\
=& T_V^\top V^\top A(D(PT_P + QM_Q))' + (T_V^\top V^\top + M_W^\top W^\top)B(PT_P + QM_Q) \\
=& T_V^\top V^\top A(DPT_P)' + T_V^\top V^\top A(DQM_Q)' + M_W^\top B_{x_1}^{\mathrm{w}} T_P + M_W^\top B_{y_1}^{\mathrm{w}} M_Q \\
&+ T_V^\top V^\top BPT_P + T_V^\top V^\top BQM_Q \\
=& T_V^\top V^\top ADPT_P' + T_V^\top V^\top ADQM_Q' + M_W^\top B_{x_1}^{\mathrm{w}} T_P + M_W^\top B_{y_1}^{\mathrm{w}} M_Q \\
&+ T_V^\top (V^\top BP + V^\top A(DP)')T_P + T_V^\top (V^\top BQ + V^\top A(DQ)')M_Q \\
=& T_V^\top B_{x_1}^{\mathrm{v}} T_P + T_V^\top G_1 T_P' + T_V^\top B_{y_1}^{\mathrm{v}} M_Q + M_W^\top B_{x_1}^{\mathrm{w}} T_P + M_W^\top B_{y_1}^{\mathrm{w}} M_Q
\end{aligned}$$

So $\bar{Q}_{y_1}^\star := T_Q^{-1} Q_{y_1}$ is a possible choice as a basis function of the next sequence of the basis chain. Hence we can find a suitable coordinate transformation $T_{Q_{y_1}}$ such that $\bar{Q}_{y_1} = T_Q^{-1} Q_{y_1} T_{Q_{y_1}}$. This procedure yields suitable coordinate transformations such that:

$$\begin{array}{rclrcl}
\bar{Q}_{y_1} &=& T_Q^{-1} Q_{y_1} T_{Q_{y_1}}, & \bar{W}_{y_1} &=& T_W^{-1} W_{y_1} T_{W_{y_1}}, \\
\bar{P}_{y_1} &=& T_Q^{-1}(P_{y_1} T_{P_{y_1}} + Q_{y_1} M_{Q_{y_1}}), & \bar{V}_{y_1} &=& T_W^{-1}(V_{y_1} T_{V_{y_1}} + W_{y_1} M_{W_{y_1}}).
\end{array}$$

We also find suitable coordinate transformations such that:

$$\bar{Q}_{x_1} = T_P^{-1} Q_{x_1} T_{Q_{x_1}}, \qquad \bar{P}_{x_1} = T_P^{-1}(P_{x_1} T_{P_{x_1}} + Q_{x_1} M_{Q_{x_1}})$$

due to

$$\bar{W}_{y_1}^\top \bar{B}_{x_1}^{\mathrm{w}} = T_{W_{y_1}}^\top W_{y_1}^\top (T_W^\top)^{-1} (T_W^\top B_{x_1}^{\mathrm{w}} T_P + T_W^\top B_{y_1}^{\mathrm{w}} M_Q) = T_{W_{y_1}}^\top W_{y_1}^\top B_{x_1}^{\mathrm{w}} T_P$$

and

$$\bar{W}_{x_1} = T_V^{-1} W_{x_1} T_{W_{x_1}}, \qquad \bar{V}_{x_1} = T_V^{-1}(V_{x_1} T_{V_{x_1}} + W_{x_1} M_{W_{x_1}})$$

due to

$$\bar{G}_1 \bar{Q}_{x_1} = T_V^\top G_1 T_P T_P^{-1} Q_{x_1} T_{Q_{x_1}} = T_V^\top G_1 Q_{x_1} T_{Q_{x_1}}.$$

Induction step $(1, ..., i \to i+1)$
We directly get the following relations:

$$
\begin{aligned}
\bar{G}_{i+1} &= \bar{V}_{x_i}^\top \bar{G}_i \bar{Q}_{x_i} \\
&= ((T_{V_{x_i}}^\top V_{x_i}^\top + M_{W_{x_i}}^\top W_{x_i}^\top) T_{V_{x_{i-1}}}^{-T})(T_{V_{x_{i-1}}}^T G_i T_{Q_{x_{i-1}}})(T_{Q_{x_{i-1}}}^{-1} Q_{x_i} T_{Q_{x_i}}) \\
&= T_{V_{x_i}}^\top V_{x_i}^\top G_i Q_{x_i} T_{Q_{x_i}} \\
&= T_{V_{x_i}}^\top G_{i+1} T_{Q_{x_i}}, \\
\bar{B}_{y_{i+1}}^{\mathrm{v}} &= \bar{V}_{x_i}^\top \bar{B}_{y_i}^{\mathrm{v}} \bar{Q}_{y_i} \\
&= ((T_{V_{x_i}}^\top V_{x_i}^\top + M_{W_{x_i}}^\top W_{x_i}^\top) T_{V_{x_{i-1}}}^{-T})(T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} T_{Q_{y_{i-1}}} + M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} T_{Q_{y_{i-1}}})(T_{Q_{y_{i-1}}}^{-1} Q_{y_i} T_{Q_{y_i}}) \\
&= (T_{V_{x_i}}^\top V_{x_i}^\top + M_{W_{x_i}}^\top W_{x_i}^\top) B_{y_i}^{\mathrm{v}} Q_{y_i} T_{Q_{y_i}} \\
&= T_{V_{x_i}}^\top V_{x_i}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} T_{Q_{y_i}} + M_{W_{x_i}}^\top W_{x_i}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} T_{Q_{y_i}} \\
&= T_{V_{x_i}}^\top B_{y_{i+1}}^{\mathrm{v}} T_{Q_{y_i}} + M_{W_{x_i}}^\top B_{y_{i+1}}^{\mathrm{w}} T_{Q_{y_i}}, \\
\bar{B}_{y_{i+1}}^{\mathrm{w}} &= \bar{W}_{x_i}^\top \bar{B}_{y_i}^{\mathrm{v}} \bar{Q}_{y_i} \\
&= (T_{W_{x_i}}^\top W_{x_i}^\top T_{V_{x_{i-1}}}^{-T})(T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} T_{Q_{y_{i-1}}} + M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} T_{Q_{y_{i-1}}})(T_{Q_{y_{i-1}}}^{-1} Q_{y_i} T_{Q_{y_i}}) \\
&= T_{W_{x_i}}^\top B_{y_{i+1}}^{\mathrm{w}} T_{Q_{y_i}}.
\end{aligned}
$$

To prove the induction step statements regarding $\bar{B}_{x_{i+1}}^{\mathrm{v}}$ and $\bar{B}_{x_{i+1}}^{\mathrm{w}}$ we need several preparation steps. First we obtain

$$
\begin{aligned}
&\bar{B}_{y_i}^{\mathrm{v}} \bar{P}_{y_i} (\bar{V}_{y_i}^\top \bar{B}_{y_i}^{\mathrm{w}} \bar{P}_{y_i})^{-1} \bar{V}_{y_i}^\top \bar{B}_{x_i}^{\mathrm{w}} \bar{Q}_{x_i} \\
=&T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} P_{y_i} T_{P_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{x_i}^{\mathrm{w}} Q_{x_i} T_{Q_{x_i}} \\
&+ T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} M_{Q_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{x_i}^{\mathrm{w}} Q_{x_i} T_{Q_{x_i}} \\
&+ M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{x_i}^{\mathrm{w}} Q_{x_i} T_{Q_{x_i}} \\
&+ T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} P_{y_i} T_{P_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} X_{Q_{x_{i-1}}} \bar{Q}_{x_i} \\
&+ T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} M_{Q_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} X_{Q_{x_{i-1}}} \bar{Q}_{x_i} \\
&+ M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}} (T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}})^{-1} T_{V_{y_i}}^\top V_{y_i}^\top B_{y_i}^{\mathrm{w}} X_{Q_{x_{i-1}}} \bar{Q}_{x_i}
\end{aligned}
$$

by the inductive arguments

$$
\begin{aligned}
&\bar{B}_{y_i}^{\mathrm{v}} \bar{P}_{y_i} \\
=&(T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} T_{Q_{y_{i-1}}} + M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} T_{Q_{y_{i-1}}})(T_{Q_{y_{i-1}}}^{-1} (P_{y_i} T_{P_{y_i}} + Q_{y_i} M_{Q_{y_i}})) \\
=&T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} P_{y_i} T_{P_{y_i}} + T_{V_{x_{i-1}}}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} M_{Q_{y_i}} + M_{W_{x_{i-1}}}^\top B_{y_i}^{\mathrm{w}} P_{y_i} T_{P_{y_i}}
\end{aligned}
$$

and

$$\begin{aligned}&\bar{V}_{y_i}^{\top}\bar{B}_{x_i}^{\mathrm{w}}\bar{Q}_{x_i}\\ =&(T_{V_{y_i}}^{\top}V_{y_i}^{\top}+M_{W_{y_i}}^{\top}W_{y_i}^{\top})(B_{x_i}^{\mathrm{w}}T_{Q_{x_{i-1}}}+B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}})\bar{Q}_{x_i}\\ =&(T_{V_{y_i}}^{\top}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}T_{Q_{x_{i-1}}}+T_{V_{y_i}}^{\top}V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}+M_{W_{y_i}}^{\top}W_{y_i}^{\top}B_{x_i}^{\mathrm{w}}T_{Q_{x_{i-1}}})\bar{Q}_{x_i}\\ =&T_{V_{y_i}}^{\top}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}+T_{V_{y_i}}^{\top}V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}.\end{aligned}$$

Furthermore we are able to explicitly formulate the Moore-Penrose inverses

$$(V_{y_i}^{\top})^{+}=B_{y_i}^{\mathrm{w}}P_{y_i}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}$$

and

$$P_{y_i}^{+}=(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}B_{y_i}^{\mathrm{w}},$$

which enables us to write

$$\begin{aligned}&\bar{B}_{y_i}^{\mathrm{v}}\bar{P}_{y_i}(\bar{V}_{y_i}^{\top}\bar{B}_{y_i}^{\mathrm{w}}\bar{P}_{y_i})^{-1}\bar{V}_{y_i}^{\top}\bar{B}_{x_i}^{\mathrm{w}}\bar{Q}_{x_i}\\ =&T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}P_{y_i}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}\\ &+T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}Q_{y_i}M_{Q_{y_i}}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i}T_{P_{y_i}})^{-1}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}\\ &+M_{W_{x_{i-1}}}^{\top}((V_{y_i}^{\top})^{+}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i})T_{Q_{x_i}}\\ &+T_{V_{x_{i-1}}}^{\top}(B_{y_i}^{\mathrm{v}}P_{y_i}P_{y_i}^{+})X_{Q_{x_{i-1}}}\bar{Q}_{x_i}\\ &+T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}Q_{y_i}M_{Q_{y_i}}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i}T_{P_{y_i}})^{-1}V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}\\ &+M_{W_{x_{i-1}}}^{\top}(B_{y_i}^{\mathrm{w}}P_{y_i}P_{y_i}^{+})X_{Q_{x_{i-1}}}\bar{Q}_{x_i}.\end{aligned}$$

Using $B_{y_i}^{\mathrm{w}}P_{y_i}P_{y_i}^{+}=B_{y_i}^{\mathrm{w}}$, $(V_{y_i}^{\top})^{+}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}=B_{x_i}^{\mathrm{w}}Q_{x_i}$ and $P_{y_i}P_{y_i}^{+}=I-Q_{y_i}Q_{y_i}^{+}$ we obtain

$$\begin{aligned}&\bar{B}_{y_i}^{\mathrm{v}}\bar{P}_{y_i}(\bar{V}_{y_i}^{\top}\bar{B}_{y_i}^{\mathrm{w}}\bar{P}_{y_i})^{-1}\bar{V}_{y_i}^{\top}\bar{B}_{x_i}^{\mathrm{w}}\bar{Q}_{x_i}\\ =&T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}P_{y_i}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}\\ &+T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}Q_{y_i}(M_{Q_{y_i}}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i}T_{P_{y_i}})^{-1}V_{y_i}^{\top}(B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}+B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i})-Q_{y_i}^{+}X_{Q_{x_{i-1}}}\bar{Q}_{x_i})\\ &+T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}+M_{W_{x_{i-1}}}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}+M_{W_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}\\ =&T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}P_{y_i}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}\\ &-T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}Q_{y_i}X_{Q_{x_i}}+T_{V_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{v}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}+M_{W_{x_{i-1}}}^{\top}B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}+M_{W_{x_{i-1}}}^{\top}B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i}\end{aligned}$$

with $X_{Q_{x_i}}:=-(M_{Q_{y_i}}(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i}T_{P_{y_i}})^{-1}V_{y_i}^{\top}(B_{x_i}^{\mathrm{w}}Q_{x_i}T_{Q_{x_i}}+B_{y_i}^{\mathrm{w}}X_{Q_{x_{i-1}}}\bar{Q}_{x_i})-Q_{y_i}^{+}X_{Q_{x_{i-1}}}\bar{Q}_{x_i})$. Further we see that

$$\bar{B}_{x_i}^{\mathrm{v}}\bar{Q}_{x_i}$$

$$\begin{aligned}
&=(T_{V_{x_{i-1}}}^{\top} B_{x_i}^{\mathrm{v}} T_{Q_{x_{i-1}}} + T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}}' + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} X_{Q_{x_{i-1}}} \\
&\quad + M_{W_{x_{i-1}}}^{\top} B_{x_i}^{\mathrm{w}} T_{Q_{x_{i-1}}} + M_{W_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{w}} X_{Q_{x_{i-1}}})\bar{Q}_{x_i} \\
&=T_{V_{x_{i-1}}}^{\top} B_{x_i}^{\mathrm{v}} Q_{x_i} T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}}' T_{Q_{x_{i-1}}}^{-1} Q_{x_i} T_{Q_{x_i}} \\
&\quad + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} M_{Q_{x_{i-1}}} \bar{Q}_{x_i} + M_{W_{x_{i-1}}}^{\top} B_{x_i}^{\mathrm{w}} Q_{x_i} T_{Q_{x_i}} + M_{W_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{w}} X_{Q_{x_{i-1}}} \bar{Q}_{x_i}
\end{aligned}$$

and

$$\begin{aligned}
&\bar{G}_i \bar{Q}_{x_1}' \\
=&T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}} (T_{Q_{x_{i-1}}}^{-1} Q_{x_i} T_{Q_{x_i}})' \\
=&T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}} ((T_{Q_{x_{i-1}}}^{-1})' Q_{x_i} T_{Q_{x_i}} + T_{Q_{x_{i-1}}}^{-1} Q_{x_i}' T_{Q_{x_i}} + T_{Q_{x_{i-1}}}^{-1} Q_{x_i} T_{Q_{x_i}}') \\
=&T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i}' T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T_{Q_{x_i}}' + T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}} (T_{Q_{x_{i-1}}}^{-1})' Q_{x_i} T_{Q_{x_i}}
\end{aligned}$$

hold, which yields together with

$$\begin{aligned}
&T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}}' T_{Q_{x_{i-1}}}^{-1} Q_{x_i} T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i T_{Q_{x_{i-1}}} (T_{Q_{x_{i-1}}}^{-1})' Q_{x_i} T_{Q_{x_i}} \\
=&T_{V_{x_{i-1}}}^{\top} G_i (T_{Q_{x_{i-1}}} T_{Q_{x_{i-1}}}^{-1})' Q_{x_i} T_{Q_{x_i}} = T_{V_{x_{i-1}}}^{\top} G_i (I)' Q_{x_i} T_{Q_{x_i}} = 0
\end{aligned}$$

the induction step statement for $\bar{B}_{x_{i+1}}$:

$$\begin{aligned}
\bar{B}_{x_i} =&\bar{G}_i \bar{Q}_{x_i}' + \bar{B}_{x_i}^{\mathrm{v}} \bar{Q}_{x_i} - \bar{B}_{y_i}^{\mathrm{v}} \bar{P}_{y_i} (\bar{V}_{y_i}^{\top} \bar{B}_{y_i}^{\mathrm{w}} \bar{P}_{y_i})^{-1} \bar{V}_{y_i}^{\top} \bar{B}_{x_i}^{\mathrm{w}} \bar{Q}_{x_i} \\
=&T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i}' T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} B_{x_i}^{\mathrm{v}} Q_{x_i} T_{Q_{x_i}} - T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} P_{y_i} (V_{y_i}^{\top} B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^{\top} B_{x_i}^{\mathrm{w}} Q_{x_i} T_{Q_{x_i}} \\
&+ T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T_{Q_{x_i}}' + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} Q_{y_i} X_{Q_{x_i}} \\
=&T_{V_{x_{i-1}}}^{\top} (G_i Q_{x_i}' + B_{x_i}^{\mathrm{v}} Q_{x_i} - B_{y_i}^{\mathrm{v}} P_{y_i} (V_{y_i}^{\top} B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^{\top} B_{x_i}^{\mathrm{w}} Q_{x_i}) T_{Q_{x_i}} \\
&+ T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T_{Q_{x_i}}' + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} Q_{y_i} X_{Q_{x_i}} \\
=&T_{V_{x_{i-1}}}^{\top} B_{x_i} T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T_{Q_{x_i}}' + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} Q_{y_i} X_{Q_{x_i}}.
\end{aligned}$$

Immediately we gain

$$\begin{aligned}
&\bar{B}_{x_{i+1}}^{\mathrm{v}} = \bar{V}_{x_i}^{\top} \bar{B}_{x_i} \\
=&((T_{V_{x_i}}^{\top} V_{x_i}^{\top} + M_{W_{x_i}}^{\top} W_{x_i}^{\top}) T_{V_{x_{i-1}}}^{-T})(T_{V_{x_{i-1}}}^{\top} B_{x_i} T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T_{Q_{x_i}}' + T_{V_{x_{i-1}}}^{\top} B_{y_i}^{\mathrm{v}} Q_{y_i} X_{Q_{x_i}}) \\
=&(T_{V_{x_i}}^{\top} V_{x_i}^{\top} + M_{W_{x_i}}^{\top} W_{x_i}^{\top})(B_{x_i} T_{Q_{x_i}} + G_i Q_{x_i} T_{Q_{x_i}}' + B_{y_i}^{\mathrm{v}} Q_{y_i} X_{Q_{x_i}}) \\
=&T_{V_{x_i}}^{\top} B_{x_{i+1}}^{\mathrm{v}} T_{Q_{x_i}} + T_{V_{x_i}}^{\top} G_{i+1} T_{Q_{x_i}}' + T_{V_{x_i}}^{\top} B_{y_{i+1}}^{\mathrm{v}} X_{Q_{x_i}} + M_{W_{x_i}}^{\top} B_{x_{i+1}}^{\mathrm{w}} T_{Q_{x_i}} + M_{W_{x_i}}^{\top} B_{y_i}^{\mathrm{w}} X_{Q_{x_i}}
\end{aligned}$$

and

$$\bar{B}_{x_{i+1}}^{\mathrm{w}} =\bar{W}_{x_i}^{\top} \bar{B}_{x_i}$$

$$
\begin{aligned}
&=(T_{W_{x_i}}^{\top} W_{x_i}^{\top} T_{V_{x_{i-1}}}^{-T})(T_{V_{x_{i-1}}}^{\top} B_{x_i} T_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} G_i Q_{x_i} T'_{Q_{x_i}} + T_{V_{x_{i-1}}}^{\top} B^{\mathrm{v}}_{y_i} Q_{y_i} X_{Q_{x_i}}) \\
&=T_{W_{x_i}}^{\top} B^{\mathrm{w}}_{x_{i+1}} T_{Q_{x_i}} + T_{W_{x_i}}^{\top} B^{\mathrm{w}}_{y_{i+1}} X_{Q_{x_i}}.
\end{aligned}
$$

Now we turn towards the induction statements regarding the basis functions. By the identity

$$
\bar{B}^{\mathrm{w}}_{y_{i+1}} = T_{W_{x_i}}^{\top} B^{\mathrm{w}}_{y_{i+1}} T_{Q_{y_i}}
$$

we find suitable coordinate transformations such that:

$$
\begin{aligned}
\bar{Q}_{y_{i+1}} &= T_{Q_{y_i}}^{-1} Q_{y_{i+1}} T_{Q_{y_{i+1}}}, & \bar{W}_{y_{i+1}} &= T_{W_{x_i}}^{-1} W_{y_{i+1}} T_{W_{y_{i+1}}}, \\
\bar{P}_{y_{i+1}} &= T_{Q_{y_i}}^{-1} (P_{y_{i+1}} T_{P_{y_{i+1}}} + Q_{y_{i+1}} M_{Q_{y_{i+1}}}), & \bar{V}_{y_{i+1}} &= T_{W_{x_i}}^{-1} (V_{y_{i+1}} T_{V_{y_{i+1}}} + W_{y_{i+1}} M_{W_{y_{i+1}}}).
\end{aligned}
$$

Analogously the identity

$$
\bar{W}_{y_{i+1}}^{\top} \bar{B}^{\mathrm{w}}_{x_{i+1}} = T_{W_{y_{i+1}}}^{\top} W_{y_{i+1}}^{\top} B^{\mathrm{w}}_{x_{i+1}} T_{Q_{x_i}},
$$

yields suitable coordinate transformations such that:

$$
\bar{Q}_{x_{i+1}} \quad = \quad T_{Q_{x_i}}^{-1} Q_{x_{i+1}} T_{Q_{x_{i+1}}}, \qquad \bar{P}_{x_{i+1}} \quad = \quad T_{Q_{x_i}}^{-1} (P_{x_{i+1}} T_{P_{x_{i+1}}} + Q_{x_{i+1}} M_{Q_{x_{i+1}}}).
$$

At last we find coordinate transformations such that:

$$
\bar{W}_{x_{i+1}} \quad = \quad T_{V_{x_i}}^{-1} W_{x_{i+1}} T_{W_{x_{i+1}}}, \qquad \bar{V}_{x_{i+1}} \quad = \quad T_{V_{x_i}}^{-1} (V_{x_{i+1}} T_{V_{x_{i+1}}} + W_{x_{i+1}} M_{W_{x_{i+1}}})
$$

due to

$$
\bar{G}_{i+1} \bar{Q}_{x_{i+1}} = T_{V_{x_i}}^{\top} G_{i+1} T_{Q_{x_i}} (T_{Q_{x_i}}^{-1} Q_{x_{i+1}} T_{Q_{x_{i+1}}}) = T_{V_{x_i}}^{\top} G_{i+1} Q_{x_i} T_{Q_{x_{i+1}}}
$$

and the Lemmata 4.17 and 4.18. Hence the induction step is complete. Thus we achieve $\operatorname{rk} \bar{B}^{\mathrm{w}}_{y_i} = \operatorname{rk}(T_{W_{x_{i-1}}}^{\top} B^{\mathrm{w}}_{y_i} T_{Q_{y_{i-1}}}) = \operatorname{rk} B^{\mathrm{w}}_{y_i}$, since the transformation matrices are non-singular. □

The second fundamental property is that the Dissection Index of a nonlinear DAE (4.9) has to relate to the Dissection Index of the associated linearized DAEs (2.16). The next theorem provides such a relation. But again we first formulate a technical lemma.

Lemma 4.20. (Reference function)
Let $\mathcal{G} \subset \mathbb{R}^n$ be open and convex. Let $(X^{\nu}, t_x), (Y^{\nu}, t_y) \in \mathbb{R}^{\nu \cdot n}$ with $(x, t_x), (y, t_y) \in \mathcal{G} \times \mathcal{I}$ then there is a reference function $\gamma(t)$ such that $\gamma(t_x) = x$ and $\gamma(t_y) = y$ and for all $i \leqslant \nu$ it holds that $\gamma^{(i)}(t_x) = x^i$ and $\gamma^{(i)}(t_y) = y^i$.

Proof.
First define a sequence of auxiliary functions and begin with

$$h_-(t;t_1,t_2) = -2\frac{t-t_1}{t_2-t_1}+1 \quad \Rightarrow \quad \begin{cases} h_-(t) \geqslant 1, & t \leqslant t_1 \\ h_-(t) \leqslant -1, & t \geqslant t_2 \end{cases}$$

$$h_+(t;t_1,t_2) = 2\frac{t-t_1}{t_2-t_1}-1 \quad \Rightarrow \quad \begin{cases} h_+(t) \geqslant 1, & t \geqslant t_1 \\ h_+(t) \leqslant -1, & t \leqslant t_2 \end{cases}$$

with $t_1, t_2 \in \mathcal{I}$ and $t_1 < t_2$. Next define

$$g(t) = \begin{cases} e^{-\frac{1}{t^2}}, & t > 0 \\ 0, & t \leqslant 0 \end{cases}$$

With the help of g define

$$k(t) = \frac{g(1+t)}{g(1+t)+g(1-t)} \quad \Rightarrow \quad \begin{cases} k(t) = 0, & t \leqslant -1 \\ 0 \leqslant k(t) \leqslant 1, & -1 \leqslant t \leqslant 1 \\ k(t) = 1, & t \geqslant 1 \end{cases}$$

with $k \in C^\infty(\mathbb{R})$. Let $\varepsilon_x, \varepsilon_y > 0$ and set

$$\Phi_x(t) = k(h_-(t;t_x+\varepsilon_x, t_x+2\varepsilon_x)) \quad \Rightarrow \quad \begin{cases} \Phi_x(t) = 0, & t \geqslant t_x + 2\varepsilon_x \\ \Phi_x(t) = 1, & t \leqslant t_x + \varepsilon_x. \end{cases}$$

$$\Phi_{xy}(t) = k(h_+(t;t_x+2\varepsilon_x, t_y-2\varepsilon_y)) \quad \Rightarrow \quad \begin{cases} \Phi_{xy}(t) = 0, & t \leqslant t_x + 2\varepsilon_x \\ \Phi_{xy}(t) = 1, & t \geqslant t_y - 2\varepsilon_y. \end{cases}$$

$$\Phi_y(t) = k(h_+(t;t_y-2\varepsilon_y, t_y-\varepsilon_y)) \quad \Rightarrow \quad \begin{cases} \Phi_y(t) = 0, & t \leqslant t_y - 2\varepsilon_y \\ \Phi_y(t) = 1, & t \geqslant t_y - \varepsilon_y. \end{cases}$$

And as the last auxiliary functions define the polynoms

$$p_x(t) = \sum_{i=1}^{\nu} \frac{1}{i!} x^i (t-t_x)^i \quad \text{and} \quad p_y(t) = \sum_{i=1}^{\nu} \frac{1}{i!} y^i (t-t_y)^i.$$

It is easy to see that $p_x^{(i)}(t_0) = x^i$ and $p_y^{(i)}(t_0) = y^i$ for all $1 \leqslant i \leqslant \nu$. Now define the curve $\gamma : [t_x, t_y] \to \mathbb{R}^n$

$$\begin{aligned} \gamma(t) &= x + \Phi_x(t)p_x(t) + \Phi_{xy}(t)(y-x) + \Phi_y(t)p_y(t) \\ &= \begin{cases} x + \Phi_x(t)p_x(t), & t_x \leqslant t < t_x + 2\varepsilon_x \\ \Phi_{xy}(t)y + (1-\Phi_{xy}(t))x, & t_x + 2\varepsilon_x \leqslant t < t_y - 2\varepsilon_y \\ y + \Phi_y(t)p_y(t), & t_y - 2\varepsilon_y \leqslant t \leqslant t_y \end{cases} \end{aligned}$$

There are $r_x, r_y > 0$ such that $B_{r_x}(x) \subset \mathcal{G}$ and $B_{r_y}(y) \subset \mathcal{G}$, since $\mathcal{G}$ is open. Let be

$$P_x := \max_{t\in\mathcal{I}} ||\sum_{i=1}^{\nu} \frac{1}{i!} x^i (t-t_x)^{i-1}|| \quad \text{and} \quad P_Y := \max_{t\in\mathcal{I}} ||\sum_{i=1}^{\nu} \frac{1}{i!} y^i (t-t_y)^{i-1}||$$

and choose

$$\varepsilon_x = \frac{r_x}{4\max\{P_x, 1\}} \text{ and } \varepsilon_y = \frac{r_y}{4\max\{P_y, 1\}}.$$

Now we can show that $\gamma(t) \in \mathcal{G}$ for all $t \in \mathcal{G}$. Therefore let $t \in [t_x, t_x + 2\varepsilon_x]$ hence $\gamma(t) = x + \Phi_x(t)p_x(t)$ and

$$\begin{aligned} ||\gamma(t) - x|| &= ||p_x(t)|| \cdot ||\Phi_x(t)|| \leqslant ||p_x(t)|| \\ &= ||\sum_{i=1}^{\nu} \frac{1}{i!} x^i (t-t_x)^{i-1}|| \cdot |t - t_x| \\ &\leqslant \max_{t\in\mathcal{I}} ||\sum_{i=1}^{\nu} \frac{1}{i!} x^i (t-t_x)^{i-1}|| \cdot 2\varepsilon_x < \frac{1}{2} r_x \end{aligned}$$

Let $t \in [t_x + 2\varepsilon_y, t_y - 2\varepsilon_y]$ hence $\gamma(t) = \Phi_{xy}(t)y + (1 - \Phi_{xy}(t))x$ and since $\mathcal{G}$ is convex and $\Phi_{xy}(t) \in [0, 1]$ for all $t \in \mathcal{I}$ it follows that $\gamma(t) \in \mathcal{G}$.
Let $t \in [t_y - 2\varepsilon_y, t_y]$ hence $\gamma(t) = y + \Phi_y(t)p_y(t)$ and

$$\begin{aligned} ||\gamma(t) - y|| &= ||p_y(t)|| \cdot ||\Phi_y(t)|| \leqslant ||p_y(t)|| \\ &= ||\sum_{i=1}^{\nu} \frac{1}{i!} y^i (t-t_y)^{i-1}|| \cdot |t - t_y| \\ &\leqslant \max_{t\in\mathcal{I}} ||\sum_{i=1}^{\nu} \frac{1}{i!} y^i (t-t_y)^{i-1}|| \cdot 2\varepsilon_y < \frac{1}{2} r_y \end{aligned}$$

At last we need that γ is sufficiently smooth but it even holds $\gamma \in C^\infty(\mathcal{I})$ since p_x and p_y are polynoms and $\Phi_x, \Phi_y, \Phi_{xy} \in C^\infty(\mathcal{I})$. □

Thereby it follows:

Corollary 4.21. (Reference function)
Let $\mathcal{G} \subset \mathbb{R}^n$ be open and connected. Let $(X^\nu, t_x), (Y^\nu, t_y) \in \mathbb{R}^{\nu \cdot n}$ with $(x, t_x), (y, t_y) \in \mathcal{G} \times \mathcal{I}$ then there is a reference function $\gamma(t)$ such that $\gamma(t_x) = x$ and $\gamma(t_y) = y$ and for all $i \leqslant \nu$ it holds that $\gamma^{(i)}(t_x) = x^i$ and $\gamma^{(i)}(t_y) = y^i$.

With the help of this lemma and the corollary we prove:

Theorem 4.22. (Linearization)
Let the DAE (4.9) satisfy the basic Assumptions (2.25) and let $\mathcal{G} \subset \mathcal{D}_f \times \mathcal{I}_f$ be open and connected. Let f and d be sufficiently smooth on $\mathcal{G}$. Then the two following statements hold:

1. Let the DAE (4.9) be regular with Dissection Index μ and with characteristic values $r_0, \ldots, r_\mu$. Then all linearizations (2.16) along reference functions $x_* \in C_*^\mu(\mathcal{G})$ are regular linear DAEs with uniform index μ and uniform characteristic values $r_0, \ldots, r_\mu$.
2. Let all linearizations (2.16) along reference functions $x_* \in C_*^\mu(\mathcal{G})$ be regular linear DAEs. Then they have a uniform Dissection Index μ and uniform characteristic values $r_0, \ldots, r_\mu$ and the nonlinear DAE (4.9) is regular on $\mathcal{G}$ with these characteristics and index μ.

Proof.
We proof Theorem 4.22 with the help of Corollary 4.21. Let $x_\star \in C_*^\mu(\mathcal{G})$ be an arbitrary reference function. Then there is the associated linear DAE

$$A_*(t)(D_*(t)x(t))' + B_*(t)x(t) = q_*(t), \quad t \in \mathcal{I}_*, \tag{4.14}$$

with the coefficients

$$\begin{aligned}
D_*(t) &:= d_x(x_*(t), t),\\
A_*(t) &:= f_y(d'(x_*(t), t), x_*(t), t),\\
B_*(t) &:= f_x(d'(x_*(t), t), x_*(t), t),\\
q_*(t) &:= -f(d'(x_*(t), t), x_*(t), t), \quad t \in \mathcal{I}_*.
\end{aligned}$$

Therefore the placeholder matrices are given by

$$\begin{aligned}
D_*(t) &:= d_x(x_*(t), t)\\
A_*(t) &:= f_y(D(x_*(t), t)x_*'(t) + d_t(x_*(t), t), x_*(t), t),\\
B_*(t) &:= f_x(D(x_*(t), t)x_*'(t) + d_t(x_*(t), t), x_*(t), t).
\end{aligned}$$

It holds that

$$\begin{aligned}
D_*(t) &= D(x_*(t), t)\\
A_*(t) &= A(x_*'(t), x_*(t), t),\\
B_*(t) &= B(x_*'(t), x_*(t), t)
\end{aligned}$$

with A, D and B being the placeholder matrices of the nonlinear problem. Therefore the first matrix of the chain of the linear DAE

$$G_{0,*}(t) := A_*(t)D_*(t) = A(x_*'(t), x_*(t), t)D(x_*(t), t) = G_0(x_*'(t), x_*(t), t)$$

is just the G_0 matrix of the nonlinear problem along the reference function. Hence the first sequence of basis functions $P_*(t)$, $Q_*(t)$, $V_*(t)$ and $W_*(t)$ associated to $G_{0,*}(t)$ are equal to the first sequence of basis functions of the nonlinear DAE along the reference function. With an induction we get for all $1 \leqslant i \leqslant \mu$

$$G_{i+1,*}(t), B^{\mathrm{v}}_{x_{i+1,*}}(t), B^{\mathrm{w}}_{x_{i+1,*}}(t), B^{\mathrm{v}}_{y_{i+1,*}}(t) \text{ and } B^{\mathrm{w}}_{y_{i+1,*}}(t)$$

as well as the associated basis functions of the linear DAE are equal to the matrix chain of the nonlinear DAE along the reference function. For $i = 1$ we get

$$\begin{aligned}
G_{1,*}(t) :=& V_*^\top(t) G_{0,*}(t) P_*(t) \\
=& V^\top(x'_*(t), x_*(t), t) G_0(x'_*(t), x_*(t), t) P(x_*(t), t) \\
=& G_1(x'_*(t), x_*(t), t), \\
B^{\mathrm{v}}_{x_1,*}(t) :=& V_*^\top(t) A_*(t) D_*(t) P'_*(t) + V_*^\top(t) B_*(t) P_*(t) \\
=& V^\top(x'_*(t), x_*(t), t) A(x'_*(t), x_*(t), t) D(x_*(t), t) P'(x'_*(t), x_*(t), t) \\
&+ V^\top(x'_*(t), x_*(t), t) B(x'_*(t), x_*(t), t) P(x_*(t), t) \\
=& B^{\mathrm{v}}_{x_1}(x'_*(t), x_*(t), t)
\end{aligned}$$

and analogous we get $B^{\mathrm{v}}_{y_1,*}(t) = B^{\mathrm{v}}_{y_1}(x'_*(t), x_*(t), t), B^{\mathrm{w}}_{x_1,*}(t) = B^{\mathrm{w}}_{x_1}(x'_*(t), x_*(t), t)$ and $B^{\mathrm{w}}_{y_1,*}(t) = B^{\mathrm{w}}_{y_1}(x'_*(t), x_*(t), t)$. Hence the second sequence of basis function of the linear DAE is the second sequence of basis function of the nonlinear DAE along the reference function since these basis functions only depend on $B^{\mathrm{w}}_{y_1,*}(t)$, $B^{\mathrm{w}}_{x_1,*}(t)$ and $G_{1,*}(t)$. If the statement now holds for $i < \mu$ then we obtain

$$\begin{aligned}
G_{i+1,*}(t) :=& V^\top_{x_i,*}(t) G_{i,*}(t) Q_{x_i,*}(t) \\
=& V^\top_{x_i}(X^i_*(t), t) G_i(X^{i-1}_*(t), t) Q_{x_i}(X^i_*(t), t) \\
=& G_{i+1}(X^i_*(t), t), \\
B^{\mathrm{v}}_{x_{i+1},*}(t) :=& V^\top_{x_i,*}(t) G_{i,*}(t) Q'_{x_i,*}(t) + V^\top_{x_i,*}(t) B^{\mathrm{v}}_{x_i,*}(t) Q_{x_i,*}(t) \\
&- V^\top_{x_i,*}(t) B^{\mathrm{v}}_{y_i,*}(t) P_{y_i,*}(t) (V^\top_{y_i,*}(t) B^{\mathrm{w}}_{y_i,*}(t) P_{y_i,*}(t))^{-1} V^\top_{y_i,*}(t) B^{\mathrm{w}}_{x_i,*}(t) Q_{x_i,*}(t) \\
=& (V^\top_{x_i} G_i Q'_{x_i})(X^{i+1}_*(t), t) + (V^\top_{x_i} B^{\mathrm{v}}_{x_i} Q_{x_i})(X^i_*(t), t) \\
&- (V^\top_{x_i} B^{\mathrm{v}}_{y_i} P_{y_i})(X^i_*(t), t) (V^\top_{y_i} B^{\mathrm{w}}_{y_i} P_{y_i})^{-1}(X^{i-1}_*(t), t) (V^\top_{y_i} B^{\mathrm{w}}_{x_i} Q_{x_i})(X^i_*(t), t) \\
=& B^{\mathrm{v}}_{x_{i+1}}(X^{i+1}_*(t), t)
\end{aligned}$$

with $X^i_*(t) := (x^{(i)}_*(t), ..., x'_*(t), x_*(t))$. Again we analogously get

$$\begin{aligned}
B^{\mathrm{v}}_{y_{i+1},*}(t) &= B^{\mathrm{v}}_{y_{i+1}}(X^i_*(t), t), \\
B^{\mathrm{w}}_{x_{i+1},*}(t) &= B^{\mathrm{w}}_{x_{i+1}}(X^{i+1}_*(t), t), \\
B^{\mathrm{w}}_{y_{i+1},*}(t) &= B^{\mathrm{w}}_{y_{i+1}}(X^i_*(t), t).
\end{aligned}$$

Hence the associated basis functions are again equal to those of the nonlinear DAE along the reference function since they only depend on $B^{\mathrm{w}}_{y_{i+1},*}(t)$, $B^{\mathrm{w}}_{x_{i+1},*}(t)$ and $G_{i+1,*}(t)$. Thus the chain of the linear DAE is the chain of the nonlinear DAE along the reference function and therefore the index and the characteristic values of the linear DAE have to be equal to those of the nonlinear DAE. This concludes the proof of (i).

We split the proof of (ii) into two parts. First we prove that all linearizations have a uniform index and characteristic values. Therefore let $\gamma_1(t)$ and $\gamma_2(t)$ be two arbitrary reference functions of the nonlinear DAE. Lemma 4.20 allows us to choose a third reference function $\gamma(t)$ such that for all $i \leqslant \nu$ it holds

$$\gamma^{(i)}(t_0) = \gamma_1^{(i)}(t_0) \text{ and } \gamma^{(i)}(T) = \gamma_2^{(i)}(T).$$

Hence the linearizations along $\gamma_1(t)$ and $\gamma(t)$ must have the same index and characteristics at t_0, since their matrix chains coincide at this point. Therefore these linearizations have the same index and characteristic values at every point, since these values are assumed to be constant. For the same reason the linearizations along $\gamma_2(t)$ and $\gamma(t)$ have the same index and characteristic values, which yields that the linearizations along $\gamma_1(t)$ and $\gamma_2(t)$ have the same index and characteristic values.
Next, we show that the uniform index and characteristic values of the linearizations are also the index and characteristic values of the nonlinear DAE. Construct the chain of the non-linear DAE to the level of the index of linearizations. Assume there is a point (X^μ, t) at which the index or the characteristic values of the nonlinear DAE do not equal the index or the characteristics of linearizations. Due to Lemma 4.20 we can construct a reference function $\gamma_3(t)$ with

$$\gamma_3^{(i)}(t) = x^i.$$

With the same argumentation as in (i) we get that the matrix chain of the linearization and the matrix chain of the nonlinear DAE equal at that point and therefore their characteristic values and their index equal at this point as well. Hence this assumption can never hold and the theorem is proven. □

Before we relate the Dissection Index to the other index concepts we remember the Examples 4.1-4.3. In contrast to the Tractability Index and the Strangeness Index the Dissection Index only needs a constant basis function to analyze these examples. Starting with Example 4.1 we write again

$$AD = \begin{pmatrix} 1+t^2 & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B_0 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}.$$

Therefore we can choose

$$P = V = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and gain

$$G_1 = \begin{pmatrix}1+t^2\end{pmatrix}, \quad B^{\mathrm{v}}_{y_1} = \begin{pmatrix}-1\end{pmatrix}, \quad B^{\mathrm{w}}_{x_1} = \begin{pmatrix}1\end{pmatrix} \text{ and } B^{\mathrm{w}}_{y_1} = \begin{pmatrix}0\end{pmatrix}.$$

Due to the matrix $B^{\mathrm{w}}_{y_1}$ we can choose

$$P_{y_1} = V_{y_1} = \begin{pmatrix}\ \end{pmatrix} \in \mathbb{R}^{1\times 0} \quad \text{and} \quad Q_{y_1} = W_{y_1} = \begin{pmatrix}1\end{pmatrix}.$$

This yields

$$W^\top_{y_1} B^{\mathrm{w}}_{x_1} = \begin{pmatrix}1\end{pmatrix} \quad \text{and} \quad Q_{x_1} = \begin{pmatrix}\ \end{pmatrix} \in \mathbb{R}^{1\times 0}.$$

After calculating $G_1 Q_{x_1} = \begin{pmatrix}\ \end{pmatrix} \in \mathbb{R}^{1\times 0}$ we can choose $W_{x_1} = \begin{pmatrix}1\end{pmatrix}$ which leads to $B^{\mathrm{w}}_{y_2} = \begin{pmatrix}1\end{pmatrix}$ hence the DAE is Dissection Index 2.
In the case of Example 4.2 we denote again

$$AD = \begin{pmatrix}1&0&0\\0&1&0\\0&0&0\end{pmatrix} \quad \text{and} \quad B_0 = \begin{pmatrix}0&0&1\\0&0&a'(e_1)\\1&1&0\end{pmatrix}.$$

Therefore we can choose

$$P = V = \begin{pmatrix}1&0\\0&1\\0&0\end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix}0\\0\\1\end{pmatrix}$$

and gain

$$G_1 = \begin{pmatrix}1&0\\0&1\end{pmatrix}, \quad B^{\mathrm{v}}_{y_1} = \begin{pmatrix}1\\a'(e_1)\end{pmatrix}, \quad B^{\mathrm{w}}_{x_1} = \begin{pmatrix}1&1\end{pmatrix} \text{ and } B^{\mathrm{w}}_{y_1} = \begin{pmatrix}0\end{pmatrix}.$$

Due to the matrix $B^{\mathrm{w}}_{y_1}$ we can choose

$$P_{y_1} = V_{y_1} = \begin{pmatrix}\ \end{pmatrix} \in \mathbb{R}^{1\times 0} \quad \text{and} \quad Q_{y_1} = W_{y_1} = \begin{pmatrix}1\end{pmatrix}.$$

This yields

$$W^\top_{y_1} B^{\mathrm{w}}_{x_1} = \begin{pmatrix}1&1\end{pmatrix} \quad \text{and} \quad Q_{x_1} = \begin{pmatrix}\frac{1}{2}\\-\frac{1}{2}\end{pmatrix}, P_{x_1} = \begin{pmatrix}\frac{1}{2}\\\frac{1}{2}\end{pmatrix}.$$

After calculating $G_1 Q_{x_1} = \begin{pmatrix}\frac{1}{2} & -\frac{1}{2}\end{pmatrix}^\top$ we can choose $W_{x_1} = \begin{pmatrix}1&1\end{pmatrix}^\top$ and $V_{x_1} = \begin{pmatrix}1&-1\end{pmatrix}^\top$ which leads to $B^{\mathrm{w}}_{y_2} = \begin{pmatrix}1+a'(e_1)\end{pmatrix}$ hence the DAE is also Dissection Index 2. When we introduced Example 4.2, we decoupled its equations without putting thoughts into a general decoupling procedure or an index concept. We notice that the Dissection Index allows us to choose the same transformation operators we used in this canonical decoupling.

Last we calculate the basis function of the matrix chain of Example 4.3 and again start by denoting again

$$AD = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & C & -C & 0 & 0 \\ 0 & -C & C & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad B_0 = \begin{pmatrix} 1 & -1 & 0 & 0 & -1 \\ -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Therefore we can choose

$$P = V = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ -1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$$

and gain

$$G_1 = \begin{pmatrix} 4C & 0 \\ 0 & 1 \end{pmatrix}, \quad B^{\mathrm{v}}_{y_1} = \begin{pmatrix} -1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{x_1} = \begin{pmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{y_1} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}.$$

Due to the matrix $B^{\mathrm{w}}_{y_1}$ we can choose

$$Q_{y_1} = W_{y_1} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}.$$

This yields

$$W_{y_1}^\top B^{\mathrm{w}}_{x_1} = \begin{pmatrix} 0 & -1 \end{pmatrix} \quad \text{and} \quad Q_{x_1} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

After calculating $G_1 Q_{x_1} = \begin{pmatrix} 4C & 0 \end{pmatrix}^\top$ we can choose $W_{x_1} = \begin{pmatrix} 0 & 1 \end{pmatrix}^\top$ which leads to $B^{\mathrm{w}}_{y_2} = \begin{pmatrix} 1 \end{pmatrix}$ hence the DAE is also Dissection Index 2.
The mixed index concept reflects the simple structure of the three examples by using only constant basis functions as desired.

4.2 Relations between Index Concepts

In the following we want to describe the relations of the Dissection Index to the other index concepts. We start to describe these relations on linear time dependent DAEs.

Definition 4.23. (Linear time dependent DAE)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{D}_x, \mathcal{D}_{x'} \subset \mathbb{R}^n$ be open subsets. Consider the following equation

$$A(t)(D(t)x)' + B(t)x = q(t) \tag{4.15}$$

with $D \in C^1(\mathcal{I}, \mathbb{R}^{m\times n})$, $A \in C(\mathcal{I}, \mathbb{R}^{n\times m})$ and $B \in C(\mathcal{I}, \mathbb{R}^{n\times n})$. We call (4.15) a linear time dependent DAE.

Afterwards we use Theorem 4.22 to lift the results up to nonlinear DAEs. In [LMT13] the relation between the Tractability Index and the Strangeness Index is described for linear time dependent DAEs. We follow their strategy to show that the characteristic values and the index belonging to the Tractability Index concept coincide with those of the Dissection Index concept. Therefore define the S-canonical form:

Definition 4.24. (S-canonical form)
The structured DAE with continuous coefficients

$$\begin{pmatrix} I_{m-l} & K \\ 0 & N \end{pmatrix} x' + \begin{pmatrix} W & 0 \\ 0 & I_l \end{pmatrix} x = q(t)$$

with $0 \leqslant l \leqslant m$, is said to be in S-canonical form, if $K = \begin{pmatrix} 0 & K_1 & \dots & K_\kappa \end{pmatrix}$ and

$$N = \begin{pmatrix} 0 & N_{1,2} & \dots & N_{1,\kappa} \\ & \ddots & & \vdots \\ & & \ddots & N_{\kappa-1,\kappa} \\ & & & 0 \end{pmatrix}$$

is a strictly upper bloc triangular with full row rank entries $N_{i,i+1}$ with $i = 1, \dots, \kappa - 1$. Furthermore denote the number of rows in the i-th block row by l_i.

Now we can prove the following theorem.

Theorem 4.25.
We consider a linear time dependent DAE (4.15). Let the DAE have a finite Dissection Index μ and a finite Tractability Index μ_T. Then the values of the index and the characteristic values of both concepts coincide, i.e. $\mu = \mu_T$ and $r_i = r_i^T$ for all $i = 1, \dots, \mu$.

Proof.
We assume $A(t)D(t)$ to be singular, otherwise there is nothing to show. Hence, the Dissection Index is at least 1. Analogous to the constant case we transform (4.15) into

$$G_1 x_0' + B_{x_1}^{\mathrm{v}} x_0 + B_{y_1}^{\mathrm{v}} y_0 = q_1^{\mathrm{v}} \tag{4.16a}$$

$$B_{x_1}^{\mathrm{w}} x_0 + B_{y_1}^{\mathrm{w}} y_0 = q_1^{\mathrm{w}} \tag{4.16b}$$

with $q_1^{\mathrm{v}}(t) = V^\top(t)q(t)$ and $q_1^{\mathrm{w}}(t) = W^\top(t)q(t)$.
By Theorem 2.79. on page 162 in [LMT13] it is sufficient to show that the DAE can be transformed into a DAE in S-canonical form with $l_{\mu-i} = n - r_i^\mu$. First, we show by an induction that the DAE (4.15) is equivalent to

$$\begin{aligned} G_k x'_{k-1} + M_k^{\mathrm{v}} \tilde{X}'_{k-1} + B_{x_k}^{\mathrm{v}} x_{k-1} + B_{y_k}^{\mathrm{v}} y_{k-1} &= q_k^{\mathrm{v}} \\ W_{x_{k-1}}^\top G_{k-1} P_{x_{k-1}} \tilde{x}'_{k-1} + M_k^{\mathrm{w}} \tilde{X}'_{k-2} + B_{x_k}^{\mathrm{w}} x_{k-1} + B_{y_k}^{\mathrm{w}} y_{k-1} &= q_k^{\mathrm{w}} \end{aligned}$$

and

$$\begin{aligned} \tilde{x}_i &= q_{\tilde{x}_i} - (W_{y_i}^\top B_{x_i}^{\mathrm{w}} P_{x_i})^{-1} W_{y_i}^\top W_{x_{i-1}}^\top G_{i-1} P_{x_{i-1}} \tilde{x}'_{i-1} + M_{x_i} \tilde{X}'_{i-2} \\ \tilde{y}_i &= q_{\tilde{y}_i} - (V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^\top W_{x_{i-1}}^\top G_{i-1} P_{x_{i-1}} \tilde{x}'_{i-1} + M_{y_i} \tilde{X}'_{i-2} \\ &\quad - (V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^\top B_{x_i}^{\mathrm{w}} (Q_{x_i} x_i + P_{x_i} \tilde{x}_i) \end{aligned}$$

with suitable matrices M_{x_i}, M_{y_i} and M_k^{v}, functions $q_{\tilde{x}_i}$, $q_{\tilde{y}_i}$, q_k^{v} and $\tilde{X}'_i := (\tilde{x}'_i, \ldots, \tilde{x}'_0)^\top$, $\tilde{X}'_{-1} = (\ )$, $\tilde{x}'_0 = (\ )$, $G_0 = (\ )$ and $i = 1, \ldots, k-1$ for all $2 \leqslant k \leqslant \mu$. Here the variables are recursively defined as in Equation (4.10).

Induction start: $k = 2$
In order to obtain the descriptions for $\tilde{x}_1$ and $\tilde{y}_1$ we insert the transformation

$$x_0 = P_{x_1}\tilde{x}_1 + Q_{x_1}x_1, \quad y_0 = P_{y_1}\tilde{y}_1 + Q_{y_1}y_1. \tag{4.17}$$

into Equation (4.16b) and we multiply the same equation by $V_{y_1}^\top$ and $W_{y_1}^\top$ from the left.

$$\begin{aligned} V_{y_1}^\top B_{x_1}^{\mathrm{w}} (P_{x_1}\tilde{x}_1 + Q_{x_1}x_1) + V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1} \tilde{y}_1 &= V_{y_1}^\top q_1^{\mathrm{w}} \\ W_{y_1}^\top B_{x_1}^{\mathrm{w}} P_{x_1} \tilde{x}_1 &= W_{y_1}^\top q_1^{\mathrm{w}} \end{aligned}$$

We obtain

$$\begin{aligned} \tilde{x}_1 &= q_{\tilde{x}_1} \\ \tilde{y}_1 &= q_{\tilde{y}_1} - (V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1})^{-1} V_{y_1}^\top B_{x_1}^{\mathrm{w}} (Q_{x_1}x_1 + P_{x_1}\tilde{x}_1) \end{aligned}$$

with

$$\begin{aligned} q_{\tilde{x}_1} &= (W_{y_1}^\top B_{x_1}^{\mathrm{w}} P_{x_1})^{-1} W_{y_1}^\top q_1^{\mathrm{w}} \\ q_{\tilde{y}_1} &= (V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1})^{-1} V_{y_1}^\top q_1^{\mathrm{w}} \end{aligned}$$

and $M_{x_1} = 0$ and $M_{y_1} = 0$. We conclude the induction start by transforming and factorizing Equation (4.16a). First we insert the splitting of x_0 and y_0 and make use of the descriptions for $\tilde{x}_1$ and $\tilde{y}_1$:

$$G_1 x'_0 + B_{x_1}^{\mathrm{v}} x_0 + B_{y_1}^{\mathrm{v}} y_0 = q_1^{\mathrm{v}}$$

$$\begin{aligned}
&\Leftrightarrow G_1(P_{x_1}\tilde{x}_1 + Q_{x_1}x_1)' + B^{\mathrm{v}}_{x_1}(P_{x_1}\tilde{x}_1 + Q_{x_1}x_1) + B^{\mathrm{v}}_{y_1}(P_{y_1}\tilde{y}_1 + Q_{y_1}y_1) = q^{\mathrm{v}}_1 \\
&\Leftrightarrow G_1Q_{x_1}x_1' + G_1P_{x_1}\tilde{x}_1' + (B^{\mathrm{v}}_{x_1}Q_{x_1} + G_1Q'_{x_1})x_1 + B^{\mathrm{v}}_{y_1}P_{y_1}\tilde{y}_1 + B^{\mathrm{v}}_{y_1}Q_{y_1}y_1 \\
&\qquad = q^{\mathrm{v}}_1 - G_1P'_{x_1}\tilde{x}_1 - B^{\mathrm{v}}_{x_1}P_{x_1}\tilde{x}_1 \\
&\Leftrightarrow G_1Q_{x_1}x_1' + G_1P_{x_1}\tilde{x}_1' + (B^{\mathrm{v}}_{x_1}Q_{x_1} + G_1Q'_{x_1} - B^{\mathrm{v}}_{y_1}P_{y_1}(V^{\top}_{y_1}B^{\mathrm{w}}_{y_1}P_{y_1})^{-1}V^{\top}_{y_1}B^{\mathrm{w}}_{x_1}Q_{x_1})x_1 \\
&\qquad + B^{\mathrm{v}}_{y_1}Q_{y_1}y_1 = q_2 \\
&\Leftrightarrow G_1Q_{x_1}x_1' + G_1P_{x_1}\tilde{x}_1' + B_{x_1}x_1 + B_{y_1}y_1 = q_2
\end{aligned}$$

with $q_2 := q^{\mathrm{v}}_1 - G_1P'_{x_1}q_{\tilde{x}_1} - B^{\mathrm{v}}_{x_1}P_{x_1}q_{\tilde{x}_1} - B^{\mathrm{v}}_{y_1}P_{y_1}(q_{\tilde{y}_1} - (V^{\top}_{y_1}B^{\mathrm{w}}_{y_1}P_{y_1})^{-1}V^{\top}_{y_1}B^{\mathrm{w}}_{x_1}P_{x_1}q_{\tilde{x}_1})$. Then we factorize the equation by $V^{\top}_{x_1}$ and $W^{\top}_{x_1}$ and obtain:

$$\begin{aligned}
&G_1Q_{x_1}x_1' + G_1P_{x_1}\tilde{x}_1' + B_{x_1}x_1 + B_{y_1}y_1 = q_2 \\
\Leftrightarrow &\begin{cases} V^{\top}_{x_1}G_1Q_{x_1}x_1' + V^{\top}_{x_1}G_1P_{x_1}\tilde{x}_1' + V^{\top}_{x_1}B_{x_1}x_1 + V^{\top}_{x_1}B_{y_1}y_1 = V^{\top}_{x_1}q_2 \\ W^{\top}_{x_1}G_1P_{x_1}\tilde{x}_1' + W^{\top}_{x_1}B_{x_1}x_1 + W^{\top}_{x_1}B_{y_1}y_1 = W^{\top}_{x_1}q_2 \end{cases} \\
\Leftrightarrow &\begin{cases} G_2x_1' + M^{\mathrm{v}}_2\tilde{x}_1' + B^{\mathrm{v}}_{x_2}x_1 + B^{\mathrm{v}}_{y_2}y_1 = q^{\mathrm{v}}_2 \\ W^{\top}_{x_1}G_1P_{x_1}\tilde{x}_1' + B^{\mathrm{w}}_{x_2}x_1 + B^{\mathrm{w}}_{y_2}y_1 = q^{\mathrm{w}}_2 \end{cases}
\end{aligned}$$

with $q^{\mathrm{v}}_2 = V^{\top}_{x_1}q_2$, $q^{\mathrm{w}}_2 = W^{\top}_{x_1}q_2$, $M^{\mathrm{v}}_2 := V^{\top}_{x_1}G_1P_{x_1}$ and $M^{\mathrm{w}}_2 := 0$.

Induction step: $k \to k+1$
The induction step proceeds analogously to the induction start. By the induction hypothesis we begin with the equations

$$G_kx'_{k-1} + M^{\mathrm{v}}_k\tilde{X}'_{k-1} + B^{\mathrm{v}}_{x_k}x_{k-1} + B^{\mathrm{v}}_{y_k}y_{k-1} = q^{\mathrm{v}}_k \tag{4.18a}$$

$$W^{\top}_{x_{k-1}}G_{k-1}P_{x_{k-1}}\tilde{x}'_{k-1} + M^{\mathrm{w}}_k\tilde{X}'_{k-2} + B^{\mathrm{w}}_{x_k}x_{k-1} + B^{\mathrm{w}}_{y_k}y_{k-1} = q^{\mathrm{w}}_k. \tag{4.18b}$$

Again we insert the transformation

$$x_{k-1} = P_{x_k}\tilde{x}_k + Q_{x_k}x_k, \quad y_{k-1} = P_{y_k}\tilde{y}_k + Q_{y_k}y_k \tag{4.19}$$

into Equation (4.18b) and factorize the same equation by $V^{\top}_{y_k}$ and $W^{\top}_{y_k}$.

$$\begin{aligned}
V^{\top}_{y_k}(W^{\top}_{x_{k-1}}G_{k-1}P_{x_{k-1}}\tilde{x}'_{k-1} + M^{\mathrm{w}}_k\tilde{X}'_{k-2} + B^{\mathrm{w}}_{x_k}(P_{x_k}\tilde{x}_k + Q_{x_k}x_k) + B^{\mathrm{w}}_{y_k}P_{y_k}\tilde{y}_k) &= V^{\top}_{y_k}q^{\mathrm{w}}_k \\
W^{\top}_{y_k}W^{\top}_{x_{k-1}}G_{k-1}P_{x_{k-1}}\tilde{x}'_{k-1} + W^{\top}_{y_k}M^{\mathrm{w}}_k\tilde{X}'_{k-2} + W^{\top}_{y_k}B^{\mathrm{w}}_{x_k}P_{x_k}\tilde{x}_k &= W^{\top}_{y_k}q^{\mathrm{w}}_k
\end{aligned}$$

By the Lemmata 4.14 and 4.16 we obtain

$$\tilde{x}_k = q_{\tilde{x}_k} - (W^{\top}_{y_k}B^{\mathrm{w}}_{x_k}P_{x_k})^{-1}W^{\top}_{y_k}W^{\top}_{x_{k-1}}G_{k-1}P_{x_{k-1}}\tilde{x}'_{k-1} + M_{x_k}\tilde{X}'_{k-2} \tag{4.20a}$$

$$\tilde{y}_k = q_{\tilde{y}_k} - (V^{\top}_{y_k}B^{\mathrm{w}}_{y_k}P_{y_k})^{-1}V^{\top}_{y_k}W^{\top}_{x_{k-1}}G_{k-1}P_{x_{k-1}}\tilde{x}'_{k-1} + M_{y_k}\tilde{X}'_{k-2} \tag{4.20b}$$

$$- (V^{\top}_{y_k}B^{\mathrm{w}}_{y_k}P_{y_k})^{-1}V^{\top}_{y_k}B^{\mathrm{w}}_{x_k}(Q_{x_k}x_k + P_{x_k}\tilde{x}_k) \tag{4.20c}$$

with

$$\begin{aligned} q_{\tilde{x}_k}(t) &= (W_{y_k}^\top B_{x_k}^{\mathrm{w}} P_{x_k})^{-1} W_{y_k}^\top q_k^{\mathrm{w}} \\ q_{\tilde{y}_k}(t) &= (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top q_k^{\mathrm{w}} \end{aligned}$$

and $M_{x_k} = (W_{y_k}^\top B_{x_k}^{\mathrm{w}} P_{x_k})^{-1} W_{y_k}^\top M_k^{\mathrm{w}}$ and $M_{y_k} = (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top M_k^{\mathrm{w}}$. We conclude the induction step by transforming and factorizing Equation (4.18a). Therefore we rewrite the descriptions of $\tilde{x}_k$ and $\tilde{y}_k$ in a more compact form.

$$\begin{aligned} \tilde{x}_k &= q_{\tilde{x}_k} - (W_{y_k}^\top B_{x_k}^{\mathrm{w}} P_{x_k})^{-1} W_{y_k}^\top W_{x_{k-1}}^\top G_{k-1} P_{x_{k-1}} \tilde{x}'_{k-1} + M_{x_k} \tilde{X}'_{k-2} \\ &= q_{\tilde{x}_k} + \bar{M}_{x_k} \tilde{X}'_{k-1} \\ \tilde{y}_k &= q_{\tilde{y}_k} - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top W_{x_{k-1}}^\top G_{k-1} P_{x_{k-1}} \tilde{x}'_{k-1} + M_{y_k} \tilde{X}'_{k-2} \\ &\quad - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} (Q_{x_k} x_k + P_{x_k} \tilde{x}_k) \\ &= q_{\tilde{y}_k} + \bar{M}_{y_k} \tilde{X}'_{k-1} - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} (Q_{x_k} x_k + P_{x_k} (q_{\tilde{x}_k} + \bar{M}_{x_k} \tilde{X}'_{k-1})) \\ &= \bar{q}_{\tilde{y}_k} + \bar{M}_{xy_k} \tilde{X}'_{k-1} - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} Q_{x_k} x_k \end{aligned}$$

with

$$\begin{aligned} \bar{q}_{\tilde{y}_k} &= q_{\tilde{y}_k} - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} P_{x_k} q_{\tilde{x}_k}, \\ \bar{M}_{x_k} &= \begin{pmatrix} -(W_{y_k}^\top B_{x_k}^{\mathrm{w}} P_{x_k})^{-1} W_{y_k}^\top W_{x_{k-1}}^\top G_{k-1} P_{x_{k-1}} & M_{x_k} \end{pmatrix}, \\ \bar{M}_{y_k} &= \begin{pmatrix} -(V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top W_{x_{k-1}}^\top G_{k-1} P_{x_{k-1}} & M_{y_k} \end{pmatrix}, \\ \bar{M}_{xy_k} &= \bar{M}_{y_k} - (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} \bar{M}_{x_k}. \end{aligned}$$

Now we insert the splitting of x_{k-1} and y_{k-1} and make use of the compact descriptions of $\tilde{x}_k$ and $\tilde{y}_k$:

$$\begin{aligned} & G_k x'_{k-1} + M_k^{\mathrm{v}} \tilde{X}'_{k-1} + B_{x_k}^{\mathrm{v}} x_{k-1} + B_{y_k}^{\mathrm{v}} y_{k-1} = q_k^{\mathrm{v}} \\ \Leftrightarrow\ & G_k (P_{x_k} \tilde{x}_k + Q_{x_k} x_k)' + M_k^{\mathrm{v}} \tilde{X}'_{k-1} + B_{x_k}^{\mathrm{v}} (P_{x_k} \tilde{x}_k + Q_{x_k} x_k) + B_{y_k}^{\mathrm{v}} (P_{y_k} \tilde{y}_k + Q_{y_k} y_k) = q_k^{\mathrm{v}} \\ \Leftrightarrow\ & G_k Q_{x_k} x'_k + G_k P_{x_k} \tilde{x}'_k + \bar{M}_k^{\mathrm{v}} \tilde{X}'_{k-1} \\ & \quad + (G_k Q'_{x_k} + B_{x_k}^{\mathrm{v}} Q_{x_k} - B_{y_k}^{\mathrm{v}} Q_{y_k} (V_{y_k}^\top B_{y_k}^{\mathrm{w}} P_{y_k})^{-1} V_{y_k}^\top B_{x_k}^{\mathrm{w}} Q_{x_k}) x_k + B_{y_k}^{\mathrm{v}} Q_{y_k} y_k = q_{k+1} \\ \Leftrightarrow\ & G_k Q_{x_k} x'_k + G_k P_{x_k} \tilde{x}'_k + \bar{M}_k^{\mathrm{v}} \tilde{X}'_{k-1} + B_{x_k} x_k + B_{y_k} y_k = q_{k+1} \end{aligned}$$

with $q_{k+1} := q_k^{\mathrm{v}} - (G_k P'_{x_k} + B_{x_k}^{\mathrm{v}} P_{x_k}) q_{\tilde{x}_k} - B_{y_k}^{\mathrm{v}} P_{y_k} \bar{q}_{\tilde{y}_k}$ and $\bar{M}_k^{\mathrm{v}} := M_k^{\mathrm{v}} + \bar{M}_{x_k} + \bar{M}_{xy_k}$. Then we factorize the equation by $V_{x_k}^\top$ and $W_{x_k}^\top$ and obtain:

$$\begin{aligned} & G_k Q_{x_k} x'_k + G_k P_{x_k} \tilde{x}'_k + \bar{M}_k^{\mathrm{v}} \tilde{X}'_{k-1} + B_{x_k} x_k + B_{y_k} y_k = q_{k+1} \\ \Leftrightarrow\ & \begin{cases} V_{x_k}^\top G_k Q_{x_k} x'_k + V_{x_k}^\top G_k P_{x_k} \tilde{x}'_k + V_{x_k}^\top \bar{M}_k^{\mathrm{v}} \tilde{X}'_{k-1} + V_{x_k}^\top B_{x_k} x_k + V_{x_k}^\top B_{y_k} y_k = V_{x_k}^\top q_{k+1} \\ W_{x_k}^\top G_k P_{x_k} \tilde{x}'_k + W_{x_k}^\top \bar{M}_k^{\mathrm{v}} \tilde{X}'_{k-1} + W_{x_k}^\top B_{x_k} x_k + W_{x_k}^\top B_{y_k} y_k = W_{x_k}^\top q_{k+1} \end{cases} \end{aligned}$$

$$\Leftrightarrow \begin{cases} G_{k+1}x_k' + M_{k+1}^{\mathrm{v}}\tilde{X}_k' + B_{x_{k+1}}^{\mathrm{v}}x_k + B_{y_k}^{\mathrm{v}}y_k = q_{k+1}^{\mathrm{v}} \\ W_{x_k}^{\top}G_kP_{x_k}\tilde{x}_k' + M_{k+1}^{\mathrm{w}}\tilde{X}_{k-1}' + B_{x_{k+1}}^{\mathrm{w}}x_k + B_{y_k}^{\mathrm{w}}y_k = q_{k+1}^{\mathrm{w}} \end{cases}$$

with $q_{k+1}^{\mathrm{v}} = V_{x_k}^{\top}q_{k+1}$, $q_{k+1}^{\mathrm{w}} = W_{x_k}^{\top}q_{k+1}$, $M_{k+1}^{\mathrm{v}} := V_{x_k}^{\top}\bar{M}_k^{\mathrm{v}}$ and $M_{k+1}^{\mathrm{w}} := W_{x_k}^{\top}\bar{M}_k^{\mathrm{v}}$. Hence the induction is concluded.

Then for $k = \mu$ we obtain

$$\begin{aligned} G_\mu x_{\mu-1}' + M_\mu^{\mathrm{v}}\tilde{X}_{\mu-1}' + B_{x_\mu}^{\mathrm{v}}x_{\mu-1} + B_{y_\mu}^{\mathrm{v}}y_{\mu-1} &= q_\mu^{\mathrm{v}} \\ W_{x_{\mu-1}}^{\top}G_{\mu-1}P_{x_{\mu-1}}\tilde{x}_{\mu-1}' + M_\mu^{\mathrm{w}}\tilde{X}_{\mu-2}' + B_{x_\mu}^{\mathrm{w}}x_{\mu-1} + B_{y_\mu}^{\mathrm{w}}y_{\mu-1} &= q_\mu^{\mathrm{w}}, \end{aligned}$$

which yields regarding Lemma 4.14

$$\begin{aligned} &x_\mu' + M_\mu\tilde{X}_{\mu-1}' + B_\mu x_\mu = q_\mu \\ &\tilde{y}_\mu = q_{\tilde{y}_\mu} - (B_{y_\mu}^{\mathrm{w}})^{-1}W_{x_{\mu-1}}^{\top}G_{\mu-1}P_{x_{\mu-1}}\tilde{x}_{\mu-1}' + M_{y_\mu}\tilde{X}_{\mu-2}' - (B_{y_\mu}^{\mathrm{w}})^{-1}B_{x_\mu}^{\mathrm{w}}x_\mu \end{aligned}$$

with $\tilde{y}_\mu := y_{\mu-1}$, $x_\mu := x_{\mu-1}$,

$$\begin{aligned} M_\mu &:= G_\mu^{-1}M_\mu^{\mathrm{v}} + G_\mu^{-1}B_{y_\mu}^{\mathrm{v}}\left(-(B_{y_\mu}^{\mathrm{w}})^{-1}W_{x_{\mu-1}}^{\top}G_{\mu-1}P_{x_{\mu-1}} \quad M_{y_\mu}\right), \\ B_\mu &:= G_\mu^{-1}B_{x_\mu}^{\mathrm{v}} - G_\mu^{-1}B_{y_\mu}^{\mathrm{v}}(B_{y_\mu}^{\mathrm{w}})^{-1}B_{x_\mu}^{\mathrm{w}}, \\ q_\mu &:= G_\mu^{-1}q_\mu^{\mathrm{v}} - G_\mu^{-1}B_{y_\mu}^{\mathrm{v}}q_{\tilde{y}_\mu} \end{aligned}$$

and

$$M_{y_\mu} := -(B_{y_\mu}^{\mathrm{w}})^{-1}M_\mu^{\mathrm{w}} \quad \text{and} \quad q_{\tilde{y}_\mu} := (B_{y_\mu}^{\mathrm{w}})^{-1}q_\mu^{\mathrm{w}}.$$

We introduce the transformation

$$\bar{y}_i := \tilde{y}_i + (V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}B_{x_i}^{\mathrm{w}}(Q_{x_i}x_i + P_{x_i}\tilde{x}_i)$$

and obtain the system regarding (4.20a) and (4.20c)

$$x_\mu' + M_\mu\tilde{X}_{\mu-1}' + B_\mu x_\mu = q_\mu(t) \tag{4.21a}$$

$$(W_{y_i}^{\top}B_{x_i}^{\mathrm{w}}P_{x_i})^{-1}W_{y_i}^{\top}W_{x_{i-1}}^{\top}G_{i-1}P_{x_{i-1}}\tilde{x}_{i-1}' - M_{x_i}\tilde{X}_{i-2}' + \tilde{x}_i = q_{\tilde{x}_i}(t), \quad \text{for } 1 \leqslant i \leqslant \mu - 1 \tag{4.21b}$$

$$(V_{y_i}^{\top}B_{y_i}^{\mathrm{w}}P_{y_i})^{-1}V_{y_i}^{\top}W_{x_{i-1}}^{\top}G_{i-1}P_{x_{i-1}}\tilde{x}_{i-1}' - M_{y_i}\tilde{X}_{i-2}' + \bar{y}_i = q_{\tilde{y}_i}(t), \quad \text{for } 1 \leqslant i \leqslant \mu. \tag{4.21c}$$

Define

$$z_0 := x_\mu, \quad z_1 := \bar{y}_\mu, \quad z_i := \begin{pmatrix} \tilde{x}_{\mu+1-i} \\ \bar{y}_{\mu+1-i} \end{pmatrix}, \qquad \text{for } 2 \leqslant i \leqslant \mu.$$

After transforming the variables of (4.21) into z, the system (4.21) is in S-canonical form with:

$$N_{\mu-i,\mu+1-i} := \begin{pmatrix} (W_{y_{i+1}}^\top B_{x_{i+1}}^{\mathrm{w}} P_{x_{i+1}})^{-1} & 0 \\ 0 & (V_{y_{i+1}}^\top B_{y_{i+1}}^{\mathrm{w}} P_{y_{i+1}})^{-1} \end{pmatrix} \begin{pmatrix} W_{y_{i+1}}^\top \\ V_{y_{i+1}}^\top \end{pmatrix} \begin{pmatrix} W_{x_i}^\top G_i P_{x_i} & 0 \end{pmatrix}.$$

Then $N_{\mu-i,\mu+1-i}$ has full row rank since

$$\begin{pmatrix} W_{y_{i+1}}^\top \\ V_{y_{i+1}}^\top \end{pmatrix}, \quad W_{x_i}^\top G_i P_{x_i}, \quad V_{y_{i+1}}^\top B_{y_{i+1}}^{\mathrm{w}} P_{y_{i+1}} \text{ and } W_{y_{i+1}}^\top B_{x_{i+1}}^{\mathrm{w}} P_{x_{i+1}}$$

are non-singular by the Lemmata 4.14 and 4.16. With the help of Equation (4.14) it holds

$$\begin{aligned} \operatorname{rk} N_{\mu-i,\mu+1-i} &:= \operatorname{rk} \begin{pmatrix} (W_{y_{i+1}}^\top B_{x_{i+1}}^{\mathrm{w}} P_{x_{i+1}})^{-1} & 0 \\ 0 & (V_{y_{i+1}}^\top B_{y_{i+1}}^{\mathrm{w}} P_{y_{i+1}})^{-1} \end{pmatrix} \begin{pmatrix} W_{y_{i+1}}^\top \\ V_{y_{i+1}}^\top \end{pmatrix} \begin{pmatrix} W_{x_i}^\top G_i P_{x_i} & 0 \end{pmatrix} \\ &= \operatorname{rk} \begin{pmatrix} W_{x_i}^\top G_i P_{x_i} & 0 \end{pmatrix} \\ &= \operatorname{rk} P_{x_i} = \operatorname{rk} P_{x_\mu} + \sum_{j=i+1}^{\mu} \operatorname{rk} B_{y_j}^{\mathrm{w}} = \sum_{j=i+1}^{\mu} \operatorname{rk} B_{y_j}^{\mathrm{w}} \\ &= n - (\sum_{j=1}^{i} \operatorname{rk} B_{y_j}^{\mathrm{w}} + \operatorname{rk} AD) = n - r_i^\mu. \end{aligned}$$

This yields $l_{\mu-i} = n - r_i^\mu$. □

Using Theorem 4.25 and Theorem 2.80 in [LMT13] we describe the relation between the Strangeness Index and the Dissection Index for the linear case. Relations between the already established concepts are described in [Meh12].

Theorem 4.26. (Index relations: linear DAEs)
Consider a linear time dependent DAE (4.15) in standard form. Let the DAE have a finite Dissection Index μ and a finite Strangeness Index μ_S. Then it holds: $\mu = \mu_S + 1$.

Proof.
Let $P(t)$ be an orthonormal basis of $(\ker A(t))^\perp$ which yields in particular $P^\top(t)Q(t) = 0$. Then $P(t)P^\top(t)$ is a projector with $A(t) = A(t)P(t)P^\top(t)$. Furthermore consider the systems

$$A(t)x' + B(t)x = q(t) \tag{4.22}$$

and

$$A(t)(D(t)x)' + (B(t) - A(t)D'(t))x = q(t) \tag{4.23}$$

with $D(t) = P(t)P^{\top}(t)$.
The Strangeness Index of (4.22) plus one coincides with the Tractability Index of the proper formulated DAE (4.23), cf. Theorem 2.80. on page 162 in [LMT13]. On the other hand the Tractability Index of (4.23) equals the Dissection Index of (4.23) by Theorem 4.25. At last the Dissection Index of (4.23) and (4.22) coincide since the matrices AD, G_1, $B^{\mathrm{v}}_{x_1}$, $B^{\mathrm{v}}_{y_1}$, $B^{\mathrm{w}}_{x_1}$ and $B^{\mathrm{w}}_{y_1}$ coincide for both systems □

We conclude this section by using Theorem 4.22 and 4.25 to describe the relation between the Tractability Index and the Dissection Index for the nonlinear case.

Theorem 4.27. (Index relations: nonlinear DAEs)
Consider a nonlinear DAE (4.9) which fulfills Assumption 2.25 and has sufficiently smooth functions f and d. Let the DAE have a finite Dissection Index μ and a finite Tractability Index μ_T. Then the values of the index and the characteristic values of both concepts coincide.

Proof.
The linearizations of (4.9) have the Dissection Index and the same Tractability Index, and vice versa, by Theorem 4.22 and Theorem 3.33 in [LMT13]. By Theorem 4.25 the values of the index and the characteristic values of the linearizations of both concepts coincide. □

4.3 Dissection Index for Circuit Applications

In Section 4.1 the simplicity of the Dissection Index is demonstrated for some example circuits. In this section we generalize this observation to the electric circuits, introduced in Chapter 3, without controlled elements. Therefore we write the equations of the MNA

$$\begin{aligned} A_{\mathcal{C}}(\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,t) + g_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,\zeta,\Psi)) + A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^{\top}e,q_M,t) + A_{\mathcal{L}}j_{\mathcal{L}} + A_V j_V + A_I i_s(t) &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}\phi_{\mathcal{L}}(j_{\mathcal{L}},t) - A_{\mathcal{L}}^{\top}e + \chi_{\mathcal{L}}E &= 0, \\ A_V^{\top}e - v_s(t) &= 0 \\ M_\zeta\frac{\mathrm{d}}{\mathrm{d}t}\zeta + h_\zeta(\zeta,\Psi,A_S^{\top}e) &= 0 \\ T\Psi(t) - h_\Psi(\zeta) &= 0 \\ \frac{\mathrm{d}}{\mathrm{d}t}\phi_M(q_M,t) - A_M^T e &= 0 \\ M_\varepsilon\frac{\mathrm{d}}{\mathrm{d}t}E + M_\sigma E - J - \chi_{\mathcal{L}}^T j_{\mathcal{L}} &= 0, \\ \frac{\mathrm{d}}{\mathrm{d}t}J + M_{CC}E &= 0 \end{aligned} \tag{4.24}$$

into the form of a nonlinear DAE (2.14). The Equation (4.24) is the MNA for circuits with the capacitors, inductors, resistors and independent sources of Section 3.1.1, the semiconductor devices of Section 3.1.2, the memristors of Section 3.1.3 and the electromagnetic devices of Section 3.1.4. Furthermore we assume the assumptions 3.9 and 3.10 of Section 3.1.5 to be fulfilled.
We denote

$$d(x,t) := \begin{pmatrix} q_{\mathcal{C}}(A_{\mathcal{C}}^{\top} e, t) \\ \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) \\ \zeta \\ \phi_M(q_M, t) \\ E \\ J \end{pmatrix}$$

and

$$f(y,x,t) := \left(\begin{array}{ccc} A_{\mathcal{C}} y_1 & + & A_{\mathcal{C}} g_{\mathcal{C}}(A_{\mathcal{C}}^{\top} e, \zeta, \Psi) + A_{\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^{\top} e, q_M, t) + A_{\mathcal{L}} j_{\mathcal{L}} + A_V j_V + A_I i_s(t) \\ y_2 & - & A_{\mathcal{L}}^{\top} e + \chi_{\mathcal{L}} E \\ & & A_V^{\top} e - v_s(t) \\ M_\zeta y_3 & + & h_\zeta(\zeta, \Psi, A_S^{\top} e) \\ & & T\Psi(t) - h_\varphi(\zeta) \\ y_4 & - & A_M^T e \\ M_\varepsilon y_5 & + & M_\sigma E - J - \chi_{\mathcal{L}}^T j_{\mathcal{L}} \\ y_6 & + & M_{CC} E \end{array}\right)$$

with the variables $x = \begin{pmatrix} e & j_{\mathcal{L}} & j_V & \zeta & \Psi & q_M & E & J \end{pmatrix}^{\top}$. Now we can construct the matrix chain of the Dissection Index. We start by defining

$$A := \begin{pmatrix} A_{\mathcal{C}} & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix},$$

$$D(x,t) := \begin{pmatrix} \mathcal{C}(A_{\mathcal{C}}^{\top} e, t) A_{\mathcal{C}}^{\top} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{L}(j_{\mathcal{L}}, t) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & M(q_M, t) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}$$

with

$$\mathcal{C}(u,t) := \frac{\partial}{\partial u} q_{\mathcal{C}}(u,t), \quad \mathcal{L}(j,t) := \frac{\partial}{\partial j}\phi_{\mathcal{L}}(j,t) \text{ and } M(q,t) := \frac{\partial}{\partial q}\phi_M(q,t)$$

and

$$B(x,t) := \begin{pmatrix} A_{\mathcal{C}}G_{\mathcal{C}}A_{\mathcal{C}}^\top + A_{\mathcal{R}}G_{\mathcal{R}}A_{\mathcal{R}}^\top & A_{\mathcal{L}} & A_V & A_{\mathcal{C}}G_{\mathcal{C},\zeta} & A_{\mathcal{C}}G_{\mathcal{C},\Psi} & A_{\mathcal{R}}G_{\mathcal{R},q_M} & 0 & 0 \\ -A_{\mathcal{L}}^\top & 0 & 0 & 0 & 0 & 0 & \chi_{\mathcal{L}} & 0 \\ -A_V^\top & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & H_\zeta & H_{\zeta,\Psi} & H_{\zeta,e}A_S^\top & 0 & 0 \\ 0 & 0 & 0 & H_\varphi & T & 0 & 0 & 0 \\ -A_M^T & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -\chi_{\mathcal{L}}^T & 0 & 0 & 0 & 0 & M_\sigma & -I \\ 0 & 0 & 0 & 0 & 0 & 0 & M_{CC} & 0 \end{pmatrix}$$

with

$$\begin{array}{rcl@{\qquad}rcl} G_{\mathcal{C}} & := & \partial_1 g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi), & H_\varphi & := & \partial_1 h_\varphi(\zeta), \\ G_{\mathcal{R}} & := & \partial_1 g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t), & H_\zeta & := & \partial_1 h_\zeta(\zeta, \Psi, A_S^\top e), \\ G_{\mathcal{C},\zeta} & := & \partial_2 g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi), & H_{\zeta,\Psi} & := & \partial_2 h_\zeta(\zeta, \Psi, A_S^\top e), \\ G_{\mathcal{C},\Psi} & := & \partial_3 g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi), & H_{\zeta,e} & := & \partial_3 h_\zeta(\zeta, \Psi, A_S^\top e), \\ G_{\mathcal{R},q_M} & := & \partial_2 g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t). & & & \end{array}$$

We hereby obtain

$$G_0(x,t) := AD(x,t) = \begin{pmatrix} A_C\mathcal{C}(A_C^\top e,t)A_C^\top & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{L}(j_{\mathcal{L}},t) & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & M_\zeta & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & M(q_M,t) & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}.$$

To continue the chain we need basis functions which are related to the incidence matrices. Let $P_{\mathcal{C}}$ and $Q_{\mathcal{C}}$ be the basis functions associated to the complementary kernel and the kernel of $A_{\mathcal{C}}^\top$. We then call

$$A_{\bar{\mathcal{C}}X} \quad := \quad Q_{\mathcal{C}}^\top A_X, \quad X \in \{V, \mathcal{R}, \mathcal{L}, I\}$$

the $\mathcal{C}$-reduced incidence matrix of the voltage sources, resistors and memristors, inductors and electromagnetic devices or current sources, respectively. Further denote the full set of associated basis functions of $A_{\bar{\mathcal{C}}V}^\top$ by P_V, Q_V, V_V and W_V. Analogously we call

$$A_{\bar{\mathcal{C}}\bar{V}X} \quad := \quad Q_V^\top Q_{\mathcal{C}}^\top A_X, \quad X \in \{\mathcal{R}, \mathcal{L}, I\}$$

the $\mathcal{C}V$-reduced incidence matrix of the resistors and memristors, inductors and electromagnetic devices or current sources, respectively. At last we obtain the basis functions $P_\mathcal{R}$ and $Q_\mathcal{R}$ associated to the co-kernel and the kernel of $A^\top_{\bar{\mathcal{C}}\bar{V}\mathcal{R}}$ and denote by

$$A_{\bar{\mathcal{C}}\bar{V}\bar{\mathcal{R}}X} \quad := \quad Q^\top_\mathcal{R} Q^\top_V Q^\top_\mathcal{C} A_X, \quad X \in \{\mathcal{L}, I\}$$

the CVR-reduced incidence matrix of the inductors and electromagnetic devices or current sources, respectively. Furthermore define the basis functions $P_{\mathcal{L}I}$ and $Q_{\mathcal{L}I}$ of the complementary kernel and the kernel of $A_{\bar{\mathcal{C}}\bar{V}\bar{\mathcal{R}}\mathcal{L}}$ and the basis functions $P_{\mathcal{C}V}$ and $Q_{\mathcal{C}V}$ of the complementary kernel and the kernel of $W^\top_V A^T_V P_\mathcal{C}$. Now we construct the matrix chain of the coupled problem. We start with

$$P = V := \begin{pmatrix} P_\mathcal{C} & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}, \quad Q = W := \begin{pmatrix} Q_\mathcal{C} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

and therefore we get

$$G_1(x,t) = V^\top G(x,t) P = \begin{pmatrix} P^\top_\mathcal{C} A_C \mathcal{C}(A^\top_C e, t) A^\top_C P_\mathcal{C} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mathcal{L}(j_\mathcal{L}, t) & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & M(q_M, t) & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}$$

and

$$B^{\mathrm{v}}_{y_1}(x,t) = V^\top B(x,t) Q = \begin{pmatrix} P^\top_\mathcal{C} A_\mathcal{R} G_\mathcal{R} A^\top_{\bar{\mathcal{C}}\mathcal{R}} & P^\top_\mathcal{C} A_V & P^\top_\mathcal{C} A_\mathcal{C} G_{\mathcal{C},\Psi} \\ -A^\top_{\bar{\mathcal{C}}\mathcal{L}} & 0 & 0 \\ 0 & 0 & H_{\zeta,\Psi} \\ -A^\top_{\bar{\mathcal{C}}M} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

$$B^{\mathrm{w}}_{x_1}(x,t) = W^\top B(x,t) P = \begin{pmatrix} A_{\bar{\mathcal{C}}\mathcal{R}} G_\mathcal{R} A^\top_\mathcal{R} P_\mathcal{C} & A_{\bar{\mathcal{C}}\mathcal{L}} & 0 & A_{\bar{\mathcal{C}}\mathcal{R}} G_{\mathcal{R},q_M} & 0 & 0 \\ -A^\top_V P_\mathcal{C} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & H_\varphi & 0 & 0 & 0 \end{pmatrix},$$

$$B^{\mathrm{w}}_{y_1}(x,t) = W^\top B(x,t) Q = \begin{pmatrix} A_{\bar{\mathcal{C}}\mathcal{R}} G_\mathcal{R} A^\top_{\bar{\mathcal{C}}\mathcal{R}} & A_{\bar{\mathcal{C}}V} & 0 \\ -A^\top_{\bar{\mathcal{C}}V} & 0 & 0 \\ 0 & 0 & T \end{pmatrix}$$

and obtain the basis functions

$$Q_{y_1} = W_{y_1} := \begin{pmatrix} Q_V Q_{\mathcal{R}} & 0 \\ 0 & W_V \\ 0 & 0 \end{pmatrix}$$

associated to $B^{\mathrm{w}}_{y_1}$. We compute

$$W_{y_1}^{\top} B^{\mathrm{w}}_{x_1} = -(B^{\mathrm{v}}_{y_1} Q_{y_1})^T = \begin{pmatrix} 0 & A_{\bar{\mathcal{C}}\bar{V}\bar{R}\mathcal{L}} & 0 & 0 & 0 & 0 \\ -W_V^{\top} A_V^{\top} P_{\mathcal{C}} & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

to obtain the basis functions

$$Q_{x_1} = W_y^* := \begin{pmatrix} Q_{\mathcal{C}V} & 0 & 0 & 0 & 0 & 0 \\ 0 & Q_{\mathcal{L}I} & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}.$$

We stop the calculation of the matrix chain with

$$\begin{aligned} &(W_y^*)^{\top} G_1(x,t) Q_{x_1} \\ &= \begin{pmatrix} Q_{\mathcal{C}V}^{\top} P_{\mathcal{C}}^{\top} A_{\mathcal{C}} \mathcal{C}(A_{\mathcal{C}}^{\top} e, t) A_{\mathcal{C}}^{\top} P_{\mathcal{C}} Q_{\mathcal{C}V} & 0 & 0 & 0 & 0 & 0 \\ 0 & Q_{\mathcal{L}I}^{T} \mathcal{L}(j_{\mathcal{L}}, t) Q_{\mathcal{L}I} & 0 & 0 & 0 & 0 \\ 0 & 0 & M_{\zeta} & 0 & 0 & 0 \\ 0 & 0 & 0 & M(q_M, t) & 0 & 0 \\ 0 & 0 & 0 & 0 & M_{\varepsilon} & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}. \end{aligned}$$

With the help of these three matrices we can prove the following topological index theorem for electrical circuits.

Theorem 4.28.
Under the assumptions 3.9 and 3.10 the MNA (3.34) has at most Dissection Index 2. In particular it has index

(i) 0, if and only if there is a spanning tree in the circuit consisting only of capacitors and there are neither voltage sources nor semiconductors in the circuit.

(ii) 1, or lower if and only if there are no loops consisting of capacitors and voltage sources with at least one voltage source and no cutsets consisting of inductors or electromagnetic devices and current sources.

Proof. As long as $(W_y^*)^\top G_1(x,t)Q_{x_1}$ is non-singular the index is 2 at most due to Lemma 4.15. We know that $\mathcal{C}(A_\mathcal{C}^\top e,t)$, $\mathcal{L}(j_\mathcal{L},t)$, M_ζ, $M(q_M,t)$ and M_ε are positive definite and that $A_\mathcal{C}^\top P_\mathcal{C}$, Q_{CV} and $Q_{\mathcal{L}I}$ have full column rank, hence the index is at most 2.

The topological index-1 conditions stated in (ii) are equivalent to the conditions that $A_{\bar{\mathcal{C}}V}$ has full column rank and that $\begin{pmatrix}A_\mathcal{C} & A_\mathcal{R} & A_V\end{pmatrix}$ has full row rank which is then equivalent to $A_{\bar{\mathcal{C}}\bar{V}\mathcal{R}}$ having full row rank since $Q_\mathcal{C}^\top$ and $Q_V^\top$ have full row rank. And

$$B_{y_1}^{\mathrm{w}}(x,t) = \begin{pmatrix} A_{\bar{\mathcal{C}}\mathcal{R}}G_\mathcal{R}A_{\bar{\mathcal{C}}\mathcal{R}}^\top & A_{\bar{\mathcal{C}}V} & 0 \\ -A_{\bar{\mathcal{C}}V}^\top & 0 & 0 \\ 0 & 0 & T \end{pmatrix}$$

is non-singular if and only if $A_{\bar{\mathcal{C}}V}$ has full column rank and $A_{\bar{\mathcal{C}}\bar{V}\mathcal{R}}$ has full row rank since $G_\mathcal{R}$ positive definite. Therefore (ii) holds.

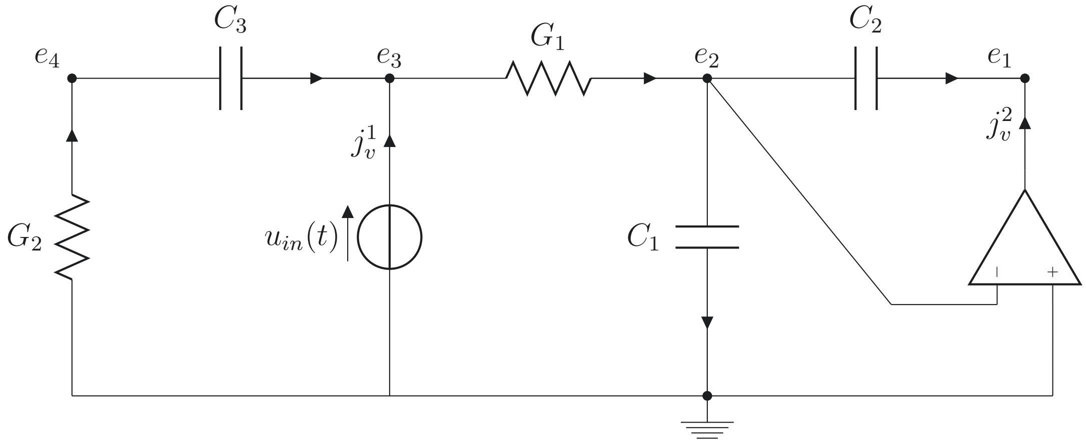
Next observe G_0 under the assumption that there are neither voltage sources nor semiconductors

$$G_0(x,t) = \begin{pmatrix} A_C C(A_C^\top e,t)A_C^\top & 0 & 0 & 0 & 0 \\ 0 & \mathcal{L}(j_\mathcal{L},t) & 0 & 0 & 0 \\ 0 & 0 & M(q_M,t) & 0 & 0 \\ 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix}.$$

The index 0 condition also states that there is a spanning tree in the circuit consisting only of capacitors and therefore $A_C^\top$ has full column rank. By $C(A_C^\top e,t)$, $\mathcal{L}(j_\mathcal{L},t)$, $M(q_M,t)$ and M_ε being positive definite again, G_0 is non-singular. □

The topological index result, with respect to the Tractability Index for circuits without semiconductor devices, memristors and electromagnetic devices, can be found in [Tis99]. In comparison to [Bau12] this result is consistent with the index result regarding the Lorenz gauge. The index result itself is new because we used another index concept, but again we noticed that we only used constant basis functions. In [Est00] a constant projector chain for circuits without semiconductor devices, memristors and electromagnetic devices can be found. But in our case the Dissection Index concept directly provides us with the constant topological basis functions.

It is well known that the Tractability Index of a circuit is no longer restricted by 2, if controlled elements are added, cf. [ET00]. A representative example is the Miller integrator 2.13, cf. [MG05, Pul12]. Example 2.13 has no inherent dynamic in the index 3 case. For demonstrative reasons we add one more resistor and one more capacitor to the example and obtain the following circuit.



If the first two capacitor capacities are non-linear, the characteristic values of the associated DAE could depend on the solution variables. Let the first two capacitors have the same constant capacity $C_1 = C_2 = C$ since we restrict ourselves to DAEs with constant characteristic values in this thesis. The other elements are allowed to be non-linear.

The corresponding equations are given by:

Example 4.29. Let $\mathcal{I} := [0, 2 \cdot 10^{-6}]$ and let $t \in \mathcal{I}$.

$$\begin{aligned}
-(C(e_2 - e_1))' - j_v^2 &= 0\\
(Ce_2)' + (C(e_2 - e_1))' - g_1(e_3 - e_2, t) &= 0\\
-q'_{C_3}(e_4 - e_3, t) + g_1(e_3 - e_2, t) - j_v^1 &= 0\\
q'_{C_3}(e_4 - e_3, t) - g_2(-e_4, t) &= 0\\
e_3 - u_{in}(t) &= 0\\
e_1 - 2e_2 &= 0
\end{aligned}$$

In the following we show that there is a constant basis chain for Example 4.29. So even for controlled sources the Dissection Index concept may provide constant basis functions. Hence the Dissection Index may provide constant basis functions for electric circuits even if they contain controlled circuits. To do so, we define $x = \begin{pmatrix} e_1 & e_2 & e_3 & e_4 & j_v^1 & j_v^2 \end{pmatrix}$, $C_3 := \frac{\partial}{\partial v} q_{C_3}(e_4 - e_3, t)$, $G_1 := \frac{\partial}{\partial v} g_1(e_3 - e_2, t)$, $G_2 := \frac{\partial}{\partial v} g_2(-e_4, t)$, the matrices and functions

$$A := \begin{pmatrix} 0 & -1 & 0\\ 1 & 1 & 0\\ 0 & 0 & -1\\ 0 & 0 & 1\\ 0 & 0 & 0 \end{pmatrix}, \quad D(x,t) := \begin{pmatrix} 0 & C & 0 & 0 & 0 & 0\\ -C & C & 0 & 0 & 0 & 0\\ 0 & 0 & -C_3 & C_3 & 0 & 0 \end{pmatrix}$$

and

$$b(x,t) := \begin{pmatrix} -j_v^2 \\ -g_1(e_3 - e_2, t) \\ g_1(e_3 - e_2, t) - j_v^1 \\ -g_2(-e_4, t) \\ e_3 - u_{in}(t) \\ e_1 - 2e_2 \end{pmatrix}.$$

Then we begin the matrix chain by

$$AD := \begin{pmatrix} C & -C & 0 & 0 & 0 & 0 \\ -C & 2C & 0 & 0 & 0 & 0 \\ 0 & 0 & C_3 & -C_3 & 0 & 0 \\ 0 & 0 & -C_3 & C_3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad B(x,t) := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & G_1 & -G_1 & 0 & 0 & 0 \\ 0 & -G_1 & G_1 & 0 & -1 & 0 \\ 0 & 0 & 0 & G_2 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -2 & 0 & 0 & 0 & 0 \end{pmatrix},$$

which allows us to choose

$$P = V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Hence we obtain

$$G_1 = \begin{pmatrix} C & -C & 0 \\ -C & 2C & 0 \\ 0 & 0 & 4C_3 \end{pmatrix}, B_{x_1}^{\mathrm{v}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & G_1 & -G_1 \\ 0 & -G_1 & G_1 + G_2 \end{pmatrix}, B_{y_1}^{\mathrm{v}} = \begin{pmatrix} 0 & -1 & 0 \\ -G_1 & 0 & 0 \\ G_1 - G_2 & 0 & -1 \end{pmatrix},$$

$$B_{x_1}^{\mathrm{w}} = \begin{pmatrix} 0 & -G_1 & G_1 - G_2 \\ 0 & 0 & 1 \\ 1 & -2 & 0 \end{pmatrix} \text{ and } B_{y_1}^{\mathrm{w}} = \begin{pmatrix} G_1 + G_2 & 0 & -1 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Due to the matrix $B_{y_1}^{\mathrm{w}}$ we can choose

$$P_{y_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}, \quad Q_{y_1} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad V_{y_1} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad W_{y_1} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Now we can calculate

$$W_{y_1}^\top B_{x_1}^{\mathrm{w}} = \begin{pmatrix} 1 & -2 & 0 \end{pmatrix}, \quad Q_{x_1} = \begin{pmatrix} 2 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad P_{x_1} = \begin{pmatrix} 1 \\ -2 \\ 0 \end{pmatrix}.$$

This yields

$$G_1 Q_{x_1} = \begin{pmatrix} C & 0 \\ 0 & 0 \\ 0 & 4C_3 \end{pmatrix}, \quad W_{x_1} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad V_{x_1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

We then obtain

$$B_{x_1} = B^{\mathrm{v}}_{x_1} Q_{x_1} - B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^\top B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^\top B^{\mathrm{w}}_{x_1} Q_{x_1} = \begin{pmatrix} 0 & 0 \\ G_1 & 0 \\ -2G_1 & 0 \end{pmatrix}$$

and thereby

$$G_2 = \begin{pmatrix} C & 0 \\ 0 & 4C_3 \end{pmatrix}, \quad B^{\mathrm{v}}_{x_2} = \begin{pmatrix} 0 & 0 \\ -2G_1 & 0 \end{pmatrix}, \quad B^{\mathrm{v}}_{y_2} = \begin{pmatrix} -1 \\ 0 \end{pmatrix},$$
$$B^{\mathrm{w}}_{x_2} = \begin{pmatrix} G_1 & 0 \end{pmatrix} \text{ and } B^{\mathrm{w}}_{y_2} = \begin{pmatrix} 0 \end{pmatrix}.$$

Due to the singularity of $B^{\mathrm{w}}_{y_2}$ we choose

$$P_{y_2} = V_{y_2} = \begin{pmatrix} \ \end{pmatrix}, \quad Q_{y_2} = W_{y_2} = \begin{pmatrix} 1 \end{pmatrix},$$

which allows us to compute

$$W_{y_2}^\top B^{\mathrm{w}}_{x_2} = \begin{pmatrix} G_1 & 0 \end{pmatrix}, \quad Q_{x_2} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad P_{x_2} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

This yields

$$G_2 Q_{x_2} = \begin{pmatrix} 0 \\ 4C_3 \end{pmatrix}, \quad W_{x_3} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad B^{\mathrm{w}}_{y_3} = \begin{pmatrix} -1 \end{pmatrix}$$

hence the DAE has Dissection Index 3. Of course it will not be possible to find a constant decoupling for all circuits including controlled sources, but Example 4.29 motivates to investigate under which conditions such a constant decoupling is possible.

4.4 Perturbation Analysis: Nonlinear DAEs

In particular the right hand side of a DAE is perturbed by the rounding error during the numerical simulation. Therefore we consider the following perturbed system.

Definition 4.30. (Perturbed DAE)
Consider a DAE (4.9) and a function $\delta \in C(\mathcal{I}, \mathbb{R}^n)$. We call

$$f(d'(x(t), t), x(t), t) = \delta(t) \tag{4.25}$$

the perturbed DAE with the perturbation δ.

The solution of a DAE (4.9) and the solution of its perturbed system (4.25) is not only influenced by the perturbation δ but also by its derivatives, in general. To measure the degree of the influence of these derivatives the Perturbation Index was defined in [HLR89]. Hence the Perturbation Index is tightly linked to numerical simulation. We saw influence the first derivative of the perturbation in Figure 2.1 when we discretized Example 2.6 with the implicit Euler.

Definition 4.31. (Perturbation Index)
A DAE (4.9) has Perturbation Index μ_P along a solution $x_\star \in C_d^1(\mathcal{I}_\star, \mathcal{D})$ on a compact interval $\mathcal{I}_\star = [t_0, T] \subset \mathcal{I}$, if μ_P is the smallest number, such that for all perturbations $\delta \in C^{\mu-1}(\mathcal{I}_\star, \mathbb{R}^n)$ with $||\delta||_\infty, ..., ||\delta^{(\mu_P-1)}||_\infty$ being sufficiently small the perturbed system has a solution $x_\delta \in C_d^1(\mathcal{I}_\star, \mathcal{D})$ and all solutions of the perturbed system fulfill

$$||x_\star(t) - x_\delta(t)|| \leqslant c(||x_\star(t_0) - x_\delta(t_0)|| + \int_{t_0}^{t} ||\delta(s)||\mathrm{d}s + \sum_{i=0}^{\mu_P-1} \max_{0 \leqslant \tilde{t} \leqslant t} ||\delta^{(i)}(\tilde{t})||) \tag{4.26}$$

if $||x_\star(t_0) - x_\delta(t_0)||$ is sufficiently small.

The major drawback of the Perturbation Index is that it does not provide any systematical way how to determine it by itself. The Tractability Index [LMT13] and the Strangeness Index [KM06] for example provide a well structured way how to determine themselves, see Section 2.4 and Section 2.3. To transfer this advantage to the Perturbation Index the Tractability and the Strangeness Index need to be related to the Perturbation Index. In the case of linear DAEs with time depending coefficients these three index concepts are already related directly to each other [LMT13, KM06]. For general nonlinear DAEs it is proven that the Tractability Index and the Perturbation Index coincide in the index one case [LMT13]. The main objective of this section is to provide a way of determining the Perturbation Index for general nonlinear higher index DAEs. To achieve this goal we use the Dissection Index.
The following lemma is crucial for the decoupling. It can be interpreted as an extension of the Implicit Function Theorem. In contrast to the Implicit Function Theorem, which provides an inverse function to an algebraic equation around a point, Lemma 4.32 provides an inverse function to an algebraic equation around the graph of a function. A similar result can be found in [OR70].

Lemma 4.32.
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{G} \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I}$ be open subsets and let $g \in C^\nu(\mathcal{G}, \mathbb{R}^m)$ for $\nu \in \mathbb{N}$. Let be $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ and let $(x_\star(t), y_\star(t))$ be a continuous solution on $\mathcal{I}_\star$ of the algebraic equation

$$g(x, y, t) = 0 \tag{4.27}$$

with $(x_\star(t), y_\star(t), t) \in \mathcal{G}$ for all $t \in \mathcal{I}_\star$. Furthermore let the Jacobian $\frac{\partial}{\partial y} g(x, y, t)$ be non-singular along $(x_\star(t), y_\star(t), t)$ for all $t \in \mathcal{I}_\star$.
Then there is an open subset $\bar{\mathcal{G}} \subset \mathbb{R}^n \times \mathcal{I}$ with $(x_\star(t), t) \in \bar{\mathcal{G}}$ for all $t \in \mathcal{I}_\star$ such that there is a unique function $\Phi \in C^\nu(\bar{\mathcal{G}}, \mathbb{R}^m)$ with $(x, \Phi(x,t), t) \in \mathcal{G}$ for all $(x,t) \in \bar{\mathcal{G}}$ and

$$g(x, \Phi(x,t), t) = 0, \quad \forall (x,t) \in \bar{\mathcal{G}}.$$

Proof.
With the Implicit Function Theorem we get for every $\bar{t} \in \mathcal{I}_\star$ three constants $r_x^{\bar{t}}, r_y^{\bar{t}}, r_t^{\bar{t}} > 0$ such that there is exactly one function

$$\Psi_{\bar{t}} \in C^\nu(B_{r_x^{\bar{t}}}(x_\star(\bar{t})) \times B_{r_t^{\bar{t}}}(\bar{t}), B_{r_y^{\bar{t}}}(y_\star(\bar{t})))$$

with

$$g(x, \Psi_{\bar{t}}(x,t), t) = 0, \quad \forall (x,t) \in B_{r_x^{\bar{t}}}(x_\star(\bar{t})) \times B_{r_t^{\bar{t}}}(\bar{t})$$

With $\mathcal{I}_\star$ being compact and $x_\star$ and $y_\star$ being continuous, both $x_\star$ and $y_\star$ have compact graphs. Therefore we can choose a finite number of time points $t_1, \cdots, t_k$ such that the associated neighborhoods $B_{r_x^{t_i}}(x_\star(t_i)), B_{r_y^{t_i}}(y_\star(t_i)), B_{r_t^{t_i}}(t_i)$ cover the sets of $\{x_\star(t) | t \in \mathcal{I}_\star\}$ and $\{y_\star(t) | t \in \mathcal{I}_\star\}$ and the time interval $\mathcal{I}_\star$, respectively. Define the combined solution function point wise by

$$\begin{aligned} \Psi : \bigcup_{i=1}^{k} (B_{r_x^{t_i}}(x_\star(t_i)) \times B_{r_t^{t_i}}(t_i)) &\to \bigcup_{i=1}^{k} B_{r_y^{t_i}}(y_\star(t_i)) \\ (x,t) &\mapsto \Psi_{t_j}(x,t) \end{aligned}$$

with j minimal such that $(x,t) \in B_{r_x^{t_j}}(x_\star(t_j)) \times B_{r_t^{t_j}}(t_j)$.
Then Ψ uniquely solves the algebraic equation (4.27) point wise since all Ψ_{t_j} solve the algebraic equation (4.27) point wise uniquely. Further we have to show that Ψ is continuously differentiable ν-times. Therefore consider an arbitrary $(x,t) \in B_{r_x^{t_j}}(x_\star(t_j)) \times B_{r_t^{t_j}}(t_j)$ with j minimal, then there is an open neighborhood $H \subset B_{r_x^{t_j}}(x_\star(t_j)) \times B_{r_t^{t_j}}(t_j)$ around (x,t) since $B_{r_x^{t_j}}(x_\star(t_j)) \times B_{r_t^{t_j}}(t_j)$ is open. We get that $\Psi_{|H}$ as well as $\Psi_{t_j|H}$ solve the algebraic equation (4.27) point wise unique on H hence $\Psi_{|H} \equiv \Psi_{t_j|H}$. Therefore we find a neighborhood around every point such that Ψ is identical to a ν-times continuously differentiable function hence Ψ is ν-times continuously differentiable itself.
Next define $d_x(t) := \operatorname{dist}(x_\star(t), \delta(\bigcup_{i=1}^k B_{r_x^{t_i}}(x_\star(t_i))))$ as the distance of the set $\{x_\star(t) | t \in \mathcal{I}_\star\}$ to the boundary of the domain $\delta(\bigcup_{i=1}^k B_{r_x^{t_i}}(x_\star(t_i))))$. It holds that $d_x(t) > 0$ for all $t \in \mathcal{I}_\star$ since $\bigcup_{i=1}^k B_{r_x^{t_i}}(x_\star(t_i))$ is open and covers $\{x_\star(t) | t \in \mathcal{I}_\star\}$. Furthermore the distance function dist and the solution $x_\star$ are continuous and $\mathcal{I}_\star$ is compact hence there is a minimum $m_x := \min\limits_{t \in \mathcal{I}_\star} d_x(t) > 0$.

Set $r_x = \frac{1}{2}m_x$ and define $\bar{\mathcal{G}} := \{(x,t)| \, ||x - x_\star(t)|| < r_x\}$ then it holds that

$$(x_\star(t), t) \in \bar{\mathcal{G}} \subset \bigcup_{i=1}^{k} (B_{r_x^{t_i}}(x_\star(t_i)) \times B_{r_t^{t_i}}(t_i)), \quad \forall t \in \mathcal{I}_\star.$$

With $\Phi := \Psi_{|\mathcal{G}}$ obtain the desired solution function. □

Besides the Lemmata 4.16 and 4.32 we need an assumption regarding the basis functions:

Assumption 4.33. (Time dependent basis chain)
Consider a DAE (4.9) with a finite Dissection Index. Assume that all basis functions except $V_{x_{\mu_{TS}-1}}$ and $W_{x_{\mu_{TS}-1}}$ depend only on the time t and let $V_{x_{\mu_{TS}-1}}$ and $W_{x_{\mu_{TS}-1}}$ only depend on the time t and the dynamical variables x_0.

Assumption 4.33 seems to be very strict when we think of the projector of the Tractability Index or the basis functions of the Strangeness Index. In contrast to these index concepts the basis functions of the Dissection Index fulfill Assumption 4.33 for a large application class, see Section 4.3. Additionally we need the parts of the functions of the DAE to be sufficiently smooth.

Assumption 4.34. (Differentiability)
Consider a DAE (4.9) with a finite Dissection Index μ_{TS}. Let the DAE fulfill Assumption 4.33. For $1 \leqslant i \leqslant \mu_{TS} - 1$ assume that the matrix valued functions:

$$B^{\mathrm{w}}_{y_i}(x^1, x, t), \quad W^\top_{y_i}(t) B^{\mathrm{w}}_{x_i}(x^1, x, t) \quad \text{and} \quad G_i(x^1, x, t) Q_{x_i}(t)$$

are $(\mu_{TS} + 2 - i)$ times continuously differentiable. Furthermore assume f to be sufficiently differentiable such that:

$$V^\top_{x_{i-2}}(t) f(d'(x,t), x, t)$$

is also $(\mu_{TS} + 2 - i)$ times continuously differentiable and let $\mathcal{G}$ be $C^{\mu_{TS}+2}$-diffeomorphic to a parallelepiped.

Furthermore we introduce some notations regarding the perturbation. Let $\delta(t)$ be a perturbation and $\delta^{(i)}(t)$ be its i-th derivative. Then we gather the perturbation and the derivatives up to order j into a vector $\Delta^{(j)}(t) = \begin{pmatrix} \delta & \delta^{(1)}(t) & \dots & \delta^{(j)}(t) \end{pmatrix}^\top$. And by $\frac{\partial}{\partial \delta^{(i)}}$ we denote the partial derivative with respect to the i-th derivative of the perturbation $\delta(t)$.
With the help of these assumptions we are able to prove the following theorem.

Theorem 4.35. (Decoupling around a solution)
Consider a DAE (4.9), let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ be compact and connected and let $x_0 \in \mathcal{G}$ be the initial value of the IVP

$$\begin{aligned} f(d'(x,t), x, t) &= 0, \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0. \end{aligned}$$

(i) Let the DAE have a finite Dissection Index μ.

(ii) Let the DAE fulfill the Assumptions 4.33 and 4.34.

(iii) Let the IVP have a global unique solution $x_\star$ on $\mathcal{I}_\star$.

Consider the perturbed DAE

$$\begin{aligned} f(d'(x,t),x,t) &= \delta(t), \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0 + \delta_0 \end{aligned}$$

and define the transformation matrix

$$T = \begin{pmatrix} Q_{x_\mu} & Q_{x_0}P_{x_1} & \dots & Q_{x_{\mu-2}}P_{x_{\mu-1}} & Q_{y_0}P_{y_1} & \dots & Q_{y_{\mu-1}}P_{y_\mu} \end{pmatrix}$$

such that the the variable x is split into:

$$\begin{aligned} x &= Q_{x_\mu}x_\mu + \sum_{i=1}^{\mu-1} Q_{x_{i-1}}P_{x_i}\tilde{x}_i + \sum_{i=1}^{\mu} Q_{y_{i-1}}P_{y_i}\tilde{y}_i \\ &= T\begin{pmatrix} x_\mu & \tilde{x}_1 & \dots & \tilde{x}_{\mu-1} & \tilde{y}_1 & \dots & \tilde{y}_\mu \end{pmatrix}^T. \end{aligned}$$

Then for $1 \leqslant i \leqslant \mu$ and $1 \leqslant j \leqslant \mu - 1$ there are neighborhoods $\mathcal{G}_{\tilde{y}_i}$ and $\mathcal{G}_{\tilde{x}_j}$ such that the solution parts $\tilde{y}_i$ and $\tilde{x}_j$ can be described by a $(\mu+1-i)$ times differentiable function $\Psi_{\tilde{y}_i}$ and by a $(\mu+2-j)$ times differentiable function $\Psi_{\tilde{x}_j}$, i.e.

$$\begin{aligned} \tilde{y}_i &= \Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)) \\ \tilde{x}_j &= \Psi_{\tilde{x}_j}(t, \Delta^{(j-1)}(t)) \end{aligned}$$

with

$$\frac{\partial}{\partial x_i}\Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)) = -(V_{y_{i-1}}^\top B_{y_{i-1}}^{\mathrm{w}} P_{y_{i-1}})^{-1} V_{y_{i-1}}^\top B_{x_{i-1}}^{\mathrm{w}} Q_{x_{i-1}}$$

and $\frac{\partial}{\partial \delta^{(j-1)}}\Psi_{\tilde{x}_j}(t, \Delta^{(j-1)}(t))$ having full row rank. Furthermore there is a function f_{x_μ} such that for the solution part x_μ holds:

$$\frac{\mathrm{d}}{\mathrm{d}t}x_\mu = f_{x_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t))$$

Proof.
Keep in mind that all basis functions except $V_{x_{\mu-1}}$ and $W_{x_{\mu-1}}$ depend only on the time t. We will drop the time argument of the basis functions for a more compact notation.

Assume that the functions $\Psi_{\tilde{x}_i}$ and $\Psi_{\tilde{y}_i}$ exist, then we recursively define functions f_i starting with

$$\begin{aligned}&f_0(x_0', x_0, y_0, t, \delta(t))\\ =&f(d_x(\hat{x}, t)(Px_0' + P'x_0 + Q'y_0) + d_t(\hat{x}, t), \hat{x}, t) - \delta(t)\end{aligned}$$

with $\hat{x} = Px_0 + Qy_0$. For all $1 \leqslant i \leqslant \mu$ define

$$f_i(x_i', x_i, y_i, t, \Delta^{(i)}(t)) = V_{x_{i-1}}^{\top} f_{i-1}(\hat{x}_i', \hat{x}_i, \hat{y}_i, t, \Delta^{(i-1)}(t))$$

with

$$\begin{aligned}\hat{x}_i' &= Q_{x_i}x_i' + Q_{x_i}'x_i + P_{x_i}'\Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)) + P_{x_i}\Psi_{\tilde{x}_i}'(t, \Delta^{(i)}(t)),\\ \hat{x}_i &= Q_{x_i}x_i + P_{x_i}\Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)),\\ \hat{y}_i &= Q_{y_i}y_i + P_{y_i}\Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)).\end{aligned}$$

The basis functions as well as the functions $\Psi_{\tilde{x}_i}$ and $\Psi_{\tilde{y}_i}$ are sufficiently smooth due to Assumption 4.34. For the Jacobians of f_0 with respect to x_0', x_0 and y_0 hold:

$$\begin{aligned}\frac{\partial}{\partial x_0'}f_0 &= ADP,\\ \frac{\partial}{\partial x_0}f_0 &= BP + A(DP)',\\ \frac{\partial}{\partial y_0}f_0 &= BQ + A(DQ)'.\end{aligned}$$

We prove the statement for $i \leqslant \mu-2$ and that for the Jacobians of f_i hold for $1 \leqslant i \leqslant \mu-1$:

$$\frac{\partial}{\partial x_i'}f_i = G_iQ_{x_i}, \quad \frac{\partial}{\partial x_i}f_i = B_{x_{i+1}}, \quad \frac{\partial}{\partial y_i}f_i = B_{y_{i+1}}$$

by a mathematical induction:
Base case: $(i = 1)$
We factorize the perturbed DAE with $V_{x_0}^{\top}$, $V_{y_1}^{\top}W_{x_0}^{\top}$ and $W_{y_1}^{\top}W_{x_0}^{\top}$:

$$\begin{aligned}&f(d'(x, t), x, t) = \delta(t)\\ \Leftrightarrow &\begin{pmatrix} V_{x_0}^{\top} \\ V_{y_1}^{\top}W_{x_0}^{\top} \\ W_{y_1}^{\top}W_{x_0}^{\top}\end{pmatrix}(f(d'(x, t), x, t) - \delta(t)) = 0\end{aligned}$$

The DAE has a proper leading term hence $\operatorname{im} AD = \operatorname{im} A$ and for this reason it holds

$$W_{x_0}^{\top}G = W^{\top}AD = 0 \Leftrightarrow W^{\top}A = 0,$$

which means that $W_{x_0}^\top(f(d'(x,t),x,t)$ is independent from its first component and we write $W_{x_0}^\top f(d'(x,t),x,t) =: W_{x_0}^\top f(x,t)$. This leads to:

$$V_{y_1}^\top W_{x_0}^\top f(x,t) - V_{y_1}^\top W_{x_0}^\top \delta(t) = 0, \tag{4.28a}$$
$$W_{y_1}^\top W_{x_0}^\top f(x,t) - W_{y_1}^\top W_{x_0}^\top \delta(t) = 0. \tag{4.28b}$$

We split $x = Px_0 + Qy_0 = Q_{x_0}x_0 + Q_{y_0}y_0$ and notice that $Q_{y_0}y_0$ vanishes in (4.28b) due to the definition of $W_{y_1}^\top$ and Lemma 4.8:

$$V_{y_1}^\top W_{x_0}^\top f(Q_{x_0}x_0 + Q_{y_0}y_0, t) - V_{y_1}^\top W_{x_0}^\top \delta(t) = 0, \tag{4.29a}$$
$$W_{y_1}^\top W_{x_0}^\top f(Q_{x_0}x_0, t) - W_{y_1}^\top W_{x_0}^\top \delta(t) = 0. \tag{4.29b}$$

As the next step we split $x_0 = P_{x_1}\tilde{x}_1 + Q_{x_1}x_1$ and $y_0 = P_{y_1}\tilde{y}_1 + Q_{y_1}y_1$ and see that $Q_{y_1}y_1$ vanishes in (4.29a) and $Q_{x_1}x_1$ vanishes in (4.29b) due to the definitions of Q_{x_1} and Q_{y_1} and Lemma 4.8 again:

$$V_{y_1}^\top W_{x_0}^\top f(Q_{x_0}P_{x_1}\tilde{x}_1 + Q_{x_0}Q_{x_1}x_1 + Q_{y_0}P_{y_1}\tilde{y}_1, t) - V_{y_1}^\top W_{x_0}^\top \delta(t) = 0 \tag{4.30a}$$
$$W_{y_1}^\top W_{x_0}^\top f(Q_{x_0}P_{x_1}\tilde{x}_1, t) - W_{y_1}^\top W_{x_0}^\top \delta(t) = 0 \tag{4.30b}$$

The Jacobian $\frac{\partial}{\partial \tilde{x}_1} W_{y_1}^\top W_{x_0}^\top f(Q_{x_0}P_{x_1}\tilde{x}_1, t) = W_{y_1}^\top W_{x_0}^\top B Q_{x_0} P_{x_1} = W_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{x_1}$ is non-singular due to Lemma 4.16. We transform the exact solution $x_\star$ with the coordinate transformation matrix T and notate the component related to $\tilde{x}_1$ with $\tilde{x}_{\star,1}$. Then by Lemma 4.32 there is an open neighborhood $\mathcal{G}_{x_1}$ around $(\mathcal{I}_\star, 0)$ such that there is a solution function for (4.30b) which describes $\tilde{x}_1$ on $\mathcal{G}_{x_1}$:

$$\tilde{x}_1 = \Psi_{\tilde{x}_1}(t,\delta)$$

with $\frac{\partial}{\partial \delta}\Psi_{\tilde{x}_1}(t,\delta)$ having full row rank since $\frac{\partial}{\partial \delta} W_{y_1}^\top W_{x_0}^\top \delta(t) = W_{y_1}^\top W_{x_0}^\top$ has full row rank. By Lemma 4.7 and Assumption 4.34 the function $\Psi_{\tilde{x}_1}$ is $(\mu+1)$ times continuously differentiable. Insert this expression into (4.30a) and obtain

$$V_{y_1}^\top W_{x_0}^\top f(Q_{x_0}P_{x_1}\Psi_{\tilde{x}_1}(t,\delta(t)) + Q_{x_0}Q_{x_1}x_1 + Q_{y_0}P_{y_1}\tilde{y}_1, t) - V_{y_1}^\top W_{x_0}^\top \delta(t) = 0 \tag{4.31}$$

with the Jacobian with respect to $\tilde{y}_1$

$$\begin{aligned}&\frac{\partial}{\partial \tilde{y}_1} V_{y_1}^\top W_{x_0}^\top f(Q_{x_0}P_{x_1}\Psi_{\tilde{x}_1}(t,\delta(t)) + Q_{x_0}Q_{x_1}x_1 + Q_{y_0}P_{y_1}\tilde{y}_1, t)\\ =&V_{y_1}^\top W_{x_0}^\top B Q_{y_0} P_{y_1} = V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1}\end{aligned}$$

being non-singular. Again by Lemma 4.32 there is an open neighborhood $\mathcal{G}_{y_1}$ around $\{(x_{\star,1}(t), t, 0) | t \in \mathcal{I}_\star\}$ such there is a solution function for (4.31) which describes $\tilde{y}_1$ on $\mathcal{G}_{y_1}$:

$$\tilde{y}_1 = \Psi_{\tilde{y}_1}(x_1, t, \delta)$$

with

$$\frac{\partial}{\partial x_1}\Psi_{\tilde{y}_1}(x_1,t,\delta) = -(V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1})^{-1} V_{y_1}^\top B_{x_1}^{\mathrm{w}} Q_{x_1}$$

which follows exactly as in the proof of the Implicit Function Theorem, c.f. [Zei86]. By Lemma 4.7 and Assumption 4.34 the function $\Psi_{\tilde{y}_1}$ is $(\mu+1)$ times continuously differentiable. Consider

$$f_1(x_1', x_1, y_1, t, \Delta^{(1)}(t)) = V_{x_0}^\top f_0(\hat{x}_1', \hat{x}_1, \hat{y}_1, t, \delta(t))$$

with

$$\begin{aligned}
\hat{x}_1' &= Q_{x_1}x_1' + Q_{x_1}'x_1 + P_{x_1}'\Psi_{\tilde{x}_1}(t,\delta(t)) + P_{x_1}\Psi_{\tilde{x}_1}'(t,\Delta^{(1)}(t)),\\
\hat{x}_1 &= Q_{x_1}x_1 + P_{x_1}\Psi_{\tilde{x}_1}(t,\delta(t)),\\
\hat{y}_1 &= Q_{y_1}y_1 + P_{y_1}\Psi_{\tilde{y}_1}(x_1,t,\delta(t)).
\end{aligned}$$

For the Jacobians of f_1 with respect to x_1', x_1 and y_1 it holds that:

$$\begin{aligned}
\frac{\partial}{\partial x_1'}f_1 &= V_{x_0}^\top(\frac{\partial}{\partial x_0'}f_0)Q_{x_1} = V_{x_0}^\top ADQ_{x_0}Q_{x_1} = G_1Q_{x_1},\\
\frac{\partial}{\partial x_1}f_1 &= V_{x_0}^\top((\frac{\partial}{\partial x_0}f_0)Q_{x_1} + (\frac{\partial}{\partial x_0'}f_0)Q_{x_1}' + (\frac{\partial}{\partial y_0}f_0)P_{y_1}(\frac{\partial}{\partial x_1}\Psi_{\tilde{y}_1}))\\
&= (V_{x_0}^\top\frac{\partial}{\partial x_0}f_0)Q_{x_1} + (V_{x_0}^\top\frac{\partial}{\partial x_0'}f_0)Q_{x_1}' + (V_{x_0}^\top\frac{\partial}{\partial y_0}f_0)P_{y_1}(\frac{\partial}{\partial x_1}\Psi_{\tilde{y}_1})\\
&= B_{x_1}^{\mathrm{v}}Q_{x_1} + G_1Q_{x_1}' - B_{y_1}^{\mathrm{v}}P_{y_1}(V_{y_1}^\top B_{y_1}^{\mathrm{w}}P_{y_1})^{-1}V_{y_1}^\top B_{x_1}^{\mathrm{w}}Q_{x_1} = B_{x_2},\\
\frac{\partial}{\partial y_1}f_1 &= (V_{x_0}^\top\frac{\partial}{\partial y_0}f_0)Q_{y_1} = B_{y_1}^{\mathrm{v}}Q_{y_1} = B_{y_2}.
\end{aligned}$$

We complete the base case by notating

$$\begin{aligned}
&f(d'(x,t),x,t) = \delta(t)\\
\Leftrightarrow &\begin{pmatrix} f_1(x_1',x_1,y_1,t,\Delta^{(1)}(t)) & = 0\\ \tilde{y}_1 - \Psi_{\tilde{y}_1}(x_1,t,\delta(t)) & = 0\\ \tilde{x}_1 - \Psi_{\tilde{x}_1}(t,\delta(t)) & = 0 \end{pmatrix}.
\end{aligned}$$

Induction step: $(i-1 \mapsto i \leqslant \mu - 1)$
By the induction hypothesis we get

$$f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = 0$$

$$\Leftrightarrow \begin{pmatrix} V_{x_{i-1}}^\top \\ V_{y_i}^\top W_{x_{i-1}}^\top \\ W_{y_i}^\top W_{x_{i-1}}^\top \end{pmatrix} f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = 0$$

with $\frac{\partial}{\partial \delta^{(i-1)}} W_{x_{i-1}}^\top f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = W_{x_{i-1}}^\top G_{i-1} P_{x_{i-1}} \frac{\partial}{\partial \delta^{(i-2)}} \Psi_{\tilde{y}_{i-1}}$ having full row rank and $\frac{\partial}{\partial x_{i-1}'} W_{x_{i-1}}^\top f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = W_{x_{i-1}}^\top G_{i-1} Q_{x_{i-1}}$ being zero due to the construction of $W_{x_{i-1}}^\top$. Hence we write

$$W_{x_{i-1}}^\top f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = W_{x_{i-1}}^\top f_{i-1}(x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t))$$

by Lemma (4.8) and obtain

$$\begin{aligned} V_{y_i}^\top W_{x_{i-1}}^\top f_{i-1}(x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) &= 0, \\ W_{y_i}^\top W_{x_{i-1}}^\top f_{i-1}(x_{i-1}, t, \Delta^{(i-1)}(t)) &= 0. \end{aligned}$$

Split $y_{i-1} = P_{y_i} \tilde{y}_i + Q_{y_i} y_i$ and $x_{i-1} = P_{x_i} \tilde{x}_i + Q_{x_i} x_i$ and obtain with the help of Lemma (4.8)

$$V_{y_i}^\top W_{x_{i-1}}^\top f_{i-1}(P_{x_i} \tilde{x}_i + Q_{x_i} x_i, P_{y_i} \tilde{y}_i, t, \Delta^{(i-1)}(t)) = 0, \tag{4.32a}$$

$$W_{y_i}^\top W_{x_{i-1}}^\top f_{i-1}(P_{x_i} \tilde{x}_i, t, \Delta^{(i-1)}(t)) = 0. \tag{4.32b}$$

Equation (4.32b) yields an explicit expression

$$\tilde{x}_i = \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t))$$

on a suitable open neighborhood $\mathcal{G}_{\tilde{x}_i}$ by Lemma 4.32 with $\Psi_{\tilde{x}_i}$ being $(\mu + 1 - i)$ times continuously differentiable by Lemma 4.7 and Assumption 4.34. Insert this expression into (4.32a) and obtain

$$V_{y_i}^\top W_{x_{i-1}}^\top f_{i-1}(P_{x_i} \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)) + Q_{x_i} x_i, P_{y_i} \tilde{y}_i, t, \Delta^{(i-1)}(t)) = 0,$$

which analogously yields an explicit expression

$$\tilde{y}_i = \Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)) \in C^{\mu - i}$$

on a suitable open neighborhood $\mathcal{G}_{\tilde{y}_i}$ by Lemma 4.32 with $\Psi_{\tilde{y}_i}$ being $(\mu + 1 - i)$ times continuously differentiable by Lemma 4.7 and Assumption 4.34. Together we achieve

$$\begin{aligned} & f_{i-1}(x_{i-1}', x_{i-1}, y_{i-1}, t, \Delta^{(i-1)}(t)) = 0 \\ \Leftrightarrow & \begin{pmatrix} f_i(x_i', x_i, y_i, t, \Delta^{(i)}(t)) & = 0 \\ \tilde{y}_i - \Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)) & = 0 \\ \tilde{x}_i - \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)) & = 0 \end{pmatrix} \end{aligned}$$

with

$$f_i(x_i', x_i, y_i, t, \Delta^{(i)}(t)) = V_{x_{i-1}}^\top f_{i-1}(\hat{x}_i', \hat{x}_i, \hat{y}_i, t, \Delta^{(i-1)}(t))$$

and

$$\begin{aligned}
\hat{x}_i' &= Q_{x_i} x_i' + Q_{x_i}' x_i + P_{x_i}' \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)) + P_{x_i} \Psi_{\tilde{x}_i}'(t, \Delta^{(i)}(t)),\\
\hat{x}_i &= Q_{x_i} x_i + P_{x_i} \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)),\\
\hat{y}_i &= Q_{y_i} y_i + P_{y_i} \Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)).
\end{aligned}$$

Such that for the Jacobians of f_i hold:

$$\begin{aligned}
\frac{\partial}{\partial x_i'} f_i &= V_{x_{i-1}}^\top \frac{\partial}{\partial x_{i-1}'} f_{i-1} Q_{x_i} = V_{x_{i-1}}^\top G_{i-1} Q_{x_{i-1}} Q_{x_i} = G_i Q_{x_i},\\
\frac{\partial}{\partial x_i} f_i &= V_{x_{i-1}}^\top ((\frac{\partial}{\partial x_{i-1}} f_{i-1}) Q_{x_i} + (\frac{\partial}{\partial x_{i-1}'} f_{i-1}) Q_{x_i}' + (\frac{\partial}{\partial y_{i-1}} f_{i-1}) P_{y_i} (\frac{\partial}{\partial x_i} \Psi_{\tilde{y}_i}))\\
&= (V_{x_{i-1}}^\top \frac{\partial}{\partial x_{i-1}} f_{i-1}) Q_{x_i} + (V_{x_{i-1}}^\top \frac{\partial}{\partial x_{i-1}'} f_{i-1}) Q_{x_i}' + (V_{x_{i-1}}^\top \frac{\partial}{\partial y_{i-1}} f_{i-1}) P_{y_i} (\frac{\partial}{\partial x_i} \Psi_{\tilde{y}_i})\\
&= B_{x_i}^{\mathrm{v}} Q_{x_i} + G_i Q_{x_i}' - B_{y_i}^{\mathrm{v}} P_{y_i} (V_{y_i}^\top B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^\top B_{x_i}^{\mathrm{w}} Q_{x_i} = B_{x_{i+1}},\\
\frac{\partial}{\partial y_1} f_i &= (V_{x_{i-1}}^\top \frac{\partial}{\partial y_{i-1}} f_{i-1}) Q_{y_i} = B_{y_i}^{\mathrm{v}} Q_{y_i} = B_{y_{i+1}}^{\mathrm{v}}.
\end{aligned}$$

The induction step is complete. Analogous to the former step, obtain by the usage of Lemma 4.32

$$\begin{aligned}
& f_{\mu-1}(x_{\mu-1}', x_{\mu-1}, y_{\mu-1}, t, \Delta^{(\mu-1)}(t))\\
\Leftrightarrow & \begin{pmatrix} V_{x_{\mu-1}}(x_{\mu-1}, t) f_{\mu-1}(x_{\mu-1}', x_{\mu-1}, y_{\mu-1}, t, \Delta^{(\mu-1)}(t)) & = 0 \\ W_{x_{\mu-1}}(x_{\mu-1}, t) f_{\mu-1}(x_{\mu-1}', x_{\mu-1}, y_{\mu-1}, t, \Delta^{(\mu-1)}(t)) & = 0 \end{pmatrix}\\
\Leftrightarrow & \begin{pmatrix} V_{x_{\mu-1}}(x_{\mu-1}, t) f_{\mu-1}(x_{\mu-1}', x_{\mu-1}, y_{\mu-1}, t, \Delta^{(\mu-1)}(t)) & = 0 \\ W_{x_{\mu-1}}(x_{\mu-1}, t) f_{\mu-1}(x_{\mu-1}, y_{\mu-1}, t, \Delta^{(\mu-1)}(t)) & = 0 \end{pmatrix}\\
\Leftrightarrow & \begin{pmatrix} x_\mu' & = f_\mu(x_\mu, t, \Delta^{(\mu-1)}(t)) \\ \tilde{y}_\mu & = \Psi_{\tilde{y}_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t)) \end{pmatrix}
\end{aligned}$$

with $\frac{\partial}{\partial \delta^{(\mu-1)}} \Psi_{\tilde{y}_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t))$ having full row rank. □

In Theorem 4.35 we obtain an inherent ODE after the decoupling. To achieve perturbation results for the DAE (4.9) we need to analyze the explicit ODE case. The next lemma covers the perturbation analysis for explicit ODEs. In the lemma we consider nonlinear perturbation to handle the nonlinear perturbations which may appear in the inherent ODE even if only the right hand side of the DAE gets perturbed.

Lemma 4.36. (Perturbed ODEs)
Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{G} \subset \mathbb{R}^n \times \mathcal{I} \times \mathbb{R}^m$ be open subsets and let $f \in C^1(\mathcal{G}, \mathbb{R}^n)$. Let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ and let $x_\star(t)$ be a continuous differentiable solution on $\mathcal{I}_\star$ of the unperturbed IVP

$$\begin{aligned} x'(t) &= f(x, t, 0), \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0. \end{aligned} \tag{4.33}$$

with $(x_\star(t), t, 0) \in \mathcal{G}$ for all $t \in \mathcal{I}_\star$. Let $\delta \in C(\mathcal{I}_\star, \mathbb{R}^n)$ be a perturbation such that $(x_\star(t), t, \delta(t)) \in \mathcal{G}$ for all $t \in \mathcal{I}_\star$. Let δ_0 be the initial perturbation then the perturbed IVP

$$\begin{aligned} x'(t) &= f(x, t, \delta(t)), \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0 + \delta_0 \end{aligned} \tag{4.34}$$

has a unique solution $x_\delta \in C^1(\mathcal{I}_\star, \mathbb{R}^n)$ and it holds

$$||x_\star(t) - x_\delta(t)||_\infty \leqslant c(||\delta_0||_\infty + \int_{t_0}^{t} ||\delta(s)||_\infty \mathrm{d}s), \tag{4.35}$$

if δ_0 and $\delta(t)$ are sufficiently small.

Proof.
We find an $r_x > 0$ with

$$\bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0) := \{(x, t, \delta) \in \mathbb{R}^n \times \mathcal{I} \times \mathbb{R}^m \mid ||x - x_\star(t)|| + ||\delta(t)|| \leqslant r_x\} \subset \mathcal{G}.$$

Let $(x_0 + \delta_0, t_0, 0) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$, i.e. $||\delta_0|| \leqslant r_x$. Define the restricted function

$$\tilde{f}(x, t, \delta) := f(x, t, \delta)_{|\bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)}$$

thus the norms of the Jacobians of $\tilde{f}$, with respect to x and δ, are bounded since $\bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$ is compact and $\tilde{f}$ is continuously differentiable.

$$\begin{aligned} ||\tilde{f}_x||_\infty &\leqslant \max_{(x,t,\delta) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)} ||f_x(x, t, \delta)||_\infty =: c_x \\ ||\tilde{f}_\delta||_\infty &\leqslant \max_{(x,t,\delta) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)} ||f_\delta(x, t, \delta)||_\infty =: c_\delta \end{aligned}$$

Notice that $x_\star$ is also a solution of

$$\begin{aligned} x'(t) &= \tilde{f}(x, t, 0), \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0 \end{aligned}$$

since $x_\star$ solves (4.33) and $(x_\star(t), t, 0) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$ for all $t \in \mathcal{I}_\star$. Assume that x_δ is a solution of the perturbed restricted IVP

$$\begin{aligned} x'(t) &= \tilde{f}(x, t, \delta(t)), \quad \forall t \in \mathcal{I}_\star \\ x(t_0) &= x_0 + \delta_0. \end{aligned} \tag{4.36}$$

We can subtract these terms from each other and achieve

$$\begin{aligned} (x_\delta(t) - x_\star(t))' =& \tilde{f}(x_\delta(t), t, \delta(t)) - \tilde{f}(x_\star(t), t, 0) \\ =& \tilde{f}(x_\delta(t), t, \delta(t)) - \tilde{f}(x_\star(t), t, \delta(t)) + \tilde{f}(x_\star(t), t, \delta(t)) - \tilde{f}(x_\star(t), t, 0) \\ =& \int_0^1 \tilde{f}_x(sx_\delta(t) + (1-s)x_\star(t), t, \delta(t))\mathrm{d}s(x_\delta(t) - x_\star(t)) \\ & + \int_0^1 \tilde{f}_\delta(x_\star(t), t, s\delta(t))\mathrm{d}s\delta(t). \end{aligned}$$

Notice that $(sx_\delta(t) + (1-s)x_\star(t), t, \delta(t)) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$ for all $(x_\delta(t), t, \delta(t)), (x_\star(t), t, \delta(t)) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$. By an integration with respect to the time we obtain

$$\begin{aligned} x_\delta(t) - x_\star(t) = \delta_0 &+ \int_{t_0}^t \int_0^1 \tilde{f}_x(sx_\delta(\tau) + (1-s)x_\star(\tau), \tau, \delta(\tau))\mathrm{d}s(x_\delta(\tau) - x_\star(\tau))\mathrm{d}\tau \\ &+ \int_{t_0}^t \int_0^1 \tilde{f}_\delta(x_\star(\tau), \tau, s\delta(\tau))\mathrm{d}s\delta(\tau)\mathrm{d}\tau. \end{aligned}$$

Finally we are able to bound the norm of $x_\delta(t) - x_\star(t)$ with the help of Gronwall's Lemma:

$$\begin{aligned} & ||x_\delta(t) - x_\star(t)||_\infty \\ \leqslant & ||\delta_0||_\infty + \int_{t_0}^t \int_0^1 ||\tilde{f}_\delta(x_\star(\tau), \tau, s\delta(\tau))||_\infty ds ||\delta(\tau)||_\infty d\tau \\ & + \int_{t_0}^t \int_0^1 ||\tilde{f}_x(sx_\delta(\tau) + (1-s)x_\star(\tau), \tau, \delta(\tau))||_\infty ds ||x_\delta(\tau) - x_\star(\tau)||_\infty d\tau \\ \leqslant & ||\delta_0||_\infty + c_\delta \int_{t_0}^t ||\delta(\tau)||_\infty d\tau + c_x \int_{t_0}^t ||x_\delta(\tau) - x_\star(\tau)||_\infty d\tau \\ \leqslant & (||\delta_0||_\infty + c_\delta \int_{t_0}^t ||\delta(\tau)||_\infty d\tau) e^{c_x(T-t_0)}. \end{aligned}$$

Let $||\delta_0||_\infty < \frac{1}{4}e^{-c_x(T-t_0)}r_x$ and $\max_{\tau \in \mathcal{I}_\star} ||\delta(\tau)||_\infty < \max(\frac{1}{4c_\delta(T-t_0)}e^{-c_x(T-t_0)}r_x, \frac{1}{4}r_x)$ be fulfilled, then $||x_\delta(t) - x_\star(t)||_\infty \leqslant \frac{1}{2}r_x$ and therefore $(x_\delta(t), t, 0) \in \bar{B}_{r_x}(x_\star, \mathcal{I}_\star, 0)$ for all $t \in \mathcal{I}_\star$. So solutions of (4.36) are bounded in $\bar{B}_{\frac{3}{4}r_x}(x_\star, \mathcal{I}_\star, 0)$ by an a-priori estimate and are therefore also solutions of (4.34). Now we find a solution interval which includes $\mathcal{I}_\star$ such that

(4.34) has a unique continuously differentiable solution on $\mathcal{I}_\star$, since $f(x,t,\delta)$ is continuous differentiable hence it is local Lipschitz continuous and the solutions of (4.34) are bounded. The inequality (4.35) holds with $c := \max(1, c_\delta)e^{c_x(T-t_0)}$. □

According to our objective to show the equivalence of the Perturbation Index and the Dissection Index it is not sufficient to show that there exists an estimation like (4.26). It also has to be shown that there is no stricter estimation, i.e. the Dissection Index μ_{TS} has to be the minimal μ such that an estimation (4.26) exists.

Lemma 4.37. Let $\mathcal{I} \subset \mathbb{R}$ and $\mathcal{G} \subset \mathbb{R}^n \times \mathcal{I} \times \mathbb{R}^{\mu \cdot n}$ be open subsets and let $g \in C^\mu(\mathcal{G}, \mathbb{R}^n)$. Let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ and let $x_\star(t)$ be a continuous solution on $\mathcal{I}_\star$ of the unperturbed algebraic equation

$$g(x, t, 0) = 0 \tag{4.37}$$

with $(x_\star(t), t, 0) \in \mathcal{G}$ for all $t \in \mathcal{I}_\star$.
Let δ be a $(\mu - 1)$-times differentiable perturbation and define $\Delta^{(\mu-1)}(t) := \big(\delta(t), \ldots, \delta^{(\mu-1)}(t)\big)$ with $(x_\star(t), t, \Delta^{(\mu-1)}(t)) \in \mathcal{G}$ for all $t \in \mathcal{I}_\star$. Let x_δ be a continuous solution on $\mathcal{I}_\star$ of the perturbed algebraic equation

$$g(x, t, \Delta^{(\mu-1)}(t)) = 0 \tag{4.38}$$

for all $t \in \mathcal{I}_\star$. Furthermore let the Jacobian $\frac{\partial}{\partial \delta^{(\mu)}} g(x, t, \Delta^{(\mu-1)}(t))$ has full row rank along $(x_\star(t), t, 0)$ for all $t \in \mathcal{I}_\star$.
Then there is no $\nu \in \mathbb{N}$ with $\nu < \mu$ such that

$$||x_\star(t) - x_\delta(t)|| \leqslant c(||x_\star(t_0) - x_\delta(t_0)|| + \int_{t_0}^{t} ||\delta(s)||\mathrm{d}s + \sum_{i=0}^{\nu-1} \max_{0 \leqslant \tilde{t} \leqslant t} ||\delta^{(i)}(\tilde{t})||), \tag{4.39}$$

for all perturbations δ with $||\delta||_\infty, ..., ||\delta^{(\mu-1)}||_\infty$ and $||x_\star(t_0) - x_\delta(t_0)||$ sufficiently small.

Proof.
We define $t_{\frac{1}{4}} := \frac{1}{4}(T - t_0) + t_0$ and $t_{\frac{3}{4}} := \frac{3}{4}(T - t_0) + t_0$ and let

$$h(t) = 4\frac{t - t_0}{T - t_0} - 2 \quad \Rightarrow \quad \begin{cases} h(t) \geqslant -1, & t \leqslant t_{\frac{1}{4}} \\ h(t) \leqslant 1, & t \geqslant t_{\frac{3}{4}}. \end{cases}$$

Next define

$$g(t) = \begin{cases} e^{-\frac{1}{t^2}}, & t > 0 \\ 0, & t \leqslant 0 \end{cases}$$

With the help of g we define

$$k(t) = \frac{g(1+t)}{g(1+t)+g(1-t)} \quad \Rightarrow \quad \begin{cases} k(t) = 0, & t \leqslant -1 \\ 0 \leqslant k(t) \leqslant 1, & -1 \leqslant t \leqslant 1 \\ k(t) = 1, & t \geqslant 1 \end{cases}$$

and define afterwards

$$a(t) = k(h(t)) \quad \Rightarrow \quad \begin{cases} a(t) = 0, & t \leqslant t_{\frac{1}{4}} \\ 0 \leqslant a(t) \leqslant 1, & t_{\frac{1}{4}} < t < t_{\frac{3}{4}} \\ a(t) = 1, & t \geqslant t_{\frac{3}{4}} \end{cases}$$

with $a \in C^\infty(\mathbb{R}, \mathbb{R})$ since $a(t)$ is a combination of C^∞ functions. With $\mathcal{I}_\star$ being compact we get that there is a constant $c_a > 0$ with

$$\max_{t \in \mathcal{I}_\star} |a^{(i)}(t)| \leqslant c_a, \quad \forall i \leqslant \mu - 1.$$

We define

$$\begin{aligned} \bar{\delta} : \mathcal{I}_\star &\to \mathbb{R} \\ t &\mapsto \eta \varepsilon^{\mu-1} \sin(\varepsilon^{-1} t) a(t) \end{aligned}$$

with $\eta, \varepsilon > 0$ and show that

1. $\bar{\delta}^{(i)}(t_0) = 0$ for all $i \leqslant \mu - 1$,

2. $\exists c_\eta > 0 : \max_{t \in \mathcal{I}_\star} |\bar{\delta}^{(i)}(t)| \leqslant c_\eta \eta$ for all $i \leqslant \mu - 1$,

3. $\exists c_\varepsilon > 0 : \max_{t \in \mathcal{I}_\star} |\bar{\delta}^{(i)}(t)| \leqslant c_\varepsilon \varepsilon$ for all $i \leqslant \mu - 2$,

4. $\forall n \in \mathbb{N} : \varepsilon = \frac{T}{2n\pi} \Rightarrow |\bar{\delta}^{(\mu-1)}(T)| = \eta$ if μ is odd.
   $\forall n \in \mathbb{N} : \varepsilon = \frac{T}{\frac{1}{2}\pi + 2n\pi} \Rightarrow |\bar{\delta}^{(\mu-1)}(T)| = \eta$ if μ is even.

First notice that $\bar{\delta} \in C^\infty(\mathcal{I}_\star, \mathbb{R})$ as a combination of C^∞ functions and that

$$\bar{\delta}^{(n)}(t) = \eta \varepsilon^{\mu-1-n} (\sum_{k=0}^{n} \binom{n}{k} \varepsilon^k a^{(k)}(t) sin^{(n-k)}(\varepsilon^{-1} t))$$

holds by a mathematical induction with $\sin^{(n)}$ the n-th derivative of the sinus function.

1. For $t \leqslant t_{\frac{1}{4}}$ holds
$$\bar{\delta}(t) = \varepsilon^{\mu-1} \sin(\varepsilon^{-1}t)a(t) = 0$$
since $a(t) = 0$ for all $t \leqslant t_{\frac{1}{4}}$ hence there is a neighborhood around t_0 in which $\bar{\delta}(t)$ vanishes completely and therefore holds $\bar{\delta}^{(i)}(t_0) = 0$ for all $i \leqslant \mu - 1$.
2. $\mathcal{I}_\star$ is a compact interval, $\bar{\delta} \in C^\infty(\mathcal{I}_\star, \mathbb{R})$ and η is a constant factor.
3. $\mathcal{I}_\star$ is a compact interval, $\bar{\delta} \in C^\infty(\mathcal{I}_\star, \mathbb{R})$ and ε is a constant factor in $\bar{\delta}^{(i)}(t)$ for $i < \mu - 1$ and $|sin^{(i)}(\varepsilon^{-1}t)| \leqslant 1$.
4. At last it holds:
$$\begin{aligned}
|\bar{\delta}^{(\mu-1)}(T)| &= \eta|(\sum_{k=0}^{\mu-1} \binom{\mu-1}{k} \varepsilon^k a^{(k)}(T) sin^{(\mu-1-k)}(\varepsilon^{-1}T))| \\
&= \eta|(a(T)sin^{(\mu-1)}(\varepsilon^{-1}T) + \sum_{k=1}^{\mu-1} \binom{\mu-1}{k} \varepsilon^k a^{(k)}(T) sin^{(\mu-1-k)}(\varepsilon^{-1}T))| \\
&= \eta|sin^{(\mu-1)}(\varepsilon^{-1}T)| \\
&= \eta \begin{cases} |sin^{(\mu-1)}(2n\pi)|, & \text{if } \mu-1 \text{ is odd} \\ |sin^{(\mu-1)}(\frac{1}{2}\pi + 2n\pi)|, & \text{if } \mu-1 \text{ is even} \end{cases} \\
&= \eta
\end{aligned}$$

Let $P_\delta(x, t, \Delta(t))$ be a complementary kernel basis function of $\frac{\partial}{\partial \delta^{(\mu)}} g(x, t, \Delta(t))$. Then we define $\delta_\star(t) = P_\delta(x_\star(T), T, 0) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \bar{\delta}(t)$, this yields $x_\star(t_0) = x_\delta(t_0)$ and with Equation (4.38) and the Implicit Function Theorem that there is a neighborhood around $(x_\star(T), T, 0)$ such that there is a function Ψ with
$$\begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \bar{\delta}^{(\mu-1)}(t) = \Psi(x_\delta(t), t, \Delta_\star^{(\mu-2)}(t))$$
and with Equation (4.37) we obtain
$$\Psi(x_\star(t), t, 0) = 0.$$
If we assume that there is a $\nu < \mu$ such that there is an estimate (4.39) then it holds
$$|\bar{\delta}^{(\mu-1)}(t)|$$

$$
\begin{aligned}
&=||\Psi(x_\delta(t), t, \Delta_\star^{(\mu-2)}(t)) - \Psi(x_\star(t), t, 0)||_\infty \\
&=||\Psi(x_\delta(t), t, \Delta_\star^{(\mu-2)}(t)) - \Psi(x_\star(t), t, \Delta_\star^{(\mu-1)}(t)) \\
&\quad + \Psi(x_\star(t), t, \Delta_\star^{(\mu-2)}(t)) - \Psi(x_\star(t), t, 0)||_\infty \\
&=||\int_0^1 \Psi_x(sx_\delta(t) + (1-s)x_\star(t), t, \Delta_\star^{(\mu-2)}(t))\mathrm{d}s(x_\delta(t) - x_\star(t)) \\
&\quad + \int_0^1 \Psi_{\Delta_\star^{(\mu-2)}}(x_\star(t), t, s\Delta_\star^{(\mu-2)}(t))\mathrm{d}s\Delta_\star^{(\mu-2)}(t)||_\infty.
\end{aligned}
$$

and we find constants $c_x, c_\delta, c > 0$ such that

$$
\begin{aligned}
|\bar{\delta}^{(\mu-1)}(t)| &\leqslant c_x||x_\delta(t) - x_\star(t)||_\infty + c_\delta||\Delta^{(\mu-2)}(t)||_\infty && \text{(4.40a)} \\
&\leqslant c(||x_\star(t_0) - x_\delta(t_0)||_\infty + ||\Delta^{(\mu-2)}(t)||_\infty) && \text{(4.40b)} \\
&= c||\Delta^{(\mu-2)}(t)||_\infty && \text{(4.40c)} \\
&\leqslant cc_\varepsilon\varepsilon. && \text{(4.40d)}
\end{aligned}
$$

But this would yield

$$
\eta = |\bar{\delta}^{(\mu-1)}(T)| \leqslant cc_\varepsilon \frac{T}{2n\pi}
$$

for all $n \in \mathbb{N}$ and therefore it would hold that $0 < \eta = 0$. By this contradiction there can not be a $\nu < \mu$ such that there is an estimate (4.39). □

We present the main theorem of this section which will be proven with the help of Theorem 4.35, Lemma 4.36 and Lemma 4.37.

Theorem 4.38. (Relation between the Dissection Index and the Perturbation Index) Consider a DAE (4.9), let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ be compact and connected and let $x_0 \in \mathcal{G}$ be the initial value of the IVP

$$
\begin{aligned}
f(d'(x,t), x, t) &= 0, \quad \forall t \in \mathcal{I}_\star \\
x(t_0) &= x_0.
\end{aligned}
$$

(i) Let the DAE have a finite Dissection Index μ.

(ii) Let the DAE fulfill the Assumptions 4.33 and 4.34.

(iii) Let the IVP have a global solution $x_\star$ on $\mathcal{I}_\star$.

Then for all perturbations δ, with $||\delta||_\infty, ..., ||\delta^{(\mu-1)}||_\infty$ being sufficiently small, it holds that:

(i) Each perturbed system has a unique global solution.

(ii) The Perturbation Index exists.

(iii) The Perturbation Index μ_P is equal to the Dissection Index μ.

Proof.
We remember the splitting

$$x = Q_{x_\mu} x_\mu + \sum_{i=1}^{\mu-1} Q_{x_{i-1}} P_{x_i} \tilde{x}_i + \sum_{i=1}^{\mu} Q_{y_{i-1}} P_{y_i} \tilde{y}_i.$$

By the Decoupling Theorem 4.35 there is a neighborhood around the exact solution such that we obtain:

$$\begin{aligned} & f(d'(x,t),x,t) = \delta(t) \\ \Leftrightarrow & \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} x_\mu & = & f_{x_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t)) \\ \tilde{y}_\mu & = & \Psi_{\tilde{y}_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t)) \\ \tilde{y}_i & = & \Psi_{\tilde{y}_i}(x_i, t, \Delta^{(i-1)}(t)) \\ \tilde{x}_i & = & \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t)) \end{pmatrix} \end{aligned}$$

with $\frac{\partial}{\partial \delta^{(\mu-1)}} \Psi_{\tilde{x}_\mu}(t, \Delta^{(\mu-1)}(t))$ having full row rank and f_{x_μ}, $\Psi_{\tilde{y}_i}$, $\Psi_{\tilde{x}_i}$ and $\Psi_{\tilde{y}_\mu}$ being continuously differentiable for $1 \leqslant i \leqslant \mu - 1$.
By Lemma 4.36 there is a unique global solution $x_{\delta,\mu}$ of the perturbed inherent ODE with an estimate:

$$\begin{aligned} ||x_{\star,\mu}(t) - x_{\delta,\mu}(t)||_\infty &\leqslant c(||\delta_0||_\infty + \int_{t_0}^{t} ||\Delta^{(\mu-1)}(t)||_\infty \mathrm{d}s) \\ &\leqslant c(||\delta_0||_\infty + \sum_{i=0}^{\mu-1} \max_{\tilde{t} \in [t_0,t]} ||\delta^{(i)}(\tilde{t})||_\infty \mathrm{d}s) \end{aligned}$$

The unique global solution of the perturbed solution components $\tilde{x}_i$ follows immediately from the decoupling since we can express $\tilde{x}_i$ by functions that only depend on the time and the derivatives of the perturbation. Additionally, by the Mean Value Theorem and $\Delta^{(i-1)}(t)$ being sufficiently small there is a constant c such that:

$$\begin{aligned} ||\tilde{x}_{\star,i}(t) - \tilde{x}_{\delta,i}(t)||_\infty &= ||\Psi_{\tilde{x}_i}(t,0) - \Psi_{\tilde{x}_i}(t, \Delta^{(i-1)}(t))||_\infty \\ &\leqslant c||\Delta^{(i-1)}(t)||_\infty \\ &\leqslant c \sum_{j=0}^{i-1} \max_{\tilde{t} \in [t_0,t]} ||\delta^{(j)}(\tilde{t})||_\infty \mathrm{d}s \end{aligned}$$

The unique existence of $x_{\delta,\mu}$ and $\tilde{x}_{\delta,i}$ for all $1 \leqslant i \leqslant \mu - 1$ yield the unique existence of $\tilde{y}_{\delta,i}$ for all $1 \leqslant i \leqslant \mu$ plus the estimate

$$||\tilde{y}_{\star,i}(t) - \tilde{y}_{\delta,i}(t)||_\infty$$

$$
\begin{aligned}
&= ||\Psi_{\tilde{y}_i}(x_{\star,i}(t),t,0) - \Psi_{\tilde{y}_i}(x_{\delta,i}(t),t,\Delta^{(i-1)}(t))||_\infty \\
&\leqslant c(||x_{\star,\mu}(t) - x_{\delta,\mu}(t)||_\infty + \sum_{j=i}^{\mu-1} ||\tilde{x}_{\star,j}(t) - \tilde{x}_{\delta,j}(t)||_\infty + ||\Delta^{(i-1)}(t)||_\infty) \\
&\leqslant \tilde{c}(||\delta_0||_\infty + \sum_{i=0}^{\mu-1} \max_{\tilde{t}\in[t_0,t]} ||\delta^{(i)}(\tilde{t})||_\infty \mathrm{d}s))
\end{aligned}
$$

We showed the unique global solvability for the perturbed problem. By the estimations we already showed the existence of Perturbation Index and additionally we showed that the Dissection Index is an upper bound of the Perturbation Index. By Lemma 4.37 the equation

$$\tilde{y}_\mu = \Psi_{\tilde{y}_\mu}(x_\mu, t, \Delta^{(\mu-1)}(t))$$

yields that there is no $\nu < \mu$ such that there is a constant c with:

$$||\tilde{y}_{\star,\mu}(t) - \tilde{y}_{\delta,\mu}(t)||_\infty \leqslant c(||\delta_0||_\infty + \sum_{i=0}^{\nu-1} \max_{\tilde{t}\in[t_0,t]} ||\delta^{(i)}(\tilde{t})||_\infty \mathrm{d}s))$$

This shows that the Dissection Index is also a lower bound of the Perturbation Index hence they coincide. □

4.5 Dissection Index for DAEs in Hessenberg Form

To analyze the mechanical applications of Section 3.2 we introduce DAEs in Hessenberg form:

Definition 4.39. (Differential-Algebraic Equation in Hessenberg form; [KM06], p. 172) Let $\mathcal{I} \subset \mathbb{R}$, $\mathcal{D}^{n_{x_i}} \subset \mathbb{R}^{n_{x_i}}$ be open subsets. Consider the following equation

$$
\begin{aligned}
x_1' &= f_1(x_1, x_2, t) \\
&\vdots \\
x_{\mu-1}' &= f_{\mu-1}(x_1, \ldots, x_{\mu-1}, x_\mu, t) \\
0 &= f_0(x_1, t)
\end{aligned}
\tag{4.41}
$$

with $f_i \in C^1(\mathcal{D}^{n_{x_1}} \times \ldots \times \mathcal{D}^{n_{x_{i+1}}} \times \mathcal{I}, \mathcal{D}^{n_{x_i}})$. We call (4.41) a DAE in Hessenberg form with μ stages, if

$$\frac{\partial}{\partial x_1} f_0 \frac{\partial}{\partial x_2} f_1 \cdots \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} \frac{\partial}{\partial x_\mu} f_{\mu-1} \tag{4.42}$$

is non-singular.

In [LMT13] it is shown that the Tractability Index always coincides with the number of the Hessenberg stages. In the following we provide this result also for the Dissection Index. Before we formulate such a theorem we split (4.42) with the help of orthonormal basis functions. Equation (4.42) yields that

$$\frac{\partial}{\partial x_1} f_0$$

has full row rank. Furthermore choose orthonormal basis functions $P_{H_1}(x_1, t)$ of $(\ker \frac{\partial}{\partial x_1} f_0)^\perp$ and Q_{H_1} of $\ker \frac{\partial}{\partial x_1} f_0$. Then it holds

$$\begin{aligned}
&\frac{\partial}{\partial x_1} f_0 \frac{\partial}{\partial x_2} f_1 \dots \frac{\partial}{\partial x_\mu} f_{\mu-1} \frac{\partial}{\partial y} f_\mu \\
=&\frac{\partial}{\partial x_1} f_0 P_{H_1} P_{H_1}^\top \frac{\partial}{\partial x_2} f_1 \dots \frac{\partial}{\partial x_\mu} f_{\mu-1} \frac{\partial}{\partial y} f_\mu \\
=&(\frac{\partial}{\partial x_1} f_0 P_{H_1})(P_{H_1}^\top \frac{\partial}{\partial x_2} f_1 \dots \frac{\partial}{\partial x_\mu} f_{\mu-1} \frac{\partial}{\partial y} f_\mu)
\end{aligned}$$

with

$$\frac{\partial}{\partial x_1} f_0 P_{H_1} \quad \text{and} \quad P_{H_1}^\top \frac{\partial}{\partial x_2} f_1 \dots \frac{\partial}{\partial x_\mu} f_{\mu-1} \frac{\partial}{\partial y} f_\mu$$

being non-singular. We successively define $P_{H_{i+1}}(x_i, \dots, x_1, t)$ as orthonormal basis functions of $(\ker P_{H_i}^\top \frac{\partial}{\partial x_{i+1}} f_i)^\perp$ and $Q_{H_{i+1}}(x_i, \dots, x_1, t)$ as orthonormal basis functions of $\ker P_{H_i}^\top \frac{\partial}{\partial x_{i+1}} f_i$ for $1 \leqslant i \leqslant \mu - 1$. By an induction

$$P_{H_i}^\top \frac{\partial}{\partial x_{i+1}} f_i$$

has full row rank and

$$P_{H_{i+1}}^\top \frac{\partial}{\partial x_{i+2}} f_{i+1} \dots \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} \frac{\partial}{\partial x_\mu} f_{\mu-1}$$

is non-singular. This particularly yields in the last step that

$$P_{H_\mu}^\top \frac{\partial}{\partial x_\mu} f_{\mu-1}$$

is non-singular.

Theorem 4.40.
Any DAE in Hessenberg form (4.41) with μ stages and sufficiently smooth functions f_i has Dissection Index μ.

Proof. We prove

$$G_i = \begin{pmatrix} I & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I \end{pmatrix},$$

$$B^{\mathrm{v}}_{x_i} = \begin{pmatrix} * & \dots & * & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & 0 & & \vdots \\ \vdots & & \vdots & \frac{\partial}{\partial x_{i+1}} f_i & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ \vdots & & \vdots & * & & \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} \\ * & \dots & * & \dots & \dots & * \end{pmatrix}, \quad B^{\mathrm{v}}_{y_i} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial}{\partial y} f_{\mu-1} \end{pmatrix}$$

and

$$B^{\mathrm{w}}_{x_i} = \begin{pmatrix} b^1_{x_i} & \dots & b^{i-1}_{x_i} & P^\top_{H_{i-1}} \frac{\partial}{\partial x_i} f_{i-1} & 0 & \dots & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{y_i} = \begin{pmatrix} 0 \end{pmatrix}$$

for $i < \mu$ by an induction.
Base case: $(i = 1)$
We gain

$$AD = \begin{pmatrix} I & & & \\ & \ddots & & \\ & & I & \\ & & & 0 \end{pmatrix}, \quad B = \begin{pmatrix} * & \frac{\partial}{\partial x_2} f_1 & 0 & \dots & 0 \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} & 0 \\ * & \dots & \dots & * & \frac{\partial}{\partial y} f_{\mu-1} \\ \frac{\partial}{\partial x_1} f_0 & 0 & \dots & 0 & 0 \end{pmatrix}$$

with

$$A := \begin{pmatrix} I & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix}, \quad D := \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & I & 0 \end{pmatrix}.$$

This allows us to choose

$$P = V = \begin{pmatrix} I & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ I \end{pmatrix}.$$

These basis functions yield

$$G_1 = V^\top ADP = \begin{pmatrix} I & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I \end{pmatrix}$$

and

$$B^{\mathrm{v}}_{x_1} = V^\top BP = \begin{pmatrix} * & \frac{\partial}{\partial x_2} f_1 & 0 & \dots \\ \vdots & & \ddots & \\ \vdots & & & \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} \\ * & \dots & \dots & * \end{pmatrix}, \quad B^{\mathrm{v}}_{y_1} = V^\top BQ = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial}{\partial y} f_{\mu-1} \end{pmatrix}$$

$$B^{\mathrm{w}}_{x_1} = W^\top BP = \begin{pmatrix} \frac{\partial}{\partial x_1} f_0 & 0 & \dots & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{y_1} = W^\top BQ = \begin{pmatrix} 0 \end{pmatrix}.$$

Induction step: $(1 \leqslant i \mapsto i+1 \leqslant \mu - 1)$
We get

$$B^{\mathrm{v}}_{x_i} = \begin{pmatrix} * & \dots & * & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ \vdots & & \vdots & 0 & & \vdots \\ \vdots & & \vdots & \frac{\partial}{\partial x_{i+1}} f_i & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ \vdots & & \vdots & * & & \frac{\partial}{\partial x_{\mu-1}} f_{\mu-2} \\ * & \dots & * & \dots & \dots & * \end{pmatrix}, \quad B^{\mathrm{v}}_{y_i} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial}{\partial y} f_{\mu-1} \end{pmatrix},$$

$$B^{\mathrm{w}}_{x_i} = \begin{pmatrix} b^1_{x_i} & \dots & b^{i-1}_{x_i} & P^\top_{H_{i-1}} \frac{\partial}{\partial x_i} f_{i-1} & 0 & \dots & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{y_i} = \begin{pmatrix} 0 \end{pmatrix}$$

by the induction statement. $B^{\mathrm{w}}_{y_i} = \begin{pmatrix} 0 \end{pmatrix}$ yields

$$Q_{y_i} = W_{y_i} = \begin{pmatrix} I \end{pmatrix}, \quad P_{y_i} = V_{y_i} = \begin{pmatrix} \end{pmatrix}.$$

Hence we get $W^\top_{y_i} B^{\mathrm{w}}_{x_i} = B^{\mathrm{w}}_{x_i}$, which allows us to choose

$$Q_{x_i} = \begin{pmatrix} I & & & & & & \\ & \ddots & & & & & \\ & & I & & & & \\ -(P^\top_{H_{i-1}} \frac{\partial}{\partial x_i} f_{i-1})^+ b^1_{x_i} & \dots & -(P^\top_{H_{i-1}} \frac{\partial}{\partial x_i} f_{i-1})^+ b^{i-1}_{x_i} & Q_{H_i} & & & \\ & & & & I & & \\ & & & & & \ddots & \\ & & & & & & I \end{pmatrix}$$

and $P_{x_i} = \begin{pmatrix} 0 & \vdots & 0 & P_{H_i}^\top & 0 & \vdots & 0 \end{pmatrix}^\top$. We obtain $G_i Q_{x_i} = Q_{x_i}$ and therefore we can choose

$$V_{x_i} = \begin{pmatrix} I &&&&&& \\ & \ddots &&&&& \\ && I &&&& \\ &&& Q_{H_i} &&& \\ &&&& I && \\ &&&&& \ddots & \\ &&&&&& I \end{pmatrix}, \quad W_{x_i} = \begin{pmatrix} (P_{H_{i-1}}^\top \frac{\partial}{\partial x_i} f_{i-1} P_{H_i}^\top)^{-1} b_{x_i}^1 \\ \vdots \\ (P_{H_{i-1}}^\top \frac{\partial}{\partial x_i} f_{i-1} P_{H_i}^\top)^{-1} b_{x_i}^{i-1} \\ P_{H_i} \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

which leads to

$$G_{i+1} = V_{x_i}^\top G_i Q_{x_i} = V_{x_i}^\top Q_{x_i} = \begin{pmatrix} I & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & I \end{pmatrix}.$$

We close the induction step by

$$\begin{array}{rclrcl} B_{x_{i+1}}^{\mathrm{v}} & = & V_{x_i}^\top B_{x_i}^{\mathrm{v}} Q_{x_i} + Q_{x_i}' & B_{y_{i+1}}^{\mathrm{v}} & = & V_{x_i}^\top B_{y_i}^{\mathrm{v}} Q_{y_i} \\ B_{x_{i+1}}^{\mathrm{w}} & = & W_{x_i}^\top B_{x_i}^{\mathrm{v}} Q_{x_i} + Q_{x_i}' & B_{y_{i+1}}^{\mathrm{w}} & = & W_{x_i}^\top B_{y_i}^{\mathrm{v}} Q_{y_i}. \end{array}$$

We can choose $Q_{y_{\mu-1}} = I$ and

$$W_{x_{\mu-1}} = \begin{pmatrix} (P_{H_{i-1}}^\top \frac{\partial}{\partial x_i} f_{i-1} P_{H_i}^\top)^{-1} b_{x_i}^1 \\ \vdots \\ (P_{H_{i-1}}^\top \frac{\partial}{\partial x_i} f_{i-1} P_{H_i}^\top)^{-1} b_{x_i}^{i-1} \\ P_{H_\mu} \end{pmatrix}$$

which leads to $B_{y_i}^{\mathrm{w}} = \begin{pmatrix} P_{H_\mu}^\top \frac{\partial}{\partial y} f_\mu \end{pmatrix}$. Hence the DAE has Dissection Index μ. □

The multibody systems of model level 1 from Section 3.2

$$p' = Z(p)v, \tag{4.43a}$$

$$M(p,t)v' = f(p,v,r,\lambda,t) - Z^\top(p)G^\top(p,t)\lambda, \tag{4.43b}$$

$$r' = b(p,v,r,\lambda,t), \tag{4.43c}$$

$$0 = g(p,t) \tag{4.43d}$$

are in Hessenberg form, under Assumption 3.11, if we multiply Equation (4.43b) by $M^{-1}(p,t)$. In the following we calculate the Dissection Index for (3.36). Therefore we define $x = \begin{pmatrix} p & v & r & \lambda \end{pmatrix}$ and write

$$B(x,t) := \begin{pmatrix} * & -Z(p) & 0 & 0 \\ * & * & * & M^{-1}(p,t)G_\lambda(p,v,r,\lambda,t) \\ * & * & * & 0 \\ G(p,t) & 0 & 0 & 0 \end{pmatrix}$$

with $G_\lambda(p,v,r,\lambda,t) := Z(p)^\top G(p,t)^\top - \frac{\partial}{\partial\lambda} f(p,v,r,\lambda,t)$. Furthermore we obtain

$$AD := \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

with

$$A := \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \\ 0 & 0 & 0 \end{pmatrix}, \quad D := \begin{pmatrix} I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \end{pmatrix},$$

which allows us to choose

$$P = V = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}.$$

These basis functions yield

$$G_1 = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \quad B^{\mathrm{v}}_{x_1} = \begin{pmatrix} * & -Z(p) & 0 \\ * & * & * \\ * & * & * \end{pmatrix},$$

$$B^{\mathrm{v}}_{y_1} = \begin{pmatrix} 0 \\ M^{-1}(p,t)G_\lambda(p,v,r,\lambda,t) \\ 0 \end{pmatrix},$$

$$B^{\mathrm{w}}_{x_1} = \begin{pmatrix} G(p,t) & 0 & 0 \end{pmatrix} \text{ and } B^{\mathrm{w}}_{y_1} = \begin{pmatrix} 0 \end{pmatrix}.$$

Due to the matrix $B^{\mathrm{w}}_{y_1}$ we can choose

$$Q_{y_1} = W_{y_1} = \begin{pmatrix} I \end{pmatrix}$$

This yields

$$W_{y_1}^\top B^{\mathrm{w}}_{x_1} = \begin{pmatrix} G(p,t) & 0 & 0 \end{pmatrix}, \quad Q_{x_1} = \begin{pmatrix} Q_G & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}.$$

Hence we obtain

$$G_1Q_{x_1} = \begin{pmatrix} Q_G & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \quad W_{x_1} = \begin{pmatrix} G^\top(p,t) \\ 0 \\ 0 \end{pmatrix}, \quad V_{x_1} = \begin{pmatrix} Q_G & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}.$$

In the next step we get

$$\begin{aligned} G_2 = \begin{pmatrix} I & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & I \end{pmatrix}, \quad B^{\mathrm{v}}_{y_2} &= \begin{pmatrix} 0 \\ M^{-1}(p,t)G_\lambda(p,v,r,\lambda,t) \\ 0 \end{pmatrix}, \\ B^{\mathrm{w}}_{x_2} &= \begin{pmatrix} b_1 & -G(p,t)Z(p) & 0 \end{pmatrix}, \\ B^{\mathrm{w}}_{y_2} &= \begin{pmatrix} 0 \end{pmatrix}. \end{aligned}$$

Due to the matrix $B^{\mathrm{w}}_{y_2}$ we can choose

$$P_{y_2} = V_{y_1} = \begin{pmatrix} \quad \end{pmatrix}, \quad Q_{y_1} = W_{y_1} = \begin{pmatrix} I \end{pmatrix}$$

Again this yields

$$W^\top_{y_2} B^{\mathrm{w}}_{x_2} = \begin{pmatrix} b_1 & -G(p,t)Z(p) & 0 \end{pmatrix}, \quad Q_{x_2} = \begin{pmatrix} I & 0 & 0 \\ (GZ)^+ b_1 & Q_{GZ} & 0 \\ 0 & 0 & I \end{pmatrix}.$$

Such that we achieve

$$G_2Q_{x_2} = Q_{x_2}, \quad W_{x_2} = \begin{pmatrix} -((GZ)^+ b_1)^\top Z^\top(p) G^\top(p,t) \\ Z^\top(p) G^\top(p,t) \\ 0 \end{pmatrix},$$

which finally yields $B^{\mathrm{w}}_{y_2} = \begin{pmatrix} Z(p)G(p,t)M^{-1}(p,t)G_\lambda(p,v,r,\lambda,t) \end{pmatrix}$. Hence the DAE has Dissection Index 3, which equals the number of the Hessenberg stages.

4.6 Perturbation Analysis: Hessenberg DAEs

In Section 4.5 the mechanical applications of Section 3.2 are presented as DAEs in Hessenberg form. Additionally the basis chain of the Dissection Index of DAEs in Hessenberg form is calculated in 4.5, yielding that the number of Hessenberg stages coincides with the Dissection Index. In this section we extend Theorem 4.38 by DAEs in Hessenberg form. Therefore we formulate the following theorem:

Theorem 4.41. (Perturbation Index of DAEs in Hessenberg form)
Consider a DAE (4.41), let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ be compact and connected and let $x^0 \in \mathcal{G}$ be the initial value of the IVP

$$\begin{aligned} x_1' &= f_1(x_1, x_2, t) \\ &\vdots \\ x_{\mu-1}' &= f_{\mu-1}(x_1, \dots, x_{\mu-1}, x_\mu, t) \\ 0 &= f_0(x_1, t) \end{aligned}$$

- Let f_i be $\mu - i$ times continuously differentiable.
- Let the IVP have a global solution $x_\star$ on $\mathcal{I}_\star$ with $x_\star(t_0) = x^0$.

Further for all perturbation δ with $||\delta||_\infty, ..., ||\delta^{(\mu-1)}||_\infty$ sufficiently small it holds that:

(i) Each perturbed system has a unique global solution.

(ii) The Perturbation Index exists.

(iii) The Perturbation Index μ_P is equal to μ.

Proof.
Consider the perturbed system:

$$\begin{aligned} x_1' &= f_1(x_1, x_2, t) + \delta_1(t) && \text{(4.44a)} \\ x_2' &= f_2(x_1, x_2, x_3, t) + \delta_2(t) && \text{(4.44b)} \\ &\vdots \\ x_{\mu-1}' &= f_{\mu-1}(x_1, \dots, x_{\mu-1}, x_\mu, t) + \delta_{\mu-1}(t) && \text{(4.44c)} \\ 0 &= f_0(x_1, t) + \delta_0(t) && \text{(4.44d)} \end{aligned}$$

We differentiate Equation (4.44d) and obtain

$$0 = \frac{\partial}{\partial x_1} f_0(x_1, t) x_1' + \frac{\partial}{\partial t} f_0(x_1, t) + \delta_0'(t).$$

This yields together with (4.44a)

$$0 = \frac{\partial}{\partial x_1} f_0(x_1, t) f_1(x_2, x_1, t) + \frac{\partial}{\partial x_1} f_0(x_1, t)\delta_1(t) + \frac{\partial}{\partial t} f_0(x_1, t) + \delta_0'(t). \tag{4.45}$$

We define

$$F_1(x_1, x_2, t, \Delta^{(0)}(t)) = \frac{\partial}{\partial x_1} f_0(x_1, t) f_1(x_1, x_2, t) + \frac{\partial}{\partial x_1} f_0(x_1, t)\delta_1(t) + \frac{\partial}{\partial t} f_0(x_1, t)$$

with

$$\Delta^{(0)}(t) := \big(\delta(t)\big) \qquad \text{and} \qquad \frac{\partial}{\partial x_2}F_1 = \frac{\partial}{\partial x_1}f_0\frac{\partial}{\partial x_2}f_1.$$

The global solution $x_\star$ with $x_\star(t_0) = x^0$ is still a solution of (4.44) if we exchange (4.44d) with (4.45) and vice versa, cf. Lemma 3.5.10 in [Ste06]. This procedure can be repeated by differentiating (4.44d) a second time and inserting (4.44a) and (4.44b), such that we obtain a function

$$F_2(x_1, x_2, x_3, t, \Delta^{(1)}(t))$$

with

$$\Delta^{(1)}(t) := \big(\delta(t) \quad \delta^{(1)}(t)\big) \qquad \text{and} \qquad \frac{\partial}{\partial x_3}F_1 = \frac{\partial}{\partial x_1}f_0\frac{\partial}{\partial x_2}f_1\frac{\partial}{\partial x_3}f_2$$

and the equation

$$0 = F_2(x_1, x_2, x_3, t, \Delta^{(1)}(t)) + \delta_0^{(2)}(t).$$

We repeat this step $\mu - 2$ times and obtain a system

$$\begin{aligned} x_1' &= f_1(x_1, x_2, t) + \delta_1(t) && (4.46a)\\ &\vdots \\ x_{\mu-1}' &= f_{\mu-1}(x_1, \ldots, x_{\mu-1}, x_\mu, t) + \delta_{\mu-1}(t) && (4.46b)\\ 0 &= F_{\mu-1}(x_1, \ldots, x_{\mu-1}, x_\mu, t, \Delta^{(\mu-2)}(t)) + \delta_0^{(\mu-1)}(t) && (4.46c) \end{aligned}$$

with $\frac{\partial}{\partial x_\mu}F_{\mu-1} = \frac{\partial}{\partial x_1}f_0\frac{\partial}{\partial x_2}f_1 \cdots \frac{\partial}{\partial x_{\mu-1}}f_{\mu-2}\frac{\partial}{\partial x_\mu}f_{\mu-1}$. By Lemma 3.5.10 in [Ste06] it is sufficient to (i)-(iii) for (4.46). There is a function Ψ_0 such that

$$x_\mu = \Psi_\mu(x_1, \ldots, x_{\mu-1}, t, \Delta^{(\mu-1)}(t)) \tag{4.47}$$

with

$$\frac{\partial}{\partial \delta^{(\mu-1)}}\Psi_\mu = \Big(\tfrac{\partial}{\partial \delta_0^{(\mu-1)}}\Psi_\mu \quad 0 \quad \ldots \quad 0\Big)$$

having full row rank by Lemma 4.35. We define

$$\tilde{f}_{\mu-1}(x_1, \ldots, x_{\mu-1}, t, \Delta^{(\mu-1)}(t)) := f_{\mu-1}(x_1, \ldots, x_{\mu-1}, \Psi_\mu(x_1, \ldots, x_{\mu-1}, t, \Delta^{(\mu-1)}(t)), t)$$

and obtain

$$x_{\mu-1}' = \tilde{f}_{\mu-1}(x_1, \ldots, x_{\mu-1}, t, \Delta^{(\mu-1)}(t)) + \delta_{\mu-1}(t) \tag{4.48a}$$

$$x'_{\mu-2} = f_{\mu-2}(x_{\mu-1}, \ldots, x_1, t) + \delta_{\mu-2}(t) \tag{4.48b}$$

$$\vdots$$

$$x'_1 = f_1(x_2, x_1, t) + \delta_1(t) \tag{4.48c}$$

$$0 = F_{\mu-1}(x_1, \ldots, x_{\mu-1}, x_\mu, t, \Delta^{(\mu-2)}(t)) + \delta_0^{(\mu-1)}(t). \tag{4.48d}$$

We apply Lemma 4.36 to the Equations (4.48a)-(4.48c), which yield together with Equation (4.47) the unique global solution of the perturbed system, the existence of the Perturbation Index and an upper bound $\mu_P \leqslant \mu$. At last we apply Lemma 4.37 to Equation (4.46c) and obtain the lower bound $\mu_P \geqslant \mu$. □

4.7 Summary and Outlook

In this chapter we introduced the concept of the Dissection Index. The Dissection Index combines the strengths of the Strangeness Index concepts and Tractability Index concepts to improve the following issues:

(i) The non-linearity of the projectors and matrices.

(ii) The differentiability assumptions regarding the involved functions.

(iii) The independence between the stages of the step-by-step analysis.

We defined the Dissection Index on semi-proper formulated DAEs, a class of DAEs which includes proper formulated DAEs as well as DAEs in standard form. The main result of this chapter is that the Dissection Index coincides with the Perturbation Index for nonlinear semi-proper formulated DAEs with an arbitrarily high Dissection Index if the basis function chain of the Dissection Index is state independent. A similar result for nonlinear proper formulated DAEs with Tractability Index 1 can be found in [LMT13]. The assumptions of this theorem hold in particular for electric circuit including the semiconductor devices, the memristors and the electromagnetic devices from Chapter 3.

For Hessenberg systems and thereby for a class of mechanical applications we also proofed a connection between the Dissection Index and the Perturbation Index. In the following chapters we will no longer consider Hessenberg systems nor mechanical applications. The Strangeness Index concept is well suited for DAEs in Hessenberg form. We will focus on electric circuits as an application.

5 Solvability and Uniqueness

At the end of Chapter 4 results about the sensitivity of the solution of a DAE, with regard to perturbations, are presented. Chapter 6 deals with the convergence of numerically calculated solutions against the exact solution of a DAE. In both cases we assume that the considered DAE has a unique solution on a fixed time interval. Under which circumstances this global solution assumption is fulfilled, will be discussed in this chapter.

Most of the solvability results for nonlinear DAEs are local. Local solvability results deal with the existence and uniqueness of solutions in a neighborhood of a given initial value, cf. [KM06, LMT13]. These results are usually obtained by a combination of a transformation of the DAE and the Implicit Function Theorem. The few global solvability results, which can be found in the literature, require strong smoothness assumptions and uniform bounds of certain inverse matrices, see [GM86, RK04, CC07, Rei91]. In particular, the uniform bounds of the inverse matrices are difficult to check for applications. Additionally these results deal only with index 1 DAEs.

This chapter is split into five sections. We start by introducing the concept of strong monotonicity. With the help of this concept we present global solvability results for nonlinear algebraic equations. Afterwards we provide criteria for the global solvability for a class of implicit ODEs. Combining the global solvability results for the algebraic equations and those for the implicit ODEs by the decoupling procedure of the Dissection Index, we obtain the global solvability for semi-linear DAEs with an arbitrary index. The chapter is concluded by applying this result to the electric circuit applications of Chapter 3.

Strong Monotonicity

We start this section by collecting more or less well-known definitions and tools for solving nonlinear algebraic equations and explicit ordinary differential equations.

In $\mathbb{R}^n$ the Euclidean scalar product and its induced norm is denoted by

$$\langle x, y\rangle := x^\top y, \quad \|x\| := \sqrt{x^\top x} \quad x, y \in \mathbb{R}^n.$$

A linear function $A : \mathbb{R}^n \to \mathbb{R}^m$ can be measured by its natural operator norm which we denote by

$$\|A\|_* := \sup_{x\in\mathbb{R}^n, x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{x\in\mathbb{R}^n, \|x\|=1} \|Ax\|\,.$$

A criterion for the global solvability of an explicit ODE (2.2) is the Lipschitz continuity of the function f, cf. [GJ09]. We define the Lipschitz continuity for a function with two arguments:

Definition 5.1 (Lipschitz continuity)**.**
Let a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^k$, $n, m, k \in \mathbb{N}$, be given. Then f is Lipschitz continuous with respect to the first argument x if there is a constant $L_f > 0$ such that

$$\|f(x_2, y) - f(x_1, y)\| \leqslant L_f \|x_2 - x_1\|, \quad \forall x_1, x_2 \in \mathbb{R}^n, y \in \mathbb{R}^m.$$

Notice that L_f is independent from the arguments x and y of the function f. We use the Lipschitz continuity to show the solvability of the differential parts of the DAE. For the algebraic parts we will use the strong monotonicity as a solvability criterion:

Definition 5.2 (Monotonicity)**.**
Let a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be given. Then f is monotone with respect to the second argument y if

$$\langle f(x, y_2) - f(x, y_1), y_2 - y_1 \rangle \geqslant 0, \quad \forall x \in \mathbb{R}^n, y_1, y_2 \in \mathbb{R}^m.$$

We call f strict monotone with respect to the second argument y if

$$\langle f(x, y_2) - f(x, y_1), y_2 - y_1 \rangle > 0, \quad \forall x \in \mathbb{R}^n, y_1 \neq y_2 \in \mathbb{R}^m.$$

At last f is strongly monotone with respect to the second argument y if there is a scalar $\mu_f > 0$ such that

$$\langle f(x, y_2) - f(x, y_1), y_2 - y_1 \rangle \geqslant \mu_f \|y_2 - y_1\|^2, \quad \forall x \in \mathbb{R}^n, y_1, y_2 \in \mathbb{R}^m.$$

Again notice that μ_f is independent from the arguments of f. Strong monotonicity can be interpreted as the counterpart of Lipschitz continuity since Lipschitz continuity bounds the rate of change from above while strong monotonicity bounds it from below. The following corollary illustrates this relation.

Corollary 5.3.
Let a function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be given. If f is strongly monotone with respect to y, then there is a $\mu_f > 0$ such that

$$\|f(x, y_2) - f(x, y_1)\| \geqslant \mu_f \|y_2 - y_1\|, \quad \forall x \in \mathbb{R}^n, y_1, y_2 \in \mathbb{R}^m.$$

Proof. For all $(x, y_1), (x, y_2) \in \mathbb{R}^n \times \mathbb{R}^m$ there is a $\mu_f > 0$ such that it holds

$$\mu_f \|y_2 - y_1\|^2 \leqslant \langle f(x, y_2) - f(x, y_1), y_2 - y_1 \rangle \leqslant \|f(x, y_2) - f(x, y_1)\| \|y_2 - y_1\|$$

by the Cauchy-Schwarz inequality. □

If f is continuously differentiable with respect to x then f is Lipschitz continuous with respect to x if and only if there is an $L > 0$ such that $\|f_x(x,y)\|_* \leqslant L$ for all $(x,y) \in \mathbb{R}^n \times \mathbb{R}^m$. A similar result holds for a strong monotone function.

Lemma 5.4.
Let the continuous function $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ be continuously differentiable with respect to y and let $f_y(x,y)$ be the Jacobian of f with respect to y at the point (x,y). Then it holds:

(i) f is strongly monotone with respect to y if and only if the map $z \mapsto f_y(x,y)z$ is strongly monotone with respect to z, i.e. there is a $\mu > 0$ such that

$$\langle z, f_y(x,y)z\rangle \geqslant \mu \|z\|^2, \quad \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^m \; \forall z \in \mathbb{R}^m.$$

(ii) In the case of (i) $f_y(x,y)$ is bounded from below by μ, i.e.

$$\|f_y(x,y)\|_* \geqslant \mu, \quad \forall (x,y) \in \mathbb{R}^n \times \mathbb{R}^m.$$

The proof can be found in [OR70, p. 142]. In the following section we use the strong monotonicity to describe the solution of a parameter depending algebraic equation.

5.1 Algebraic equations

We start this subsection by stretching the differences between monotonicity, strict monotonicity and strong monotonicity. Therefore we discuss necessity and sufficiency of these monotonicity concepts for the solvability of the equation

$$f(x) = y \tag{5.1}$$

with $f : \mathbb{R}^n \to \mathbb{R}^n$ being continuous. The following solvability theorem for strongly monotone functions can be found in [OR70, Theorem 6.4.4].

Theorem 5.5.
Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuous and strongly monotone. Then the equation

$$f(x) = y$$

has a unique solution $x \in \mathbb{R}^n$ for each $y \in \mathbb{R}^n$.
Furthermore the inverse function $f^{-1} : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous.

Hence strong monotonicity is sufficient for the solvability of Equation (5.1). Let $f(x) = x^3$, then f is monotone but neither strict monotone nor strongly monotone. Equation (5.1) is globally unique solvable for this particular f, since $f^{-1} = \sqrt[3]{x}$ is the inverse function

of f. Hence strong monotonicity is not necessary for the solvability of Equation (5.1), but strong monotonicity, in contrast to monotonicity and strict monotonicity, provides the Lipschitz continuity of the inverse function f^{-1}. The Theorem 5.5 is a special case of the Browder-Minty Theorem, see [Zei90]. The Browder-Minty Theorem only needs monotonicity, amongst other assumptions, to provide the solvability of (5.1) in its general case. But without the strong monotonicity the Lipschitz continuity of the inverse function f^{-1} is not guaranteed as we can see for $f(x) = x^3$. At last we consider the exponential function $f(x) = e^x$. The exponential function is strictly monotone but not strongly monotone and Equation (5.1) is not globally unique solvable for this particular f. Hence strict monotonicity is also not sufficient for the solvability of (5.1).
We extend Theorem 5.5 to parameter depending nonlinear equations.

Lemma 5.6. ([JMT13])
Let $\mathcal{I} \subseteq \mathbb{R}$ be an interval, $k \geqslant 0$ and $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I} \to \mathbb{R}^m$ be a k-times continuously differentiable function. The $k = 0$ case means that f is assumed to be continuous. Then for all $(x, t) \in \mathbb{R}^n \times \mathcal{I}$ the equation

$$f(x, y, t) = 0 \tag{5.2}$$

has a unique solution $y \in \mathbb{R}^m$ if f is strongly monotone with respect to y and Lipschitz continuous with respect to x. The solution depends on (x, t) and we write $y = \psi(x, t)$ with the k-times continuously differentiable function $\psi : \mathbb{R}^n \times \mathcal{I} \to \mathbb{R}^m$ which is Lipschitz continuous with respect to x.

Proof. The unique solvability is derived from Theorem 5.5 for all but fixed (x, t) with the solution $y_{x,t}$. By setting $\psi(x, t) := y_{x,t}$ we obtain the solution function ψ. The continuity of ψ, as well as the Lipschitz continuity with respect to x, have to be checked.
We start with the continuity. Let $(x_n, t_n) \in \mathbb{R}^n \times \mathcal{I}$ be a sequence with

$$(x_n, t_n) \to (x, t) \in \mathbb{R}^n \times \mathcal{I} \quad \text{as } n \to \infty$$

and hence

$$f(x_n, \psi(x_n, t_n), t_n) = 0 = f(x, \psi(x, t), t).$$

We obtain with the strong monotonicity (scalar $\mu_f > 0$)

$$\begin{aligned} & \|\psi(x, t) - \psi(x_n, t_n)\| \\ \leqslant \; & \frac{1}{\mu_f} \|f(x_n, \psi(x, t), t_n) - f(x_n, \psi(x_n, t_n), t_n)\| \\ = \; & \frac{1}{\mu_f} \|f(x_n, \psi(x, t), t_n) - f(x, \psi(x, t), t)\| \to 0 \quad \text{as } n \to \infty \end{aligned}$$

because f is continuous. So ψ is continuous. Next we verify the Lipschitz continuity. Let $x_1, x_2 \in \mathbb{R}^n,\ t \in \mathcal{I}$ and hence

$$f(x_1, \psi(x_1,t), t) = 0 = f(x_2, \psi(x_2,t), t).$$

Similar as before it follows by the strong monotonicity of f that

$$\begin{aligned} \|\psi(x_2,t) - \psi(x_1,t)\| &\leqslant \frac{1}{\mu_f} \|f(x_2, \psi(x_2,t), t) - f(x_2, \psi(x_1,t), t)\| \\ &= \frac{1}{\mu_f} \|f(x_1, \psi(x_1,t), t) - f(x_2, \psi(x_1,t), t)\| \\ &\leqslant \frac{L_f}{\mu_f} \|x_2 - x_1\| . \end{aligned}$$

The last line was obtained by using the Lipschitz continuity of f $(L_f > 0)$.
Lemma 5.6 can be seen as the global version of the Implicit Function Theorem. The k-times continuously differentiability follows as in the proof of the Implicit Function Theorem, which can be found in [Zei86](p. 153, Theorem 4.B (d)). □

A function ψ, as in Lemma 5.6, will be called a solution function. This means generally that there is a unique function ψ satisfying

$$y = \psi(x,t) \Leftrightarrow f(x,y,t) = 0.$$

The following lemma is a preparation for the solvability Theorem 5.20 regarding the circuit applications:

Lemma 5.7.
Let $\mathcal{I} \subseteq \mathbb{R}$ be an interval and $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I} \to \mathbb{R}^m$ be a continuously differentiable function. Then for all $(x,t) \in \mathbb{R}^n \times \mathcal{I}$ the equation

$$f(x,y,t) = 0 \tag{5.3}$$

has a unique solution $y \in \mathbb{R}^m$ if

$$\frac{\partial}{\partial y} f(x,y,t) = \begin{pmatrix} L_2(x,y,t) & L_3(x,y,t) & C_2 \\ L_1(x,y,t) & M(x,y,t) & 0 \\ C_1 & 0 & 0 \end{pmatrix}$$

with $L_i(x,y,t)$ being bounded for all $(x,y,t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I}$, $M(x,y,t)$ being strongly monotone, C_i being constant and non-singular and $\frac{\partial}{\partial x} f(x,y,t)$ being bounded for all $(x,y,t) \in \mathbb{R}^n \times \mathbb{R}^m \times \mathcal{I}$. The solution depends on (x,t) and we write $y = \psi(x,t)$ with the continuous function $\psi : \mathbb{R}^n \times \mathcal{I} \to \mathbb{R}^m$ which is Lipschitz continuous with respect to x.

Proof. We split $y = \begin{pmatrix} y_1 & y_2 & y_3 \end{pmatrix}^\top$. Then there are functions f^1, f^2 and f^3 by Lemma 4.8 such that we can write

$$f(x, y, t) = \begin{pmatrix} f^3(y_1, y_2, x, t) + C_2 y_3 \\ f^2(y_1, y_2, x, t) \\ C_1 y_1 + f^1(x, t) \end{pmatrix}.$$

With the help of Lemma 5.6 we obtain a function $\Psi_2(y_1, x, t)$, which is Lipschitz continuous in y_1 and x, with $y_2 = \Psi_2(y_1, x, t)$. Hence, we obtain

$$y = \Psi(x, t) := \begin{pmatrix} -C_2^{-1} f^3(-C_1^{-1} f^1(x, t), \Psi_2(-C_1^{-1} f^1(x, t), x, t), x, t) \\ \Psi_2(-C_1^{-1} f^1(x, t), x, t) \\ -C_1^{-1} f^1(x, t) \end{pmatrix}$$

with Ψ being Lipschitz continuous in x as a composition of Lipschitz continuous functions.

□

5.2 Implicit ODEs

In this subsection we provide criteria for the global unique solvability of an implicit ODE of the following form:

$$\frac{\mathrm{d}}{\mathrm{d}t} m(x, t) = f(x, t) \tag{5.4}$$

with $f \in C(\mathbb{R}^n \times \mathcal{I}, \mathbb{R}^n)$, $m \in C^1(\mathbb{R}^n \times \mathcal{I}, \mathbb{R}^n)$ and $\mathcal{I} \subseteq \mathbb{R}$ being a compact interval. Implicit ODE of this kind may occur after the decoupling of the dissection concept.

Theorem 5.8. (Global Solvability,[JMT13])
Consider an implicit ODE (5.4) with f being continuous, m being continuously differentiable and $\mathcal{I}$ being a time interval. If

(i) m is strongly monotone with respect to x,

(ii) f is Lipschitz continuous with respect to x

then (5.4) has a unique solution $x_\star \in C^1(\mathcal{I}, \mathbb{R}^n)$ for every initial value $x_\star(t_0) = x^0 \in \mathbb{R}^{n_x}$.

Proof.
First we prove the existence of a solution and afterwards we will show its uniqueness. We show an a priori estimate for any solution $x : J \to \mathbb{R}^{n_x}$ to (5.4) on an arbitrary subinterval $J := [t_0, T_J] \subseteq \mathcal{I}$. If x solves (5.4) we can integrate over $[t_0, t] \subseteq J$ and obtain

$$m(x(t), t) = m(x^0, t_0) + \int_{t_0}^{t} f(x(s), s)\mathrm{d}s$$

with $x(t_0) = x^0$. Using the strong monotonicity of m we get

$$\begin{aligned}
&\mu \left\| x(t) - x^0 \right\| \\
\leqslant &\left\| m(x(t), t) - m(x^0, t) \right\| \\
\leqslant &\left\| m(x^0, t_0) - m(x^0, t) \right\| + \int_{t_0}^{t} \| f(x(s), s) \| \,\mathrm{d}s \\
\leqslant &\left\| m(x^0, t_0) - m(x^0, t) \right\| + \int_{t_0}^{t} \left\| f(x^0, s) \right\| \mathrm{d}s + \int_{t_0}^{t} \left\| f(x(s), s) - f(x^0, s) \right\| \mathrm{d}s \\
\leqslant &(T - t_0) \max_{\tau_f, \tau_m \in [t_0, T]} (\left\| m_t(x^0, \tau_m) \right\| + \left\| f(x^0, \tau_f) \right\|) + L_f \int_{t_0}^{t} \left\| x(s) - x^0 \right\| \mathrm{d}s
\end{aligned}$$

with $L_f > 0$. The last line is a consequence of the mean value theorem and the Lipschitz continuity of f. We conclude that there are constants $c_1, c_2 > 0$, independent of t, such that

$$\left\| x(t) - x^0 \right\| \leqslant c_1 + c_2 \int_{t_0}^{t} \left\| x(s) - x^0 \right\| \mathrm{d}s.$$

Applying the Gronwall Lemma gives us the desired a priori estimate

$$\left\| x(t) - x^0 \right\| \leqslant c_1 e^{c_2 (T - t_0)} =: C$$

with $C > 0$ being independent of t. It is

$$\frac{d}{dt} m(x, t) = m_x(x, t) x' + m_t(x, t)$$

for x being continuously differentiable. Using Lemma 5.4 and the fact that m is strongly monotone the map $z \mapsto m_x(x, t) z$ is continuous and strongly monotone with respect to z, hence $m_x(x, t)$ is non-singular . Furthermore $m_x(x, t)$ is continuous in x and t since $m \in C^1(\mathbb{R}^n \times \mathcal{I}, \mathbb{R}^n)$ and so the inverse $m_x^{-1}(x, t)$ is also continuous in x and t.
Then (5.4) can be reformulated as

$$x' = m_x(x, t)^{-1} \left(f(x, t) - m_t(x, t) \right) =: \tilde{f}(x, t)$$

for $t \in \mathcal{I}$ with initial value $x(t_0) = x^0 \in \mathbb{R}^n$. The function $\tilde{f}$ is continuous as a combination of continuous functions. Hence we can apply the Peano Theorem, cf. [Zei86, Theorem 3.B]. We obtain a local solution $x \in C^1(J, \mathbb{R}^n)$ on a subinterval $J \subseteq \mathcal{I}$ which can be extended to the whole interval $\mathcal{I}$ because of the a priori estimate above, cf. [Zei90, p.801 (iii)]. So there is a solution $x_\star \in C^1(\mathcal{I}, \mathbb{R}^n)$ of (5.4).

Now we prove the uniqueness of the solution. Therefore let x_1, x_2 be two solutions which fulfill (5.4). Therefore we have on $\mathcal{I}$:

$$\frac{d}{dt} m(x_1(t), t) - \frac{d}{dt} m(x_2(t), t) = f(x_1(t), t) - f(x_2(t), t).$$

We have $x_1(t_0) = x^0 = x_2(t_0)$ and integration over $[t_0, t]$, $t \in \mathcal{I}$ yields

$$m(x_1(t), t) - m(x_2(t), t) = \int_{t_0}^{t} f(x_1(s), s) - f(x_2(s), s)\mathrm{d}s.$$

Using the strong monotonicity of m and the Lipschitz continuity of f we see that

$$\begin{aligned}\|x_1(t) - x_2(t)\| &\leqslant \frac{1}{\mu} \|m(x_1(t), t) - m(x_2(t), t)\| \\ &\leqslant \frac{1}{\mu} \int_{t_0}^{t} \|f(x_1(s), s) - f(x_2(s), s)\| \, \mathrm{d}s \\ &\leqslant \frac{L_f}{\mu} \int_{t_0}^{t} \|x_1(s) - x_2(s)\| \, \mathrm{d}s.\end{aligned}$$

with $\mu > 0$. Gronwall's Lemma now reveals that $x_1(t) = x_2(t)$ for all $t \in \mathcal{I}$. □

Now we have solvability criteria for both algebraic and differential equations.

5.3 General DAEs

Next we obtain a global solvability result for quasilinear DAEs by combining the results of the last two sections.

Definition 5.9. (Quasilinear DAE)
Let $\mathcal{I} \subset \mathbb{R}$ be a compact time interval. Consider the following set of equations

$$A \frac{\mathrm{d}}{\mathrm{d}t} d(x(t), t) + b(x(t), t) = 0 \tag{5.5}$$

with $A \in \mathbb{R}^{n \times m}$, $d \in C^1(\mathbb{R}^n \times \mathcal{I}, \mathbb{R}^m)$ and $b \in C(\mathbb{R}^n \times \mathcal{I}, \mathbb{R}^n)$ with a continuous partial derivative $\frac{\partial}{\partial x} b(x, t)$. We call (5.5) a quasilinear DAE.

Analogous to Section 4.4 our next objective is to decouple the DAE. In contrast to Theorem 4.35 the decoupling in this section will be global. Before formulating such a global decoupling we present the matrix chain of a quasilinear DAE with a semi-properly stated derivative term and constant basis functions, since these restrictions essentially simplify the matrix chain. The definition of a semi-properly stated derivative term simplifies to the conditions

$$\operatorname{im} A = \operatorname{im} A \frac{\partial}{\partial x} d(x, t), \quad \forall (x, t) \in \mathbb{R}^n \times \mathcal{I}, \tag{5.6}$$

and that $\operatorname{im} \frac{\partial}{\partial x} d(x, t)$ has a basis continuously depending on x and t in the case of a quasilinear DAE.

We define the matrix functions

$$\begin{aligned} D(x,t) &= \frac{\partial}{\partial x} d(x,t) \\ B(x,t) &= \frac{\partial}{\partial x} b(x,t) \end{aligned}$$

and notice that the matrix A in Equation (5.5) equals the matrix A at the beginning of the matrix chain. We obtain the next sequence of matrix functions

$$\begin{aligned} G_1(x,t) =& V^\top AD(x,t)P, \\ B^{\mathrm{v}}_{x_1}(x^1,x,t) =& V^\top B(x,t)P + V^\top A(D(x,t)P)' = B^{\mathrm{v}}_{x_{1,*}}(x,t) + (G_1(x,t))', \\ B^{\mathrm{v}}_{y_1}(x,t) =& V^\top B(x,t)Q + V^\top A(D(x,t)Q)' = V^\top B(x,t)Q, \\ B^{\mathrm{w}}_{x_1}(x,t) =& W^\top B(x,t)P, \\ B^{\mathrm{w}}_{y_1}(x,t) =& W^\top B(x,t)Q \end{aligned}$$

with $B^{\mathrm{v}}_{x_{1,*}}(x,t) := V^\top B(x,t)P$. Next we denote the other stages of the matrix chain

$$\begin{array}{rclrclrcl} G_i &=& V^\top_{x_{i-1}} G_{i-1} Q_{x_{i-1}} & B^{\mathrm{v}}_{x_i} &=& V^\top_{x_{i-1}} B_{x_{i-1}} & B^{\mathrm{v}}_{y_i} &=& V^\top_{x_{i-1}} B_{y_{i-1}} \\ &&& B^{\mathrm{w}}_{x_i} &=& W^\top_{x_{i-1}} B_{x_{i-1}} & B^{\mathrm{w}}_{y_i} &=& W^\top_{x_{i-1}} B_{y_{i-1}} \end{array}$$

with

$$\begin{aligned} B_{y_{i-1}} =& B^{\mathrm{v}}_{y_{i-1}} Q_{y_{i-1}} \\ B_{x_{i-1}} =& B^{\mathrm{v}}_{x_{i-1}} Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}} \\ =& (B^{\mathrm{v}}_{x_{i-1,*}} + G'_{i-1}) Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}}. \end{aligned}$$

We want to stress the fact that the terms $(G_i(x,t))'$ only appears in the $B^{\mathrm{v}}_{x_i}$ matrices and consequently only these matrices depend on a jet variable.

$$\begin{aligned} B^{\mathrm{v}}_{x_i} &= V^\top_{x_{i-1}} B_{x_{i-1}} \\ &= V^\top_{x_{i-1}} ((B^{\mathrm{v}}_{x_{i-1,*}} + G'_{i-1}) Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}}) \\ &= V^\top_{x_{i-1}} (B^{\mathrm{v}}_{x_{i-1,*}} Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}}) + G'_i \\ &= B^{\mathrm{v}}_{x_{i,*}} + G'_i \\ B^{\mathrm{w}}_{x_i} &= W^\top_{x_{i-1}} B_{x_{i-1}} \\ &= W^\top_{x_{i-1}} (B^{\mathrm{v}}_{x_{i-1,*}} Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}}) \\ &= W^\top_{x_{i-1}} B_{x_{i-1},*} \end{aligned}$$

with

$$B_{x_{i-1},*} := B^{\mathrm{v}}_{x_{i-1,*}} Q_{x_{i-1}} - B^{\mathrm{v}}_{y_{i-1}} P_{y_{i-1}} (V^\top_{y_{i-1}} B^{\mathrm{w}}_{y_{i-1}} P_{y_{i-1}})^{-1} V^\top_{y_{i-1}} B^{\mathrm{w}}_{x_{i-1}} Q_{x_{i-1}}$$

$$B^{\mathrm{v}}_{x_{i,*}} := V^{\top}_{x_{i-1}} B_{x_{i-1},*}.$$

Similar to Section 4.4 the basis functions have to meet the following assumption:

Assumption 5.10. (Constant basis chain)
Consider a DAE (5.5) with a finite Dissection Index. Assume that all basis functions except $V_{x_{\mu-1}}$ and $W_{x_{\mu-1}}$ are constant. Furthermore let the alternative basis functions V^*_y and W^*_y, see Definition 4.15 , be constant.

Assumption 5.10 seems to be very strict when we think of the projector of the Tractability Index or the basis functions of the Strangeness Index. In contrast to these index concepts the basis functions of the Dissection Index fulfill Assumption 5.10 for a large application class, i.e. electric circuits including semiconductor devices, memristors and electromagnetic devices. Additionally we need the DAE to be sufficiently smooth and certain matrices of the matrix chain need to be strongly monotone.

Assumption 5.11. (Differentiability and strong monotonicity)
Consider a DAE (5.5) with a semi-properly stated derivative term and a finite Dissection Index μ. Let the DAE fulfill Assumption 5.10. For $1 \leqslant i \leqslant \mu - 1$ assume that the matrix valued functions:

$$\begin{aligned} V^{\top}_{y_i} B^{\mathrm{w}}_{y_i}(x,t) P_{y_i} \quad &\text{and} \quad W^{\top}_{y_i} B^{\mathrm{w}}_{x_i}(x,t) P_{x_i}, \\ (W^*_y)^{\top} G_{\mu-1}(x,t) Q_{x_{\mu-1}} \quad &\text{and} \quad (V^*_y)^{\top} B^{\mathrm{v}}_{y_{\mu-1}}(x,t) Q_{y_{\mu-1}} \end{aligned}$$

are strongly monotone. Furthermore assume that b is continuously differentiable and that:

$$\begin{aligned} W^{\top}_{x_{i-1}} V^{\top}_{x_{i-2}} \dots V^{\top}_{x_0} b(x,t), \quad &\text{for } i = 1, \dots, \mu - 1 \\ W^{\top}_{x_i} V^{\top}_{x_{i-1}} \dots V^{\top}_{x_0} A d(x,t), \quad &\text{for } i = 1, \dots, \mu - 2 \end{aligned}$$

are $(\mu - i)$ times continuously differentiable with $V^{\top}_{x_{i-2}} \dots V^{\top}_{x_0} := I$ for $i = 1$.

With the help of these preparations we formulate the following theorem.

Theorem 5.12. (Global decoupling)
Consider a DAE (5.5) with $\mathcal{I}_{\star} := [t_0, T] \subset \mathcal{I}$:

$$A \frac{\mathrm{d}}{\mathrm{d}t} d(x(t),t) + b(x(t),t) = 0, \quad \forall t \in \mathcal{I}_{\star}.$$

(i) Let the DAE have a finite Dissection Index μ.

(ii) Let the DAE fulfill the Assumptions 5.10 and 5.11.

Then there are $(\mu+1-i)$ times differentiable functions $\Psi_{\tilde{y}_i}$ and $(\mu+2-j)$ times differentiable functions $\Psi_{\tilde{x}_j}$ with $1 \leqslant i \leqslant \mu$ and $1 \leqslant j \leqslant \mu-1$ such that the DAE can be decoupled into

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} d_\mu(x_\mu, t) + b_\mu(x_\mu, t) = 0 \\
&\qquad\qquad\qquad \tilde{y}_\mu = \Psi_{\tilde{y}_\mu}(x_\mu, t, \frac{\mathrm{d}}{\mathrm{d}t} d_{\mu-1}(x_\mu, t)) \\
&\qquad\qquad\qquad \tilde{y}_i = \Psi_{\tilde{y}_i}(x_i, t), \qquad\qquad 1 \leqslant i \leqslant \mu-1 \\
&\qquad\qquad\qquad \tilde{x}_j = \Psi_{\tilde{x}_j}(t), \qquad\qquad 1 \leqslant j \leqslant \mu-1.
\end{aligned}$$

Proof.
We define the transformation matrix

$$T = \begin{pmatrix} Q_{x_\mu} & Q_{x_0}P_{x_1} & \dots & Q_{x_{\mu-2}}P_{x_{\mu-1}} & Q_{y_0}P_{y_1} & \dots & Q_{y_{\mu-1}}P_{y_\mu} \end{pmatrix}$$

such that the variable x is split into:

$$\begin{aligned}
x &= Q_{x_\mu} x_\mu + \sum_{i=1}^{\mu-1} Q_{x_{i-1}} P_{x_i} \tilde{x}_i + \sum_{i=1}^{\mu} Q_{y_{i-1}} P_{y_i} \tilde{y}_i \\
&= T \begin{pmatrix} x_\mu & \tilde{x}_1 & \dots & \tilde{x}_{\mu-1} & \tilde{y}_1 & \dots & \tilde{y}_\mu \end{pmatrix}^\top.
\end{aligned}$$

We will prove by an induction that for $1 \leqslant i \leqslant \mu$ and $1 \leqslant j \leqslant \mu-1$ the solution parts $\tilde{y}_i$ and $\tilde{x}_j$ can be described by a $(\mu+1-i)$ times differentiable function $\Psi_{\tilde{y}_i}$ and by a $(\mu+2-j)$ times differentiable function $\Psi_{\tilde{x}_j}$, i.e.

$$\tilde{y}_i = \Psi_{\tilde{y}_i}(x_i, t) \quad \text{and} \quad \tilde{x}_j = \Psi_{\tilde{x}_j}(t)$$

with

$$\frac{\partial}{\partial x_i} \Psi_{\tilde{y}_i}(x_i, t) = -(V_{y_{i-1}}^\top B_{y_{i-1}}^{\mathrm{w}} P_{y_{i-1}})^{-1} V_{y_{i-1}}^\top B_{x_{i-1}}^{\mathrm{w}} Q_{x_{i-1}}.$$

Furthermore there are functions d_μ and b_μ such that for the solution part x_μ holds:

$$\frac{\mathrm{d}}{\mathrm{d}t} d_\mu(x_\mu, t) + b_\mu(x_\mu, t) = 0$$

with $\frac{\partial}{\partial x_\mu} d_\mu(x_\mu, t) = (W_y^*)^\top G_{\mu-1}(x^1, x, t) Q_{x_{\mu-1}}$.
We assume, for a moment, the existence of the functions $\Psi_{\tilde{x}_i}$ and $\Psi_{\tilde{y}_i}$. Based on these functions we recursively define functions d_i and b_i starting with

$$d_0(x_0, t) = Ad(Px_0, t) \quad \text{and} \quad b_0(x_0, y_0, t) = b(Px_0 + Qy_0, t)$$

For the Jacobians of d_0 and b_0 with respect to x_0 and y_0 it holds that:

$$\frac{\partial}{\partial x_0} d_0 = ADP, \quad \frac{\partial}{\partial x_0} b_0 = BP \text{ and } \frac{\partial}{\partial y_0} b_0 = BQ.$$

For all $1 \leqslant i \leqslant \mu$ we define

$$\begin{aligned} d_i(x_i, t) &= V_{x_{i-1}}^\top d_{i-1}(Q_{x_i} x_i + P_{x_i} \Psi_{\tilde{x}_i}(t), t), \\ b_i(x_i, y_i, t) &= V_{x_{i-1}}^\top b_{i-1}(Q_{x_i} x_i + P_{x_i} \Psi_{\tilde{x}_i}(t), Q_{y_i} y_i + P_{y_i} \Psi_{\tilde{y}_i}(x_i, t), t) \end{aligned}$$

First we prove the statement of the theorem for $1 \leqslant i \leqslant \mu - 2$ by a mathematical induction. Additionally we show that for the Jacobians of b_i it holds that

$$\frac{\partial}{\partial x_i} d_i = G_i Q_{x_i}, \quad \frac{\partial}{\partial x_i} b_i = B_{x_i,*}, \quad \frac{\partial}{\partial y_i} b_i = B_{y_i}$$

for $1 \leqslant i \leqslant \mu - 1$.

Base case: $(i = 1)$

We factorize the DAE with $V_{x_0}^\top$, $V_{y_1}^\top W_{x_0}^\top$ and $W_{y_1}^\top W_{x_0}^\top$:

$$\begin{aligned} & A \frac{\mathrm{d}}{\mathrm{d}t} d(x(t), t) + b(x(t), t) = 0 \\ \Leftrightarrow & \begin{pmatrix} V_{x_0}^\top \\ V_{y_1}^\top W_{x_0}^\top \\ W_{y_1}^\top W_{x_0}^\top \end{pmatrix} (A \frac{\mathrm{d}}{\mathrm{d}t} d(x(t), t) + b(x(t), t)) = 0 \end{aligned}$$

The DAE has a semi-properly stated derivative term hence $\operatorname{im} AD = \operatorname{im} A$ and for this reason it holds

$$W_{x_0}^\top G = W^\top AD = 0 \Leftrightarrow W^\top A = 0$$

with $W_{x_0}^\top = W^\top$. This leads to:

$$V_{y_1}^\top W_{x_0}^\top b(x, t) = 0, \tag{5.7a}$$
$$W_{y_1}^\top W_{x_0}^\top b(x, t) = 0. \tag{5.7b}$$

We split $x = Px_0 + Qy_0 = Q_{x_0} x_0 + Q_{y_0} y_0$ and notice that $Q_{y_0} y_0$ vanishes in (5.7b) due to the definition of $W_{y_1}^\top$ and Lemma (4.8):

$$V_{y_1}^\top W_{x_0}^\top b(Q_{x_0} x_0 + Q_{y_0} y_0, t) = 0, \tag{5.8a}$$
$$W_{y_1}^\top W_{x_0}^\top b(Q_{x_0} x_0, t) = 0. \tag{5.8b}$$

As the next step we split $x_0 = P_{x_1}\tilde{x}_1 + Q_{x_1}x_1$ and $y_0 = P_{y_1}\tilde{y}_1 + Q_{y_1}y_1$ and see that $Q_{y_1}y_1$ vanishes in (5.8a) and $Q_{x_1}x_1$ vanishes in (5.8b) due to the definitions of Q_{x_1} and Q_{y_1} and Lemma (4.8) again:

$$V_{y_1}^\top W_{x_0}^\top b(Q_{x_0}P_{x_1}\tilde{x}_1 + Q_{x_0}Q_{x_1}x_1 + Q_{y_0}P_{y_1}\tilde{y}_1, t) = 0 \tag{5.9a}$$

$$W_{y_1}^\top W_{x_0}^\top b(Q_{x_0}P_{x_1}\tilde{x}_1, t) = 0. \tag{5.9b}$$

The Jacobian $\frac{\partial}{\partial \tilde{x}_1} W_{y_1}^\top W_{x_0}^\top b(Q_{x_0}P_{x_1}\tilde{x}_1, t) = W_{y_1}^\top W_{x_0}^\top B Q_{x_0} P_{x_1} = W_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{x_1}$ is strongly monotone due to Assumption 5.11. Then by Lemma 5.6 there is a solution function which globally describes $\tilde{x}_1$:

$$\tilde{x}_1 = \Psi_{\tilde{x}_1}(t).$$

Also by Lemma 5.6 and Assumption 5.11 the function $\Psi_{\tilde{x}_1}$ is $(\mu + 1)$ times continuously differentiable. Insert this expression into (5.9a) and obtain

$$V_{y_1}^\top W_{x_0}^\top b(Q_{x_0}P_{x_1}\Psi_{\tilde{x}_1}(t) + Q_{x_0}Q_{x_1}x_1 + Q_{y_0}P_{y_1}\tilde{y}_1, t) = 0 \tag{5.10a}$$

$$\Leftrightarrow V_{y_1}^\top W_{x_0}^\top b_0(P_{x_1}\Psi_{\tilde{x}_1}(t) + Q_{x_1}x_1, P_{y_1}\tilde{y}_1, t) = 0. \tag{5.10b}$$

The partial derivative of Equation (5.10b) with respect to $\tilde{y}_1$

$$\frac{\partial}{\partial \tilde{y}_1} V_{y_1}^\top W_{x_0}^\top b_0(P_{x_1}\Psi_{\tilde{x}_1}(t) + Q_{x_1}x_1, P_{y_1}\tilde{y}_1, t) = V_{y_1}^\top W_{x_0}^\top B Q_{y_0} P_{y_1} = V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1}$$

is strongly monotone due to Assumption 5.11. Again by Lemma 5.6 there is a solution function which globally describes $\tilde{y}_1$:

$$\tilde{y}_1 = \Psi_{\tilde{y}_1}(x_1, t)$$

with

$$\frac{\partial}{\partial x_1}\Psi_{\tilde{y}_1} = -(V_{y_1}^\top B_{y_1}^{\mathrm{w}} P_{y_1})^{-1} V_{y_1}^\top B_{x_1}^{\mathrm{w}} Q_{x_1}$$

which follows exactly as in the proof of the Implicit Function Theorem, c.f. [Zei86](p. 153). By Lemma 5.6 and Assumption 5.11 the function $\Psi_{\tilde{y}_1}$ is μ times continuously differentiable. We consider

$$\begin{aligned} d_1(x_1, t) &= V_{x_0}^\top d_0(Q_{x_1}x_1 + P_{x_1}\Psi_{\tilde{x}_1}(t), t), \\ b_1(x_1, y_1, t) &= V_{x_0}^\top b_0(Q_{x_1}x_1 + P_{x_1}\Psi_{\tilde{x}_1}(t), Q_{y_1}y_1 + P_{y_1}\Psi_{\tilde{y}_1}(x_1, t), t). \end{aligned}$$

For the Jacobians of d_1 and b_1 with respect to x_1 and y_1 it holds that:

$$\frac{\partial}{\partial x_1} d_1 = V_{x_0}^\top (\frac{\partial}{\partial x_0} d_0) Q_{x_1} = V_{x_0}^\top A D Q_{x_0} Q_{x_1} = G_1 Q_{x_1},$$

$$
\begin{aligned}
\frac{\partial}{\partial x_1} b_1 &= V_{x_0}^{\top}((\frac{\partial}{\partial x_0} b_0) Q_{x_1} + (\frac{\partial}{\partial y_0} b_0) P_{y_1} (\frac{\partial}{\partial x_1} \Psi_{\tilde{y}_1})) \\
&= (V_{x_0}^{\top} \frac{\partial}{\partial x_0} b_0) Q_{x_1} + (V_{x_0}^{\top} \frac{\partial}{\partial y_0} b_0) P_{y_1} (\frac{\partial}{\partial x_1} \Psi_{\tilde{y}_1}) \\
&= B^{\mathrm{v}}_{x_1,*} Q_{x_1} - B^{\mathrm{v}}_{y_1} P_{y_1} (V_{y_1}^{\top} B^{\mathrm{w}}_{y_1} P_{y_1})^{-1} V_{y_1}^{\top} B^{\mathrm{w}}_{x_1} Q_{x_1} \\
&= B_{x_1,*}, \\
\frac{\partial}{\partial y_1} b_1 &= (V_{x_0}^{\top} \frac{\partial}{\partial y_0} b_0) Q_{y_1} = B^{\mathrm{v}}_{y_1} Q_{y_1} = B_{y_1}.
\end{aligned}
$$

We complete the base case by notating

$$
\begin{aligned}
& A \frac{\mathrm{d}}{\mathrm{d}t} d(x,t) + b(x,t) = 0 \\
\Leftrightarrow & \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} d_1(x_1,t) + b_1(x_1,y_1,t) & = 0 \\ \tilde{y}_1 - \Psi_{\tilde{y}_1}(x_1,t) & = 0 \\ \tilde{x}_1 - \Psi_{\tilde{x}_1}(t) & = 0 \end{pmatrix}.
\end{aligned}
$$

Induction step: $(i-1 \mapsto i \leqslant \mu - 1)$
By the induction hypothesis we got

$$
\begin{aligned}
& \frac{\mathrm{d}}{\mathrm{d}t} d_{i-1}(x_{i-1},t) + b_{i-1}(x_{i-1},y_{i-1},t) = 0 \\
\Leftrightarrow & \begin{pmatrix} V_{x_{i-1}}^{\top} \\ V_{y_i}^{\top} W_{x_{i-1}}^{\top} \\ W_{y_i}^{\top} W_{x_{i-1}}^{\top} \end{pmatrix} (\frac{\mathrm{d}}{\mathrm{d}t} d_{i-1}(x_{i-1},t) + b_{i-1}(x_{i-1},y_{i-1},t)) = 0
\end{aligned}
$$

with $\frac{\partial}{\partial x_{i-1}} W_{x_{i-1}}^{\top} d_{i-1}(x_{i-1},t) = W_{x_{i-1}}^{\top} G_{i-1} Q_{x_{i-1}}$ being zero due to the construction of $W_{x_{i-1}}^{\top}$. Hence we write

$$
W_{x_{i-1}}^{\top} d_{i-1}(x_{i-1},t) = W_{x_{i-1}}^{\top} d_{i-1}(t)
$$

by Lemma (4.8) and obtain

$$
\begin{aligned}
V_{y_i}^{\top} W_{x_{i-1}}^{\top} b_{i-1}(x_{i-1},y_{i-1},t) + \frac{\mathrm{d}}{\mathrm{d}t}(V_{y_i}^{\top} W_{x_{i-1}}^{\top} d_{i-1}(t)) &= 0, \\
W_{y_i}^{\top} W_{x_{i-1}}^{\top} b_{i-1}(x_{i-1},t) + \frac{\mathrm{d}}{\mathrm{d}t}(W_{y_i}^{\top} W_{x_{i-1}}^{\top} d_{i-1}(t)) &= 0.
\end{aligned}
$$

Split $y_{i-1} = P_{y_i}\tilde{y}_i + Q_{y_i} y_i$ and $x_{i-1} = P_{x_i}\tilde{x}_i + Q_{x_i} x_i$ and obtain with the help of Lemma (4.8)

$$
V_{y_i}^{\top} W_{x_{i-1}}^{\top} b_{i-1}(P_{x_i}\tilde{x}_i + Q_{x_i} x_i, P_{y_i}\tilde{y}_i, t) + \frac{\mathrm{d}}{\mathrm{d}t}(V_{y_i}^{\top} W_{x_{i-1}}^{\top} d_{i-1}(t)) = 0, \tag{5.11a}
$$

$$W_{y_i}^{\top} W_{x_{i-1}}^{\top} b_{i-1}(P_{x_i}\tilde{x}_i, t) + \frac{\mathrm{d}}{\mathrm{d}t}(W_{y_i}^{\top} W_{x_{i-1}}^{\top} d_{i-1}(t)) = 0. \tag{5.11b}$$

Equation (5.11b) yields an explicit global expression

$$\tilde{x}_i = \Psi_{\tilde{x}_i}(t)$$

with $\Psi_{\tilde{x}_i}$ being $(\mu + 1 - i)$ times continuously differentiable by Assumption 5.11. Insert this expression into (5.11a) and obtain

$$V_{y_i}^{\top} W_{x_{i-1}}^{\top} b_{i-1}(P_{x_i}\Psi_{\tilde{x}_i}(t) + Q_{x_i}x_i, P_{y_i}\tilde{y}_i, t) + \frac{\mathrm{d}}{\mathrm{d}t}(V_{y_i}^{\top} W_{x_{i-1}}^{\top} d_{i-1}(t)) = 0, \tag{5.12}$$

which analogously yields an explicit global expression

$$\tilde{y}_i = \Psi_{\tilde{y}_i}(x_i, t)$$

with $\Psi_{\tilde{y}_i}$ being $(\mu+1-i)$ times continuously differentiable by Assumption 5.11. Together we achieve

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t} d_{i-1}(x_{i-1}, t) + b_{i-1}(x_{i-1}, y_{i-1}, t) = 0 \\
\Leftrightarrow &\begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t} d_i(x_i, t) + b_i(x_i, y_i, t) & = 0 \\ \tilde{y}_i - \Psi_{\tilde{y}_i}(x_i, t) & = 0 \\ \tilde{x}_i - \Psi_{\tilde{x}_i}(t) & = 0 \end{pmatrix}
\end{aligned}$$

with

$$b_i(x_i, y_i, t) = V_{x_{i-1}}^{\top} b_{i-1}(Q_{x_i}x_i + P_{x_i}\Psi_{\tilde{x}_i}(t), Q_{y_i}y_i + P_{y_i}\Psi_{\tilde{y}_i}(x_i, t), t)$$

and $d_i(x_i, t) := V_{x_{i-1}}^{\top} d_{i-1}(Q_{x_i}x_i + P_{x_i}\Psi_{\tilde{x}_i}(t), t)$. Furthermore we obtain:

$$\begin{aligned}
\frac{\partial}{\partial x_i} d_i &= V_{x_{i-1}}^{\top}(\frac{\partial}{\partial x_{i-1}} d_{i-1}) Q_{x_{i-1}} = V_{x_{i-1}}^{\top} G_{i-1} Q_{x_{i-1}} Q_{x_i} = G_i Q_{x_i}, \\
\frac{\partial}{\partial x_i} b_i &= V_{x_{i-1}}^{\top}((\frac{\partial}{\partial x_{i-1}} b_{i-1}) Q_{x_i} + (\frac{\partial}{\partial y_{i-1}} b_{i-1}) P_{y_i}(\frac{\partial}{\partial x_i}\Psi_{\tilde{y}_i})) \\
&= (V_{x_{i-1}}^{\top} \frac{\partial}{\partial x_{i-1}} b_{i-1}) Q_{x_i} + (V_{x_{i-1}}^{\top} \frac{\partial}{\partial y_{i-1}} b_{i-1}) P_{y_i}(\frac{\partial}{\partial x_i}\Psi_{\tilde{y}_i}) \\
&= B_{x_{i,*}}^{\mathrm{v}} Q_{x_i} - B_{y_i}^{\mathrm{v}} P_{y_i}(V_{y_i}^{\top} B_{y_i}^{\mathrm{w}} P_{y_i})^{-1} V_{y_i}^{\top} B_{x_i}^{\mathrm{w}} Q_{x_i} = B_{x_{i,*}}, \\
\frac{\partial}{\partial y_i} b_i &= (V_{x_{i-1}}^{\top} \frac{\partial}{\partial y_{i-1}} b_{i-1}) Q_{y_i} = B_{y_i}^{\mathrm{v}} Q_{y_i} = B_{y_i}.
\end{aligned}$$

The induction step is completed. Analogous to the previous steps, we obtain under the usage of Lemma 5.6 and Assumption 5.11

$$\frac{\mathrm{d}}{\mathrm{d}t} d_{\mu-1}(x_{\mu-1}, t) + b_{\mu-1}(x_{\mu-1}, y_{\mu-1}, t) = 0$$

$$\Leftrightarrow \begin{pmatrix} V_{x_{\mu-1,*}}^\top(\frac{\mathrm{d}}{\mathrm{d}t}d_{\mu-1}(x_{\mu-1},t)+b_{\mu-1}(x_{\mu-1},y_{\mu-1},t)) & =0 \\ W_{x_{\mu-1,*}}^\top(\frac{\mathrm{d}}{\mathrm{d}t}d_{\mu-1}(x_{\mu-1},t)+b_{\mu-1}(x_{\mu-1},y_{\mu-1},t)) & =0 \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} V_{x_{\mu-1,*}}^\top\frac{\mathrm{d}}{\mathrm{d}t}d_{\mu-1}(x_{\mu-1},t)+V_{x_{\mu-1,*}}^\top b_{\mu-1}(x_{\mu-1},y_{\mu-1},t) & =0 \\ \frac{\mathrm{d}}{\mathrm{d}t}W_{x_{\mu-1,*}}^\top d_{\mu-1}(x_{\mu-1},t)+W_{x_{\mu-1,*}}^\top b_{\mu-1}(x_{\mu-1},t) & =0 \end{pmatrix}$$

$$\Leftrightarrow \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}d_\mu(x_\mu,t)+b_\mu(x_\mu,t)=0 \\ \tilde{y}_\mu=\Psi_{\tilde{y}_\mu}(x_\mu,t,\frac{\mathrm{d}}{\mathrm{d}t}d_{\mu-1}(x_\mu,t)) \end{pmatrix}.$$

□

Now we obtain the global solvability result with the help of Theorem 5.8 and Theorem 5.12.

Theorem 5.13. (Global solvability)
Consider a DAE (5.5), let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ be compact and connected and let $x^0 \in \mathbb{R}^n$ be the initial value of the IVP

$$A\frac{\mathrm{d}}{\mathrm{d}t}d(x(t),t)+b(x(t),t)=0, \quad \forall t\in\mathcal{I}_\star$$
$$x(t_0)=x^0.$$

(i) Let the DAE have a finite Dissection Index μ.

(ii) Let the DAE fulfill the Assumptions 5.10 and 5.11.

(iii) Let b be Lipschitz continuous with respect to x.

Then there is at least one consistent initial value x^0 and for each consistent initial value x^0 there is a unique solution on $\mathcal{I}_\star$.

Proof.
With the help of Theorem 5.12 we can decouple the DAE into

$$\frac{\mathrm{d}}{\mathrm{d}t}d_\mu(x_\mu,t)+b_\mu(x_\mu,t)=0 \tag{5.13a}$$

$$\tilde{y}_\mu=\Psi_{\tilde{y}_\mu}(x_\mu,t,\frac{\mathrm{d}}{\mathrm{d}t}d_{\mu-1}(x_\mu,t)) \tag{5.13b}$$

$$\tilde{y}_i=\Psi_{\tilde{y}_i}(x_i,t), \qquad 1\leqslant i\leqslant \mu-1 \tag{5.13c}$$

$$\tilde{x}_j=\Psi_{\tilde{x}_j}(t), \qquad 1\leqslant j\leqslant \mu-1 \tag{5.13d}$$

with $\frac{\partial}{\partial x_\mu}d_\mu(x_\mu,t)=(W_y^*)^\top G_{\mu-1}(x^1,x,t)Q_{x_{\mu-1}}$ being strongly monotone and consequently non-singular and

$$x=T\begin{pmatrix}x_\mu & \tilde{x}_1 & \dots & \tilde{x}_{\mu-1} & \tilde{y}_1 & \dots & \tilde{y}_\mu\end{pmatrix}^\top$$

with T being non-singular. We mainly have to show that the Lipschitz continuity of b is inherited by b_μ. Therefore we proof by an induction that b_i is Lipschitz continuous with respect to x_i and y_i and that $\Psi_{\tilde{y}_i}$ is Lipschitz continuous with respect to y_i.
Base case: $(i = 1)$
With b being Lipschitz continuous and P and Q being constant, we obtain the Lipschitz continuity of

$$b_0(x_0, y_0, t) = b(Px_0 + Qy_0, t)$$

with respect to x_0 and y_0. We remember Equation (5.10b):

$$V_{y_1}^\top W_{x_0}^\top b_0(P_{x_1}\Psi_{\tilde{x}_1}(t) + Q_{x_1}x_1, P_{y_1}\tilde{y}_1, t) = 0.$$

The function on the left hand side is Lipschitz continuous with respect to x_1 while it is strongly monotone with respect to $\tilde{y}_1$. By Lemma 5.6 the global solution function $\Psi_{\tilde{y}_1}(x_1, t)$ is Lipschitz continuous with respect to x_1. Due to the constant basis functions we achieve the Lipschitz continuity of b_1 with respect to x_1 and y_1, since b_1 is a composition of Lipschitz continuous functions:

$$b_1(x_1, y_1, t) = V_{x_0}^\top b_0(Q_{x_1}x_1 + P_{x_1}\Psi_{\tilde{x}_1}(t), Q_{y_1}y_1 + P_{y_1}\Psi_{\tilde{y}_1}(x_1, t), t).$$

Induction step: $(i - 1 \mapsto i \leqslant \mu - 1)$
For the induction step we remember Equation (5.12):

$$V_{y_i}^\top W_{x_{i-1}}^\top b_{i-1}(P_{x_i}\Psi_{\tilde{x}_i}(t) + Q_{x_i}x_i, P_{y_i}\tilde{y}_i, t) + \frac{\mathrm{d}}{\mathrm{d}t}(V_{y_i}^\top W_{x_{i-1}}^\top d_{i-1}(t)) = 0.$$

Again the function on the left hand side is Lipschitz continuous with respect to x_i while it is strongly monotone with respect to $\tilde{y}_i$. Hence, $\Psi_{\tilde{y}_i}(x_i, t)$ is Lipschitz continuous with respect to x_i and therefore

$$b_i(x_i, y_i, t) = V_{x_{i-1}}^\top b_{i-1}(Q_{x_i}x_i + P_{x_i}\Psi_{\tilde{x}_i}(t), Q_{y_i}y_i + P_{y_i}\Psi_{\tilde{y}_i}(x_i, t), t)$$

is Lipschitz continuous with respect to x_i and y_i as a composition of Lipschitz continuous functions. The induction is concluded.

At last we obtain the Lipschitz continuity of b_μ, since the alternative basis ending is constant:

$$b_\mu(x_\mu, t) = W_{x_{\mu-1,*}}^\top b_{\mu-1}(x_{\mu-1}, t) = W_{x_{\mu-1,*}}^\top b_{\mu-1}(x_{\mu-1}, y_{\mu-1}, t).$$

For every x_μ we obtain a unique global solution on a fixed interval of the implicit ODE (5.13a) by Theorem 5.8. Inserting this solution and the Equations (5.13d) into the Equations (5.13b) and (5.13c) concludes the proof. □

We generalize the solvability result by relaxing the Assumption 5.11.

Assumption 5.14. Consider a DAE (5.5) with a semi-properly stated derivative term and a finite Dissection Index μ. Let the DAE fulfill Assumption 5.10. For $1 \leqslant i \leqslant \mu - 1$ assume that the matrix valued functions:

$$\begin{aligned} V_{y_i}^\top B_{y_i}^{\mathrm{w}}(x,t)P_{y_i} \quad &\text{and} \quad W_{y_i}^\top B_{x_i}^{\mathrm{w}}(x,t)P_{x_i}, \\ (W_y^*)^\top G_{\mu-1}(x,t)Q_{x_{\mu-1}} \quad &\text{and} \quad (V_y^*)^\top B_{y_{\mu-1}}^{\mathrm{v}}(x,t)Q_{y_{\mu-1}} \end{aligned}$$

fulfill the assumptions of the Jacobian in Lemma 5.7. Furthermore assume that b is continuously differentiable and that:

$$\begin{aligned} W_{x_{i-1}}^\top V_{x_{i-2}}^\top \dots V_{x_0}^\top b(x,t), \quad &\text{for } i = 1, \dots, \mu - 1 \\ W_{x_i}^\top V_{x_{i-1}}^\top \dots V_{x_0}^\top Ad(x,t), \quad &\text{for } i = 1, \dots, \mu - 2 \end{aligned}$$

are $(\mu - i)$ times continuously differentiable with $V_{x_{i-2}}^\top \dots V_{x_0}^\top := I$ for $i = 1$.

We notice that a strong monotone Jacobian fulfills the assumptions in Lemma 5.7. Hence Assumption 5.11 is stricter than Assumption 5.14. With the help of Assumption 5.14 we formulate the following corollary:

Corollary 5.15. Consider a DAE (5.5), let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$ be compact and connected and let $x^0 \in \mathbb{R}^n$ be the initial value of the IVP

$$\begin{aligned} A\frac{\mathrm{d}}{\mathrm{d}t}d(x(t),t) + b(x(t),t) = 0, \quad \forall t \in \mathcal{I}_\star \\ x(t_0) = x^0. \end{aligned}$$

(i) Let the DAE have a finite Dissection Index μ.

(ii) Let the DAE fulfill the Assumptions 5.10 and 5.14.

(iii) Let b be Lipschitz continuous with respect to x.

Then there is at least one consistent initial value x^0 and for each consistent initial value x^0 there is a unique solution on $\mathcal{I}_\star$.

To prove Corollary 5.15 we would need to mimic the proof of Theorem 5.12 and 5.13 under Assumption 5.14 instead of Assumption 5.11.

5.4 Circuit Equations

In this section we apply Theorem 5.13 to the equations of the circuit applications of Section 3.1. First we write the extended MNA equations (3.34) as a semi-linear DAE. Therefore we use the matrix and the function

$$A := \begin{pmatrix} A_{\mathcal{C}} & 0 & 0 & 0 & 0 & 0 \\ 0 & I & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix} \quad \text{and} \quad d(x,t) := \begin{pmatrix} q_{\mathcal{C}}(A_{\mathcal{C}}^\top e, t) \\ \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) \\ \zeta \\ \phi_M(q_M, t) \\ E \\ J \end{pmatrix}$$

from Section 4.3 and define

$$b(x,t) := \begin{pmatrix} A_{\mathcal{C}} g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) + A_{\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t) + A_{\mathcal{L}} j_{\mathcal{L}} + A_V j_V + A_I i_s(t) \\ -A_{\mathcal{L}}^\top e + \chi_{\mathcal{L}} E \\ A_V^\top e - v_s(t) \\ h_\zeta(A_S^\top e, \zeta, \Psi) \\ T\Psi(t) - h_\Psi(\zeta) \\ -A_M^T e \\ M_\sigma E - J - \chi_{\mathcal{L}}^T j_{\mathcal{L}} \\ M_{CC} E \end{pmatrix}$$

with the variables $x = \begin{pmatrix} e & j_{\mathcal{L}} & j_V & \zeta & \Psi & q_M & E & J \end{pmatrix}^\top$. This enables us to write the extended MNA in the form $Ad'(x,t) + b(x,t) = 0$. To prove the global solvability of the extended MNA we assume the capacitors, inductors and resistors to be globally passive.

Assumption 5.16 (Global passivity of the simple elements)**.**
The element relations q_C, ϕ_L and g_R are strongly monotone with respect to their first arguments and g_R is Lipschitz continuous with respect to its first argument.

The matrices C_S and L_E are positive definite for the models of the semiconductor and electromagnetic devices investigated in Section 3. Therefore they are also strongly monotone since these matrices are constant. Hence the functions

$$q_{\mathcal{C}}(A_{\mathcal{C}}^\top e, t) = \begin{pmatrix} q_C(A_C^\top e, t) \\ C_S A_S^\top e \end{pmatrix} \quad \text{and} \quad \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) = \begin{pmatrix} \phi_L(j_L, t) \\ L_E j_E \end{pmatrix}$$

are strongly monotone with respect to their first argument if we assume 5.16. The functions ϕ_M and g_M are not strongly monotone with respect to their first argument, respectively, for the example given in Section 3.1.3. Furthermore the functions

$$g_M(A_M^\top e, q_M, t), \quad g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) \quad \text{and} \quad h_\zeta(A_S^\top e, \zeta, \Psi)$$

are not Lipschitz continuous in general. At the first glance this seems to be a problem since the strong monotonicity and the Lipschitz continuity are needed for the global solvability of the extended MNA. But since these functions only describe the physical behavior correctly inside a physically reasonable domain, it does not weaken the models if we alter these functions outside this region. The same strategy is used in [Bar04] with respect to the temperature of a thermal resistor and in [JMT13] for circuits including memristor. By the next two lemmata we introduce two cut-off strategies.

Lemma 5.17. (Lipschitz cut-off)
Let $k \geqslant 1$ and define the box $Q = [a_i, b_i]^n$ with $a_i < b_i$ for $1 \leqslant i \leqslant n$. Consider a function $f \in C^k(\mathbb{R}^n, \mathbb{R}^m)$ which is not Lipschitz continuous. Then there exists a function $\tilde{f} \in C^k(\mathbb{R}^n, \mathbb{R}^m)$ with $\tilde{f}$ being Lipschitz continuous and $\tilde{f}(x) = f(x)$ for all $x \in Q$.

Proof. Let there be a small real number $\delta > 0$ and a slightly bigger box $Q_\delta = [a_i - \delta, b_i + \delta]^n$. We define the auxiliary functions

$$g(t) = \begin{cases} e^{-\frac{1}{t^2}}, & \text{for } t > 0 \\ 0, & \text{for } t \leqslant 0 \end{cases}$$

and

$$k(t) = \frac{g(1+t)}{g(1+t) + g(1-t)} \quad \Rightarrow \quad \begin{cases} k(t) = 0, & \text{for } t \leqslant -1 \\ 0 \leqslant k(t) \leqslant 1, & \text{for } -1 \leqslant t \leqslant 1 \\ k(t) = 1, & \text{for } t \geqslant 1. \end{cases}$$

Additionally we define the functions

$$h_{i1}(t) = \frac{2}{\delta}(t - a_i) + 1 \quad \Rightarrow \quad \begin{cases} h_{i1}(t) \leqslant -1, & \text{for } t \leqslant a_i - \delta \\ h_{i1}(t) \geqslant 1, & \text{for } t \geqslant a_i \end{cases}$$

and

$$h_{i2}(t) = \frac{2}{\delta}(t - b_i - \delta) + 1 \quad \Rightarrow \quad \begin{cases} h_{i2}(t) \leqslant -1, & \text{for } t \leqslant b_i \\ h_{i2}(t) \geqslant 1, & \text{for } t \geqslant b_i + \delta. \end{cases}$$

Combining these functions we define

$$k_{i1}(t) = k(h_{i1}(t)) \quad \Rightarrow \quad \begin{cases} k_{i1}(t) = 0, & \text{for } t \leqslant a_i - \delta \\ 0 \leqslant k_{i1}(t) \leqslant 1, & \text{for } a_i - \delta < t < a_i \\ k_{i1}(t) = 1, & \text{for } t \geqslant a_i \end{cases}$$

and

$$k_{i2}(t) = k(h_{i2}(t)) \quad \Rightarrow \quad \begin{cases} k_{i2}(t) = 0, & \text{for } t \leqslant b_i \\ 0 \leqslant k_{i2}(t) \leqslant 1, & \text{for } b_i < t < b_i + \delta \\ k_{i2}(t) = 1, & \text{for } t \geqslant b_i + \delta \end{cases}$$

with $k_{i1}, k_{i2} \in C^\infty(\mathbb{R}, \mathbb{R})$ since both k_{i1} and k_{i2} are combinations of C^∞ functions. Now we can notate our cut-off function component-wise

$$l_i(x_i) := (a_i - \delta)(1 - k_{i1}(x_i)) + x_i k_{i1}(x_i)(1 - k_{i2}(x_i)) + (b_i + \delta)k_{i2}(x_i)$$

with

$$\begin{aligned} l_i(x_i) &= a_i - \delta, & &\text{for } x_i \leqslant a_i - \delta \\ l_i(x_i) &= x_i, & &\text{for } a_i < x_i < b_i \\ l_i(x_i) &= b_i + \delta, & &\text{for } x_i \geqslant b_i + \delta. \end{aligned}$$

This yields for the complete cut-off function that $l \in C^\infty(\mathbb{R}^n, \mathbb{R}^n)$, $l(x) = x$ for all $x \in Q$ and $l(x) \in Q_\delta$ for all $x \in \mathbb{R}$.
We define $\tilde{f}(x) := f(l(x))$ and it directly follows that $\tilde{f}(x) = f(x)$ for all $x \in Q$. The Lipschitz continuity follows if we consider the following: The Jacobian of l has a compact supporter since l is constant except for a compact region. Therefore there is a constant L such that $\max_{x\in\mathbb{R}^n} \left\| \frac{\partial}{\partial x} l(x) \right\| \leqslant L$. And so it holds:

$$\begin{aligned} \left\| \frac{\mathrm{d}}{\mathrm{d}x} \tilde{f}(x) \right\| &= \left\| \frac{\mathrm{d}}{\mathrm{d}x} f(l(x)) \right\| \\ &= \left\| \frac{\partial}{\partial x} f(l(x)) \frac{\partial}{\partial x} l(x) \right\| \\ &\leqslant \left\| \frac{\partial}{\partial x} f(l(x)) \right\| \left\| \frac{\partial}{\partial x} l(x) \right\| \\ &\leqslant \max_{x \in Q_\delta} \left\| \frac{\partial}{\partial x} f(x) \right\| L. \end{aligned}$$

Therefore we can create a Lipschitz continuous auxiliary function which is identical to the original function on a compact region. □

In the case of the function $g_{\mathcal{R}}$ we need a cut-off strategy which creates Lipschitz continuity while it preserves strong monotonicity. The following lemma provides such a cut-off strategy.

Lemma 5.18. (Monotonicity preserving cut-off)
Consider a function $M : \mathbb{R} \to \mathbb{R}$ with $M_1, M_2 > 0$ such that $M_1 \leqslant M(y) \leqslant M_2$ and $|\frac{\partial}{\partial y} M(y)| \leqslant M_3$ for all $y \in \mathbb{R}$ and an interval $Q = [-c, c]$ with $c > 0$. The function $g(x, y) := M(y)x$ is both strongly monotone and Lipschitz continuous in x, but not Lipschitz continuous in y. Then there exists a function $\tilde{g}$ which is Lipschitz continuous in y while it preserves the strong monotonicity and the Lipschitz continuity in x while it fulfills $\tilde{g}(x, y) = g(x, y)$ for all $x \in Q$.

Proof.
Let $\delta > 0$, $a = -c - \delta$ and $b = c + \delta$. We define two functions k_1 and k_2 as before depending on a and b. Then we define

$$\tilde{g}(x, y) := (M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)))x,$$

for $x \in Q$ it holds

$$\begin{aligned}\tilde{g}(x, y) &= (M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)))x \\ &= (M_2 \cdot 0 + M(y) \cdot 1 \cdot 1)x \\ &= g(x, y).\end{aligned}$$

Hence $\tilde{g}$ is a suitable auxiliary function if we are only interested in g for $x \in Q$. First we check preservation of the strong monotonicity and the Lipschitz continuity with respect to x. Therefore we calculate the partial derivative by using $k_2(x) = 0$ for all x with $k_1'(x) \neq 0$ and $k_1(x) = 1$ for all x with $k_2'(x) \neq 0$:

$$\begin{aligned}\frac{\partial}{\partial x}\tilde{g}(x, y) =& M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)) \\ &+ (M_2(k_2'(x) - k_1'(x)) + M(y)k_1'(x)(1 - k_2(x)) - M(y)k_1(x)k_2'(x))x \\ =& M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)) \\ &+ (M_2(k_2'(x) - k_1'(x)) - M(y)(k_2'(x) - k_1'(x)))x \\ =& M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)) \\ &+ (M_2 - M(y))(k_2'(x) - k_1'(x))x.\end{aligned}$$

We obtain two inequalities

$$\begin{aligned}M_1 &\leqslant M_1((1 - k_1(x))k_2(x) + 1) \\ &= M_1(k_2(x) + 1 - k_1(x) + k_1(x)(1 - k_2(x))) \\ &\leqslant M_2(k_2(x) + 1 - k_1(x)) + M(y)k_1(x)(1 - k_2(x)) \\ &\leqslant M_2(k_2(x) + 1 - k_1(x) + k_1(x)(1 - k_2(x))) \\ &= M_2((1 - k_1(x))k_2(x) + 1) \\ &\leqslant 2M_2\end{aligned}$$

and

$$0 \leqslant (M_2 - M(y))(k_2'(x) - k_1'(x))x \leqslant M_4$$

with the help of the description

$$(M_2 - M(y))(k_2'(x) - k_1'(x))x = \begin{cases} -(M_2 - M(y))k_1'(x)x & , x \in [-c-\delta, -c], \\ (M_2 - M(y))k_2'(x)x & , x \in [c, c+\delta], \\ 0 & , \text{ else.} \end{cases}$$

This leads to

$$M_1 \leqslant \frac{\partial}{\partial x}\tilde{g}(x,y) \leqslant 2M_2 + M_4$$

which yields the strong monotonicity and the Lipschitz continuity since M_1, M_2 and M_4 are positive. The Lipschitz continuity with respect to y follows by

$$|\frac{\partial}{\partial y}\tilde{g}(x,y)| = |\frac{\partial}{\partial y}M(y)k_1(x)(1-k_2(x))x| = |\frac{\partial}{\partial y}M(y)||k_1(x)(1-k_2(x))x| \leqslant M_3(c+\delta).$$

□

We use the cut-off strategy of Lemma 5.17 to obtain substitute functions for

$$g_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e, \zeta, \Psi) \quad \text{and} \quad h_{\zeta}(A_{\mathcal{S}}^{\top}e, \zeta, \Psi).$$

Therefore we have to choose reasonable compact regions for $A_{\mathcal{C}}^{\top}e$, ζ and Ψ. We start with ζ which represents the electron densities and hole densities in the semiconductor material. Since the exact solutions of these densities are never negative we choose $a_i = 0$ for all a_i belonging to ζ. Depending on the material coefficients of the semiconductor and the oxide, we have to choose the absolute values of the remaining a_i and b_i sufficiently large. It is possible to choose these boundaries sufficiently large since in reality the semiconductor device only works properly if it is neither too hot nor too cold. Therefore the functions $g_{\mathcal{C}}$ and h_{ζ} represent the real behavior of the semiconductor only if $A_{\mathcal{C}}^{\top}e$ and Ψ do not grow too large and so it does not matter if we cut-off the function after this point.

For the memristor function g_M we provide a specific auxiliary function for the HP memristor of [SSSW08]. The following steps can also be found in [JMT13]. But here we present them more detailed, in particular we use Lemma 5.18. In this case we have $g_M(u,q) = M(q)^{-1}u$ with the Jacobian $M(q) = R_{\mathit{off}}(1 - \frac{\mu_V R_{on}}{d^2}q)$. We introduce the parameter $\alpha = 1 - \frac{1}{360}$ and the cut-off function $l_M(q)$ with $a = -\alpha\frac{d^2}{\mu_V R_{on}}$ and $b = \alpha\frac{d^2}{\mu_V R_{on}}$ which is constructed as in Lemma 5.17. With the help of $l_M(q)$ we are able to define $\tilde{M}(q) = M(l_M(q)) = R_{\mathit{off}}(1 - \frac{\mu_V R_{on}}{d^2}l_M(q))$. Then $\tilde{M}(q) = M(q)$ as long as $R_{on} \leqslant M(q) \leqslant R_{\mathit{off}}$ and there are $M_1, M_2 > 0$ with $M_1 \leqslant \tilde{M}(q) \leqslant M_2$ and $|\frac{\partial}{\partial q}\tilde{M}(q)| = R_{\mathit{off}} \frac{\mu_V R_{on}}{d^2} |\frac{\partial}{\partial q} l_M(q)| \leqslant M_3$. This leads to the functions

$$\tilde{\phi}_M(q) = \int_0^q \tilde{M}(p)\mathrm{d}p \quad \text{and} \quad \tilde{g}_M(u,q) = \tilde{M}(q)^{-1}u$$

with $\tilde{\phi}_M(q)$ being strongly monotone and $\tilde{g}_M(u,q)$ being strongly monotone and Lipschitz continuous in u. By Lemma 5.18 we obtain a function $\bar{g}_M(u,q) = \bar{M}(u,q)^{-1}u$ which is both strongly monotone and Lipschitz continuous in u and Lipschitz continuous in q and coincides with $\tilde{\phi}_M(q)$ for u in an arbitrary but fixed box.
Hence, we are able to provide auxiliary functions for

$$\phi_M(q_M,t), \quad g_M(A_M^\top e, q_M, t), \quad g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) \quad \text{and} \quad h_\zeta(A_{\mathcal{S}}^\top e, \zeta, \Psi)$$

with ϕ_M and g_M being strongly monotone with respect to their first argument and g_M, $g_{\mathcal{C}}$ and h_ζ being Lipschitz continuous in all their arguments except the time t. Therefore we assume without any loss of generality:

Assumption 5.19 (Global passivity)**.**
The functions $q_{\mathcal{C}}$, $\phi_{\mathcal{L}}$, ϕ_M and $g_{\mathcal{R}}$ are strongly monotone with respect to their first arguments. Furthermore the functions $g_{\mathcal{C}}$, $g_{\mathcal{R}}$, h_ζ and h_Ψ are Lipschitz continuous with respect to all their arguments but the time.

Under this assumption we formulate the following theorem.

Theorem 5.20. Let Assumption 5.19 be fulfilled and let $v_s(t)$ and $i_s(t)$ be continuously differentiable. Then the extended MNA equations (3.34) have at least one initial value and for each initial value there is a global unique solution.

Proof. We have to check the requirements of Theorem 5.13. By the results of Section 4.3 we already know that the extended MNA equations have a Dissection Index two or lower with a constant basis chain. Therefore requirement (i) is fulfilled. To check requirement (ii) we choose the basis functions

$$P_{y_1} = \begin{pmatrix} 0 & P_V & Q_V P_{\mathcal{R}} & 0 \\ 0 & 0 & 0 & V_V \\ I & 0 & 0 & 0 \end{pmatrix}, \quad V_{y_1} = \begin{pmatrix} P_V & Q_V P_{\mathcal{R}} & 0 & 0 \\ 0 & 0 & V_V & 0 \\ 0 & 0 & 0 & I \end{pmatrix}$$

and

$$P_{x_1} = V_y^* = \begin{pmatrix} P_{\mathcal{C}V} & 0 \\ 0 & P_{\mathcal{L}I} \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$$

to obtain

$$V_{y_1}^\top B^{\mathrm{w}}_{y_1}(x,t)P_{y_1} = \begin{pmatrix} 0 & P_V^\top A_{\bar{\mathcal{C}}\mathcal{R}} G_{\mathcal{R}} A^\top_{\bar{\mathcal{C}}\mathcal{R}} P_V & P_V^\top A_{\bar{\mathcal{C}}\mathcal{R}} G_{\mathcal{R}} A^\top_{\bar{\mathcal{C}}\bar{V}\mathcal{R}} P_{\mathcal{R}} & P_V^\top A_{\bar{\mathcal{C}}V} V_V \\ 0 & P_{\mathcal{R}}^\top A_{\bar{\mathcal{C}}\bar{V}\mathcal{R}} G_{\mathcal{R}} A^\top_{\bar{\mathcal{C}}\mathcal{R}} P_V & P_{\mathcal{R}}^\top A_{\bar{\mathcal{C}}\bar{V}\mathcal{R}} G_{\mathcal{R}} A^\top_{\bar{\mathcal{C}}\bar{V}\mathcal{R}} P_{\mathcal{R}} & 0 \\ 0 & -V_V^\top A^\top_{\bar{\mathcal{C}}V} P_V & 0 & 0 \\ T & 0 & 0 & 0 \end{pmatrix},$$

$$W_{y_1}^\top B^{\mathrm{v}}_{x_1}(x,t)P_{x_1} = \begin{pmatrix} 0 & A_{\bar{\mathcal{C}}\bar{V}\bar{\mathcal{R}}\mathcal{L}} P_{\mathcal{L}I} \\ -W_V^\top A_V^\top P_{\mathcal{C}} P_{\mathcal{C}V} & 0 \end{pmatrix}$$

$$(V_y^*)^\top B^{\mathrm{v}}_{y_1}(x,t)Q_{y_1} = \begin{pmatrix} 0 & P_{\mathcal{C}V}^\top P_{\mathcal{C}}^\top A_V W_V \\ -P_{\mathcal{L}I}^\top A^\top_{\bar{\mathcal{C}}\bar{V}\bar{\mathcal{R}}\mathcal{L}} & 0 \end{pmatrix},$$

and

$$\begin{aligned}&(W_y^*)^\top G_1(x,t)Q_{x_1} \\ &= \begin{pmatrix} Q_{\mathcal{C}V}^\top P_{\mathcal{C}}^\top A_{\mathcal{C}} \mathcal{C}(A_{\mathcal{C}}^\top e,t) A_{\mathcal{C}}^\top P_{\mathcal{C}} Q_{\mathcal{C}V} & 0 & 0 & 0 & 0 & 0 \\ 0 & Q_{\mathcal{L}I}^T \mathcal{L}(j_{\mathcal{L}},t) Q_{\mathcal{L}I} & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & M(q_M,t) & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}.\end{aligned}$$

The matrices $W_{y_1}^\top B^{\mathrm{v}}_{x_1}(x,t)P_{x_1}$ and $(V_y^*)^\top B^{\mathrm{v}}_{y_1}(x,t)Q_{y_1}$ are constant hence they fulfill the requirements of Lemma 5.7 and $(W_y^*)^\top G_1(x,t)Q_{x_1}$ fulfills these requirements since it is strongly monotone. The matrix $V_{y_1}^\top B^{\mathrm{w}}_{y_1}(x,t)P_{y_1}$ fulfills the requirements of Lemma 5.7 since $G_{\mathcal{R}}$ is strongly monotone and Lipschitz continuous and

$$\begin{pmatrix} 0 & -V_V^\top A^\top_{\bar{\mathcal{C}}V} P_V \\ T & 0 \end{pmatrix} \text{ and } P_V^\top A_{\bar{\mathcal{C}}V} V_V$$

are constant.
The third condition, namely the Lipschitz continuity of b, follows directly from Assumption 5.19 since b is a composition of Lipschitz continuous functions. □

5.5 Summary and Outlook

In this chapter we provided sufficient criteria for the global unique solvability of semilinear DAEs. This has been done in two major steps. First we derived criteria for the solvability of an implicit ODE. Afterwards we showed under which assumption it is possible to decouple a semi-linear DAE such that the implicit inherent ODE fulfills these criteria. We emphasize at this point that the solvability results in this chapter are not limited to index 1 DAEs but hold for DAEs with an arbitrary index with the monotonicity properties proposed in Assumption 5.11.

One important tool for the decoupling is the concept of the strong monotonicity. To use the concept of strong monotonicity to obtain global solvability results for DAEs, or abstract DAEs, was introduced by Michael Matthes in [Mat13]. The Dissection concept cooperates very well with the strong monotonicity concept since it is able to preserve the strong monotonicity during the decoupling. In [JMT13] this cooperation has already been used for the index 1 circuits including memristors.

The results regarding the circuit application in this chapter can be seen as a generalization of the results in [JMT13]. Here we showed the global unique solvability of index 2 circuits including a semiconductor model, memristors and an electromagnetic model.

6 Convergence Analysis

This chapter deals with DAE related convergence issues. In Section 2.1 we have already witnessed that classical ODE methods, like the implicit Euler, may fail when applied to DAEs. This problem is presented by Example 2.12 which also plays an important role in this section. In particular we discuss the influence of the terms $Q'_{x_0}(X^2, t)$ and $Q'_{x_i}(X^{i+1}, t)$ which appear in the matrix chain of the Dissection Index. These derivative terms may lead to numerical instabilities which are the main topic of this chapter.
After presenting more examples for the problems caused by the derivative terms we provide sufficient criteria to avoids these problems. In the final section of this chapter we show that there is a class of collocation methods which are unharmed by these instabilities. In [KM07] these convergence issues are also tackled. The strategy in [KM07] is to use a new stabilization method which is used together with the classical integration methods to obtain convergence. The combination of the methods consists of three steps: In every integration step the DAE is transformed by the stabilization method, then the classical integration method is applied and afterwards the transformation is undone again by the stabilization method. In contrast to [KM07] we are interested in methods which do not need a transformation of the DAE.

Regularity and Characteristic Values

We consider again Example 2.12 from Section 2.1: Let $\mathcal{I} := [0, 3]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} x_1' + \eta t x_2' &= -(1 + \eta) x_2 \\ x_1 + \eta t x_2 &= e^{-t} \end{aligned}$$

with $\eta \in \mathbb{R}$. We compute a basis chain of Example 2.12, before we further discuss its numerical problems. We denote

$$A = \begin{pmatrix} 1 & \eta t \\ 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 1 + \eta \\ 1 & \eta t \end{pmatrix}.$$

Therefore we can choose

$$P = \begin{pmatrix} 1 \\ \eta t \end{pmatrix}, \quad Q = \begin{pmatrix} -\eta t \\ 1 \end{pmatrix}, \quad V = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

and gain

$$G_1 = \big(1 + (\eta t)^2\big)\,, \quad B^{\mathrm{w}}_{x_1} = \big(1 + (\eta t)^2\big)\,, \quad B^{\mathrm{w}}_{y_1} = \big(0\big)$$

and

$$B^{\mathrm{v}}_{y_1} = V^\top B(t)Q(t) + V^\top A(t)Q'(t) = \big(1+\eta\big) + \big(-\eta\big) = \big(1\big)\,.$$

Hence, the DAE has Dissection Index 2. We notice that the derivative term $V^\top A(t)Q'(t)$ is needed for reflecting the regularity of the DAE correctly for $\eta = -1$.
In Section 2.1 it is mentioned that the numerical solution provided by the implicit Euler does not converge to the exact solution

$$x_1(t) = (1 - \eta t)e^{-t}, \quad x_2(t) = e^{-t}$$

of Example 2.12 if $\eta < -\frac{1}{2}$. In particular the implicit Euler is not able to provide any numerical solution values if $\eta = -1$. In this case we obtain the equations

$$\begin{aligned} x_1' - t x_2' &= 0 \\ x_1 - t x_2 &= e^{-t}, \end{aligned}$$

which leads to the discretized system

$$\begin{aligned} x_{1,n} - t_n x_{2,n} &= x_{1,n-1} - t_n x_{2,n-1} \\ x_{1,n} - t_n x_{2,n} &= e^{-t_n}. \end{aligned}$$

This system is not solvable with respect to $x_{1,n}$ and $x_{2,n}$. It seems that the implicit Euler does not put the derivative term $V^\top A(t)Q'(t)$ into action, which in this case leads to an singular system. The next example has the same problem, but this time the problem is more hidden.

Example 6.1.
Let $\mathcal{I} := [0,3]$ and let $t \in \mathcal{I}$.

$$\begin{aligned} sin(t)x_1' + cos(t)x_2' &= -x_3 \\ x_3' &= -cos(t)x_1 + sin(t)x_2 \\ 0 &= 1 - sin(t)x_1 - cos(t)x_2 \end{aligned}$$

The exact solution of the problem is given by $x_1(t) = sin(t), x_2(t) = cos(t)$ and $x_3(t) = 0$.

Again we provide a basis chain of the Dissection Index first. We denote

$$A = \begin{pmatrix} sin(t) & cos(t) & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 & 0 & 1 \\ cos(t) & -sin(t) & 0 \\ sin(t) & cos(t) & 0 \end{pmatrix}.$$

Therefore we can choose

$$P = \begin{pmatrix} sin(t) & 0 \\ cos(t) & 0 \\ 0 & 1 \end{pmatrix}, \quad Q = \begin{pmatrix} cos(t) \\ -sin(t) \\ 0 \end{pmatrix}, \quad V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad W = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

and gain

$$G_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad B^{\mathrm{w}}_{x_1} = \begin{pmatrix} 1 & 0 \end{pmatrix}, \quad B^{\mathrm{w}}_{y_1} = \begin{pmatrix} 0 \end{pmatrix}$$

and

$$B^{\mathrm{v}}_{y_1} = V^\top B(t)Q(t) + V^\top A(t)Q'(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} + \begin{pmatrix} -1 \\ 0 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \end{pmatrix}.$$

Due to the matrices $B^{\mathrm{w}}_{y_1}$, $B^{\mathrm{w}}_{x_1}$ we can choose $Q_{x_1} = \begin{pmatrix} 0 & 1 \end{pmatrix}^\top$. After calculating $G_1 Q_{x_1} = \begin{pmatrix} 0 & 1 \end{pmatrix}^\top$ we can choose $W_{x_1} = \begin{pmatrix} 1 & 0 \end{pmatrix}^\top$ and obtain $B^{\mathrm{w}}_{y_2} = \begin{pmatrix} 1 \end{pmatrix}$. Hence, the DAE has Dissection Index 2. We notice again the derivative term $V^\top A(t)Q'(t)$ is crucial for the DAE to have index 2. If we would calculate the matrix chain without the derivative term, the chain would end after the third stage instead of the second. We calculate

$$\tilde{B}^{\mathrm{v}}_{y_1} = V^\top B(t)Q(t) = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{and} \quad \tilde{B}^{\mathrm{v}}_{x_1} = V^\top B(t)P(t) = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$$

instead of $B^{\mathrm{v}}_{y_1}$ and $B^{\mathrm{v}}_{x_1}$. Then we obtain

$$\tilde{G}_2 = \begin{pmatrix} 1 \end{pmatrix} \quad \text{and} \quad \tilde{B}^{\mathrm{v}}_{y_2} = V^\top_{x_1} \tilde{B}^{\mathrm{v}}_{y_1} Q_{y_1} = \begin{pmatrix} 1 \end{pmatrix}$$

which leads to

$$\tilde{B}^{\mathrm{w}}_{y_2} = W^\top_{x_1} \tilde{B}^{\mathrm{v}}_{y_1} Q_{y_1} = \begin{pmatrix} 0 \end{pmatrix} \quad \text{and} \quad \tilde{B}^{\mathrm{w}}_{x_2} = W^\top_{x_1} \tilde{B}^{\mathrm{v}}_{x_1} Q_{x_1} = \begin{pmatrix} 1 \end{pmatrix}.$$

We try to simulate Example 6.1 with the implicit Euler. Though the discretized system is regular, simulating this DAE by the implicit Euler does not provide satisfying results, see Figure 6.1. From these examples we deduce that the implicit Euler is not suitable for integrating DAEs with characteristic values depending on the derivative term $V^\top A(t)Q'(t)$.

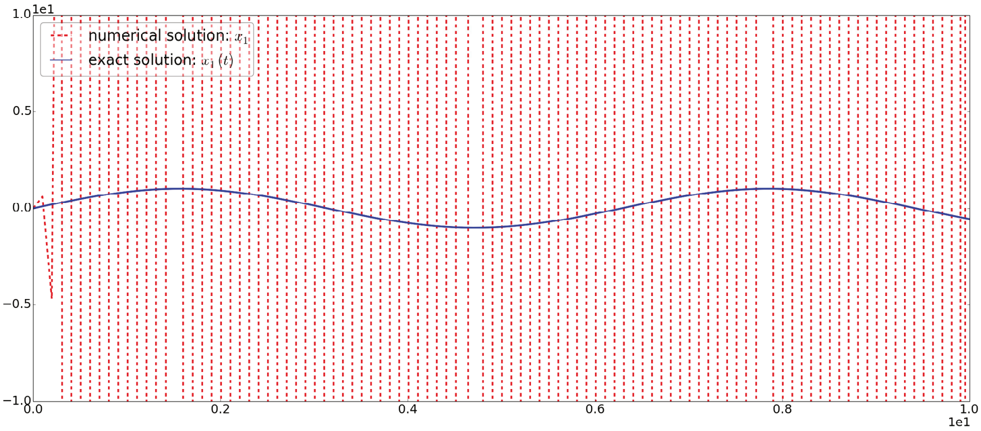


Figure 6.1: Numerical and exact solution of Example 6.1 simulated with the implicit Euler and a time step size $h = 0.1$. The magnitude of the numerical solution is around 10^{100}.

This problem invokes two questions: Is there for each DAE a numerical method which preserves the index behavior during the discretization? Or is there even a numerical method which preserves the index behavior during the discretization for each DAE?

Artificial Dynamics

Example 2.12 in Section 2.1 shows us that the pure appearance of the derivative term $V^{\top}A(t)Q'(t)$ in the matrix chain is enough for the numerical solution to diverge. For $\eta = -0.55$ neither the index nor the characteristic values of Example 2.12 depend on the derivative terms but the numerical solution, generated by the implicit Euler, grow unbounded as the time step size decreases, see Figure 6.2.

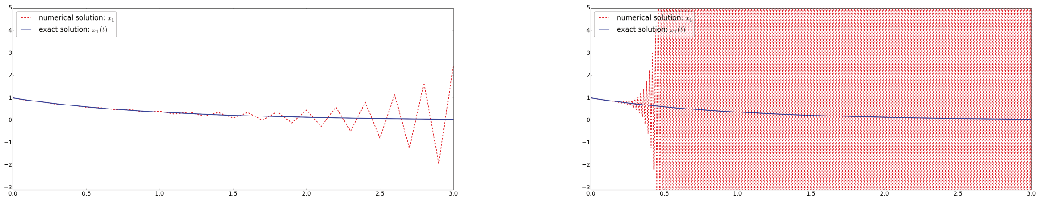


Figure 6.2: Numerical and exact solution of Example 2.12 simulated with the implicit Euler and the time step sizes $h = 0.1$ (left) and $h = 0.01$ (right).

This problem is not confined to the implicit Euler method. Also the BDF2 Method

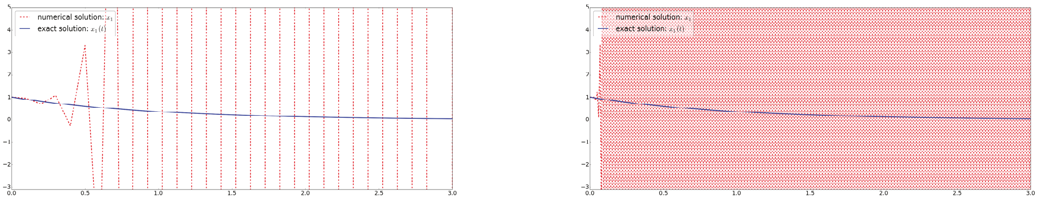


Figure 6.3: Numerical and exact solution of Example 2.12 simulated with the BDF2 method and the time step sizes $h = 0.1$ (left) and $h = 0.01$ (right).

as well as the RadauIIA Method show the same problem.

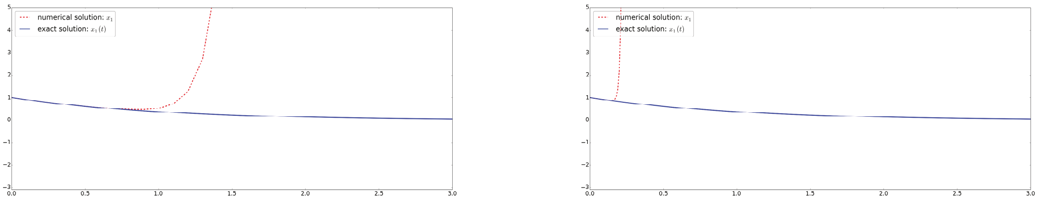


Figure 6.4: Numerical and exact solution of Example 2.12 simulated with the RadauIIA Method with 3 stages and the time step sizes $h = 0.1$(left) and $h = 0.01$(right).

These simulations show us that not only the implicit Euler is affected by this instability but also the BDF and Radau IIA methods. Two of the best known DAE solver packages are DASSL and RADAU, see [Pet82] and [HNW02]. While RADAU is based on the Radau IIA method, DASSL uses BDF-methods to solve a DAE. In the following we explain the underling problem of these methods on the basis of implicit Euler.

The basic idea behind the implicit Euler is the approximation of the derivative by a difference quotient: $(h > 0)$

$$x'(t) \approx \frac{x(t) - x(t-h)}{h}$$

Difference schemes like the difference quotient have a fundamental flaw. They do not commutate with the product rule. We use two differentiable functions $f, g \in C^1(\mathbb{R}, \mathbb{R})$ to explain this problem.

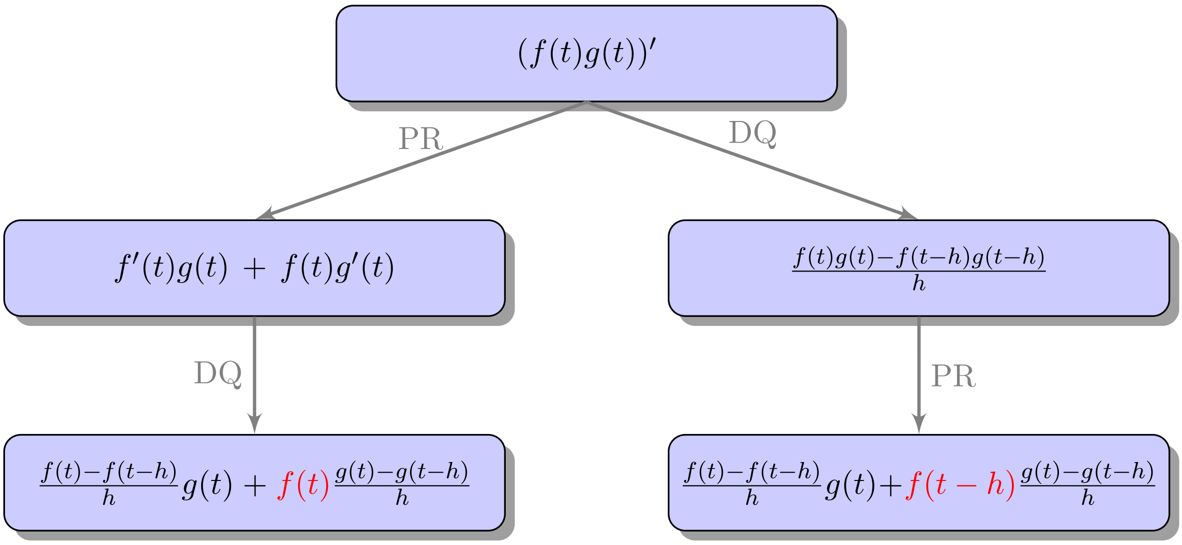

$(f(t)g(t))'$
PR
DQ
$f'(t)g(t) + f(t)g'(t)$
$\frac{f(t)g(t)-f(t-h)g(t-h)}{h}$
DQ
PR
$\frac{f(t)-f(t-h)}{h}g(t) + f(t)\frac{g(t)-g(t-h)}{h}$
$\frac{f(t)-f(t-h)}{h}g(t)+f(t-h)\frac{g(t)-g(t-h)}{h}$

Figure 6.5: The difference quotient and the product rule do not commutate.

In Figure 6.5 we see that the function f is evaluated at different time points. These evaluation points depend on the order of the usage of the product rule and the difference quotient.
With the help of Example 2.12 we demonstrate where this effect appears during the simulation of a DAE. First we consider the continuous DAE:

$$\begin{pmatrix} 1 & \eta t \\ 0 & 0 \end{pmatrix} x' + \begin{pmatrix} 0 & 1+\eta \\ 1 & \eta t \end{pmatrix} x = \begin{pmatrix} 0 \\ e^{-t} \end{pmatrix}$$

and we insert the variable splitting induced by the basis functions P and Q

$$x = \begin{pmatrix} 1 \\ \eta t \end{pmatrix} x_0 + \begin{pmatrix} -\eta t \\ 1 \end{pmatrix} y_0$$

in order to see

$$\begin{aligned} &\begin{pmatrix} 1 & \eta t \\ 0 & 0 \end{pmatrix} \left(\begin{pmatrix} 1 \\ \eta t \end{pmatrix} x_0 + \begin{pmatrix} -\eta t \\ 1 \end{pmatrix} y_0 \right)' \\ + &\begin{pmatrix} 0 & 1+\eta \\ 1 & \eta t \end{pmatrix} \left(\begin{pmatrix} 1 \\ \eta t \end{pmatrix} x_0 + \begin{pmatrix} -\eta t \\ 1 \end{pmatrix} y_0 \right) = \begin{pmatrix} 0 \\ e^{-t} \end{pmatrix} . \end{aligned} \tag{6.1}$$

We use the product rule to get

$$\begin{pmatrix} 1+(\eta t)^2 \\ 0 \end{pmatrix} x_0' + \begin{pmatrix} (1+2\eta)\eta t \\ 1+(\eta t)^2 \end{pmatrix} x_0 + \left(\begin{pmatrix} 1+\eta \\ 0 \end{pmatrix} + \begin{pmatrix} -\eta \\ 0 \end{pmatrix} \right) y_0 = \begin{pmatrix} 0 \\ e^{-t} \end{pmatrix} , \tag{6.2}$$

which leads to an explicit description of the numerical solution if we discretize Equation (6.2) with the implicit Euler method

$$\begin{aligned} x_{0,n} &= \frac{1}{1+(\eta t_n)^2} e^{-t_n}, \\ y_{0,n} &= -(1+2\eta)\eta t_n x_{0,n} - (1+(\eta t_n)^2) \frac{x_{0,n} - x_{0,n-1}}{h}. \end{aligned}$$

We notice that there are no dynamics in these equations, i.e. the solutions at a time point t_n are independent from the initial values at t_0. This behavior is expected since we have to deal with an index two DAE with two components. Next we apply the implicit Euler method before using the product rule and the decoupling of the Dissection Index and obtain

$$\begin{pmatrix} 1 & \eta t_n \\ 0 & 0 \end{pmatrix} \frac{x_n - x_{n-1}}{h} + \begin{pmatrix} 0 & 1+\eta \\ 1 & \eta t_n \end{pmatrix} x_n = \begin{pmatrix} 0 \\ e^{-t_n} \end{pmatrix}.$$

In the discrete case we use the variable splitting in the time point t_n

$$x_n = \begin{pmatrix} 1 \\ \eta t_n \end{pmatrix} x_{0,n} + \begin{pmatrix} -\eta t_n \\ 1 \end{pmatrix} y_{0,n}$$

and in the time point t_{n-1}

$$x_{n-1} = \begin{pmatrix} 1 \\ \eta t_{n-1} \end{pmatrix} x_{0,n-1} + \begin{pmatrix} -\eta t_{n-1} \\ 1 \end{pmatrix} y_{0,n-1}$$

to obtain the system

$$\begin{aligned} &\begin{pmatrix} 1 & \eta t_n \\ 0 & 0 \end{pmatrix} \frac{\begin{pmatrix} 1 \\ \eta t_n \end{pmatrix} x_{0,n} + \begin{pmatrix} -\eta t_n \\ 1 \end{pmatrix} y_{0,n} - \left(\begin{pmatrix} 1 \\ \eta t_{n-1} \end{pmatrix} x_{0,n-1} + \begin{pmatrix} -\eta t_{n-1} \\ 1 \end{pmatrix} y_{0,n-1} \right)}{h} \\ &\qquad + \begin{pmatrix} 0 & 1+\eta \\ 1 & \eta t_n \end{pmatrix} \left(\begin{pmatrix} 1 \\ \eta t_n \end{pmatrix} x_{0,n} + \begin{pmatrix} -\eta t_n \\ 1 \end{pmatrix} y_{0,n} \right) = \begin{pmatrix} 0 \\ e^{-t_n} \end{pmatrix}. \end{aligned}$$

We rearrange these equations into the form of Equation (6.1).

$$\begin{aligned} \begin{pmatrix} 1+(\eta t_n)^2 \\ 0 \end{pmatrix} \frac{x_{0,n} - x_{0,n-1}}{h} &+ \begin{pmatrix} (1+\eta)\eta t_n \\ 1+(\eta t_n)^2 \end{pmatrix} x_{0,n} + \begin{pmatrix} \eta^2 t_n \\ 0 \end{pmatrix} x_{0,n-1} \\ &+ \begin{pmatrix} 1+\eta \\ 0 \end{pmatrix} y_{0,n} + \begin{pmatrix} -\eta \\ 0 \end{pmatrix} y_{0,n-1} = \begin{pmatrix} 0 \\ e^{-t_n} \end{pmatrix} \end{aligned}$$

The product rule and the difference quotient do not commutate and as a consequence we have to deal with the terms

$$\begin{pmatrix} -\eta \\ 0 \end{pmatrix} y_{0,n-1} \quad \text{and} \quad \begin{pmatrix} \eta^2 t_n \\ 0 \end{pmatrix} x_{0,n-1}.$$

The evaluation shift in the function f in the Figure 6.5 is reflected by the evaluation shift of x_0 and y_0. In this case the numerical solution must be described dependent of $y_{0,n-1}$

$$\begin{aligned}
x_{0,n} &= \frac{1}{1+(\eta t_n)^2} e^{-t_n} \\
y_{0,n} &= \frac{\eta}{1+\eta} y_{0,n-1} - \frac{1}{1+\eta} \left((1+\eta)\eta t_n x_{0,n} + \eta^2 t_n x_{0,n-1} + (1+(\eta t_n)^2) \frac{x_{0,n} - x_{0,n-1}}{h} \right).
\end{aligned}$$

Hence we deal with a dynamical behavior, which should not be the case. Therefore we call this dynamics artificial. In particular this artificial dynamic is unstable if $\eta < -0.5$, since it holds

$$\left| \frac{\eta}{1+\eta} \right| \leqslant 1 \Leftrightarrow \eta \geqslant -0.5.$$

These artificial dynamics can also manipulate the inherent dynamics. We demonstrate this with the following example

Example 6.2. ([HMM98, LMT13])
Let $\mathcal{I} := [0,3]$ and let $t \in \mathcal{I}$. We consider the DAE

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} x' - \begin{pmatrix} -10 & 1 & 1 \\ \eta(\eta t^2 - t + 1) & -10 & \eta t \\ (1-\eta t) & 1 & 0 \end{pmatrix} x = 0 \tag{6.3}$$

with the exact solution

$$x(t) = \begin{pmatrix} e^{-10t} & -(1-\eta t)e^{-10t} & (1-\eta t)e^{-10t} \end{pmatrix}^\top$$

for the initial value $x^0 = \begin{pmatrix} 1 & -1 & 1 \end{pmatrix}^\top$.

By the canonical choice

$$P = V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad Q = W = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

we obtain

$$W^\top B P x_0 = \begin{pmatrix} 1-\eta t & 1 \end{pmatrix} x_0 = 0 \tag{6.4}$$

with $x = Px_0 + Qy_0$. Hence, we can choose

$$P_{x_1} = \begin{pmatrix} 1-\eta t \\ 1 \end{pmatrix} \quad \text{and} \quad Q_{x_1} = \begin{pmatrix} 1 \\ \eta t - 1 \end{pmatrix}$$

and define $x_0 = P_{x_1}\tilde{x}_1 + Q_{x_1}x_1$ which yields $\tilde{x}_1 = 0$ with the help of Equation (6.4). Thereby we obtain

$$x_0 = Q_{x_1}x_1 = \begin{pmatrix} 1 \\ \eta t - 1 \end{pmatrix} x_1$$

and the derivative of x_0 can be continuously described by

$$x_0' = (Q_{x_1}x_1)' = Q_{x_1}x_1' + Q_{x_1}'x_1 = \begin{pmatrix} 1 \\ \eta t - 1 \end{pmatrix} x_1' + \begin{pmatrix} 0 \\ \eta \end{pmatrix} x_1. \tag{6.5}$$

By factorizing the equations of Example 6.2 by $V^\top$ and transforming them by P and Q we get

$$x_0' = \begin{pmatrix} -10 & 1 \\ \eta(\eta t^2 - t + 1) & -10 \end{pmatrix} x_0 + \begin{pmatrix} 1 \\ \eta t \end{pmatrix} y_0$$

which yields together with Equation (6.7) and $x_0 = Q_{x_1}x_1$:

$$\begin{pmatrix} 1 \\ \eta t - 1 \end{pmatrix} x_1' + \begin{pmatrix} 0 \\ \eta \end{pmatrix} x_1 = \begin{pmatrix} \eta t - 11 \\ \eta t(\eta t - 11) + \eta + 10 \end{pmatrix} x_1 + \begin{pmatrix} 1 \\ \eta t \end{pmatrix} y_0 \tag{6.6}$$

We multiply one of the basis functions of the alternative basis ending from Lemma 4.15

$$W_y^* = \begin{pmatrix} \eta t \\ -1 \end{pmatrix}$$

to the left of (6.6) and thereby we obtain

$$x_1' = -10x_1$$

which describes the inherent dynamic. If we discretize Example 6.3 by the implicit Euler method, before we use the product rule and the decoupling, we obtain:

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \frac{x_n - x_{n-1}}{h} - \begin{pmatrix} -10 & 1 & 1 \\ \eta(\eta t^2 - t + 1) & -10 & \eta t \\ (1 - \eta t) & 1 & 0 \end{pmatrix} x_n = 0.$$

By factorizing these equations by $V^\top$ and $W^\top$ and transforming them by P and Q we obtain

$$\begin{aligned} \frac{x_{0,n} - x_{0,n-1}}{h} &= \begin{pmatrix} -10 & 1 \\ \eta(\eta t_n^2 - t + 1) & -10 \end{pmatrix} x_{0,n} + \begin{pmatrix} 1 \\ \eta t_n \end{pmatrix} y_{0,n} \\ 0 &= \begin{pmatrix} 1 - \eta t_n & 1 \end{pmatrix} x_{0,n}. \end{aligned}$$

We use the transformation $x_{0,n} = P_{x_{1,n}}\tilde{x}_{1,n} + Q_{x_{1,n}}x_{1,n}$ and get again $\tilde{x}_{1,n} = 0$ and thereby $x_{0,n} = Q_{x_{1,n}}x_{1,n}$. But after the time discretization this yields

$$\frac{x_{0,n} - x_{0,n-1}}{h} = \begin{pmatrix} 1 \\ \eta t_n - 1 \end{pmatrix} \frac{x_{1,n} - x_{1,n-1}}{h} + \begin{pmatrix} 0 \\ \eta \end{pmatrix} x_{1,n-1} \tag{6.7}$$

instead of (6.5). Hence we obtain

$$\begin{pmatrix} 1 \\ \eta t_n - 1 \end{pmatrix} \frac{x_{1,n} - x_{1,n-1}}{h} + \begin{pmatrix} 0 \\ \eta \end{pmatrix} x_{1,n-1} = \begin{pmatrix} \eta t_n - 11 \\ \eta t_n(\eta t_n - 11) + \eta + 10 \end{pmatrix} x_{1,n} + \begin{pmatrix} 1 \\ \eta t_n \end{pmatrix} y_{0,n}$$

instead of (6.6). By a multiplication with

$$W^*_{y,n} = \begin{pmatrix} \eta t_n \\ -1 \end{pmatrix}$$

from the left we obtain

$$\frac{x_{1,n} - x_{1,n-1}}{h} - \eta x_{1,n-1} = -(\eta + 10)x_{1,n}$$

which describes the discrete version of the inherent dynamic. By rearranging this equation to

$$x_{1,n} = \frac{1 + h\eta}{1 + h(\eta + 10)} x_{1,n-1}$$

we see that $|\frac{1+h\eta}{1+h(\eta+10)}| \leqslant 1$ is necessary for the convergence of $x_{1,n}$, if we apply the implicit Euler method. Due to

$$\begin{aligned} \left|\frac{1 + h\eta}{1 + h(\eta + 10)}\right| < 1 &\Rightarrow \left|1 - \frac{10h}{1 + h(\eta + 10)}\right| < 1 \\ &\Rightarrow \frac{10h}{1 + h(\eta + 10)} > 0 \\ &\Rightarrow 1 + h(\eta + 10) > 0, \quad \eta < -10 \\ &\Rightarrow h < -\frac{1}{\eta + 10} \end{aligned}$$

the condition $h < -\frac{1}{\eta+10}$ is necessary for the convergence of $x_{1,n}$. Hence the appearance of an artificial dynamic in an inherent dynamic can invoke additional time step size restrictions as we see in Figure 6.6.

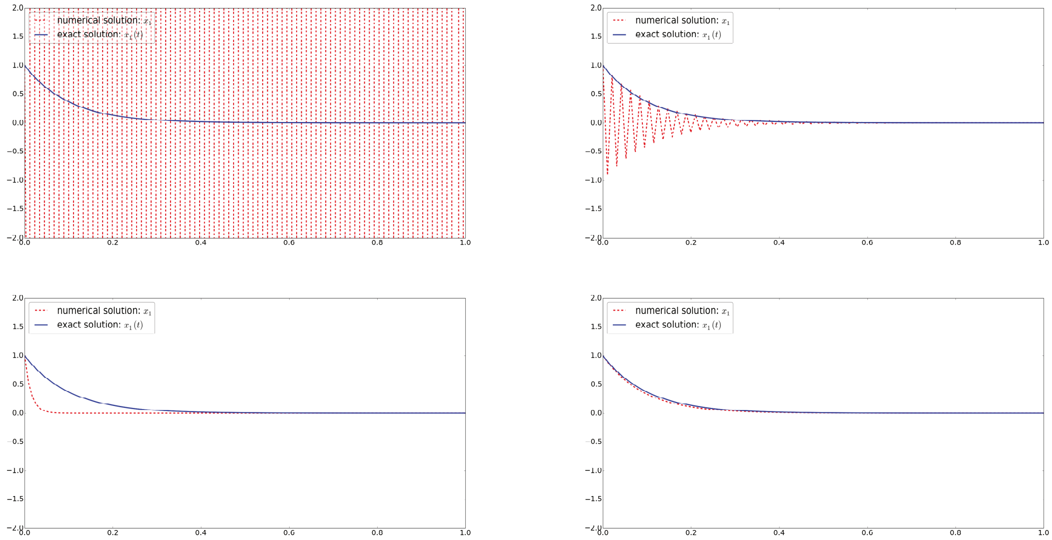


Figure 6.6: Numerical and exact solution of Example 6.2 simulated with the implicit Euler and a time step sizes $h = 11 \cdot 10^{-3}$(upper left), $h = 10.5 \cdot 10^{-3}$(upper right), $h = 9 \cdot 10^{-3}$(lower left) and $h = 1 \cdot 10^{-3}$(lower right).

The implicit Euler method, applied to Example 2.12, is unconditionally unstable for $\eta < -0.5$. When we apply the implicit Euler method to Example 6.2 we only have to choose the time step size sufficiently small for the Euler method to converge. For proper formulated index 2 DAEs the implicit Euler method might have to fulfill additional time step size restrictions but it never becomes unconditionally unstable. As the next example shows this is no longer the case if we deal with index 3 DAEs.

Example 6.3. Let $\mathcal{I} := [0, 3]$ and let $t \in \mathcal{I}$. We consider the DAE

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} x' = \begin{pmatrix} sin(t)cos(t) & -sin(t)sin(t) & cos(t) \\ cos(t)cos(t) & -sin(t)cos(t) & -sin(t) \\ sin(t) & cos(t) & 0 \end{pmatrix} x + \begin{pmatrix} cos(t) - cos(t)sin(t) \\ -sin(t) + sin(t)sin(t) \\ 1 \end{pmatrix}$$

with the exact solution

$$x(t) = \begin{pmatrix} sin(t) & cos(t) & sin(t) \end{pmatrix}^\top$$

for the initial value $x^0 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}^\top$.

Again we split the DAE by the canonical choice of the first basis functions and obtain:

$$x_0' = \begin{pmatrix} sin(t)cos(t) & -sin(t)sin(t) \\ cos(t)cos(t) & -sin(t)cos(t) \end{pmatrix} x_0 + \begin{pmatrix} cos(t) \\ -sin(t) \end{pmatrix} y_0 + \begin{pmatrix} cos(t) - cos(t)sin(t) \\ -sin(t) + sin(t)sin(t) \end{pmatrix}$$

$$0 = \begin{pmatrix} sin(t) & cos(t) \end{pmatrix} x_0 - 1$$

We choose

$$P_{x_1} = \begin{pmatrix} sin(t) \\ cos(t) \end{pmatrix} \quad \text{and} \quad Q_{x_1} = \begin{pmatrix} cos(t) \\ -sin(t) \end{pmatrix}$$

and split $x_0 = P_{x_1}(t)\tilde{x}_1 + Q_{x_1}(t)x_1$ which yields $\tilde{x}_1 = 1$. Then we can write $x_0 = Q_{x_1}(t)x_1 + P_{x_1}(t)$ and get

$$x_0' = Q_{x_1}'(t)x_1 + Q_{x_1}(t)x_1' + P_{x_1}'(t) \tag{6.8}$$

By inserting these relations we get

$$\begin{pmatrix} cos(t) \\ -sin(t) \end{pmatrix} x_1' = 2\begin{pmatrix} sin(t) \\ cos(t) \end{pmatrix} x_1 + \begin{pmatrix} cos(t) \\ -sin(t) \end{pmatrix} y_0 + \begin{pmatrix} -cos(t)sin(t) \\ sin(t)sin(t) \end{pmatrix}$$

which enables us to choose

$$V_{x_1} = \begin{pmatrix} cos(t) \\ -sin(t) \end{pmatrix} \quad \text{and} \quad W_{x_1} = \begin{pmatrix} sin(t) \\ cos(t) \end{pmatrix}.$$

By factorizing with $V_{x_1}^\top$ and $W_{x_1}^\top$ we achieve

$$\begin{aligned} y_0 &= x_1' + sin(t) \\ 2x_1 &= 0. \end{aligned} \tag{6.9}$$

Now we decouple the discretized system. We start with the system

$$\begin{aligned} \frac{x_{0,n} - x_{0,n-1}}{h} &= \begin{pmatrix} sin(t_n)cos(t_n) & -sin(t_n)sin(t_n) \\ cos(t_n)cos(t_n) & -sin(t_n)cos(t_n) \end{pmatrix} x_{0,n} + \begin{pmatrix} cos(t_n) \\ -sin(t_n) \end{pmatrix} y_{0,n} \\ &+ \begin{pmatrix} cos(t_n) - cos(t_n)sin(t_n) \\ -sin(t_n) + sin(t_n)sin(t_n) \end{pmatrix} \\ 0 &= \begin{pmatrix} sin(t_n) & cos(t_n) \end{pmatrix} x_{0,n} - 1 \end{aligned} \tag{6.10}$$

and define $P_{x_{1,n}} := P_{x_1}(t_n)$, $Q_{x_{1,n}} := Q_{x_1}(t_n)$, $V_{x_{1,n}} := V_{x_1}(t_n)$ and $W_{x_{1,n}} := W_{x_1}(t_n)$. Analogous to the continuous case we obtain $\tilde{x}_{1,n} = 1$ and analog to (6.8) we get

$$\frac{x_{0,n} - x_{0,n-1}}{h} = \frac{P_{x_{1,n}} - P_{x_{1,n-1}}}{h} + \frac{Q_{x_{1,n}} - Q_{x_{1,n-1}}}{h} x_{1,n-1} + Q_{x_{1,n}} \frac{x_{1,n} - x_{1,n-1}}{h}. \tag{6.11}$$

With the help of the Taylor expansions

$$sin(t_{n-1}) = sin(t_n) - cos(t_n)h - \frac{1}{2}sin(t_n)h^2 + \mathcal{O}(h^3)$$

$$cos(t_{n-1}) = cos(t_n) + sin(t_n)h - \frac{1}{2}cos(t_n)h^2 + \mathcal{O}(h^3)$$

we get

$$\frac{P_{x_{1,n}} - P_{x_{1,n-1}}}{h} = Q_{x_{1,n}} + \frac{1}{2}P_{x_{1,n}}h + \mathcal{O}(h^2) \tag{6.12}$$

and

$$\frac{Q_{x_{1,n}} - Q_{x_{1,n-1}}}{h} = -P_{x_{1,n}} + \frac{1}{2}Q_{x_{1,n}}h + \mathcal{O}(h^2). \tag{6.13}$$

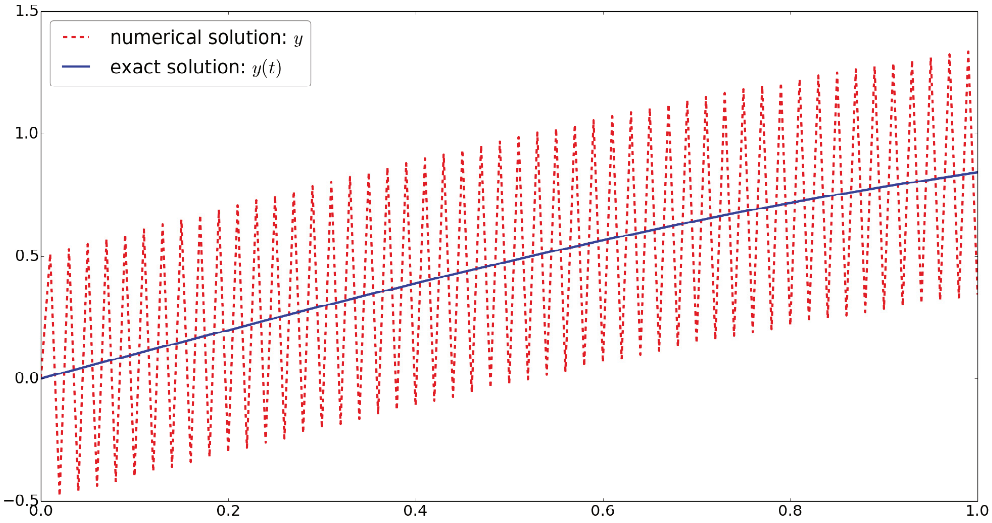


Figure 6.7: Numerical and exact solutions of the third component of Example 6.3 with $h = 0.01$ using the implicit Euler.

With the help of (6.11), (6.12) and (6.13) we can transform (6.10) into

$$\begin{aligned} x_{1,n} &= -x_{1,n-1} + \frac{1}{2}h + \mathcal{O}(h^2) \\ y_{0,n} &= sin(t_n) + \frac{x_{1,n} - x_{1,n-1}}{h} + \mathcal{O}(h). \end{aligned}$$

This yields

$$x_{1,2n-1} = \frac{1}{2}h + \mathcal{O}(h^2)$$

$$x_{1,2n} = \mathcal{O}(h^2)$$

and

$$\begin{aligned} y_{0,2n-1} &= sin(t_{2n-1}) + \frac{1}{2} + \mathcal{O}(h) \\ y_{0,2n} &= sin(t_{2n}) - \frac{1}{2} + \mathcal{O}(h) \end{aligned}$$

which coincides with the numerical simulation results in Figure 6.7.
In the following section we amplify the source of the problems presented in this section. Furthermore we present a class of methods which overcomes the problems regarding the artificial dynamics.

6.1 Implicit Methods

The previous examples show that classical numerical methods may lose their convergence properties when applied to DAEs. In this section sufficient convergence criteria for nonlinear DAEs in standard form are presented. The mentioned convergence problems already occur in the index 2 case and therefore we restrict ourselves to index 2 DAEs in this section. The main aim of this section is to amplify the source of the convergence issues. In particular we want to exclude the non-linearity of the DAE as such a source and draw focus to the basis functions, such that it is sufficient to consider linear DAEs when analyzing these convergence problems. In this section we use BDF methods for time discretization. Similar results for index 2 DAEs discretized by other methods can be found in [Voi06]. In contrast to these results we do not need any structural assumption of the equation of the DAE, but we need some of the basis function to be constant.

Definition 6.4.
Consider a sufficiently smooth nonlinear DAE in standard form with initial conditions on a time interval $\mathcal{I} = [t_0, T]$:

$$f(x', x, t) = 0, \quad x(t_0) = x^0.$$

We define the difference quotient with k-steps and the BDF-coefficients α_j

$$x'_n := \frac{1}{h} \sum_{j=0}^{k} \alpha_j x_{n-j}$$

and therewith we formulate the BDF method with k-steps

$$f(x'_n, x_n, t_n) = \delta_n$$

with x_n being the solution approximation at t_n and δ_n being the perturbation caused by the rounding errors and the used nonlinear solver.

For the convergence analysis we need to define the consistency error of a BDF:

Definition 6.5. (Consistency error) The consistency error of a BDF method is defined by

$$L_n(x) := \frac{1}{h}\sum_{j=0}^{k}\alpha_j x(t_{n-j}) - x'(t_n).$$

We say a BDF method has a consistency error of order k if $L_n(x) = \mathcal{O}(h^k)$.

It is common knowledge that a BDF method with k-steps has consistency order k, if the exact solution is smooth enough. The following theorem shows that the only possible problem source is the non-linearity or even time dependence of the basis functions Q and Q_{x_1}. In particular the equations of the extended MNA fulfill the assumptions of Theorem 6.6, see Section 4.3.

Theorem 6.6.
Consider a sufficiently smooth nonlinear DAE in standard form with Dissection Index 2 and a global unique solution which is sufficiently smooth. Assume that there exist constant basis functions Q and Q_{x_1}. Furthermore let the errors in the first k steps and the rounding errors be sufficiently small. Then the BDF method with k-steps converges with order k for $2 \leqslant k \leqslant 6$, i.e.

$$\|e_n\| = \|x(t_n) - x_n\| \leqslant ch^k$$

with $c > 0$ being independent of h and e_n being the global error.

Proof.
First we show the feasibility of the methods by an induction over the time steps. In particular we show that the methods produce numerical values x_i in a sufficiently small neighborhood around the solution. We assumed that the error in the first k steps fulfills

$$\|x(t_i) - x_i\| \leqslant \mathcal{O}(h^k), \qquad \text{for } i \leqslant k-1.$$

Therefore the induction start is automatically fulfilled. Hence we assume that the errors of the previous steps are of order k and show that also the error in the n-th step is of order k.
Both basis functions Q and Q_{x_1} are assumed to be constant which enables us to also choose constant basis functions P and P_{x_1}. The remaining basis functions may depend on time, the solution function and its derivative.
We define the following notation

$$V_n^\top := V^\top(x'(t_n), x(t_n), t_n) \quad V_{x1,n}^\top := V_{x_1}^\top(x'(t_n), x(t_n), t_n)$$

$$\begin{aligned} W_n^\top &:= W^\top(x'(t_n), x(t_n), t_n) & W_{x_1,n}^\top &:= W_{x_1}^\top(x'(t_n), x(t_n), t_n) \\ P_{y_1,n} &:= P_{y_1}(x'(t_n), x(t_n), t_n) & V_{y_1,n}^\top &:= V_{y_1}^\top(x'(t_n), x(t_n), t_n) \\ Q_{y_1,n} &:= Q_{y_1}(x'(t_n), x(t_n), t_n) & W_{y_1,n}^\top &:= W_{y_1}^\top(x'(t_n), x(t_n), t_n) \end{aligned}$$

to decouple the discretized DAE

$$f(x_n', x_n, t_n) = \delta_n.$$

Before we start the decoupling we split the exact solution into

$$x(t) = PP_{x_1}\tilde{x}_1(t) + PQ_{x_1}x_1(t) + Qy_0(t) \tag{6.14}$$

and the global error into

$$e_n = PP_{x_1}\tilde{e}_{1,n}^x + PQ_{x_1}e_{1,n}^x + Qe_{0,n}^y. \tag{6.15}$$

First we insert the splitting $x_n =: Px_{0,n} + Qy_{0,n}$ and apply a factorization by multiplying $V_n^\top$ and $W_n^\top$ from the left side

$$V_n^\top f(Px_{0,n}', Px_{0,n} + Qy_{0,n}, t_n) = V_n^\top \delta_n \tag{6.16a}$$

$$W_n^\top f(Px_{0,n}', Px_{0,n} + Qy_{0,n}, t_n) = W_n^\top \delta_n. \tag{6.16b}$$

Notice that it holds $(Px_{0,n})' = Px_{0,n}'$ since P is constant. Hence, we do not deal with the product stability problem here. Next we insert transformations and factorizations at once. We factorize (6.16a) by $V_{x_1,n}^\top$ and $W_{x_1,n}^\top$ and (6.16b) by $V_{y_1,n}^\top$ and $W_{y_1,n}^\top$, while we insert the transformations $x_{0,n} =: P_{x_1}\tilde{x}_{1,n} + Q_{x_1}x_{1,n}$ and $y_{0,n} =: P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}\tilde{y}_{2,n}$.

$$V_{x_1,n}^\top V_n^\top f(PP_{x_1}\tilde{x}_{1,n}' + PQ_{x_1}x_{1,n}', PP_{x_1}\tilde{x}_{1,n} + PQ_{x_1}x_{1,n} + QQ_{y_1,n}\tilde{y}_{2,n} + QP_{y_1,n}\tilde{y}_{1,n}, t_n) = V_{x_1,n}^\top V_n^\top \delta_n \tag{6.17a}$$

$$W_{x_1,n}^\top V_n^\top f(PP_{x_1}\tilde{x}_{1,n}' + PQ_{x_1}x_{1,n}', PP_{x_1}\tilde{x}_{1,n} + PQ_{x_1}x_{1,n} + QQ_{y_1,n}\tilde{y}_{2,n} + QP_{y_1,n}\tilde{y}_{1,n}, t_n) = W_{x_1,n}^\top V_n^\top \delta_n \tag{6.17b}$$

$$V_{y_1,n}^\top W_n^\top f(PP_{x_1}\tilde{x}_{1,n}' + PQ_{x_1}x_{1,n}', PP_{x_1}\tilde{x}_{1,n} + PQ_{x_1}x_{1,n} + QQ_{y_1,n}\tilde{y}_{2,n} + QP_{y_1,n}\tilde{y}_{1,n}, t_n) = V_{y_1,n}^\top W_n^\top \delta_n \tag{6.17c}$$

$$W_{y_1,n}^\top W_n^\top f(PP_{x_1}\tilde{x}_{1,n}' + PQ_{x_1}x_{1,n}', PP_{x_1}\tilde{x}_{1,n} + QQ_{y_1,n}\tilde{y}_{2,n} + QP_{y_1,n}\tilde{y}_{1,n}, t_n) = W_{y_1,n}^\top W_n^\top \delta_n \tag{6.17d}$$

Again we notice that we avoid the product stability problem due to $(Q_{x_1}x_{1,n})' = Q_{x_1}x_{1,n}'$. The next objective is to reduce the system component down to $\tilde{x}_{1,n}$ and $x_{1,n}$. Therefore we apply the Lemma 4.32 on (6.17a) around the transformed exact solutions $\tilde{x}_1(t), x_1(t)$, $\tilde{y}_1(t), \tilde{y}_2(t)$ and the derivatives $\tilde{x}_1'(t), x_1'(t)$ and obtain

$$x_{1,n}' = \bar{\Psi}_{x_1'}(\tilde{x}_{1,n}', x_{1,n}, \tilde{x}_{1,n}, \tilde{y}_{1,n}, \tilde{y}_{2,n}, t_n, \delta_n). \tag{6.18}$$

We also use the Lemma 4.32 to achieve an explicit description for $\tilde{y}_{1,n}$ and $\tilde{y}_{2,n}$. In particular we apply the Lemma 4.32 on (6.17b) and (6.17c) around the transformed exact solution and its derivative. Together with (6.18) this yields

$$\tilde{y}_{1,n} = \Psi_{\tilde{y}_1}(\tilde{x}_{1,n}', x_{1,n}, \tilde{x}_{1,n}, t_n, \delta_n) \tag{6.19a}$$

$$\tilde{y}_{2,n} = \Psi_{\tilde{y}_2}(\tilde{x}'_{1,n}, x_{1,n}, \tilde{x}_{1,n}, t_n, \delta_n) \tag{6.19b}$$

with $\frac{\partial}{\partial \tilde{x}'_{1,n}} \Psi_{\tilde{y}_1}(\tilde{x}'_1(t), x_1(t), \tilde{x}_1(t), t, 0) = 0$ for all $t \in \mathcal{I}$. At last we apply the Lemma 4.32 on (6.17d) and together with (6.18) and (6.19) we obtain

$$\tilde{x}_{1,n} = \Psi_{\tilde{x}_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n) \tag{6.20}$$

with $\frac{\partial}{\partial \tilde{x}'_{1,n}} \Psi_{\tilde{x}_1}(\tilde{x}'_1(t), x_1(t), t, 0) = 0$ and $\frac{\partial}{\partial x_{1,n}} \Psi_{\tilde{x}_1}(\tilde{x}'_1(t), x_1(t), t, 0) = 0$ for all $t \in \mathcal{I}$. Now we can combine (6.18) with (6.20) and (6.19) and achieve

$$x'_{1,n} = \Psi_{x'_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n), \tag{6.21}$$

which finally yields the system

$$x_{1,n} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{1,n-j} + h \frac{1}{\alpha_0} \Psi_{x'_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n) \tag{6.22a}$$

$$\tilde{x}_{1,n} = \Psi_{\tilde{x}_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n). \tag{6.22b}$$

Before we estimate the error between the numerical and the exact solution we need to prove that (6.22) has a solution in a neighborhood of the exact solution. Therefore we consider the function

$$\Phi(x_{1,n}, \tilde{x}_{1,n}) = \begin{pmatrix} \Phi_1(x_{1,n}, \tilde{x}_{1,n}) \\ \Phi_2(x_{1,n}, \tilde{x}_{1,n}) \end{pmatrix} = \begin{pmatrix} -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{1,n-j} + h \frac{1}{\alpha_0} \Psi_{x'_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n) \\ \Psi_{\tilde{x}_1}(\tilde{x}'_{1,n}, x_{1,n}, t_n, \delta_n) \end{pmatrix}$$

and notice that a fixpoint of Φ is a solution of (6.22). Let $r > 0$ and $z \in \mathbb{R}^n$. We define

$$B_r(z) := \{x \in \mathbb{R}^n \quad | \quad \|x - z\|_2 \leqslant r\}$$

as the closed sphere around z with the radius r. We remember that the consistency error is of order k, i.e. $||L_n(x)|| \leqslant c_L h^k$ and choose

$$\Omega_n = B_{h^k}\left(x_1(t_n) + \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} e^x_{1,n-j} \right) \times B_{h^{k+1}}(\tilde{x}_1(t_n))$$

and

$$\Omega'_n = B_{h^{k-1}}(\tilde{x}'_1(t_n)) \times B_{h^{k-1}}(x_1(t_n)).$$

Then Φ has a unique fixpoint in Ω_n with h being sufficiently small. We will prove this by the Schauder Fixed Point Theorem. We only have to show that $\Phi(x) \in \Omega_n$ for all $x \in \Omega_n$, since Φ is continuous on Ω_n. Therefore we define the four closed convex hulls

$$\Omega_{\partial \Psi_{x'_1}} := \overline{\operatorname{conv}} \left\{ \frac{\partial}{\partial(\tilde{x}'_1, x_1, \delta)} \Psi_{x'_1}(z) | z \in \bigcup_i \Omega'_i \times \mathcal{I} \times B_{h^{k+1}}(0) \right\},$$

$$
\begin{aligned}
\Omega_{\partial\delta,\Psi_{x_1'}} &:= \overline{\mathrm{conv}}\left\{ \frac{\partial}{\partial\delta}\Psi_{x_1'}(z) | z \in \bigcup_i \Omega_i' \times \mathcal{I} \times B_{h^{k+1}}(0) \right\}, \\
\Omega_{\partial\delta,\Psi_{\tilde{x}_1}} &:= \overline{\mathrm{conv}}\left\{ \frac{\partial}{\partial\delta}\Psi_{\tilde{x}_1}(z) | z \in \bigcup_i \Omega_i' \times \mathcal{I} \times B_{h^{k+1}}(0) \right\}, \\
\Omega_{\partial^2\Psi_{\tilde{x}_1}} &:= \overline{\mathrm{conv}}\left\{ \frac{\partial^2}{\partial(\tilde{x}_1', x_1)^2}\Psi_{\tilde{x}_1}(z) | z \in \bigcup_i \Omega_i' \times \mathcal{I} \times B_{h^{k+1}}(0) \right\}
\end{aligned}
$$

and the constants

$$
\begin{aligned}
c_{\partial\Psi_{x_1'}} &:= \max_{x\in\Omega_{\partial\Psi_{x_1'}}} \|x\|, \quad c_{\partial\delta,\Psi_{x_1'}} := \max_{x\in\Omega_{\partial\delta,\Psi_{x_1'}}} \|x\|, \\
c_{\partial\delta,\Psi_{\tilde{x}_1}} &:= \max_{x\in\Omega_{\partial\delta,\Psi_{\tilde{x}_1}}} \|x\|, \quad c_{\partial^2\Psi_{\tilde{x}_1}} := \max_{x\in\Omega_{\partial^2\Psi_{\tilde{x}_1}}} \|x\|
\end{aligned}
$$

which exist since $\bigcup_i \Omega_i' \times \mathcal{I} \times B_{h^{k+1}}(0)$ is compact. By the Mean Value Theorem it holds for all $(x_{1,n}, \tilde{x}_{1,n}) \in \Omega_n$ that there is a $J_1 \in \Omega_{\partial\Psi_{x_1'}}$ such that

$$
\begin{aligned}
&\Phi_1(x_{1,n}, \tilde{x}_{1,n}) - \left(x_1(t_n) + \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} e^x_{1,n-j} \right) \\
&= -\left(x_1(t_n) + \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} e^x_{1,n-j} \right) - \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{1,n-j} + h\frac{1}{\alpha_0}\Psi_{x_1'}(\tilde{x}_{1,n}', x_{1,n}, t_n, \delta_n) \\
&= -\sum_{j=0}^{k} \frac{\alpha_j}{\alpha_0} x_1(t_{n-j}) + h\frac{1}{\alpha_0}\Psi_{x_1'}(\tilde{x}_{1,n}', x_{1,n}, t_n, \delta_n) \\
&= -h\frac{1}{\alpha_0}\left(\frac{1}{h}\sum_{j=0}^{k} \alpha_j x_1(t_{n-j}) - \Psi_{x_1'}(\tilde{x}_{1,n}', x_{1,n}, t_n, \delta_n) \right) \\
&= -h\frac{1}{\alpha_0}\left(x_1'(t_n) + L_n(x) - \Psi_{x_1'}(\tilde{x}_{1,n}', x_{1,n}, t_n, \delta_n) \right) \\
&= -h\frac{1}{\alpha_0}\left(L_n(x) - J_1 \begin{pmatrix} (\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)) \\ (x_{1,n} - x_1(t_n)) \\ \delta_n \end{pmatrix} \right)
\end{aligned}
$$

and that there is a $J_2 \in \Omega_{\partial\delta,\Psi_{\tilde{x}_1}}$ and a $H \in \Omega_{\partial^2\Psi_{\tilde{x}_1}}$ such that

$$
\begin{aligned}
\Phi_2(x_{1,n}, \tilde{x}_{1,n}) - \tilde{x}_1(t_n) &= \Psi_{\tilde{x}_1}(\tilde{x}_{1,n}', x_{1,n}, t_n, \delta_n) - \tilde{x}_1(t_n) \\
&= J_2\delta + \begin{pmatrix} (\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)) \\ (x_{1,n} - x_1(t_n)) \end{pmatrix}^\top H \begin{pmatrix} (\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)) \\ (x_{1,n} - x_1(t_n)) \end{pmatrix}.
\end{aligned}
$$

We obtain with the help of the splitting of the exact solution (6.14) and the induction statement

$$\begin{aligned}\tilde{x}_{1,n}' - \tilde{x}_1'(t_n) &= \frac{1}{h}\sum_{j=0}^{k}\alpha_j\tilde{x}_{1,n-j} - \tilde{x}_1'(t_n) = \frac{1}{h}\sum_{j=0}^{k}\alpha_j\tilde{x}_1(t_{n-j}) - \tilde{x}_1'(t_n) + \mathcal{O}(h^k)\\ &= L_n(\tilde{x}_1) + \mathcal{O}(h^k) = \mathcal{O}(h^k)\end{aligned}$$

and

$$x_{1,n} - x_1(t_n) = \sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}e^x_{1,n-j} + \mathcal{O}(h^k) = \mathcal{O}(h^k)$$

for all $(x_{1,n}, \tilde{x}_{1,n}) \in \Omega_n$. Hence we obtain for a sufficiently small step size h and for $\|\delta\| \leqslant \frac{1}{2c_{\partial\delta,\Psi_{x_1'}}}h^{k+1}$:

$$\begin{aligned}&\left\|\Phi_1(x_{1,n}, \tilde{x}_{1,n}) - \left(x_1(t_n) + \sum_{j=0}^{k}\frac{\alpha_j}{\alpha_0}e^x_{1,n-j}\right)\right\|\\ &\leqslant h\frac{1}{\alpha_0}(\|L_n(x)\| + c_{\partial\Psi_{x_1'}}(\left\|\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)\right\| + \|x_{1,n} - x_1(t_n)\| + \|\delta\|))\\ &\leqslant \mathcal{O}(h^{k+1}) \leqslant h^k\end{aligned}$$

and

$$\begin{aligned}&\|\Phi_2(x_{1,n}, \tilde{x}_{1,n}) - \tilde{x}_1(t_n)\|\\ &\leqslant c_{\partial\delta,\Psi_{x_1'}}\|\delta\| + (\left\|\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)\right\| + \|x_{1,n} - x_1(t_n)\|)c_{\partial^2\Psi_{\tilde{x}_1}}(\left\|\tilde{x}_{1,n}' - \tilde{x}_1'(t_n)\right\| + \|x_{1,n} - x_1(t_n)\|)\\ &\leqslant \frac{1}{2}h^{k+1} + \mathcal{O}(h^{2k}) \leqslant h^{k+1}.\end{aligned}$$

Thereby it holds $\Phi(x) \in \Omega_n$ for all $x \in \Omega_n$ and we get a solution of (6.22) by the Schauder Fixed Point Theorem. At this point the induction is concluded and the feasibility of the methods is shown.

To show convergence we use the knowledge that $x_{1,n}$ and $\tilde{x}_{1,n}$ are solutions of (6.22) in Ω_n. Then Equation (6.22a) yields

$$x_{1,n} = -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}x_{1,n-j} + h\frac{1}{\alpha_0}\Psi_{x_1'}(\tilde{x}_1'(t_n) + \mathcal{O}(h^k), x_{1,n}, t_n, \delta_n)$$

which can be written as

$$x_{1,n} = -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}x_{1,n-j} + h\frac{1}{\alpha_0}\Psi_{x_1'}(\tilde{x}_1'(t_n), x_{1,n}, t_n, 0) + \mathcal{O}(h^{k+1}) + J_3\delta_n$$

by a Taylor expansion and the Mean Value Theorem and a $J_3 \in \Omega_{\partial\delta,\Psi_{x_1'}}$. For a more compact notation we define

$$\tilde{f}(x_{1,n}, t_n) := \Psi_{x_1'}(\tilde{x}_1'(t_n), x_{1,n}, t_n, 0)$$

and get

$$x_{1,n} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{1,n-j} + h\frac{1}{\alpha_0}\tilde{f}(x_{1,n}, t_n) + \mathcal{O}(h^{k+1}) + J_3\delta_n. \tag{6.23}$$

Analogous to the discretized system we decouple the DAE

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

along the exact solution $x(t)$. We obtain the equations

$$V_{x_1,n}^\top V_n^\top f(PP_{x_1}\tilde{x}_1'(t) + PQ_{x_1}x_1'(t), PP_{x_1}\tilde{x}_1(t) + PQ_{x_1}x_1(t) + QQ_{y_1,n}\tilde{y}_2(t) + QP_{y_1,n}\tilde{y}_1(t), t) = 0 \tag{6.24a}$$
$$W_{x_1,n}^\top V_n^\top f(PP_{x_1}\tilde{x}_1'(t) + PQ_{x_1}x_1'(t), PP_{x_1}\tilde{x}_1(t) + PQ_{x_1}x_1(t) + QQ_{y_1,n}\tilde{y}_2(t) + QP_{y_1,n}\tilde{y}_1(t), t) = 0 \tag{6.24b}$$
$$V_{y_1,n}^\top W_n^\top f(PP_{x_1}\tilde{x}_1'(t) + PQ_{x_1}x_1'(t), PP_{x_1}\tilde{x}_1(t) + PQ_{x_1}x_1(t) + QQ_{y_1,n}\tilde{y}_2(t) + QP_{y_1,n}\tilde{y}_1(t), t) = 0 \tag{6.24c}$$
$$W_{y_1,n}^\top W_n^\top f(PP_{x_1}\tilde{x}_1'(t) + PQ_{x_1}x_1'(t), PP_{x_1}\tilde{x}_1(t) + QQ_{y_1,n}\tilde{y}_2(t) + QP_{y_1,n}\tilde{y}_1(t), t) = 0 \tag{6.24d}$$

which yield the same inverse functions $\bar{\Psi}_{x_1'}$, $\Psi_{\tilde{y}_1}$, $\Psi_{\tilde{y}_2}$, $\Psi_{\tilde{x}_1}$ and $\Psi_{x_1'}$ as the discretized system by the Lemma 4.32. Hence we obtain

$$x_1'(t) = \bar{\Psi}_{x_1'}(\tilde{x}_1'(t), x_1(t), \tilde{x}_1(t), \tilde{y}_1(t), \tilde{y}_2(t), t, 0)$$

and thereby

$$\tilde{y}_1(t) = \Psi_{\tilde{y}_1}(\tilde{x}_1'(t), x_1(t), \tilde{x}_1(t), t, 0)$$
$$\tilde{y}_2(t) = \Psi_{\tilde{y}_2}(\tilde{x}_1'(t), x_1(t), \tilde{x}_1(t), t, 0)$$

which yield

$$\tilde{x}_1(t) = \Psi_{\tilde{x}_1}(\tilde{x}_1'(t), x_1(t), t, 0)$$

and we finally get

$$x_1'(t) = \Psi_{x_1'}(\tilde{x}_1'(t), x_1(t), t, 0) = \tilde{f}(x_1(t), t).$$

Together with Equation (6.23) we obtain

$$x_{1,n} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{1,n-j} + h\frac{1}{\alpha_0}\tilde{f}(x_{1,n}, t_n) + \mathcal{O}(h^{k+1}) + c_\delta(\cdot)\delta_n$$

$$x_1(t_n) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_1(t_{n-j}) + h\frac{1}{\alpha_0}\tilde{f}(x_1(t_n), t_n) + \mathcal{O}(h^{k+1}).$$

We subtract these equations from each other and get

$$x_1(t_n) - x_{1,n} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0}(x_1(t_{n-j}) - x_{1,n-j}) + h\frac{1}{\alpha_0}(\tilde{f}(x_1(t_n), t_n) - \frac{1}{\alpha_0}\tilde{f}(x_{1,n}, t_n)) + \mathcal{O}(h^{k+1})$$

with $\delta_n \in \mathcal{O}(h^{k+1})$. We use the Mean Value Theorem to obtain

$$x_1(t_n) - x_{1,n} = (\alpha_0 I - hB_n)^{-1}(-\sum_{j=1}^{k} \alpha_j(x_1(t_{n-j}) - x_{1,n-j}) + \mathcal{O}(h^{k+1}))$$

with $B_n := \int_0^1 \frac{\partial}{\partial x}\tilde{f}(sx_1(t_n)+(s-1)x_{1,n}, t_n)\mathrm{d}s$ bounded by a bound independent from n since $x_{1,n} \in B_{h^k}\left(x_1(t_n) + \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} e^x_{1,n-j}\right)$ and $e^x_{1,n-j} \in \mathcal{O}(h^k)$. But this equation is nothing else than the error recursion of a BDF-method applied to an ODE with a Lipschitz continuous function $\tilde{f}$. Hence, the stability of the BDF-methods for $k \leqslant 6$ yields

$$\|x_1(t_n) - x_{1,n}\| \leqslant \mathcal{O}(h^k)$$

which yields together with (6.19) and $(x_{1,n}, \tilde{x}_{1,n}) \in \Omega_n$

$$\|x(t_n) - x_n\| \leqslant \mathcal{O}(h^k).$$

□

Q and Q_{x_1} being constant is the crucial assumption in Theorem 6.6. This assumption can be compared to the definition of numerical qualified DAEs in [HMT03]. In the next section we will present methods which converge independently of this assumption.

6.2 Left-discontinuous Collocation Methods

In the last section we identified the non-linearity or even the time dependence of the basis functions Q and Q_{x_1} as the source of the numerical instabilities while the non-linearity of the functions of the DAE proved itself harmless. By this reason we restrict ourselves to linear time dependent DAEs in this section for the purpose of simplicity.

Definition 6.7. (Linear time dependent DAEs in standard form)
Let $\mathcal{I} \subset \mathbb{R}$ be an open subset. Let $A, B \in C(\mathcal{I}, \mathbb{R}^{n\times n})$ be continuous with A being singular for all $t \in \mathcal{I}$. We call

$$A(t)x'(t) + B(t)x(t) = q(t), \quad x(t_0) = x^0 \tag{6.25}$$

a linear time dependent DAE with $q \in C(\mathcal{I}, \mathbb{R}^n)$. Let $\mathcal{I}_\star := [t_0, T] \subset \mathcal{I}$. We call $x_\star \in C^1(\mathcal{I}_\star, \mathbb{R}^n)$ a solution of (6.25) on $\mathcal{I}_\star$ if the initial conditions are fulfilled, i.e. $x_\star(t_0) = x^0$, and

$$A(t)x_\star'(t) + B(t)x_\star(t) = q(t) \quad \forall t \in \mathcal{I}_\star.$$

We introduce the left-discontinuous collocation methods for linear DAEs. These methods will be able to handle the convergence issues regarding the basis functions Q and Q_{x_1} without a transformation of the DAE in contrast to the methods used in literature, see [KM07].

Definition 6.8. (Left-discontinuous collocation methods)
Let $0 = c_1 < \ldots < c_s = 1$ be real numbers, and let $b \neq 0$ an arbitrary real number. The corresponding left-discontinuous collocation method is then defined via a polynomial of degree $s-1$ satisfying

$$A(t_{n-1})\left(\frac{u(t_{n-1}) - x_{n-1}}{hb} + u'(t_{n-1})\right) + B(t_{n-1})u(t_{n-1}) = q(t_{n-1}) \tag{6.26a}$$

$$A(t_{ni})u'(t_{ni}) + B(t_{ni})u(t_{ni}) = q(t_{ni}), i = 2, \ldots, s \tag{6.26b}$$

$$x_n = u(t_n) \tag{6.26c}$$

with $t_{ni} = t_{n-1} + c_i h$.

A left-discontinuous collocation method can be written as a Runge-Kutta method. The analogous result for ODEs can be found in [HWL06].

Theorem 6.9.
Denote the Lagrange-polynomials by $\ell_i(\tau) = \prod_{l=2, l\neq i}^{s} \frac{\tau - c_l}{c_i - c_l}$ and define

$$(\mathcal{A})_{ij} := a_{ij} := \begin{cases} b, & j = 1, \\ \int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau - b\ell_j(0), & \text{else.} \end{cases}$$

The matrix $\mathcal{A}$ is non-singular and we notate $(\mathcal{A}^{-1})_{ij} := \alpha_{ij}$. Then the Equations (6.26) are equivalent to

$$A(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{u(t_{nj}) - x_{n-1}}{h} + B(t_{ni})u(t_{ni}) = q(t_{ni}), \qquad i = 1, \ldots, s$$

$$x_n = u(t_n).$$

Proof.

First we show that $\mathcal{A}$ is non-singular. Therefore we define the k-th power of the vector of the nodes c_i by $c^k := \begin{pmatrix} \dots & c_i^k & \dots \end{pmatrix}^\top$. Then it holds for every discontinuous collocation method, cf. [HWL06] Theorem 1.8,

$$\mathcal{A}c^{k-1} := \frac{1}{k}c^k, \qquad\qquad k = 1, \dots, s-1.$$

Together with $\mathcal{A}\begin{pmatrix} 1 & 0 & \dots & 0 \end{pmatrix}^\top = b\begin{pmatrix} 1 & \dots & 1 \end{pmatrix}^\top$ this yields $\{c^0, c^1, \dots, c^{s-1}\} \subset \operatorname{im}\mathcal{A}$ and since $\begin{pmatrix} c^0 & c^1 & \dots & c^{s-1} \end{pmatrix}$ is a Vandermonde-Matrix due to $c_i < c_j$ for $i < j$ we obtain

$$\mathbb{R}^s = \operatorname{im}\begin{pmatrix} c^0 & c^1 & \dots & c^{s-1} \end{pmatrix} \subset \operatorname{im}\mathcal{A},$$

which yields the regularity of $\mathcal{A}$.
We define $k_1 := \frac{u(t_{n-1}) - x_{n-1}}{hb} + u'(t_{n-1})$ and $k_i := u'(t_{ni})$ for $i = 2, \dots, s$, such that (6.26) can be written as:

$$\begin{aligned} A(t_{ni})k_i + B(t_{ni})u(t_{ni}) &= q(t_{ni}), \quad i = 1, \dots, s \\ x_n &= u(t_n). \end{aligned}$$

We can represent the derivative of the polynomial u as

$$u'(t_{n-1} + \tau h) = \sum_{j=2}^{s} u'(t_{nj})\ell_j(\tau) = \sum_{j=2}^{s} k_j\ell_j(\tau) \tag{6.27}$$

with the help of the Lagrange-polynomials $\ell_i(\tau)$. This yields in particular $u'(t_{n-1}) = \sum_{j=2}^{s} k_j\ell_j(0)$, while the definition of k_1 can be rearranged into

$$\begin{aligned} u(t_{n-1}) &= x_{n-1} + hbk_1 - hbu'(t_{n-1}) \\ &= x_{n-1} + hbk_1 - hb\sum_{j=2}^{s} k_j\ell_j(0). \end{aligned}$$

We integrate (6.27) from 0 to c_i for $i = 2, \dots, s$ and obtain

$$\begin{aligned} u(t_{ni}) &= u(t_{n-1}) + h\sum_{j=2}^{s} k_j \int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau \\ &= x_{n-1} + hbk_1 + h\sum_{j=2}^{s} k_j \left(\int_0^{c_i} \ell_j(\tau)\mathrm{d}\tau - b\ell_j(0) \right) \\ &= x_{n-1} + h\sum_{j=1}^{s} a_{ij}k_j. \end{aligned}$$

Hence we get $\frac{u(t_{ni})-x_{n-1}}{h} = \sum_{j=1}^{s} a_{ij}k_j$ for $i = 1, \dots, s$, which yields

$$k_i = \sum_{j=1}^{s} \alpha_{ij} \frac{u(t_{nj}) - x_{n-1}}{h},$$

since $\mathcal{A}$ is non-singular. □

We integrate (6.27) from 0 to 1 and obtain

$$u(t_{ns}) = x_{n-1} + hbk_1 + h \sum_{j=2}^{s} a_{sj}k_j.$$

By Equation (6.26c) we achieve

$$x_n = x_{n-1} + hbk_1 + h \sum_{j=2}^{s} a_{sj}k_j$$

hence the collocation method (6.26) is a Runge-Kutta method with the Butcher-tableau

$$\begin{array}{c|cccc} 0 & b & a_{12} & \dots & a_{1s} \\ c_2 & & & & \\ \vdots & \vdots & \vdots & & \vdots \\ c_{s-1} & & & & \\ 1 & b & a_{s2} & \dots & a_{ss} \\ \hline & b & a_{s2} & \dots & a_{ss} \end{array}$$

Therefore the Runge-Kutta method is stiffly accurate with $\mathcal{A} := (a_{ij})$ being non-singular, but additionally we have the property

$$a_{i1} = b \text{ for } 1 \leqslant i \leqslant s \text{ and } c_1 = 0, \tag{6.28}$$

which will be crucial for the convergence of the method. We define the consistency error as in [LMT13].

Definition 6.10. (Consistency error) We define the consistency error of a discontinuous collocation method by

$$L_{ni}(x) := \frac{1}{h} \sum_{j=1}^{s} \alpha_{ij}(x(t_{nj}) - x(t_{n-1})) - x'(t_{ni}).$$

We say a discontinuous collocation method has a consistency error of order k if $L_{ni}(x) = \mathcal{O}(h^k)$ for $i = 1, \dots, s$.

With the help of a standard Taylor expansion approach we describe the consistency order of the collocation methods (6.26).

Theorem 6.11. (Consistency)
All discontinuous collocation methods (6.26) have a consistency error of order $s-1$.

Proof.
First we show that for every left-discontinuous collocation method holds:

$$c_1 = 0 \quad \text{and} \quad \sum_{j=1}^{s} \alpha_{ij} = 0, \qquad i = 2, \dots, s \tag{6.29}$$

and

$$\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i) = 1 \text{ and } \sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i)^k = 0, \qquad i = 1, \dots, s. \tag{6.30}$$

for $k = 2, \dots, s-1$. It holds $c_1 = 0$ by definition and $\sum_{j=1}^{s} \alpha_{ij} = 0$ holds due to

$$\mathcal{A}^{-1}\mathcal{A} = I \quad \Rightarrow \quad \mathcal{A}^{-1}\mathcal{A}\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \quad \Rightarrow \quad \mathcal{A}^{-1}b\begin{pmatrix}1\\\vdots\\1\end{pmatrix} = \begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} \quad \Rightarrow \quad \sum_{j=1}^{s} \alpha_{ij} = 0$$

for $i = 2, \dots, s$ and $b \neq 0$. To prove (6.30) we need again the collocation property from [HWL06]

$$\begin{aligned}
\mathcal{A}c^{k-1} &= \frac{1}{k}c^k, && k = 1, \dots, s-1\\
\Rightarrow kc^{k-1} &= \mathcal{A}^{-1}c^k, && k = 1, \dots, s-1\\
\Rightarrow kc_i^{k-1} &= \sum_{j=1}^{s} \alpha_{ij}c_j^k, && i = 1, \dots, s, \quad k = 1, \dots, s-1.
\end{aligned}$$

By (6.29) for $k = 1$ it follows:

$$\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i) = -c_i\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij} + \sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}c_j = \alpha_{ii}c_i + \sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}c_j = \sum_{j=1}^{s} \alpha_{ij}c_j = 1$$

and for $k = 2, \dots, s$ and $i = 1$ there follows:

$$\sum_{\substack{j=1\\j\neq 1}}^{s} \alpha_{1j}(c_j - c_1)^k = \sum_{j=2}^{s} \alpha_{1j}c_j^k = \sum_{j=1}^{s} \alpha_{1j}c_j^k = kc_1^{k-1} = 0$$

while for $i = 2, \dots, s$ there follows:

$$\begin{aligned}
&\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i)^k \\
=&\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij} \sum_{l=0}^{k} \binom{k}{l} c_j^{k-l} c_i^l (-1)^l \\
=&\sum_{l=0}^{k} \binom{k}{l} c_i^l (-1)^l \sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij} c_j^{k-l} \\
=&\sum_{l=0}^{k} \binom{k}{l} c_i^l (-1)^l ((k-l)c_i^{k-l-1} - \alpha_{ii} c_i^{k-l}) \\
=&c_i^{k-1} \sum_{l=0}^{k} \binom{k}{l} (-1)^l (k - l - \alpha_{ii} c_i) \\
=&c_i^{k-1} \sum_{l=0}^{k} \binom{k}{l} (-1)^l (k - \alpha_{ii} c_i) + c_i^{k-1} \sum_{l=0}^{k} \binom{k}{l} (-1)^l (-l) \\
=&(k - a_{ii} c_i) c_i^{k-1} \sum_{l=0}^{k} \binom{k}{l} (-1)^l + k c_i^{k-1} \sum_{l=1}^{k} \binom{k-1}{l-1} (-1)^{l-1} \\
=&(k - a_{ii} c_i) c_i^{k-1} (1-1)^k + k c_i^{k-1} (1-1)^{k-1} = 0.
\end{aligned}$$

With the help of the Taylor expansions

$$x(t_{nj}) := x(t_{ni}) + x'(t_{ni})(c_j - c_i)h + \sum_{k=2}^{s-1} \frac{1}{k!} x^{(k)}(t_{ni})(c_j - c_i)^k h^k + \mathcal{O}(h^s)$$

we obtain the desired result:

$$\begin{aligned}
L_{ni}(x) :=&\frac{1}{h}(x(t_{ni}) - x(t_{n-1})) \sum_{j=1}^{s} \alpha_{ij} + (\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i) - 1) x'(t_{ni}) \\
&+ \sum_{k=2}^{s-1} \frac{1}{k!} x^{(k)}(t_{ni}) (\sum_{\substack{j=1\\j\neq i}}^{s} \alpha_{ij}(c_j - c_i)^k) h^{k-1} + \mathcal{O}(h^{s-1}) \\
=&\mathcal{O}(h^{s-1}).
\end{aligned}$$

□

We chose the discontinuous collocation methods (6.26) to overcome the instabilities created by artificial dynamics as in Example 2.12. The singular matrix problem of Example 2.12 however is not eliminated by this choice. At least not for DAEs in standard formulation. We formulate the singular matrix problem in terms of the applicability of a discontinuous collocation method:

Definition 6.12.
Consider a s-stage discontinuous collocation method (6.26) with $s \geqslant 2$ and an index 2 DAE $A(t)x' + B(t)x = q(t)$. We call the collocation method applicable to the DAE if the matrix

$$\frac{1}{h}\begin{pmatrix} \ddots & & \\ & A(t_i) & \\ & & \ddots \end{pmatrix}(\mathcal{A} \otimes I) + \begin{pmatrix} \ddots & & \\ & B(t_i) & \\ & & \ddots \end{pmatrix} \tag{6.31}$$

is non-singular, with $t_i = t + c_i h$ for $t \in \mathcal{I}$.

Furthermore we define the matrix

$$(M)_{ij} := \begin{cases} 0, & i = j, \\ \alpha_{ij}(c_j - c_i), & j \neq i \end{cases} \tag{6.32}$$

and formulate the following lemma which will be needed to provide an equivalent criterion for the applicability of a collocation method:

Lemma 6.13.
Consider a s-stage discontinuous collocation method (6.26) with $s \geqslant 2$ and two matrices $A, B \in \mathbb{R}^{n \times n}$. Let $A + B$ be non-singular let M be the matrix as described above. It follows that

$$M \otimes A + I \otimes B$$

is non-singular if $A - \frac{1}{s-1}B$ is non-singular.

Proof.
For the entries of M holds

$$\sum_{\substack{j=1\\ j\neq i}}^{s} m_{ij} = 1, \qquad\qquad i = 1, \ldots, s$$

$$\sum_{\substack{j=1\\ j\neq i}}^{s} m_{ij}(c_j - c_i)^k = 0, \qquad k = 1, \ldots, s-2, \quad i = 1, \ldots, s$$

by (6.30). These conditions can be written as

$$C_i\hat{m}_i = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \quad \text{with} \quad (C_i)_{kj} := \begin{cases} (c_j - c_i)^{k-1}, & j < i, \\ (c_{j+1} - c_i)^{k-1}, & j \geqslant i \end{cases}$$

and $\hat{m}_i$ the i-th row of M without the diagonal entry. Again we deal with a Vandermonde matrix, hence we can write

$$m_{ij} = \prod_{\substack{k=1 \\ k\neq i,j}}^{s} \frac{c_k - c_i}{c_k - c_j} \quad \text{with } j \neq i \tag{6.33}$$

as the first column of each inverse of the matrices C_i. With the help of this explicit description (6.33) it follows

$$m_{1j}m_{j1} = \prod_{\substack{k=1 \\ k\neq 1,j}}^{s} \frac{c_k - c_1}{c_k - c_j} \prod_{\substack{k=1 \\ k\neq j,1}}^{s} \frac{c_k - c_j}{c_k - c_1} = 1 \tag{6.34}$$

for $j > 1$ and

$$\begin{aligned} m_{1l}m_{lj} &= \prod_{\substack{k=1 \\ k\neq 1,l}}^{s} \frac{c_k - c_1}{c_k - c_l} \prod_{\substack{k=1 \\ k\neq l,j}}^{s} \frac{c_k - c_l}{c_k - c_j} \\ &= \left(\prod_{\substack{k=1 \\ k\neq 1,l,j}}^{s} \frac{c_k - c_1}{c_k - c_l} \right) \left(\prod_{\substack{k=1 \\ k\neq l,j,1}}^{s} \frac{c_k - c_l}{c_k - c_j} \right) \frac{c_j - c_1}{c_j - c_l} \frac{c_1 - c_l}{c_1 - c_j} \\ &= - \left(\prod_{\substack{k=1 \\ k\neq 1,l,j}}^{s} \frac{c_k - c_1}{c_k - c_j} \right) \frac{c_l - c_1}{c_l - c_j} \\ &= - \left(\prod_{\substack{k=1 \\ k\neq 1,j}}^{s} \frac{c_k - c_1}{c_k - c_j} \right) \\ &= -m_{1j} \end{aligned} \tag{6.35}$$

for $j \neq l > 1$. After these preparations we prove that

$$M \otimes A + I \otimes B = \begin{pmatrix} B & m_{12}A & m_{13}A & \dots & m_{1s}A \\ m_{21}A & B & m_{23}A & \dots & m_{2s}A \\ m_{31}A & m_{32}A & \ddots & & \\ \vdots & \vdots & & \ddots & \\ m_{s1}A & m_{s2}A & & & B \end{pmatrix}$$

is non-singular if $A + B$ and $B - (s-1)A$ are non-singular. To do so we transform $M{\otimes}A{+}I{\otimes}B$ into a matrix which is obviously non-singular and we only use transformation which preserve the non-singularity. We start by multiplying the k-th row block by m_{1k} for $k > 1$ and obtain:

$$\begin{pmatrix} B & m_{12}A & m_{13}A & \dots & m_{1s}A \\ A & m_{12}B & -m_{13}A & \dots & -m_{1s}A \\ A & -m_{12}A & \ddots & & \\ \vdots & \vdots & & \ddots & \\ A & -m_{12}A & & & m_{1s}B \end{pmatrix}$$

by (6.34) and (6.35). We add the first row block to all other rows and afterwards multiply all but the first rows by $(A + B)^{-1}$ to achieve:

$$\begin{pmatrix} B & m_{12}A & m_{13}A & \dots & m_{1s}A \\ I & m_{12}I & 0 & \dots & 0 \\ I & 0 & \ddots & & \\ \vdots & \vdots & & \ddots & \\ I & 0 & & & m_{1s}I. \end{pmatrix}$$

By multiplying all rows but the first row with A and subtracting them from the first row we get

$$\begin{pmatrix} B - (s-1)A & 0 & 0 & \dots & 0 \\ I & m_{12}I & 0 & \dots & 0 \\ I & 0 & \ddots & & \\ \vdots & \vdots & & \ddots & \\ I & 0 & & & m_{1s}I \end{pmatrix}$$

and the proof is concluded. □

Now we can state a lemma which helps to exclude the regularity problem.

Lemma 6.14.
Consider a s-stage discontinuous collocation method (6.26) with $s \geqslant 2$ and a DAE $A(t)x' + B(t)x = q(t)$ with Dissection Index 2 with $A, B \in C^2(\mathcal{I}, \mathbb{R}^{n\times n})$.
Then there is a $H > 0$ such that the collocation method is applicable for all $h \leqslant H$ if $A(t)x' - \frac{1}{s-1}B(t)x = q(t)$ is a DAE with Dissection Index 2.

Proof. First of all we mention that the basis functions P, Q, V and W as well as the basis functions $P_{y_1}, Q_{y_1}, V_{y_1}, W_{y_1}, P_{x_1}, Q_{x_1}, V_{x_1}$ and W_{x_1} are two times continuously differentiable due to Lemma 4.7 and $A, B \in C^2(\mathcal{I}, \mathbb{R}^{n\times n})$.
In the following we show that the matrix (6.31) has a trivial kernel if $A(t)x' - \frac{1}{s-1}B(t)x = q(t)$ is a DAE with Dissection Index 2. Therefore let $X = (X_1, \dots, X_s)$ be an element of the kernel of (6.31):

$$A(t_i)\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}X_j + B(t_i)X_i = 0, \qquad \forall 1 \leqslant i \leqslant s,$$

which can be written as

$$V^\top(t_i)A(t_i)\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}X_j + V^\top(t_i)B(t_i)X_i = 0 \tag{6.36a}$$

$$W^\top(t_i)B(t_i)X_i = 0. \tag{6.36b}$$

For $1 \leqslant i \leqslant s$ we transform X_i with the help of the basis functions $P, Q, P_{x_1}, Q_{x_1}, P_{y_1}$ and Q_{y_1}:

$$\begin{aligned} X_i &= P(t_i)X_i^{x_0} + Q(t_i)X_i^{y_0} \\ &= P(t_i)(Q_{x_1}(t_i)X_i^{x_2} + P_{x_1}(t_i)X_i^{\tilde{x}_1}) + Q(t_i)(Q_{y_1}(t_i)X_i^{\tilde{y}_2} + P_{y_1}(t_i)X_i^{\tilde{y}_1}). \end{aligned}$$

We factorize Equation (6.36b) by $V_{y_1}^\top$ and $W_{y_1}^\top$ and insert the variable transformation to obtain

$$\begin{aligned} V_{y_1}^\top(t_i)B_{x_1}^{\mathrm{w}}(t_i)P_{x_1}(t_i)X_i^{\tilde{x}_1} + V_{y_1}^\top(t_i)B_{x_1}^{\mathrm{w}}(t_i)Q_{x_1}(t_i)X_i^{x_2} + V_{y_1}^\top(t_i)B_{y_1}^{\mathrm{w}}(t_i)P_{y_1}(t_i)X_i^{\tilde{y}_1} &= 0, \\ W_{y_1}^\top(t_i)B_{x_1}^{\mathrm{w}}(t_i)P_{x_1}(t_i)X_i^{\tilde{x}_1} &= 0. \end{aligned}$$

This yields

$$X_i^{\tilde{y}_1} = -(V_{y_1}^\top(t_i)B_{y_1}^{\mathrm{w}}(t_i)P_{y_1}(t_i))^{-1}V_{y_1}^\top(t_i)B_{x_1}^{\mathrm{w}}(t_i)Q_{x_1}(t_i)X_i^{x_2} = M_{x_2,y_1}^{i}X_i^{x_2}, \tag{6.37a}$$

$$X_i^{\tilde{x}_1} = 0. \tag{6.37b}$$

We remember the definition of $t_i = t + c_i h$ as in Definition 6.12 and notate the Taylor series of the basis function P and Q at t_i evaluated in t_j

$$\begin{aligned} P(t_j) &= P(t_i) + P'(t_i)(t_j - t_i) + R_P(t_i, t_j)(t_j - t_i)^2 \\ &= P(t_i) + P'(t_i)(c_j - c_i)h + R_P(t_i, t_j)(c_j - c_i)^2h^2, \\ Q(t_j) &= Q(t_i) + Q'(t_i)(c_j - c_i)h + R_Q(t_i, t_j)(c_j - c_i)^2h^2 \end{aligned} \tag{6.38}$$

for $1 \leqslant i, j \leqslant s$ with $R_P(t_i, t_j)$ and $R_Q(t_i, t_j)$ being the residual term of the Taylor series expansion. Next we insert the variable transformations into Equation (6.36a)

$$V^\top(t_i)A(t_i)\frac{1}{h}\sum_{j=1}^{s}\alpha_{ij}(P(t_j)X_j^{x_0} + Q(t_j)X_j^{y_0}) + V^\top(t_i)B(t_i)X_i = 0$$

and make use of the Taylor series to obtain

$$\begin{aligned}0 =&G_1(t_i)\sum_{j=1}^{s}\alpha_{ij}\frac{X_j^{x_0}}{h}\\&+\sum_{\substack{j=1\\j\neq i}}^{s}\alpha_{ij}(c_j - c_i)V^\top(t_i)A(t_i)P'(t_i)X_j^{x_0} + V^\top(t_i)B(t_i)P(t_i)X_i^{x_0} + h\sum_{\substack{j=1\\j\neq i}}^{s}M_P(t_i, t_j)X_j^{x_0}\\&+\sum_{\substack{j=1\\j\neq i}}^{s}\alpha_{ij}(c_j - c_i)V^\top(t_i)A(t_i)Q'(t_i)X_j^{y_0} + V^\top(t_i)B(t_i)Q(t_i)X_i^{y_0} + h\sum_{\substack{j=1\\j\neq i}}^{s}M_Q(t_i, t_j)X_j^{y_0}\end{aligned}$$

with

$$M_P(t_i, t_j) := \alpha_{ij}(c_j - c_i)^2V^\top(t_i)A(t_i)R_P(t_i, t_j) \tag{6.39a}$$

$$M_Q(t_i, t_j) := \alpha_{ij}(c_j - c_i)^2V^\top(t_i)A(t_i)R_Q(t_i, t_j). \tag{6.39b}$$

We split the variable into

$$\begin{aligned}X_i^{x_0} &= Q_{x_1}(t_i)X_i^{x_2}\\X_i^{y_0} &= Q_{y_1}(t_i)X_i^{\tilde{y}_2} - P_{y_1}(t_i)(V_{y_1}^\top(t_i)B_{y_1}^{\mathrm{w}}(t_i)P_{y_1}(t_i))^{-1}V_{y_1}^\top(t_i)B_{x_1}^{\mathrm{w}}(t_i)Q_{x_1}(t_i)X_i^{x_2}\end{aligned}$$

with the help of (6.37). Again we make use of a Taylor series expansion:

$$\begin{aligned}Q_{y_1}(t_j) &= Q_{y_1}(t_i) + R_{Q_{y_1}}(t_i, t_j)(c_j - c_i)h,\\Q_{x_1}(t_j) &= Q_{x_1}(t_i) + R_{Q_{x_1}}(t_i, t_j)(c_j - c_i)h\end{aligned} \tag{6.40}$$

to obtain

$$\begin{aligned}0 =&G_1(t_i)Q_{x_1}(t_i)\sum_{j=1}^{s}\alpha_{ij}\frac{X_j^{x_2}}{h} + M_{x_2}^{i}X^{x_2} \qquad (6.41)\\&+\sum_{\substack{j=1\\j\neq i}}^{s}\alpha_{ij}(c_j - c_i)V^\top(t_i)A(t_i)Q'(t_i)Q_{y_1}(t_i)X_j^{\tilde{y}_2} + V^\top(t_i)B(t_i)Q(t_i)Q_{y_1}(t_i)X_i^{\tilde{y}_2}\\&+h\sum_{\substack{j=1\\j\neq i}}^{s}\left(M_Q(t_i, t_j)Q_{y_1}(t_j) + \alpha_{ij}(c_j - c_i)V^\top(t_i)A(t_i)Q'(t_i)R_{Q_{y_1}}(t_i, t_j)(c_j - c_i)\right)X_j^{\tilde{y}_2}\end{aligned}$$

with certain matrices $M^i_{x_2}$. In the following we will deal with the $X^{\tilde{y}_2}$ component. Therefore we multiply (6.41) by $W^\top_{x_1}(t_i)$ and obtain

$$\begin{aligned}0 = &W^\top_{x_1}(t_i)M^i_{x_2}X^{x_2}\\ &+\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)W^\top_{x_1}(t_i)V^\top(t_i)A(t_i)Q'(t_i)Q_{y_1}(t_i)X^{\tilde{y}_2}_j + W^\top_{x_1}(t_i)V^\top(t_i)B(t_i)Q(t_i)Q_{y_1}(t_i)X^{\tilde{y}_2}_i\\ &+h\sum_{\substack{j=1\\ j\neq i}}^{s}W^\top_{x_1}(t_i)\left(M_Q(t_i,t_j)Q_{y_1}(t_j)+\alpha_{ij}(c_j-c_i)V^\top(t_i)A(t_i)Q'(t_i)R_{Q_{y_1}}(t_i,t_j)(c_j-c_i)\right)X^{\tilde{y}_2}_j\end{aligned}$$

With the help of two additional Taylor series expansions

$$\begin{aligned}(W^\top_{x_1}V^\top AQ'Q_{y_1})(t_i) &= (W^\top_{x_1}V^\top AQ'Q_{y_1})(t) + R_A(t_i,t)c_ih\\ (W^\top_{x_1}V^\top BQQ_{y_1})(t_i) &= (W^\top_{x_1}V^\top BQQ_{y_1})(t) + R_B(t_i,t)c_ih\end{aligned}\tag{6.42}$$

we then get

$$\begin{aligned}0 =&W^\top_{x_1}(t_i)M^i_{x_2}X^{x_2}\\ &+\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)(W^\top_{x_1}V^\top AQ'Q_{y_1})(t)X^{\tilde{y}_2}_j + (W^\top_{x_1}V^\top BQQ_{y_1})(t)X^{\tilde{y}_2}_i\\ &+h\sum_{\substack{j=1\\ j\neq i}}^{s}M_{Q_{y_1}}(t_i,t_j,t)X^{\tilde{y}_2}_j\end{aligned}$$

by denoting the matrices in front of the $hX^{\tilde{y}_2}_j$ terms by $M_{Q_{y_1}}(t_i,t_j,t)$ and get

$$\begin{aligned}&M^i_{x_2}X^{x_2}\\ =&\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)W^\top_{x_1}(t)V^\top(t)A(t)Q'(t)Q_{y_1}(t)X^{\tilde{y}_2}_j + W^\top_{x_1}(t)V^\top(t)B(t)Q(t)Q_{y_1}(t)X^{\tilde{y}_2}_i\\ +&h\sum_{\substack{j=1\\ j\neq i}}^{s}M_{Q_{y_1}}(t_i,t_j,t)X^{\tilde{y}_2}_j.\end{aligned}$$

This equation can be written as

$$(M\otimes(W^\top_{x_1}V^\top AQ'Q_{y_1})(t) + I\otimes(W^\top_{x_1}V^\top BQQ_{y_1})(t) + h\bar{M}_{Q_{y_1}})X^{\tilde{y}_2} =M_{x_2}X^{x_2}$$

with the help of the matrix tensor product $\otimes$ and the matrix

$$(M)_{ij} := \begin{cases}0, & i=j,\\ \alpha_{ij}(c_j-c_i), & j\neq i.\end{cases}.$$

By Lemma 6.13 the matrix

$$M \otimes (W_{x_1}^\top V^\top AQ'Q_{y_1})(t) + I \otimes (W_{x_1}^\top V^\top BQQ_{y_1})(t)$$

is non-singular because

$$\begin{aligned}
&(W_{x_1}^\top V^\top AQ'Q_{y_1})(t) + (W_{x_1}^\top V^\top BQQ_{y_1})(t)\\
=&(W_{x_1}^\top V^\top AQ'Q_{y_1} + W_{x_1}^\top V^\top BQQ_{y_1})(t)\\
=&(W_{x_1}^\top (V^\top AQ' + V^\top BQ)Q_{y_1})(t)\\
=&(W_{x_1}^\top B_{y_1}^{\mathrm{w}} Q_{y_1})(t)\\
=&B_{y_2}^{\mathrm{w}}(t)
\end{aligned}$$

is non-singular since we deal with an index 2 DAE and

$$\begin{aligned}
&(W_{x_1}^\top V^\top AQ'Q_{y_1})(t) - \frac{1}{s-1}(W_{x_1}^\top V^\top BQQ_{y_1})(t)\\
=&(W_{x_1}^\top V^\top AQ'Q_{y_1})(t) + \left(W_{x_1}^\top V^\top \left(-\frac{1}{s-1}B\right) QQ_{y_1}\right)(t)\\
=&\left(W_{x_1}^\top V^\top AQ'Q_{y_1} + W_{x_1}^\top V^\top \left(-\frac{1}{s-1}B\right) QQ_{y_1}\right)(t)\\
=&\left(W_{x_1}^\top \left(V^\top AQ' + V^\top \left(-\frac{1}{s-1}B\right) Q\right) Q_{y_1}\right)(t)\\
=&\tilde{B}_{y_2}^{\mathrm{w}}(t)
\end{aligned}$$

is non-singular since $A(t)x' - \frac{1}{s-1}B(t)x = q(t)$ is an index 2 DAE, here $\tilde{B}_{y_2}^{\mathrm{w}}$ belongs to the matrix chain of $A(t)x' - \frac{1}{s-1}B(t)x = q(t)$. By the Banach Perturbation Lemma we get

$$X^{\tilde{y}_2} = M_{x_2,\tilde{y}_2} X^{x_2} \tag{6.43}$$

with $M_{x_2,\tilde{y}_2} := (M \otimes (W_{x_1}^\top V^\top AQ'Q_{y_1})(t) + I \otimes (W_{x_1}^\top V^\top BQQ_{y_1})(t) + h\bar{M}_{Q_{y_1}})^{-1} M_{x_2}$. Therefore we only have to deal with the X^{x_2} component. Again we start with Equation (6.41)

$$\begin{aligned}
0 =& G_1(t_i)Q_{x_1}(t_i) \sum_{j=1}^{s} \alpha_{ij} \frac{X_j^{x_2}}{h} + M_{x_2}^i X^{x_2}\\
&+ \sum_{\substack{j=1\\ j\neq i}}^{s} \alpha_{ij}(c_j - c_i) V^\top(t_i)A(t_i)Q'(t_i)Q_{y_1}(t_i)X_j^{\tilde{y}_2} + V^\top(t_i)B(t_i)Q(t_i)Q_{y_1}(t_i)X_i^{\tilde{y}_2}\\
&+ h \sum_{\substack{j=1\\ j\neq i}}^{s} \left(M_Q(t_i,t_j)Q_{y_1}(t_j) + \alpha_{ij}(c_j - c_i)V^\top(t_i)A(t_i)Q'(t_i)R_{Q_{y_1}}(t_i,t_j)(c_j - c_i)\right) X_j^{\tilde{y}_2}
\end{aligned}$$

By (6.37a), (6.37b) and (6.43) there exist matrices $K^i_{x_2}$ such that it holds:

$$G_1(t_i)Q_{x_1}(t_i)\sum_{j=1}^{s}\alpha_{ij}X_j^{x_2} + hK^i_{x_2}X^{x_2} = 0.$$

With the help of a multiplication by $G_2^{-1}(t_i)V_{x_1}^{\top}(t_i)$ from the left we obtain

$$\sum_{j=1}^{s}\alpha_{ij}X_j^{x_2} + hG_2^{-1}(t_i)V_{x_1}^{\top}(t_i)K^i_{x_2}X^{x_2} = 0$$

which can be written as

$$(\mathcal{A}^{-1}\otimes I + h\bar{K}_{x_2})X^{x_2} = 0$$

with the help of the matrix tensor product and

$$\bar{K}_{x_2} = \begin{pmatrix} \vdots \\ G_2^{-1}(t_i)V_{x_1}^{\top}(t_i)K^i_{x_2} \\ \vdots \end{pmatrix}.$$

Finally the Banach Perturbation Lemma yields

$$X^{x_2} = 0$$

which concludes the proof. □

Finally we show that a discontinuous collocation method converges if it is applicability.

Theorem 6.15.
Consider a s-stage discontinuous collocation method (6.26) and a DAE

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

with Dissection Index 2, $A, B \in C^2(\mathcal{I}, \mathbb{R}^{n\times n})$ and q being continuously differentiable. Furthermore let the DAE be uniquely solvable by a global solution which is sufficiently smooth. Let x^0 fulfill the obvious constraints, i.e. $W^{\top}(t_0)B(t_0)x^0 = W^{\top}(t_0)q(t_0)$, and let $A(t)x' - \frac{1}{s-1}B(t)x = q(t)$ have Dissection Index 2. Then for $h > 0$ and $\delta_n \in \mathcal{O}(h^s)$ it holds:

$$\exists C > 0: \qquad ||e_n||_\infty \leqslant C(||e_0||_\infty + h^{s-1})$$

with $e_n := x(t_n) - x_n$ being the global error at t_n and δ_n being the rounding and solver errors in the n-th step.

Proof.
By Definition 6.10 it holds

$$A(t_{ni})\sum_{j=1}^{s}\alpha_{ij}\frac{x(t_{nj})-x(t_{n-1})}{h}+B(t_{ni})x(t_{ni})=q(t_{ni})+A(t_{ni})L_{ni}(x),\qquad \forall 1\leqslant i\leqslant s,$$
$$x(t_n)=x(t_n).$$

Due to Lemma 6.14 there are X_{ni} which fulfill

$$A(t_{ni})\sum_{j=1}^{s}\alpha_{ij}\frac{X_{nj}-x_{n-1}}{h}+B(t_{ni})X_{ni}=q(t_{ni})-\delta_{ni},\qquad \forall 1\leqslant i\leqslant s,$$
$$x_n=X_{ns}$$

with $\delta_n := \begin{pmatrix}\ldots & \delta_{ni} & \ldots\end{pmatrix}^\top$. We call δ_n the global perturbation and δ_{ni} the stage perturbations. Then x_n is the solution of the discontinuous collocation method 6.26 regarding these perturbations. We define $e_{ni} := x(t_{ni}) - X_{ni}$ and $e_n := x(t_n) - x_n$ and obtain

$$A(t_{ni})\sum_{j=1}^{s}\alpha_{ij}\frac{e_{nj}-e_{n-1}}{h}+B(t_{ni})e_{ni}=A(t_{ni})L_{ni}(x)+\delta_{ni},\qquad \forall 1\leqslant i\leqslant s,\tag{6.44a}$$
$$e_n=e_{ns}.\tag{6.44b}$$

We split the stage errors into

$$\begin{aligned}e_{ni}&=P(t_{ni})e_{ni}^{x_0}+Q(t_{ni})e_{ni}^{y_0}\\&=P(t_{ni})(Q_{x_1}(t_{ni})e_{ni}^{x_2}+P_{x_1}(t_{ni})e_{ni}^{\tilde{x}_1})+Q(t_{ni})(Q_{y_1}(t_{ni})e_{ni}^{\tilde{y}_2}+P_{y_1}(t_{ni})e_{ni}^{\tilde{y}_1})\end{aligned}\tag{6.45}$$

and the global step errors into

$$\begin{aligned}e_{n}&=P(t_{n})e_{n}^{x_0}+Q(t_{n})e_{n}^{y_0}\\&=P(t_{n})(Q_{x_1}(t_{n})e_{n}^{x_2}+P_{x_1}(t_{n})e_{n}^{\tilde{x}_1})+Q(t_{n})(Q_{y_1}(t_{n})e_{n}^{\tilde{y}_2}+P_{y_1}(t_{n})e_{n}^{\tilde{y}_1}).\end{aligned}\tag{6.46}$$

We multiply the i-th equation of (6.44a) by $W^\top(t_{ni})$ and get

$$W^\top(t_{ni})B(t_{ni})P(t_{ni})e_{ni}^{x_0}+W^\top(t_{ni})B(t_{ni})Q(t_{ni})e_{ni}^{y_0}=W^\top(t_{ni})\delta_{ni}$$

for all $1\leqslant i\leqslant s$. Using the matrix chain notation we obtain

$$B_{x_1}^{\mathrm{w}}(t_{ni})e_{ni}^{x_0}+B_{y_1}^{\mathrm{w}}(t_{ni})e_{ni}^{y_0}=W^\top(t_{ni})\delta_{ni},$$

which can be further decoupled to

$$V_{y_1}^\top B_{x_1}^{\mathrm{w}}P_{x_1}e_{ni}^{\tilde{x}_1}+V_{y_1}^\top B_{x_1}^{\mathrm{w}}Q_{x_1}e_{ni}^{x_2}+V_{y_1}^\top B_{y_1}^{\mathrm{w}}P_{y_1}e_{ni}^{\tilde{y}_1}=V_{y_1}^\top W^\top\delta_{ni},$$

$$W_{y_1}^\top B_{x_1}^{\mathrm{w}} P_{x_1} e_{ni}^{\tilde{x}_1} = W_{y_1}^\top W^\top \delta_{ni}$$

evaluated in t_{ni}. Hence we obtain

$$\begin{aligned} e_{ni}^{\tilde{y}_1} &= \delta_{ni}^{y_1} - (V_{y_1}^\top(t_{ni}) B_{y_1}^{\mathrm{w}}(t_{ni}) P_{y_1}(t_{ni}))^{-1} V_{y_1}^\top(t_{ni}) B_{x_1}^{\mathrm{w}}(t_{ni}) Q_{x_1}(t_{ni}) e_{ni}^{x_2}, \\ e_{ni}^{\tilde{x}_1} &= \delta_{ni}^{x_1} \end{aligned} \tag{6.47}$$

with

$$\begin{aligned} \delta_{ni}^{x_1} &:= (W_{y_1}^\top(t_{ni}) B_{x_1}^{\mathrm{w}}(t_{ni}) P_{x_1}(t_{ni}))^{-1} W_{y_1}^\top(t_{ni}) W^\top(t_{ni}) \delta_{ni}, \\ \delta_{ni}^{y_1} &:= (V_{y_1}^\top(t_{ni}) B_{y_1}^{\mathrm{w}}(t_{ni}) P_{y_1}(t_{ni}))^{-1} (V_{y_1}^\top(t_{ni}) W^\top(t_{ni}) \delta_{ni} - V_{y_1}^\top(t_{ni}) B_{x_1}^{\mathrm{w}}(t_{ni}) P_{x_1}(t_{ni}) \delta_{ni}^{x_1}). \end{aligned}$$

Next we multiply the i-th equation of (6.44a) by $V^\top(t_{ni})$ and obtain

$$V^\top(t_{ni}) A(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{e_{nj} - e_{n-1}}{h} + V^\top(t_{ni}) B(t_{ni}) e_{ni} = V^\top(t_{ni})(A(t_{ni}) L_{ni}(x) + \delta_{ni})$$

for all $1 \leqslant i \leqslant s$. Inserting the splitting of the errors then yields

$$\begin{aligned} &V^\top(t_{ni}) A(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{P(t_{nj}) e_{nj}^{x_0} + Q(t_{nj}) e_{nj}^{y_0} - P(t_{n-1}) e_{n-1}^{x_0}}{h} + V^\top(t_{ni}) B(t_{ni}) e_{ni} \\ &= V^\top(t_{ni})(A(t_{ni}) L_{ni}(x) + \delta_{ni}). \end{aligned}$$

with the help of (6.29) for all $1 \leqslant i \leqslant s$. By the Taylor expansions (6.38) in t_{ni} for $P(t_{nj})$ and $Q(t_{nj})$ and (6.29) we obtain

$$\begin{aligned} &V^\top(t_{ni})(A(t_{ni}) L_{ni}(x) + \delta_{ni}) \\ &= G_1(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{e_{nj}^{x_0} - e_{n-1}^{x_0}}{h} \\ &\quad + \sum_{\substack{j=1 \\ j \neq i}}^{s} \alpha_{ij}(c_j - c_i) V^\top(t_{ni}) A(t_{ni}) P'(t_{ni}) e_{nj}^{x_0} + V^\top(t_{ni}) B(t_{ni}) P(t_{ni}) e_{ni}^{x_0} \\ &\quad + \sum_{\substack{j=1 \\ j \neq i}}^{s} \alpha_{ij}(c_j - c_i) V^\top(t_{ni}) A(t_{ni}) Q'(t_{ni}) e_{nj}^{y_0} + V^\top(t_{ni}) B(t_{ni}) Q(t_{ni}) e_{ni}^{y_0} \\ &\quad + h \sum_{\substack{j=1 \\ j \neq i}}^{s} M_P(t_{ni}, t_{nj}) e_{nj}^{x_0} + h \sum_{\substack{j=1 \\ j \neq i}}^{s} M_Q(t_{ni}, t_{nj}) e_{nj}^{y_0}. \end{aligned}$$

By (6.47) the error splitting can be described by

$$e_{n-1}^{x_0} = Q_{x_1}(t_{n-1}) e_{n-1}^{x_2} + P_{x_1}(t_{n-1}) \delta_{n-1}^{x_1}$$

$$
\begin{aligned}
e_{ni}^{x_0} &= Q_{x_1}(t_{ni})e_{ni}^{x_2} + P_{x_1}(t_{ni})\delta_{ni}^{x_1}\\
e_{ni}^{y_0} &= Q_{y_1}(t_{ni})e_{ni}^{\tilde{y}_2} + P_{y_1}(t_{ni})\delta_{ni}^{y_1}\\
&\quad - P_{y_1}(t_{ni})(V_{y_1}^\top(t_{ni})B_{y_1}^{\mathrm{w}}(t_{ni})P_{y_1}(t_{ni}))^{-1}V_{y_1}^\top(t_{ni})B_{x_1}^{\mathrm{w}}(t_{ni})Q_{x_1}(t_{ni})e_{ni}^{x_2}.
\end{aligned}
$$

Here the global perturbation, as well as the stage perturbations, are split analogously to the global error (6.45) and the stage error (6.46). This leads, combined with Taylor expansions (6.40) in t_{ni} for $Q_{x_1}(t_{nj})$ and $Q_{y_1}(t_{nj})$, to

$$
\begin{aligned}
0 = G_1(t_{ni})Q_{x_1}(t_{ni})\sum_{j=1}^{s}\alpha_{ij}\frac{e_{nj}^{x_2}-e_{n-1}^{x_2}}{h} + M_{x_2,n}^{i}e_n^{X_2} - V^\top(t_{ni})A(t_{ni})L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h}M_{\delta,2}\delta_n\\
+\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)V^\top(t_{ni})A(t_{ni})Q'(t_{ni})Q_{y_1}(t_{ni})e_{nj}^{\tilde{y}_2} + V^\top(t_{ni})B(t_{ni})Q(t_{ni})Q_{y_1}(t_{ni})e_{ni}^{\tilde{y}_2}\\
+h\sum_{\substack{j=1\\ j\neq i}}^{s}\left(M_Q(t_{ni},t_{nj})Q_{y_1}(t_{nj}) + \alpha_{ij}(c_j-c_i)V^\top(t_{ni})A(t_{ni})Q'(t_{ni})R_{Q_{y_1}}(t_{ni},t_{nj})(c_j-c_i)\right)e_{nj}^{\tilde{y}_2}
\end{aligned}
$$

with $e_n^{X_2} := \begin{pmatrix}\dots & e_{ni}^{x_2} & \dots\end{pmatrix}^\top$ and certain matrices $M_{x_2,n}^{i}$, $M_{\delta,1}$ and $M_{\delta,2}$. We multiply each of these equations by $W_{x_1}^\top(t_{ni})$ respectively and obtain

$$
\begin{aligned}
0 &= W_{x_1}^\top(t_{ni})(M_{x_2,n}^{i}e_n^{X_2} + V^\top(t_{ni})A(t_{ni})L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h}M_{\delta,2}\delta_n)\\
&+\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)W_{x_1}^\top(t_{ni})V^\top(t_{ni})A(t_{ni})Q'(t_{ni})Q_{y_1}(t_{ni})e_{nj}^{\tilde{y}_2}\\
&+W_{x_1}^\top(t_{ni})V^\top(t_{ni})B(t_{ni})Q(t_{ni})Q_{y_1}(t_{ni})e_{ni}^{\tilde{y}_2}\\
&+h\sum_{\substack{j=1\\ j\neq i}}^{s}\left(M_Q(t_{ni},t_{nj})Q_{y_1}(t_{nj}) + \alpha_{ij}(c_j-c_i)V^\top(t_{ni})A(t_{ni})Q'(t_{ni})R_{Q_{y_1}}(t_{ni},t_{nj})(c_j-c_i)\right)e_{nj}^{\tilde{y}_2}
\end{aligned}
$$

which will describe the error components $e_{ni}^{\tilde{y}_2}$. With the help of the Taylor expansions (6.42) we get

$$
\begin{aligned}
0 =& W_{x_1}^\top(t_{ni})(M_{x_2,n}^{i}e_n^{X_2} + V^\top(t_{ni})A(t_{ni})L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h}M_{\delta,2}\delta_n)\\
&+\sum_{\substack{j=1\\ j\neq i}}^{s}\alpha_{ij}(c_j-c_i)(W_{x_1}^\top V^\top AQ'Q_{y_1})(t_{n-1})e_{nj}^{\tilde{y}_2} + (W_{x_1}^\top V^\top BQQ_{y_1})(t_{n-1})e_{ni}^{\tilde{y}_2}\\
&+h\sum_{\substack{j=1\\ j\neq i}}^{s}M_{Q_{y_1}}(t_{ni},t_{nj},t_{n-1})e_{nj}^{\tilde{y}_2}.
\end{aligned}
$$

This equation can be written as

$$\begin{aligned}&(M \otimes (W_{x_1}^\top V^\top A Q' Q_{y_1})(t_{n-1}) + I \otimes (W_{x_1}^\top V^\top B Q Q_{y_1})(t_{n-1}) + h\bar{M}_{Q_{y_1}}) e_n^{\tilde{Y}_2} \\ =&W_{x_1}^\top(t_{ni})(M^i_{x_2,n} e_n^{X_2} + V^\top(t_{ni}) A(t_{ni}) L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n)\end{aligned}$$

with the help of the matrix tensor product $\otimes$, $e_n^{\tilde{Y}_2} := \begin{pmatrix} \dots & e_{ni}^{\tilde{y}_2} & \dots \end{pmatrix}^\top$ and the matrix M as in (6.32). By the Banach Perturbation Lemma we get

$$\begin{aligned}e_n^{\tilde{Y}_2} =&(M \otimes (W_{x_1}^\top V^\top A Q' Q_{y_1})(t_{n-1}) + I \otimes (W_{x_1}^\top V^\top B Q Q_{y_1})(t_{n-1}) + h\bar{M}_{Q_{y_1}})^{-1} \\ &W_{x_1}^\top(t_{ni})(M^i_{x_2,n} e_n^{X_2} + V^\top(t_{ni}) A(t_{ni}) L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n).\end{aligned} \tag{6.48}$$

To deal with the error terms $e_{ni}^{x_2}$ we again consider the system

$$\begin{aligned}0 =&G_1(t_{ni}) Q_{x_1}(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{e_{nj}^{x_2} - e_{n-1}^{x_2}}{h} + M^i_{x_2,n} e_n^{X_2} + V^\top(t_{ni}) A(t_{ni}) L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n \\ &+ \sum_{\substack{j=1 \\ j\neq i}}^{s} \alpha_{ij}(c_j - c_i) V^\top(t_{ni}) A(t_{ni}) Q'(t_{ni}) Q_{y_1}(t_{ni}) e_{nj}^{\tilde{y}_2} + V^\top(t_{ni}) B(t_{ni}) Q(t_{ni}) Q_{y_1}(t_{ni}) e_{ni}^{\tilde{y}_2} \\ &+ h \sum_{\substack{j=1 \\ j\neq i}}^{s} \left(M_Q(t_{ni}, t_{nj}) Q_{y_1}(t_{nj}) + \alpha_{ij}(c_j - c_i) V^\top(t_{ni}) A(t_{ni}) Q'(t_{ni}) R_{Q_{y_1}}(t_{ni}, t_{nj})(c_j - c_i) \right) e_{nj}^{\tilde{y}_2}\end{aligned}$$

which can be written as

$$\begin{aligned}0 =&G_1(t_{ni}) Q_{x_1}(t_{ni}) \sum_{j=1}^{s} \alpha_{ij} \frac{e_{nj}^{x_2} - e_{n-1}^{x_2}}{h} + M^i_{x_2,n} e_n^{X_2} \\ &+ V^\top(t_{ni}) A(t_{ni}) L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n\end{aligned}$$

with the help of Equation (6.48). We multiply each of these equations by $G_2^{-1}(t_{ni}) V_{x_1}^\top(t_{ni})$ and achieve

$$\begin{aligned}0 =& \sum_{j=1}^{s} \alpha_{ij} \frac{e_{nj}^{x_2} - e_{n-1}^{x_2}}{h} \\ &+ G_2^{-1}(t_{ni}) V_{x_1}^\top(t_{ni}) (M^i_{x_2,n} e_n^{X_2} + V^\top(t_{ni}) A(t_{ni}) L_{ni}(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n)\end{aligned}$$

This can be written as

$$0 =(\mathcal{A}^{-1} \otimes I) \frac{e_n^{X_2} - E_{n-1}^{x_2}}{h} + M_{x_2,n} e_n^{X_2} + L_n(x) + M_{\delta,1}\delta_n + \frac{1}{h} M_{\delta,2}\delta_n$$

which finally leads to

$$e_n^{X_2} =(I + hM_{x_2,n})^{-1}(E_{n-1}^{x_2} - h(L_n(x) + M_{\delta,1}\delta_n + \frac{1}{h}M_{\delta,2}\delta_n))$$

with $E_{n-1}^{x_2} := \begin{pmatrix} e_{n-1}^{x_2} & \dots & e_{n-1}^{x_2} \end{pmatrix}^\top$. Now we obtain the estimation

$$\begin{aligned}\|e_n^{x_2}\| \leqslant \left\|e_n^{X_2}\right\| &\leqslant \left\|(I + hM_{x_2,n})^{-1}\right\| (\left\|E_{n-1}^{x_2}\right\| + h\|L_n(x) + M_{\delta,1}\delta_n + \frac{1}{h}M_{\delta,2}\delta_n\|) \\ &\leqslant \left\|(I + hM_{x_2,n})^{-1}\right\| (\left\|e_{n-1}^{x_2}\right\| + \mathcal{O}(h^s)) \\ &\leqslant (1 + 2\left\|M_{x_2,n}\right\| h)(\left\|e_{n-1}^{x_2}\right\| + \mathcal{O}(h^s))\end{aligned}$$

which leads by standard ODE estimations to the existence of a constant $C_{x_2} > 0$ such that

$$\|e_n^{x_2}\| \leqslant C_{x_2}(\|e_0^{x_2}\| + \mathcal{O}(h^{s-1})).$$

With the help of (6.48) and (6.47) this yields a constant $C > 0$ such that

$$\|e_n\| \leqslant C(\|e_0\| + \mathcal{O}(h^{s-1}))$$

and the proof is concluded. □

The most important feature of Theorem 6.15 is that it does not need any restriction regarding the basis functions, except differentiability and that their ranks are constant. According to this the simulation of the Examples 6.2 and 2.12 by an left-discontinuous collocation method should converge against the exact solution without any additional step size restriction. As an example of the class of left-discontinuous collocation method we choose the Lobatto IIIC methods. While the implicit Euler cannot provide satisfying results for Example 6.2 by using a step size $h = 10^{-1}$ due to the new step size restriction, see Figure 6.6, the Lobatto IIIC method with two stages manages to do so.

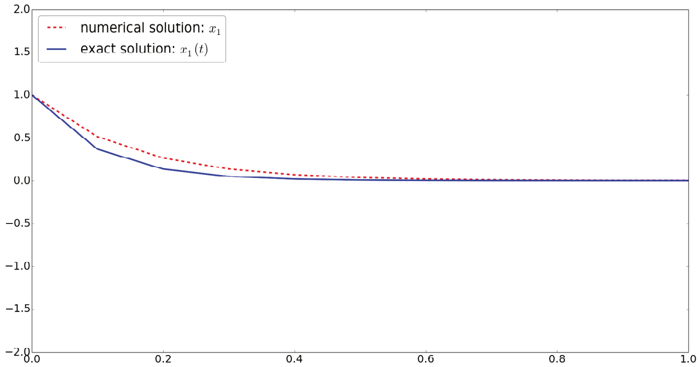


Figure 6.8: Numerical and exact solution of Example 6.2 simulated with the Lobatto IIIC method with two stages and the time step size $h = 10^{-1}$.

Also the manipulation of an algebraic variable by artificial dynamics is no problem for the Lobatto IIIC methods, in contrast to the results we saw for the implicit Euler in Figure 2.4.

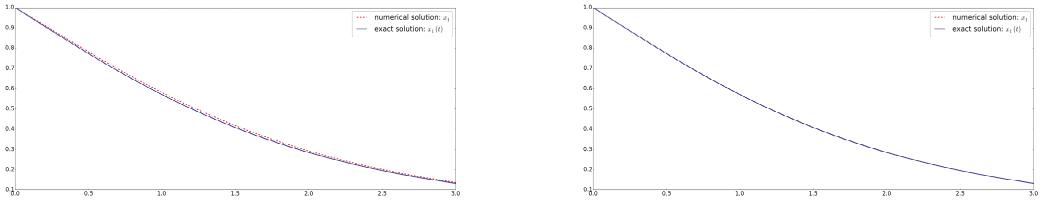


Figure 6.9: Numerical and exact solution of Example 2.12 simulated with the Lobatto IIIC method with two stages and the time step sizes $h = 10^{-1}$(left) and $h = 10^{-2}$(right).

6.3 Summary and Outlook

In this chapter we saw that widely used methods like the BDF-methods, as well as the RAUDAU IIA, methods may not converge against the exact solution of a DAE if the basis functions $Q_{x_0}(X^1, t)$ and $Q_{x_i}(X^i, t)$ are not constant.

We showed that the BDF methods, with at least two steps, converge for a general non-linear index 2 DAE if $Q_{x_0}(X^1, t)$ and $Q_{x_1}(X^1, t)$ are constant. Thereby we isolated the source for the convergence problems to these basic functions.

At last we presented a class of collocation methods and proved their convergence for the time dependent linear case even if $Q_{x_0}(t)$ and $Q_{x_1}(t)$ are not constant. But we still needed a regularity assumption.

Hence, the left-discontinuous collocation methods overcome the problems regarding the artificial dynamics but not those regarding the regularity. We close this chapter with the following question: Is there a method which is convergent and applicable to every sufficiently smooth DAE with Dissection Index 2 and constant characteristic values?

7 Half-explicit Methods

In Section 4.3 we showed that the Dissection Index concept provides a constant basis chain for the extended MNA without controlled sources (4.24). Even with controlled sources we may be able to obtain a constant basis chain as we saw in Example 4.29. Perturbation estimations and global solvability results are obtained in the Sections 4.4 and 5.4 by using the constant basis chain of the MNA. In this chapter we use the constant basis chain to decouple the MNA, including controlled sources into a semi-explicit DAE to accelerate its simulation, by using half-explicit methods. Half-explicit methods can be found in literature for various kinds of DAE systems. In [ASW93] half-explicit methods for semi-explicit index 1 DAEs are analyzed. A more general class of index 1 DAEs is studied in [LM14]. In [LM14] only the splitting of the equation is explicitly given such that the DAE can be written as

$$\begin{aligned} f(x', x, t) &= 0 \\ g(x, t) &= 0. \end{aligned}$$

In [BH93, Arn98, Mur97] half-explicit methods for index 2 DAEs in Hessenberg-form

$$\begin{aligned} x' &= f(x, y, t) \\ 0 &= g(x, t) \end{aligned}$$

are presented and studied and [Ost93] even considers Hessenberg systems up to index 3. The structure

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

of the MNA prevents us from using half-explicit methods. Therefore we use the constant basis chain to decouple the MNA into a semi-explicit DAE and afterwards we present half-explicit methods arising from a mix of the BDF-methods and the Adams-Bashford methods. The index of the resulting semi-explicit DAEs can exceed two since we consider circuits including controlled sources. For the convergence proof of our half-explicit methods, we restrict ourselves to DAEs with a Dissection Index of three or lower.

7.1 Topological Decoupling

In this section we will transform the extended MNA (3.34) into a DAE with the structure

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

$$0 = g(x, y, t)$$

with $M(x)$ being sparse, positive definite and symmetric. Therefore we exploit the constant basis functions. To do so, we need to provide a cheap way to calculate these basis functions. The basis functions may be constant but for large systems their calculation could get troublesome nevertheless. Considering that it can be difficult to calculate the kernel of a very large matrix. To avoid the calculation of the basis functions we present a way to describe them directly by the topology of the electric circuit.

7.1.1 Topological basis functions

For the network topological description of the basis functions we consider two arbitrary element classes and notate the sets of elements in these classes by X and Y. Let the mass node be connected to at least one element in X and let A_X be the associated incidence matrix. The element set X may decompose into n_X connected components $C_{X,i}$ with $1 \leqslant i \leqslant n_X$. We renumber the nodes with respect to the X-connected components and number the mass node at the end to achieve a more elegant notation. Hence C_{X,n_X} includes the mass node. Then the basis functions of the kernel and the complementary kernel of $A_X^\top$ can be chosen as

$$Q_X = \begin{pmatrix} \mathbb{1}_{|C_{X,1}|} & & \\ & \ddots & \\ & & \mathbb{1}_{|C_{X,n_X-1}|} \\ & & 0 \end{pmatrix} \text{ and } P_X = \begin{pmatrix} I_{|C_{X,1}|-1} & & & \\ 0^\top_{|C_{X,1}|-1} & & & \\ & \ddots & & \\ & & I_{|C_{X,n_X-1}|-1} & \\ & & 0^\top_{|C_{X,n_X-1}|-1} & \\ & & & I_{|C_{X,n_X}|-1} \end{pmatrix}$$

with $\mathbb{1}_N := \begin{pmatrix} 1 & \dots & 1 \end{pmatrix}^\top \in \mathbb{R}^N$. Now $A_{\bar{X}Y} = A_Y^\top Q_X$ is the incidence matrix of the graph shrunk by the edges of the elements in X. Notice that there may be elements in Y which are now linked to only one node in the shrunken graph. Next we can choose the matrices Q_Y and P_Y with respect to the shrunken graph analogously to Q_X and P_X. This process can be continued successively.

For the basis functions of the transposed kernel and the transposed complementary kernel of $A_X^\top$ we need to provide the definition a spanning tree:

Definition 7.1. (Spanning tree)
A tree of a graph G is a connected undirected graph with no loop. It is a spanning tree of a graph G if it includes every node of G. A spanning tree of a connected graph G can also be defined as a maximal set of edges of G that contains no loops, or as a minimal set of edges that connect all nodes.

and the definition a fundamental loop:

Definition 7.2. (Fundamental loop)
We consider a graph $G = (N, E)$ and a spanning tree $T = (N_T, E_T)$ of the graph G. Adding one edge of $E \backslash E_T$ to the spanning tree will create a loop. Such a loop is called a fundamental loop.

There is a distinct fundamental loop for each edge in $E \backslash E_T$. Thus, there is a one-to-one correspondence between fundamental loops and edges not in the spanning tree. For a connected graph with $|N|$ nodes, any spanning tree will have $|N| - 1$ edges, and thus, a graph of $|E|$ edges and one of its spanning trees will have $|E| - |N| + 1$ fundamental loops. We define a set of m_X X-fundamental loops $L_{X,i}$ with $1 \leqslant i \leqslant m_X$ to describe the transposed kernel and the transposed complementary kernel of $A_X^\top$. For every fundamental loop we choose an arbitrary loop direction. Furthermore, we choose a spanning tree T and renumber the edges such that the first edges do not belong to the spanning tree. With the help of these preparations we can describe each of the fundamental loops by a vector

$$(\mathbb{L}_{X,i})_j = \begin{cases} 1, & \text{if the } j\text{-th edge of the } i\text{-th loop has the } i\text{-th loop's direction} \\ -1, & \text{if the } j\text{-th edge of the } i\text{-th loop has not the } i\text{-th loop's direction} \\ 0, & \text{else.} \end{cases}$$

Then the basis functions of the transposed kernel and the transposed complementary kernel of $A_X^\top$ can be chosen as

$$W_X = \begin{pmatrix} \mathbb{L}_{X,1} & \dots & \mathbb{L}_{X,m_X} \end{pmatrix} \quad \text{and} \quad V_X = \begin{pmatrix} 0_{m_X} \\ I_{|X|-m_X} \end{pmatrix}.$$

This strategy works for an arbitrary graph. Therefore it also works for a shrunken graph. In the next subsection we make use of these topological basis functions to transform the extended MNA, including controlled sources.

7.1.2 Extended MNA with controlled sources

In the previous chapters we always assume the sources in an electric circuit to be independent, with a few exceptions. This is a reasonable assumption as long as all semiconductor and electromagnetic devices are modeled with the help of a PDE. If we approximate some of these devices by an equivalent circuit, there usually appear many controlled sources in the electric circuit. Hence it becomes interesting to efficiently simulate electric circuits including controlled sources, as well as distributed devices. The extended MNA equations including controlled sources are given by

$$
\begin{aligned}
A_{\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,t)+g_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,\zeta,\Psi)\right)+A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^{\top}e,q_M,t)+A_{\mathcal{L}}j_{\mathcal{L}}&\\
+A_Vj_V+A_Ii_s(A^{\top}e,j_{\mathcal{L}},j_V,t)&=0, &&(7.1\text{a})\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\mathcal{L}}(j_{\mathcal{L}},t)-A_{\mathcal{L}}^{\top}e+\chi_{\mathcal{L}}E&=0, &&(7.1\text{b})\\
A_V^{\top}e-v_s(A^{\top}e,j_{\mathcal{L}},j_V,t)&=0, &&(7.1\text{c})\\
M_{\zeta}\frac{\mathrm{d}}{\mathrm{d}t}\zeta+h_{\zeta}(A_S^{\top}e,\zeta,\Psi)&=0, &&(7.1\text{d})\\
T\Psi-h_{\Psi}(\zeta)&=0, &&(7.1\text{e})\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_M(q_M,t)-A_M^Te&=0, &&(7.1\text{f})\\
M_{\varepsilon}\frac{\mathrm{d}}{\mathrm{d}t}E+M_{\sigma}E-J-\chi_{\mathcal{L}}^Tj_{\mathcal{L}}&=0, &&(7.1\text{g})\\
\frac{\mathrm{d}}{\mathrm{d}t}J+M_{CC}E&=0. &&(7.1\text{h})
\end{aligned}
$$

In order to transform these equations into the following form

$$
\begin{aligned}
M(x)x'&=f(x,y,t) &&(7.2\text{a})\\
0&=g(x,y,t) &&(7.2\text{b})
\end{aligned}
$$

we have to deal with the term $A_{\mathcal{C}}\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,t)$. Therefore let Q_{cs} and P_{cs} be the basis functions with respect to the kernel and the complementary kernel of $A_{\mathcal{C}}^{\top}$. These basis functions can be described in a topological way and are thereby suitable for a fast simulation. We split the node potentials with the help of Q_{cs} and P_{cs} and obtain

$$e=P_{cs}e_x+Q_{cs}e_y.$$

By inserting this splitting and factorizing (7.1a) by $Q_{cs}^{\top}$ and $P_{cs}^{\top}$ we get

$$
\begin{aligned}
P_{cs}^{\top}A_{\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}P_{cs}e_x,t)+g_{\mathcal{C}}(A_{\mathcal{C}}^{\top}P_{cs}e_x,\zeta,\Psi)\right)+P_{cs}^{\top}A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^{\top}(P_{cs}e_x+Q_{cs}e_y),q_M,t)+P_{cs}^{\top}A_{\mathcal{L}}j_{\mathcal{L}}&\\
+P_{cs}^{\top}A_Vj_V+P_{cs}^{\top}A_Ii_s(A^{\top}(P_{cs}e_x+Q_{cs}e_y),j_{\mathcal{L}},j_V,t)&=0,\\
Q_{cs}^{\top}A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^{\top}(P_{cs}e_x+Q_{cs}e_y),q_M,t)+Q_{cs}^{\top}A_{\mathcal{L}}j_{\mathcal{L}}+Q_{cs}^{\top}A_Vj_V+Q_{cs}^{\top}A_Ii_s(A^{\top}(P_{cs}e_x+Q_{cs}e_y),j_{\mathcal{L}},j_V,t)&=0.
\end{aligned}
$$

We use the chain rule to obtain

$$P_{cs}^{\top}A_{\mathcal{C}}\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}P_{cs}e_x,t)=P_{cs}^{\top}A_{\mathcal{C}}\frac{\partial}{\partial v}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}P_{cs}e_x,t)A_{\mathcal{C}}^{\top}P_{cs}\frac{\mathrm{d}}{\mathrm{d}t}e_x+P_{cs}^{\top}A_{\mathcal{C}}\frac{\partial}{\partial t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}P_{cs}e_x,t)$$

which leads to

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

$$0 = g(x, y, t)$$

with the variables $x := \begin{pmatrix} e_x & j_{\mathcal{L}} & \zeta & q_M & E & J \end{pmatrix}^\top$ and $y := \begin{pmatrix} e_y & j_V & \Psi \end{pmatrix}^\top$, the matrix

$$M(x) = \begin{pmatrix} P_{cs}^\top A_{\mathcal{C}} \frac{\partial}{\partial v} q_{\mathcal{C}}(A_{\mathcal{C}}^\top P_{cs} e_x, t) A_{\mathcal{C}}^\top P_{cs} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\partial}{\partial j} \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial}{\partial q} \phi_M(q_M, t) & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}$$

and the functions

$$f(x, y, t) := \begin{pmatrix} f_1(x, y, t) \\ A_{\mathcal{L}}^\top e - \chi_{\mathcal{L}} E - \frac{\partial}{\partial t} \phi_{\mathcal{L}}(j_{\mathcal{L}}, t) \\ -h_\zeta(A_S^\top e, \zeta, \Psi) \\ A_M^T e - \frac{\partial}{\partial t} \phi_M(q_M, t) \\ J + \chi_{\mathcal{L}}^T j_{\mathcal{L}} - M_\sigma E \\ -M_{CC} E \end{pmatrix} \text{ and } g(x, y, t) := \begin{pmatrix} g_1(x, y, t) \\ A_V^\top e - v_s(A^\top e, j_{\mathcal{L}}, j_V, t) \\ T\Psi - h_\varphi(\zeta) \end{pmatrix}$$

with

$$\begin{aligned} & f_1(x, y, t) \\ := & -(P_{cs}^\top A_{\mathcal{C}} \left(\frac{\partial}{\partial t} q_{\mathcal{C}}(A_{\mathcal{C}}^\top P_{cs} e_x, t) + g_{\mathcal{C}}(A_{\mathcal{C}}^\top P_{cs} e_x, \zeta, \Psi) \right) + P_{cs}^\top A_{\mathcal{L}} j_{\mathcal{L}} + P_{cs}^\top A_V j_V \\ & + P_{cs}^\top A_{\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top (P_{cs} e_x + Q_{cs} e_y), q_M, t) + P_{cs}^\top A_I i_s(A^\top (P_{cs} e_x + Q_{cs} e_y), j_{\mathcal{L}}, j_V, t)) \end{aligned}$$

and

$$\begin{aligned} g_1(x, y, t) := & \, Q_{cs}^\top A_{\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top (P_{cs} e_x + Q_{cs} e_y), q_M, t) + Q_{cs}^\top A_{\mathcal{L}} j_{\mathcal{L}} + Q_{cs}^\top A_V j_V \\ & + Q_{cs}^\top A_I i_s(A^\top (P_{cs} e_x + Q_{cs} e_y), j_{\mathcal{L}}, j_V, t) \end{aligned}$$

with $M(x)$ being positive definite and symmetric since

$$\frac{\partial}{\partial v} q_{\mathcal{C}}(v, t), \frac{\partial}{\partial j} \phi_{\mathcal{L}}(j, t), M_\zeta, \frac{\partial}{\partial q} \phi_M(q, t) \text{ and } M_\varepsilon$$

are positive definite and symmetric. We can transform (7.2) into a semi-explicit DAE since $M(x)$ is non-singular:

$$\begin{aligned} x' &= M(x)^{-1} f(x, y, t) \\ 0 &= g(x, y, t). \end{aligned}$$

Naturally in practice we do not actually calculate the inverse. In our particular case the matrix $M(x)$ is positive definite and symmetric and therefore we can use iterative methods like the CG-method.

7.1.3 Extended MNA without controlled sources

When we exclude the controlled sources we deal with the following DAE:

$$A_{\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,t)+g_{\mathcal{C}}(A_{\mathcal{C}}^{\top}e,\zeta,\Psi)\right)+A_{\mathcal{R}}g_{\mathcal{R}}(A_{\mathcal{R}}^{\top}e,q_M,t)+A_{\mathcal{L}}j_{\mathcal{L}}$$

$$+A_V j_V+A_I i_s(t)=0, \tag{7.3a}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\mathcal{L}}(j_{\mathcal{L}},t)-A_{\mathcal{L}}^{\top}e+\chi_{\mathcal{L}}E=0, \tag{7.3b}$$

$$A_V^{\top}e-v_s(t)=0, \tag{7.3c}$$

$$M_{\zeta}\frac{\mathrm{d}}{\mathrm{d}t}\zeta+h_{\zeta}(A_S^{\top}e,\zeta,\Psi)=0,$$

$$T\Psi-h_{\Psi}(\zeta)=0,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_M(q_M,t)-A_M^Te=0,$$

$$M_{\varepsilon}\frac{\mathrm{d}}{\mathrm{d}t}E+M_{\sigma}E-J-\chi_{\mathcal{L}}^Tj_{\mathcal{L}}=0,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}J+M_{CC}E=0.$$

In this case, it is possible to extract an index 1 DAE, in semi-explicit form, from the equations. Therefore we define a sequence of topological motivated basis functions. Let Q_V and P_V be the basis function associated to the kernel and the complementary kernel of $A_V^{\top}$. Then we call

$$A_{\bar{V}X} \;:=\; Q_V^{\top}A_X, \quad X\in\{\mathcal{C},\mathcal{R},\mathcal{L},I\}$$

the V-reduced incidence matrix of the capacitor-like elements, resistor-like elements, inductor-like elements or current sources, respectively. Further let $Q_{\mathcal{C}}$ and $P_{\mathcal{C}}$ be the basis function associated to the kernel and the complementary kernel of $A_{\bar{V}\mathcal{C}}^{\top}$. Analogously we call

$$A_{\bar{V}\bar{\mathcal{C}}X} \;:=\; Q_{\mathcal{C}}^{\top}Q_V^{\top}A_X, \quad X\in\{\mathcal{R},\mathcal{L},I\}$$

the V$\mathcal{C}$-reduced incidence matrix of the resistor-like elements, inductor-like elements or current sources, respectively. At last we obtain the basis function $Q_{\mathcal{R}}$ and $P_{\mathcal{R}}$ associated to the kernel and the complementary kernel of $A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^{\top}$ and denote by

$$A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}X} \;:=\; Q_{\mathcal{R}}^{\top}Q_{\mathcal{C}}^{\top}Q_V^{\top}A_X, \quad X\in\{\mathcal{L},I\}$$

the V$\mathcal{CR}$-reduced incidence matrix of the inductor-like elements or current sources, respectively.

We consider an arbitrary electric circuit. We remove all voltage sources and identify all nodes which were connected by voltage sources. We call this new circuit the V-reduced circuit. The V-reduced incidence matrices, defined above, are the incidence matrices of the V-reduced circuit if we choose the basis function according to Section 7.1.1. Analogously we can interpret the V$\mathcal{C}$-reduced and the V$\mathcal{CR}$-reduced incidence matrices.
Successively we split the potential variable e into

$$\begin{aligned} e &= P_V e_V + Q_V(P_{\mathcal{C}} e_{\mathcal{C}} + Q_{\mathcal{C}}(P_{\mathcal{R}} e_{\mathcal{R}} + Q_{\mathcal{R}} e_{\mathcal{L}})) \\ &= P_V e_V + Q_V P_{\mathcal{C}} e_{\mathcal{C}} + Q_V Q_{\mathcal{C}} P_{\mathcal{R}} e_{\mathcal{R}} + Q_V Q_{\mathcal{C}} Q_{\mathcal{R}} e_{\mathcal{L}} \end{aligned}$$

with the help of the basis splitting approach. The equations of (7.3) will also be split successively in order (7.3c),(7.3a) and (7.3b).
Equation (7.3c) provides

$$A_V^T e = v_s(t) \quad \Rightarrow \quad A_V^T P_V e_V = v_s(t) \quad \Rightarrow \quad e_V = (A_V^T P_V)^{-1} v_s(t) =: v_s^*(t)$$

and therefore e_V can be written as a known time depending function. Next we split equation (7.3a) by multiplying $P_V^\top$, $P_{\mathcal{C}}^\top Q_V^\top$, $P_{\mathcal{R}}^\top Q_{\mathcal{C}}^\top Q_V^\top$ and $Q_{\mathcal{R}}^\top Q_{\mathcal{C}}^\top Q_V^\top$ from the left and obtain an explicit description of the currents through the voltage sources

$$j_V = -(P_V^\top A_V)^{-1} P_V^\top (A_{\mathcal{C}} \left(\frac{\mathrm{d}}{\mathrm{d}t} q_{\mathcal{C}}(A_{\mathcal{C}}^\top e, t) + g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) \right) + A_{\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t) + A_{\mathcal{L}} j_{\mathcal{L}} + A_I i_s(t))$$

and a system which does not depend on these currents

$$P_{\mathcal{C}}^\top (A_{\bar{V}\mathcal{C}} \left(\frac{\mathrm{d}}{\mathrm{d}t} q_{\mathcal{C}}(A_{\mathcal{C}}^\top e, t) + g_{\mathcal{C}}(A_{\mathcal{C}}^\top e, \zeta, \Psi) \right) + A_{\bar{V}\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t) + A_{\bar{V}\mathcal{L}} j_{\mathcal{L}} + A_{\bar{V}I} i_s(t)) = 0 \tag{7.4a}$$

$$P_{\mathcal{R}}^\top (A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}} g_{\mathcal{R}}(A_{\mathcal{R}}^\top e, q_M, t) + A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}} j_{\mathcal{L}} + A_{\bar{V}\bar{\mathcal{C}}I} i_s(t)) = 0 \tag{7.4b}$$

$$A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}} j_{\mathcal{L}} + A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}I} i_s(t) = 0. \tag{7.4c}$$

Let $Q_{\mathcal{L}I}$ and $P_{\mathcal{L}I}$ be the associated basis functions of the kernel and the complementary kernel of $A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}}$. Then we split the currents along the inductors into

$$j_{\mathcal{L}} = P_{\mathcal{L}I} j_{\mathcal{L}I} + Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}}.$$

Equation (7.4c) then provides a explicit formula for $j_{\mathcal{L}I}$ by

$$\begin{aligned} & A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}} j_{\mathcal{L}} = -A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}I} i_s(t) \\ \Rightarrow\ & A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}} P_{\mathcal{L}I} j_{\mathcal{L}I} = -A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}I} i_s(t) \\ \Rightarrow\ & j_{\mathcal{L}I} = -(A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}} P_{\mathcal{L}I})^{-1} A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}I} i_s(t) =: i_s^*(t). \end{aligned}$$

With the help of $v_s^*(t)$ and $i_s^*(t)$ we define the functions

$$q_{\bar{V}\mathcal{C}}(x, t) := q_{\mathcal{C}}(x + A_{\mathcal{R}}^\top P_V v_s^*(t), t)$$

$$
\begin{aligned}
g_{\bar{V}\mathcal{C}}(e_{\mathcal{C}},\zeta,\Psi) &:= g_{\mathcal{C}}(A_{\bar{V}\mathcal{R}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}} + A_{\mathcal{R}}^{\top}P_V v_s^*(t),\zeta,\Psi)\\
\phi_{\bar{I}\mathcal{L}}(x,t) &:= \phi_{\mathcal{L}}(x + P_{\mathcal{L}I}i_s^*(t),t)\\
g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(x,e_{\mathcal{C}},q_M,t) &:= g_{\mathcal{R}}(x + A_{\bar{V}\mathcal{R}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}} + A_{\mathcal{R}}^{\top}P_V v_s^*(t),q_M,t).
\end{aligned}
$$

Then we insert the variable splitting of the potentials and the current of the inductors into (7.4a), (7.4b) and (7.3b) to obtain

$$
\begin{aligned}
P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\bar{V}\mathcal{C}}(A_{\bar{V}\mathcal{C}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}},t) + g_{\bar{V}\mathcal{C}}(e_{\mathcal{C}},\zeta,\Psi)\right) + P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{R}}g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}},e_{\mathcal{C}},q_M,t)&\\
+P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{L}}Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}} + i_C(t) &= 0\\
\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}},t) - A_{\bar{V}\mathcal{L}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}} - A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}} - A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}}^{\top}e_{\mathcal{L}} - A_L^{\top}P_V v_s^*(t) + \chi_{\mathcal{L}}E &= 0\\
P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}},e_{\mathcal{C}},q_M,t) + P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}} + i_R(t) &= 0
\end{aligned}
$$

with

$$
\begin{aligned}
i_C(t) &:= P_{\mathcal{C}}^{\top}A_{\bar{V}I}i_s(t) + P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{L}}P_{\mathcal{L}I}i_s^*(t)\\
i_R(t) &:= P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}I}i_s(t) + P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}P_{\mathcal{L}I}i_s^*(t).
\end{aligned}
$$

Next we split (7.3b) by multiplying $P_{\mathcal{L}I}^{\top}, Q_{\mathcal{L}I}^{\top}$ from the left and obtain a reduced system in $e_{\mathcal{C}}, e_{\mathcal{R}}, j_{\mathcal{L}\bar{I}}$:

$$
\begin{aligned}
P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{C}}\left(\frac{\mathrm{d}}{\mathrm{d}t}q_{\bar{V}\mathcal{C}}(A_{\bar{V}\mathcal{C}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}},t) + g_{\bar{V}\mathcal{C}}(e_{\mathcal{C}},\zeta,\Psi)\right) + P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{R}}g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}},e_{\mathcal{C}},q_M,t)&\\
+P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{L}}Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}} + i_C(t) &= 0\\
Q_{\mathcal{L}I}^{\top}\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}},t) - Q_{\mathcal{L}I}^{\top}A_{\bar{V}\mathcal{L}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}} - Q_{\mathcal{L}I}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}} + Q_{\mathcal{L}I}^{\top}\chi_{\mathcal{L}}E + v_L(t) &= 0\\
P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}},e_{\mathcal{C}},q_M,t) + P_{\mathcal{R}}^{\top}A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}} + i_R(t) &= 0
\end{aligned}
$$

with $v_L(t) := -Q_{\mathcal{L}I}^{\top}A_L^{\top}P_V v_s^*(t)$ and an explicit presentation for

$$
e_{\mathcal{L}} = (P_{\mathcal{L}I}^{\top}A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}}^{\top})^{-1}P_{\mathcal{L}I}^{\top}(\frac{\mathrm{d}}{\mathrm{d}t}\phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}},t) - A_{\bar{V}\mathcal{L}}^{\top}\ P_{\mathcal{C}}e_{\mathcal{C}} - A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}^{\top}P_{\mathcal{R}}e_{\mathcal{R}} - A_L^{\top}P_V v_s^*(t) + \chi_{\mathcal{L}}E).
$$

We define the matrices

$$
\begin{aligned}
M_{\mathcal{C}}(e_{\mathcal{C}},t) &:= P_{\mathcal{C}}^{\top}A_{\bar{V}\mathcal{C}}\frac{\mathrm{d}}{\mathrm{d}x}q_{\bar{V}\mathcal{C}}(A_{\bar{V}\mathcal{C}}^{\top}P_{\mathcal{C}}e_{\mathcal{C}},t)A_{\bar{V}\mathcal{C}}^{\top}P_{\mathcal{C}}\\
M_{\mathcal{L}}(j_{\mathcal{L}\bar{I}},t) &:= Q_{\mathcal{L}I}^{\top}\frac{\mathrm{d}}{\mathrm{d}x}\phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I}j_{\mathcal{L}\bar{I}},t)Q_{\mathcal{L}I}
\end{aligned}
$$

which leads to

$$
M(x)x' = f(x,y,t) \tag{7.5a}
$$

$$0 = g(x, y, t) \tag{7.5b}$$
$$z = h(x', x, y, t) \tag{7.5c}$$

with the variables $x := \begin{pmatrix} e_{\mathcal{C}} & j_{\mathcal{L}\bar{I}} & \zeta & q_M & E & J \end{pmatrix}^\top$ and $y := \begin{pmatrix} e_{\mathcal{R}} & \Psi \end{pmatrix}^\top$, the matrix

$$M(x) = \begin{pmatrix} M_{\mathcal{C}}(e_{\mathcal{C}}, t) & 0 & 0 & 0 & 0 & 0 \\ 0 & M_{\mathcal{L}}(j_{\mathcal{L}\bar{I}}, t) & 0 & 0 & 0 & 0 \\ 0 & 0 & M_\zeta & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\partial}{\partial q}\phi_M(q_M, t) & 0 & 0 \\ 0 & 0 & 0 & 0 & M_\varepsilon & 0 \\ 0 & 0 & 0 & 0 & 0 & I \end{pmatrix}$$

and the functions

$$f(x, y, t) := \begin{pmatrix} -f_1(x, y, t) \\ -f_2(x, y, t) \\ -h_\zeta(A_S^\top(P_V v_s^*(t) + Q_V P_{\mathcal{C}} e_{\mathcal{C}}), \zeta, \Psi) \\ A_M^T(P_V v_s^*(t) + Q_V P_{\mathcal{C}} e_{\mathcal{C}} + Q_V Q_{\mathcal{C}} P_{\mathcal{R}} e_{\mathcal{R}}) - \frac{\partial}{\partial t}\phi_M(q_M, t) \\ J + \chi_{\mathcal{L}}^T(P_{\mathcal{L}I} i_s^*(t) + Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}}) - M_\sigma E \\ -M_{CC} E \end{pmatrix}$$

with

$$\begin{aligned} f_1(x, y, t) :=& P_{\mathcal{C}}^\top A_{\bar{V}\mathcal{C}} \left(\frac{\partial}{\partial t} q_{\bar{V}\mathcal{C}}(A_{\bar{V}\mathcal{C}}^\top P_{\mathcal{C}} e_{\mathcal{C}}, t) + g_{\bar{V}\mathcal{C}}(e_{\mathcal{C}}, \zeta, \Psi) \right) + P_{\mathcal{C}}^\top A_{\bar{V}\mathcal{L}} Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}} \\ & + P_{\mathcal{C}}^\top A_{\bar{V}\mathcal{R}} g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^\top P_{\mathcal{R}} e_{\mathcal{R}}, e_{\mathcal{C}}, q_M, t) + i_C(t) \\ f_2(x, y, t) :=& Q_{\mathcal{L}I}^\top \frac{\partial}{\partial t} \phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}}, t) - Q_{\mathcal{L}I}^\top A_{\bar{V}\mathcal{L}}^\top P_{\mathcal{C}} e_{\mathcal{C}} - Q_{\mathcal{L}I}^\top A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}^\top P_{\mathcal{R}} e_{\mathcal{R}} + Q_{\mathcal{L}I}^\top \chi_{\mathcal{L}} E + v_L(t) \end{aligned}$$

and

$$g(x, y, t) := \begin{pmatrix} P_{\mathcal{R}}^\top A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}} g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^\top P_{\mathcal{R}} e_{\mathcal{R}}, e_{\mathcal{C}}, q_M, t) + P_{\mathcal{R}}^\top A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}} Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}} + i_R(t) \\ T\Psi - h_\varphi(\zeta) \end{pmatrix}$$

with $M(x)$ and $\frac{\partial}{\partial y} g(x, y, t)$ being positive definite and symmetric. Furthermore we define $z := \begin{pmatrix} j_V & e_{\mathcal{L}} & j_{\mathcal{L}I} & e_V \end{pmatrix}^\top$ and

$$h(x', x, y, t) :=$$
$$\begin{pmatrix} -(P_V^\top A_V)^{-1} P_V^\top h_1(x, y, t) \\ (P_{\mathcal{L}I}^\top A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}}^\top)^{-1} P_{\mathcal{L}I}^\top (\frac{\mathrm{d}}{\mathrm{d}t} \phi_{\bar{I}\mathcal{L}}(Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}}, t) - A_{\bar{V}\mathcal{L}}^\top \; P_{\mathcal{C}} e_{\mathcal{C}} - A_{\bar{V}\bar{\mathcal{C}}\mathcal{L}}^\top P_{\mathcal{R}} e_{\mathcal{R}} - A_L^\top P_V v_s^*(t) + \chi_{\mathcal{L}} E) \\ -(A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}\mathcal{L}} P_{\mathcal{L}I})^{-1} A_{\bar{V}\bar{\mathcal{C}}\bar{\mathcal{R}}I} i_s(t) \\ (A_V^T P_V)^{-1} v_s(t) \end{pmatrix}$$

with

$$\begin{aligned} h_1(x,y,t) :=& (A_{\bar{V}\mathcal{C}} \left(\frac{\mathrm{d}}{\mathrm{d}t} q_{\bar{V}\mathcal{C}}(A_{\bar{V}\mathcal{C}}^\top P_{\mathcal{C}} e_{\mathcal{C}}, t) + g_{\bar{V}\mathcal{C}}(e_{\mathcal{C}}, \zeta, \Psi) \right) + A_{\bar{V}\mathcal{L}} Q_{\mathcal{L}I} j_{\mathcal{L}\bar{I}} \\ &+ A_{\bar{V}\mathcal{R}} g_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}(A_{\bar{V}\bar{\mathcal{C}}\mathcal{R}}^\top P_{\mathcal{R}} e_{\mathcal{R}}, e_{\mathcal{C}}, q_M, t) + A_{\bar{V}I} i_s(t) + A_{\bar{V}\mathcal{L}} P_{\mathcal{L}I} i_s^*(t). \end{aligned}$$

We drop the last equation of (7.5) and obtain

$$\begin{aligned} x' &= M(x)^{-1} f(x,y,t) \\ 0 &= g(x,y,t) \end{aligned}$$

by multiplying the inverse of $M(x)$ from the left. Again we can use iterative solvers instead of actually calculating the inverse of M.

7.2 Explicit Methods

In this section we introduce a new class of half-explicit methods and prove their convergence. Our class of half-explicit methods will be defined on semi-explicit DAEs. We repeat the definition of a semi-explicit DAE at this point.

Definition 7.3. (Semi-explicit DAE)
Let $\mathcal{I} \subset \mathbb{R}$, $\mathcal{D}_x \subset \mathbb{R}^{n_x}$ and $\mathcal{D}_y \subset \mathbb{R}^{n_y}$ be open subsets. Consider the following set of equations

$$x' = f(x,y,t) \tag{7.6a}$$

$$0 = g(x,y,t) \tag{7.6b}$$

with $f \in C(\mathcal{D}_x \times \mathcal{D}_y \times \mathcal{I}, \mathbb{R}^{n_x})$ and $g \in C(\mathcal{D}_x \times \mathcal{D}_y \times \mathcal{I}, \mathbb{R}^{n_y})$. Further, let the partial derivatives of f and g, with respect to x and y, be continuous. We call (7.6) a semi-explicit DAE.

We restrict ourselves to a subclass of semi-explicit DAEs by the following set of assumptions:

Assumption 7.4.
Consider a semi-explicit DAE (7.6). Let the Dissection Index be 3 at most. Furthermore we assume that there are constant basis functions Q_{x_i} for $i \leqslant 2$ and that it is possible to choose all the other basis functions, including the alternative basis function ending from Lemma 4.15, state independent.

The basis function Q_{x_0} is always constant for a semi-explicit DAE since the solution variable is already split into dynamic and algebraic components. The Assumption 7.4

excludes the mechanical applications but includes the circuits applications without controlled sources. In the case of circuits applications with controlled sources Assumption 7.4 may not hold. Before we formulate the half-explicit methods, we calculate the basis chain for a semi-explicit DAE since we can profit from its structure and Assumption 7.4. In particular we express the matrix chain directly in terms of the partial derivatives of f and g. As always we start the matrix chain by

$$AD = A = D = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B(x,y,t) = \begin{pmatrix} -\frac{\partial}{\partial x} f(x,y,t) & -\frac{\partial}{\partial y} f(x,y,t) \\ \frac{\partial}{\partial x} g(x,y,t) & \frac{\partial}{\partial y} g(x,y,t) \end{pmatrix}.$$

The first set of basis functions is given by:

$$P = V = \begin{pmatrix} I \\ 0 \end{pmatrix} \quad \text{and } Q = W = \begin{pmatrix} 0 \\ I \end{pmatrix}.$$

Thereby we obtain the set of matrices:

$$\begin{aligned} G_1 = I, \quad B^{\mathrm{v}}_{x_1}(x,y,t) &= -\frac{\partial}{\partial x} f(x,y,t), \quad & B^{\mathrm{v}}_{y_1}(x,y,t) &= -\frac{\partial}{\partial y} f(x,y,t), \\ B^{\mathrm{w}}_{x_1}(x,y,t) &= \frac{\partial}{\partial x} g(x,y,t), \quad & B^{\mathrm{w}}_{y_1}(x,y,t) &= \frac{\partial}{\partial y} g(x,y,t). \end{aligned}$$

Let $P_{y_1}(t), Q_{y_1}(t), V_{y_1}(t)$ and $W_{y_1}(t)$ be the associated basis function of the partial derivative

$$\frac{\partial}{\partial y} g(x,y,t) = B^{\mathrm{w}}_{y_1}(x,y,t).$$

Next we obtain P_{x_1} and Q_{x_1} as the basis functions of the complementary kernel and the kernel of

$$W^{\top}_{y_1}(t) \frac{\partial}{\partial x} g(x,y,t) = W^{\top}_{y_1}(t) B^{\mathrm{w}}_{x_1}(x,y,t).$$

Here P_{x_1} and Q_{x_1} are chosen such that $\begin{pmatrix} P_{x_1} & Q_{x_1} \end{pmatrix}$ is orthonormal. This enables us to choose

$$V_{x_1} = Q_{x_1} \quad \text{and} \quad W_{x_1} = P_{x_1}$$

due to $G_1 Q_{x_1} = I Q_{x_1} = Q_{x_1}$. This leads to

$$G_2 = V^{\top}_{x_1} G_1 Q_{x_1} = V^{\top}_{x_1} Q_{x_1} = Q^{\top}_{x_1} Q_{x_1} = I$$

and

$$B^{\mathrm{v}}_{x_2}(\cdot) = -V^{\top}_{x_1} \frac{\partial}{\partial x} f(\cdot) Q_{x_1} + V^{\top}_{x_1} \frac{\partial}{\partial y} f(\cdot) P_{y_1}(t) \left(V^{\top}_{y_1}(t) \frac{\partial}{\partial y} g(\cdot) P_{y_1}(t) \right)^{-1} V^{\top}_{y_1}(t) \frac{\partial}{\partial x} g(\cdot) Q_{x_1},$$

$$\begin{aligned}
B^{\mathrm{v}}_{y_2}(\cdot) &= -V^{\top}_{x_1}\frac{\partial}{\partial y}f(\cdot)Q_{y_1}(t),\\
B^{\mathrm{w}}_{x_2}(\cdot) &= -W^{\top}_{x_1}\frac{\partial}{\partial x}f(\cdot)Q_{x_1} + W^{\top}_{x_1}\frac{\partial}{\partial y}f(\cdot)P_{y_1}(t)\left(V^{\top}_{y_1}(t)\frac{\partial}{\partial y}g(\cdot)P_{y_1}(t)\right)^{-1}V^{\top}_{y_1}(t)\frac{\partial}{\partial x}g(\cdot)Q_{x_1},\\
B^{\mathrm{w}}_{y_2}(\cdot) &= -W^{\top}_{x_1}\frac{\partial}{\partial y}f(\cdot)Q_{y_1}(t)
\end{aligned}$$

with $(\cdot)$ short for (x, y, t). Next, let $P_{y_2}(t), Q_{y_2}(t), V_{y_2}(t)$ and $W_{y_2}(t)$ be the associated basis function of the partial derivative

$$W^{\top}_{x_1}\frac{\partial}{\partial y}f(x,y,t)Q_{y_1}(t) = B^{\mathrm{w}}_{y_2}(x,y,t).$$

Again we choose the basis functions P_{x_2} and Q_{x_2} of the complementary kernel and the kernel of

$$\begin{aligned}
&W^{\top}_{y_2}(t)(-W^{\top}_{x_1}\frac{\partial}{\partial x}f(\cdot)Q_{x_1} + W^{\top}_{x_1}\frac{\partial}{\partial y}f(\cdot)P_{y_1}(t)\left(V^{\top}_{y_1}(t)\frac{\partial}{\partial y}g(\cdot)P_{y_1}(t)\right)^{-1}V^{\top}_{y_1}(t)\frac{\partial}{\partial x}g(\cdot)Q_{x_1})\\
=&W^{\top}_{y_2}(t)B^{\mathrm{w}}_{x_2}(x,y,t)
\end{aligned}$$

such that $\begin{pmatrix}P_{x_1} & Q_{x_1}\end{pmatrix}$ is orthonormal. Again this leads to the possibility to choose

$$V_{x_2} = Q_{x_2} \quad \text{and} \quad W_{x_2} = P_{x_2}$$

due to $G_2Q_{x_2} = IQ_{x_2} = Q_{x_2}$. By Assumption 7.4 we restricted ourselves to index 3 DAEs. Hence, we conclude the matrix chain with

$$B^{\mathrm{w}}_{y_3}(x,y,t) := -W^{\top}_{x_2}W^{\top}_{x_1}\frac{\partial}{\partial y}f(\cdot)Q_{y_1}(t)Q_{y_2}(t).$$

The explicit description of the matrix chain by the partial derivatives of f and g will become useful when we prove the convergence of the half-explicit methods. In the following we define a class of half-explicit methods consisting of one Adams-Bashforth step and two BDF steps. Therefore we call this methods the ABDF methods. Multistep methods are popular in circuit simulation and by the ABDF methods there now is a half-explicit multistep method applicable to index 3 DAEs.

Definition 7.5. (ABDF Methods)
We consider a semi-explicit DAE (7.6) fulfilling Assumption 7.4. For the differentiable part (7.6a), we define the BDF with k steps by the function

$$F(Z,x,y,t) = -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0}Z_j + h\frac{1}{\alpha_0}f(x,y,t)$$

with α_i being the BDF-coefficients for $0 \leqslant i \leqslant 6$. We denote $X_{n-1} := \begin{pmatrix} x_{n-1} & \dots & x_{n-k} \end{pmatrix}^\top$ and define the ABDF method with k-steps

$$\begin{aligned} x_n^{AB} &= x_{n-1} + h \sum_{j=1}^{k} \beta_j f(x_{n-j}, y_{n-j}, t_{n-j}) && (7.7a) \\ x_n &= \quad F(X_{n-1}, F(X_{n-1}, x_n^{AB}, y_n, t_n), y_n, t_n) && (7.7b) \\ 0 &= g(F(X_{n-1}, F(X_{n-1}, x_n^{AB}, y_n, t_n), y_n, t_n), y_n, t_n) && (7.7c) \end{aligned}$$

with β_i being the AB-coefficients for $0 \leqslant i \leqslant 6$.

The following theorem guarantees the convergence of the ABDF methods under certain assumptions.

Theorem 7.6. (Convergence of the ABDF Methods)
We consider a semi-explicit DAE (7.6) fulfilling Assumption 7.4. Furthermore, we assume that f and g are $k+1$ times continuously differentiable but at least two times continuously differentiable. Let there be a global unique solution for the DAE (7.6) and let the initial errors in the first k steps be sufficiently small. Then the ABDF Methods, with k steps, converge with order $k-1$ in the y-components and with order k in the x-components.

Proof. Analog to the previous sections we split the numerical solutions

$$x_n = P_{x_1}\tilde{x}_{1,n} + Q_{x_1}P_{x_2}\tilde{x}_{2,n} + Q_{x_1}Q_{x_2}x_{3,n}$$

and

$$y_n = P_{y_1}(t_n)\tilde{y}_{1,n} + Q_{y_1}(t_n)P_{y_2}(t_n)\tilde{y}_{2,n} + Q_{y_1}(t_n)Q_{y_2}(t_n)\tilde{y}_{3,n}$$

as well as the global error of x

$$e_{x,n} = P_{x_1}e_{\tilde{x}_1,n} + Q_{x_1}P_{x_2}e_{\tilde{x}_2,n} + Q_{x_1}Q_{x_2}e_{x_3,n}$$

and the global error of y

$$e_{y,n} = P_{y_1}(t_n)e_{\tilde{y}_1,n} + Q_{y_1}(t_n)P_{y_2}(t_n)e_{\tilde{y}_2,n} + Q_{y_1}(t_n)Q_{y_2}(t_n)e_{\tilde{y}_3,n}.$$

The complete proof is an induction over the time steps. In particular we show that it holds for all n:

$$\begin{aligned} e_{\tilde{x}_1,n} &= \mathcal{O}(h^{k+1}), \quad e_{\tilde{x}_2,n} = \mathcal{O}(h^k), \quad e_{x_3,n} = \mathcal{O}(h^k) \\ e_{\tilde{y}_1,n} &= \mathcal{O}(h^k), \quad e_{\tilde{y}_2,n} = \mathcal{O}(h^k), \quad e_{\tilde{y}_3,n} = \mathcal{O}(h^{k-1}) \end{aligned}$$

The induction start is automatically fulfilled since the initial errors are assumed to be sufficiently small.

Solvability of the discretized system:
First, we have to show that the Equation (7.7c) has a solution y_n. Therefore we consider

$$0 = g(F(X_{n-1}, F(X_{n-1}, x_n^{AB}, y_n, t_n), y_n, t_n), y_n, t_n) + \delta_n,$$

which can be written as

$$x_n^1 = x_{n-1} + h \sum_{j=1}^{k} \beta_j f(x_{n-j}, y_{n-j}, t_{n-j}) \tag{7.8a}$$

$$x_n^2 = F(X_{n-1}, x_n^1, y_n, t_n) \tag{7.8b}$$

$$x_n^3 = F(X_{n-1}, x_n^2, y_n, t_n) \tag{7.8c}$$

$$0 = g(x_n^3, y_n, t_n) + \delta_n \tag{7.8d}$$

with the help of the auxiliary variables x_n^1, x_n^2 and x_n^3.

Decoupling to a fix point equation:
We denote

$$\begin{array}{ll} P_{y_i,n} := P_{y_i}(t_n) & Q_{y_i,n} := Q_{y_i}(t_n) \\ V_{y_i,n}^\top := V_{y_i}^\top(t_n) & W_{y_i,n}^\top := W_{y_i}^\top(t_n) \end{array}$$

for $i \leqslant 2$ and decouple Equation (7.8d) with the help of these basis functions

$$0 = V_{y_1,n}^\top g(x_n^3, P_{y_1,n}\tilde{y}_{1,n}, t_n) + V_{y_1,n}^\top \delta_n \tag{7.9a}$$

$$0 = W_{y_1,n}^\top g(P_{x_1}(P_{x_1}^\top x_n^3), \cdot, t_n) + W_{y_1,n}^\top \delta_n. \tag{7.9b}$$

In a neighborhood around the solution we obtain by Lemma 4.32 and by the Equations (7.9a), (7.9b) and (7.8c):

$$\tilde{y}_{1,n} = \Psi_{\tilde{y}_1}(x_n^3, t_n, \delta_n)$$

$$\Psi_{\tilde{x}_1}(t_n, \delta_n) = P_{x_1}^\top x_n^3 = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} P_{x_1}^\top x_{n-j} + h \frac{1}{\alpha_0} P_{x_1}^\top f(x_n^2, y_n, t_n)$$

with $\Psi_{\tilde{x}_1}(t_n, 0) = \tilde{x}_1(t_n)$. Furthermore it holds

$$\begin{aligned} &\frac{1}{h}\left(\alpha_0 \Psi_{\tilde{x}_1}(t_n, \delta_n) + \sum_{j=1}^{k} \alpha_j \tilde{x}_{1,n-j}\right) \\ =&\tilde{x}_1'(t_n) + L_n(\tilde{x}_1) + \frac{1}{h}\left(\sum_{j=1}^{k} \alpha_j e_{n-j}^{\tilde{x}_1} + \alpha_0(\Psi_{\tilde{x}_1}(t_n, \delta_n) - \Psi_{\tilde{x}_1}(t_n, 0))\right) \end{aligned}$$

$$
\begin{aligned}
&= \tilde{x}_1'(t_n) + L_n(\tilde{x}_1) + \frac{\alpha_0}{h} J_{\Psi_{\tilde{x}_1}} \delta_n + \frac{1}{h}\sum_{j=1}^{k} \alpha_j e_{n-j}^{\tilde{x}_1} \\
&= \tilde{x}_1'(t_n) + \mathcal{O}(h^k)
\end{aligned}
$$

since $L_n(\tilde{x}_1) = \mathcal{O}(h^k)$, $\delta_n = \mathcal{O}(h^{k+1})$ and $e_{n-j}^{\tilde{x}_1} = \mathcal{O}(h^{k+1})$ due to the induction statement. Hence we obtain

$$
\tilde{x}_1'(t_n) + \mathcal{O}(h^k) = \frac{1}{h}\left(\alpha_0 \Psi_{\tilde{x}_1}(t_n, \delta_n) + \sum_{j=1}^{k} \alpha_j \tilde{x}_{1,n-j}\right) = P_{x_1}^\top f(x_n^2, y_n, t_n)
$$

which can be written as

$$
\tilde{x}_1'(t_n) + \mathcal{O}(h^k) = W_{x_1}^\top f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n}, t_n) \tag{7.10}
$$

since we are able to choose $P_{x_1}^\top = W_{x_1}^\top$. As second decoupling step we split (7.10) with the help of $V_{y_2,n}^\top$ and $W_{y_2,n}^\top$ and obtain

$$
\begin{aligned}
V_{y_2,n}^\top \tilde{x}_1'(t_n) + \mathcal{O}(h^k) &= V_{y_2,n}^\top W_{x_1}^\top f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n}, t_n) \\
W_{y_2,n}^\top \tilde{x}_1'(t_n) + \mathcal{O}(h^k) &= W_{y_2,n}^\top W_{x_1}^\top f(x_n^2, P_{y_1,n}\tilde{y}_{1,n}, t_n).
\end{aligned}
$$

By inserting the expression $\Psi_{\tilde{y}_1}(x_n^3, t_n, \delta_n)$ for $\tilde{y}_1$, we get:

$$
\begin{aligned}
V_{y_2,n}^\top \tilde{x}_1'(t_n) + \mathcal{O}(h^k) &= V_{y_2,n}^\top W_{x_1}^\top f(x_n^2, P_{y_1,n}\Psi_{\tilde{y}_1}(x_n^3, t_n, \delta_n) + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n}, t_n) \\
W_{y_2,n}^\top \tilde{x}_1'(t_n) + \mathcal{O}(h^k) &= W_{y_2,n}^\top W_{x_1}^\top f(x_n^2, P_{y_1,n}\Psi_{\tilde{y}_1}(x_n^2 + (x_n^3 - x_n^2), t_n, \delta_n), t_n)
\end{aligned}
$$

which then yields functions $\Psi_{\tilde{y}_2}$ and $\Psi_{\tilde{x}_2}$ by Lemma 4.32 such that:

$$
\begin{aligned}
\tilde{y}_{2,n} &= \Psi_{\tilde{y}_2}(x_n^2, x_n^3, t_n, \delta_n, h^k) \\
\tilde{x}_{2,n}^2 &= \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k).
\end{aligned}
$$

It holds

$$
\begin{aligned}
&\frac{1}{h}\left(\alpha_0 \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) + \sum_{j=1}^{k} \alpha_j \tilde{x}_{2,n-j}\right) \\
&= \tilde{x}_2'(t_n) + L_n(\tilde{x}_2) + \frac{1}{h}\left(\sum_{j=1}^{k} \alpha_j e_{n-j}^{\tilde{x}_2} + \alpha_0(\Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0))\right) \\
&= \tilde{x}_2'(t_n) + \mathcal{O}(h^{k-1}) + \frac{1}{h}\alpha_0\left(\Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0)\right)
\end{aligned}
$$

since $L_n(\tilde{x}_2) = \mathcal{O}(h^k)$ and $e_{n-j}^{\tilde{x}_1} = \mathcal{O}(h^k)$ due to the induction statement. Therefore we obtain

$$
\tilde{x}_2'(t_n) + \mathcal{O}(h^{k-1}) + \frac{1}{h}\alpha_0\left(\Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0)\right) = P_{x_2}^\top Q_{x_1}^\top f(x_n^1, y_n, t_n).
$$

We choose $V_{x_1} = Q_{x_1}$ and $P_{x_2} = W_{x_2}$ and write

$$\tilde{x}_2'(t_n) + \mathcal{O}(h^{k-1}) + \frac{1}{h}\alpha_0 \left(\Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0)\right) = W_{x_2}^\top V_{x_1}^\top f(x_n^1, y_n, t_n)$$

which yields

$$\tilde{y}_{3,n} = \Psi_{\tilde{y}_3}(x_n^1, x_n^2, x_n^3, \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0), t_n, \delta_n, h^{k-1}).$$

At last we need to describe $x_{3,n}^3$. Therefore we multiply (7.8c) by $(W_y^*)^\top Q_{x_1}^\top$ from the left and obtain

$$\begin{aligned}
&(W_y^*)^\top Q_{x_1}^\top x_n^3 \\
=& -\sum_{j=1}^k \frac{\alpha_j}{\alpha_0}(W_y^*)^\top Q_{x_1}^\top x_{n-j} \\
&+ h\frac{1}{\alpha_0}(W_y^*)^\top Q_{x_1}^\top f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_{3,n}, t_n) \\
=& -\sum_{j=1}^k \frac{\alpha_j}{\alpha_0}(W_y^*)^\top Q_{x_1}^\top x_{n-j} \\
&+ h\frac{1}{\alpha_0}(W_y^*)^\top Q_{x_1}^\top f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_3(t_n), t_n) \\
=&(W_y^*)^\top Q_{x_1}^\top \left(-\sum_{j=1}^k \frac{\alpha_j}{\alpha_0}x_{n-j} + h\frac{1}{\alpha_0}f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_3(t_n), t_n)\right).
\end{aligned}$$

We choose P_{x_2} such that $(W_y^*)^\top P_{x_2} = 0$ and thereby obtain

$$\begin{aligned}
&Q_{x_2}^\top Q_{x_1}^\top x_n^3 \\
=&Q_{x_2}^\top Q_{x_1}^\top \left(-\sum_{j=1}^k \frac{\alpha_j}{\alpha_0}x_{n-j} + h\frac{1}{\alpha_0}f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_3(t_n), t_n)\right)
\end{aligned}$$

which can be written as

$$\begin{aligned}
&x_3(t_n) - x_{3,n}^3 \\
=&Q_{x_2}^\top Q_{x_1}^\top \frac{1}{\alpha_0}h\left(x'(t) - f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_3(t_n), t_n)\right) + \mathcal{O}(h^k)
\end{aligned}$$

since $L_n(x) = \mathcal{O}(h^k)$ and $e_{x,n-j} = \mathcal{O}(h^k)$ due to the induction statement. Together we obtain the equations

$$\tilde{y}_{1,n} = \Psi_{\tilde{y}_1}(x_n^3, t_n, \delta_n) \tag{7.11a}$$

$$
\begin{aligned}
\tilde{y}_{2,n} &= \Psi_{\tilde{y}_2}(x_n^2, x_n^3, t_n, \delta_n, h^k) && \text{(7.11b)}\\
\tilde{y}_{3,n} &= \Psi_{\tilde{y}_3}(x_n^1, x_n^2, x_n^3, \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0), t_n, \delta_n, h^{k-1}) && \text{(7.11c)}\\
\tilde{x}_{1,n}^3 &= \Psi_{\tilde{x}_1}(t_n, \delta_n) && \text{(7.11d)}\\
\tilde{x}_{2,n}^2 &= \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) && \text{(7.11e)}
\end{aligned}
$$

and

$$
\begin{aligned}
&x_3(t_n) - x_{3,n}^3 && \text{(7.12)}\\
=&Q_{x_2}^\top Q_{x_1}^\top \frac{1}{\alpha_0} h \left(x'(t) - f(x_n^2, P_{y_1,n}\tilde{y}_{1,n} + Q_{y_1,n}P_{y_2,n}\tilde{y}_{2,n} + Q_{y_1,n}Q_{y_2,n}\tilde{y}_3(t_n), t_n) \right) + \mathcal{O}(h^k).
\end{aligned}
$$

We define

$$
\Phi(\tilde{y}_{1,n}, \tilde{y}_{2,n}, \tilde{y}_{3,n}) = \begin{pmatrix} \Phi_{\tilde{y}_1}(x_n^3(y_n), t_n, \delta_n) \\ \Phi_{\tilde{y}_2}(x_n^2(y_n), x_n^3(y_n), t_n, \delta_n, h^k) \\ \Phi_{\tilde{y}_3}(x_n^1(y_n), x_n^2(y_n), x_n^3(y_n), \Delta\Psi, t_n, \delta_n, h^{k-1}) \end{pmatrix}
$$

with $y_n = P_{y_1}\tilde{y}_{1,n} + Q_{y_1}P_{y_2}\tilde{y}_{2,n} + Q_{y_1}Q_{y_2}\tilde{y}_{3,n}$ and $\Delta\Psi := \Psi_{\tilde{x}_2}(x_n^3(y_n) - x_n^2(y_n), t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0)$. Next we have to show that Φ has a fix-point.

Fix-point Theorem of Schauder:
We prove that there are constants $c_{\tilde{y}_1}$, $c_{\tilde{y}_2}$ and $c_{\tilde{y}_3}$ such that Φ has a fix-point in

$$
\mathcal{D} := B_{c_{\tilde{y}_1}h^k}(\tilde{y}_1(t_n)) \times B_{c_{\tilde{y}_2}h^k}(\tilde{y}_2(t_n)) \times B_{c_{\tilde{y}_3}h^{k-1}}(\tilde{y}_3(t_n))
$$

by the Fix-point Theorem of Schauder. For h sufficiently small Φ is defined on the whole domain $\mathcal{D}$. Since Φ is continuous we have to show that there are constants $c_{\tilde{y}_1} > 0$, $c_{\tilde{y}_2} > 0$ and $c_{\tilde{y}_3} > 0$ such that $\Phi(x) \in \mathcal{D}$ for all $x \in \mathcal{D}$. In the following we assume that $y_n \in D$ and denote $c_y = \max\{c_{\tilde{y}_1}, c_{\tilde{y}_2}, c_{\tilde{y}_3}\}$. Then it follows $y_n \in B_{c_y h^{k-1}}(y(t_n))$ and by

$$
\begin{aligned}
x_n^1 &= x_{n-1} + h \sum_{j=1}^{s+1} \beta_j f(x_{n-j}, y_{n-j}, t_{n-j})\\
x_n^2 &= F(X_{n-1}, x_n^1, y_n, t_n)
\end{aligned}
$$

there is a constant $c_1 > 0$ with

$$
x_n^1, x_n^2 \in B_{c_1 h^k}(x(t_n)). \tag{7.13}
$$

Notice that c_1 may depend on c_y. Furthermore we obtain a constant $c_2 > 0$ by

$$
\begin{aligned}
x_n^3 - x_n^2 &= F(X_{n-1}, x_n^2, y_n, t_n) - F(X_{n-1}, x_n^1, y_n, t_n)\\
&= \frac{h}{\alpha_0}(f(x_n^2, y_n, t) - f(x_n^1, y_n, t))
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{h}{\alpha_0} \int_0^1 \frac{\partial}{\partial x} f(sx_n^2 + (1-s)x_n^1, y_n, t)\mathrm{d}s(x_n^2 - x_n^1) \\
&= \frac{h}{\alpha_0} \int_0^1 \frac{\partial}{\partial x} f(sx_n^2 + (1-s)x_n^1, y_n, t)\mathrm{d}s((x_n^2 - x(t_n)) - (x_n^1 - x1(t_n)))
\end{aligned}
$$

with $(x_n^3 - x_n^2) \in B_{c_2 h^{k+1}}(0)$ due to $x_n^1, x_n^2 \in B_{c_1 h^k}(x(t_n))$. Again c_2 may depend on c_y. But for a sufficiently small h we get a constant $c_3 > 0$ with

$$
(x_n^3 - x_n^2) \in B_{c_3 h^k}(x(t_n)) \tag{7.14}
$$

and c_3 being independent from c_y. Hence there is a ξ_2 with $x_n^3 = x_n^2 + \xi_2 h^k$ and $\|\xi_2\| \leqslant c_3$. By (7.11d) and $\|\delta\| \leqslant h^{k+1}$, we get

$$
\tilde{x}_{1,n}^3 \in B_{\tilde{c}_1 h^{k+1}}(\tilde{x}_1(t_n)) \tag{7.15}
$$

with $\tilde{c}_1 > 0$ being a constant independent from c_y. Furthermore, (7.11e) can now be rewritten into:

$$
\begin{aligned}
\tilde{x}_{2,n}^2 &= \Psi_{\tilde{x}_2}(x_n^3 - x_n^2, t_n, \delta_n, h^k) \\
&= \Psi_{\tilde{x}_2}(x_n^2 + \xi_2 h^k - x_n^2, t_n, \delta_n, h^k) \\
&= \Psi_{\tilde{x}_2}(\xi_2 h^k, t_n, \delta_n, h^k)
\end{aligned}
$$

which yields a constant $\tilde{c}_2 > 0$ with

$$
\tilde{x}_{2,n}^2 \in B_{\tilde{c}_2 h^k}(\tilde{x}_2(t_n)) \tag{7.16}
$$

and $\tilde{c}_2 > 0$ being a constant independent from c_y. Finally we obtain a constant $\tilde{c}_3 > 0$ with $\tilde{x}_{3,n}^3 \in B_{\tilde{c}_3 h^k}(\tilde{x}_3(t_n))$ and $\tilde{c}_3 > 0$ being a constant independent from c_y by (7.12) and $(\tilde{y}_{1,n}, \tilde{y}_{2,n}) \in B_{c_{\tilde{y}_1} h^k}(\tilde{y}_1(t_n)) \times B_{c_{\tilde{y}_2} h^k}(\tilde{y}_2(t_n))$. Together with (7.13) and (7.14) the three constants $\tilde{c}_1, \tilde{c}_2$ and $\tilde{c}_3$ yield a constant $\tilde{c}_x > 0$ with

$$
x_n^1 \in B_{\tilde{c}_x h^{k-1}}(x(t_n)) \quad \text{and} \quad x_n^2, x_n^3 \in B_{\tilde{c}_x h^k}(x(t_n)) \tag{7.17}
$$

and $\tilde{c}_x$ being a constant independent from c_y. Thereby we find constant $\tilde{c}_{\tilde{y}_i}$ for $i = 1, 2, 3$ with

$$
\begin{aligned}
&\Phi(\tilde{y}_{1,n}, \tilde{y}_{2,n}, \tilde{y}_{3,n}) \\
&= \begin{pmatrix} \Phi_{\tilde{y}_1}(x_n^3(y_n), t_n, \delta_n) \\ \Phi_{\tilde{y}_2}(x_n^2(y_n), x_n^3(y_n), t_n, \delta_n, h^k) \\ \Phi_{\tilde{y}_3}(x_n^1(y_n), x_n^2(y_n), x_n^3(y_n), \Delta\Psi, t_n, \delta_n, h^{k-1}) \end{pmatrix} \\
&= \begin{pmatrix} \Phi_{\tilde{y}_1}(x(t_n), t_n, 0) + \tilde{c}_{\tilde{y}_1} h^k \\ \Phi_{\tilde{y}_2}(x(t_n), x(t_n), t_n, 0, 0) + \tilde{c}_{\tilde{y}_2} h^k \\ \Phi_{\tilde{y}_3}(x(t_n), x(t_n), x(t_n), \Psi_{\tilde{x}_2}(0, t_n, 0, 0) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0), t_n, 0, 0) + \tilde{c}_{\tilde{y}_3} h^{k-1} \end{pmatrix}
\end{aligned}
$$

$$= \begin{pmatrix} \tilde{y}_1(t_n) + \tilde{c}_{\tilde{y}_1} h^k \\ \tilde{y}_2(t_n) + \tilde{c}_{\tilde{y}_2} h^k \\ \tilde{y}_3(t_n) + \tilde{c}_{\tilde{y}_3} h^{k-1} \end{pmatrix}$$

with $\Delta\Psi := \Psi_{\tilde{x}_2}(x_n^3(y_n) - x_n^2(y_n), t_n, \delta_n, h^k) - \Psi_{\tilde{x}_2}(0, t_n, 0, 0)$. For h sufficiently small, we can choose $\tilde{c}_{\tilde{y}_i} = \frac{1}{2} c_{\tilde{y}_i}$ for $i = 1, 2, 3$ and get $\Phi(\tilde{y}_{1,n}, \tilde{y}_{2,n}, \tilde{y}_{3,n}) \in \mathcal{D}$.

Error recursion for x_n
In the following let y_n solve

$$0 = g(F(X_{n-1}, F(X_{n-1}, x_n^{AB}, y_n, t_n), y_n, t_n), y_n, t_n) + \delta_n.$$

Again we consider the equations of the method

$$x_n^1 = x_{n-1} + h \sum_{j=1}^{k} \beta_j f(x_{n-j}, y_{n-j}, t_{n-j}) \tag{7.18a}$$

$$x_n^2 = F(X_{n-1}, x_n^1, y_n, t_n) \tag{7.18b}$$

$$x_n = F(X_{n-1}, x_n^2, y_n, t_n). \tag{7.18c}$$

Furthermore the consistency error yields the equations:

$$x(t_n) = x(t_{n-1}) + h \sum_{j=1}^{k} \beta_j f(x(t_{n-j}), y(t_{n-j}), t_{n-j}) + \mathcal{O}(h^{k+1}) \tag{7.19a}$$

$$x(t_n) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x(t_{n-j}) + h \frac{1}{\alpha_0} f(x(t_n), y(t_n), t_n) + \mathcal{O}(h^{k+1}). \tag{7.19b}$$

Error bound for $\tilde{x}_{1,n}$ and $\tilde{x}_{2,n}$:
We notice that the constants in (7.14), (7.15) and (7.16) are not only independent of c_y but can also be chosen independently from the step n. Therefore we get $e_{\tilde{x}_{1,n}} = \mathcal{O}(h^{k+1})$ and $e_{\tilde{x}_{2,n}} = \mathcal{O}(h^k)$. Now we only have to analyze the error of the inherent dynamic $x_{3,n}$.

Error recursion for $x_{3,n}$:
We multiply $(W_y^*)^\top Q_{x_1}^\top$ to (7.18c), to obtain

$$\begin{aligned} &(W_y^*)^\top Q_{x_1}^\top x_n \\ =&(W_y^*)^\top Q_{x_1}^\top \left(-\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{n-j} + h \frac{1}{\alpha_0} f(x_n^2, y(t_n) + P_{y_1,n} c_{\tilde{y}_{1,n}} h^k + Q_{y_1,n} P_{y_2,n} c_{\tilde{y}_{2,n}} h^k, t_n) \right) \\ =&(W_y^*)^\top Q_{x_1}^\top \left(-\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{n-j} + h \frac{1}{\alpha_0} f(x_n^2, y(t_n), t_n) \right) + \mathcal{O}(h^{k+1}) \end{aligned}$$

and insert the transformation of x_n to get

$$(W_y^*)^\top Q_{x_2} x_{3,n} = -(W_y^*)^\top Q_{x_2} \sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{3,n-j} + h\frac{1}{\alpha_0}(W_y^*)^\top Q_{x_1}^\top f(x_n^2, y(t_n), t_n) + \mathcal{O}(h^{k+1}).$$

By multiplying the inverse of $(W_y^*)^\top Q_{x_2}$ we achieve

$$x_{3,n} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{3,n-j} + h\frac{1}{\alpha_0}((W_y^*)^\top Q_{x_2})^{-1}(W_y^*)^\top Q_{x_1}^\top f(x_n^2, y(t_n), t_n) + \mathcal{O}(h^{k+1})$$

and analogously we obtain

$$x_3(t_n) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_3(t_{n-j}) + h\frac{1}{\alpha_0}((W_y^*)^\top Q_{x_2})^{-1}(W_y^*)^\top Q_{x_1}^\top f(x(t_n), y(t_n), t_n) + \mathcal{O}(h^{k+1})$$

from (7.19b). By subtracting these two equations, we obtain an error recursion for $x_{3,n}$ which depends on $(x(t_n) - x_n^2)$:

$$e_{x_{3,n}} = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} e_{x_{3,n-j}} + h\partial f_n^3 (x(t_n) - x_n^2) + \mathcal{O}(h^{k+1}).$$

Here ∂f_n^3 is a suitable matrix provided by the Mean Value Theorem, as in the proof to Theorem 6.6. Hence, we have to investigate the term $x(t_n) - x_n^2$. Therefore consider (7.18b) and (7.19b). By inserting $y(t_n) + \mathcal{O}(h^{k-1})$ for y_n, we obtain:

$$x_n^2 = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x_{n-j} + h\frac{1}{\alpha_0} f(x_n^1, y(t_n), t_n) + \mathcal{O}(h^k)$$

$$x(t_n) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} x(t_{n-j}) + h\frac{1}{\alpha_0} f(x(t_n), y(t_n), t_n) + \mathcal{O}(h^k)$$

which leads to

$$(x(t_n) - x_n^2) = -\sum_{j=1}^{k} \frac{\alpha_j}{\alpha_0} (x(t_{n-j}) - x_{n-j}) + h\partial f_n^2 (x(t_n) - x_n^1) + \mathcal{O}(h^k)$$

by subtracting these two equations. Also ∂f_n^2 is a suitable matrix provided by the Mean Value Theorem, as in the proof to Theorem 6.6.
We now have to deal with the term $x(t_n) - x_n^1$. Considering (7.18a) and (7.19a), we again use $y_n = y(t_n) + \mathcal{O}(h^{k-1})$ to obtain

$$x_n^1 = x_{n-1} + h\sum_{j=1}^{k} \beta_j f(x_{n-j}, y(t_{n-j}), t_{n-j}) + \mathcal{O}(h^k)$$

$$x(t_n) = x(t_{n-1}) + h\sum_{j=1}^{k}\beta_j f(x(t_{n-j}), y(t_{n-j}), t_{n-j}) + \mathcal{O}(h^k).$$

Once again by subtracting we obtain

$$(x(t_n) - x_n^1) = (x(t_{n-1}) - x_{n-1}) + h\sum_{j=1}^{k}\beta_j \partial f_{n-j}^1 (x(t_{n-j}) - x_{n-j}^1) + \mathcal{O}(h^k).$$

Again ∂f_{n-j}^1 are obtained by the Mean Value Theorem, as in the proof to Theorem 6.6. Altogether we have the following three equations available:

$$\begin{aligned}
(x(t_n) - x_n^1) &= e_{n-1} + h\sum_{j=1}^{k}\beta_j \partial f_{n-j}^1 e_{n-j} + \mathcal{O}(h^k)\\
(x(t_n) - x_n^2) &= -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{n-j} + h\partial f_n^2 (x(t_n) - x_n^1) + \mathcal{O}(h^k)\\
e_{x_3,n} &= -\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{x_3,n-j} + h\partial f_n^3 (x(t_n) - x_n^2) + \mathcal{O}(h^{k+1})
\end{aligned}$$

which can be combined to:

$$\begin{aligned}
e_{x_3,n} = &-\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{x_3,n-j}\\
&+ h\partial f_n^3 (-\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{n-j} + h\partial f_n^2 (e_{n-1} + h\sum_{j=1}^{k}\beta_j \partial f_{n-j}^1 e_{n-j}) + \mathcal{O}(h^{k+1}).
\end{aligned}$$

By transforming the error e_n into $e_{\tilde{x}_1,n}$, $e_{\tilde{x}_2,n}$ and $e_{x_3,n}$ and by using $e_{\tilde{x}_1,n} = \mathcal{O}(h^{k+1})$ and $e_{\tilde{x}_2,n} = \mathcal{O}(h^k)$ we get:

$$\begin{aligned}
e_{x_3,n} = &-\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{x_3,n-j}\\
&+ h\partial \tilde{f}_n^3 (-\sum_{j=1}^{k}\frac{\alpha_j}{\alpha_0} e_{x_3,n-j} + h\partial \tilde{f}_n^2 (e_{x_3,n-1} + h\sum_{j=1}^{k}\beta_j \partial \tilde{f}_{n-j}^1 e_{x_3,n-j}) + \mathcal{O}(h^{k+1}).
\end{aligned}$$

with certain matrices $\tilde{f}_n^3$, $\tilde{f}_n^2$ and $\tilde{f}_{n-j}^1$. This recursion can be written as

$$\begin{aligned}
E_{\tilde{x}_3,n} &= AE_{\tilde{x}_3,n-1} + hBE_{\tilde{x}_3,n-1} + \mathcal{O}(h^{k+1})\\
&= (A + hB)E_{\tilde{x}_3,n-1} + \mathcal{O}(h^{k+1})
\end{aligned}$$

with

$$A = \begin{pmatrix} -\frac{\alpha_1}{\alpha_0} I & \cdots & \cdots & \cdots & -\frac{\alpha_k}{\alpha_0} I \\ I & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & I & 0 \end{pmatrix}, \quad E_{\tilde{x}_3,n} := \begin{pmatrix} e_{x_3,n} \\ \vdots \\ e_{x_3,n-k+1} \end{pmatrix}$$

and a matrix B with $\|B\|$ being bounded for a sufficiently small h. At this point we can proceed according to standard ODE convergence proofs and obtain $e_{x_3,n} = \mathcal{O}(h^k)$. □

We close this section by applying a $ABDF$ method to the model of the Ring Modulator, see Figure 7.2, a well known electric circuit example, cf. [Hor76] and [KRS92]. Therefore define the constant capacitance values, resistance values, inductance values

$$\begin{array}{llll} C = 1.6 \cdot 10^{-8} F & L_{s1} = 0.002 H & R = 25000 \Omega & R_p = 50 \Omega \\ C_p = 10^{-8} F & L_{s2} = 5 \cdot 10^{-4} H & R_{g1} = 36.3 \Omega & R_i = 50 \Omega \\ L_h = 4.45 H & L_{s3} = 5 \cdot 10^{-4} H & R_{g2} = R_{g3} = 17.3 \Omega & R_c = 600 \Omega \end{array}$$

the input potentials $e^1_{\text{in}}(t) = 0.5 * sin(2000\pi t)$, $e^2_{\text{in}}(t) = 2 * sin(20000\pi t)$ and the nonlinear conductance $g(u) = \gamma(e^{\delta u} - 1)$ with $\delta = 17.7493332$ and $\gamma = 40.67286402 \cdot 10^{-9}$.
The equations of the MNA of the Ring Modulator can be reduced to a semi-explicit DAE with 11 differential equations and 4 algebraic equations. We apply the $ABDF$ method with order one and a constant time step size $h = 3 \cdot 10^{-8}$ to the reduced set of equation of the Ring Modulator with the simulation interval $\mathcal{I} = [0, 10^{-3}]$. We obtain the same results as in [Hor76] and [KRS92] and present the voltages U_1, U_2 and the voltage $U_3 := e_2 - e_1$ for a qualitative comparison.

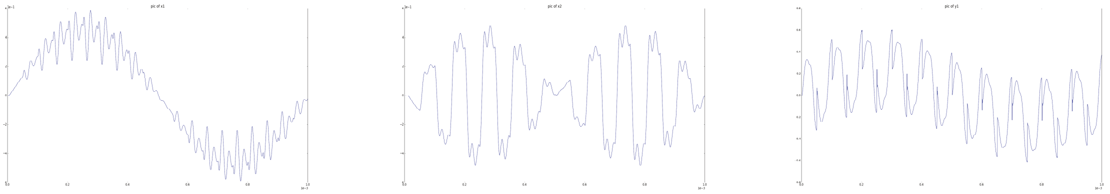


Figure 7.1: The voltages U_1, U_2 and $U_3 := e_2 - e_1$.

The circuit diagram of the Ring Modulator is given by:

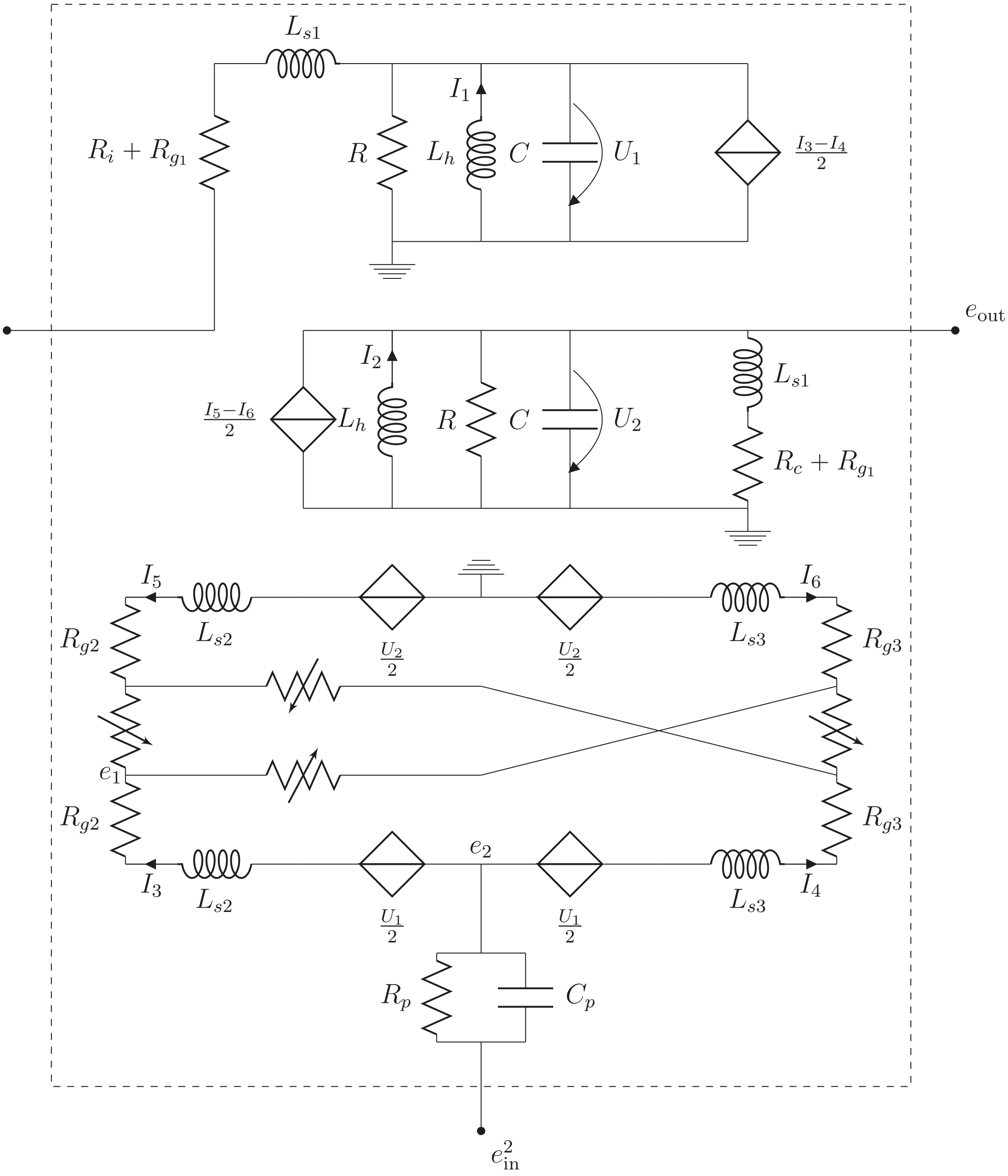


Figure 7.2: Circuit diagram of the Ring Modulator

7.3 Summery and Outlook

This chapter had two objectives. The first objective was to obtain an explicit description of the basis functions for electric circuits including semiconductor devices, memristors and electromagnetic devices by the networks topology. With the help of these basis functions we decoupled the equations of the extended MNA into a semi-explicit DAE. If we only consider independent sources the decoupled DAE has index one. In the case of a circuit with controlled sources we still obtain a semi-explicit DAE but the index may be larger than one.

While half-explicit Runge-Kutta methods are well established for higher index DAEs, half-explicit multistep methods are mostly defined for index one DAEs. The second objective of this chapter was to introduce a half-explicit multistep method for semi-explicit DAEs with an index three or lower. We constructed such methods by a mixture of BDF and Adams Bashford methods and proved their convergence for semi-explicit index three DAEs. Then we closed the chapter with a numeric example.

8 Conclusion and Outlook

This thesis addressed differential-algebraic equations. We focused our investigations to general differential-algebraic equations resulting from a spatial discretization of PDAEs. We discussed and extended existing results regarding the modeling and numerical simulation of DAEs. Furthermore, we investigated the global unique solvability and the sensitivity of solutions with respect to perturbations of DAEs.
We aimed for three main objectives:

1. A global solvability theorem which can be applied to coupled systems to mathematically justify their coupling approach.

2. Numerical methods which are stable without needing any structural assumptions.

3. A way to apply explicit methods to coupled systems to be able to handle the size of the coupled systems by parallelizing the algorithms.

The most important tool to achieve these objectives was the concept of the Dissection Index. In contrast to the the Tractability Index and the Strangeness Index, the Dissection Index fulfills the following properties:

1. The complexity of the decoupling procedure reflects the complexity of the DAE.

2. The decoupling procedure preserves properties like symmetry, monotonicity and positive definiteness.

3. The decoupling procedure is realized by a step-by-step approach with independent stages.

The Dissection Index can be interpreted as a mix of the Tractability Index and the Strangeness Index. The index arises as we use the linearization concept of the Tractability Index and the decoupling procedure of the Strangeness Index.
After introducing our new index concept and proving that it is well defined, see Theorem 4.19 and 4.22, we analyzed the sensitivity to perturbations of differential-algebraic equations. We were able to proof a connection between the Dissection Index and the Perturbation Index for DAEs with an arbitrarily high index, see Theorem 4.38. In case of the perturbation analysis and also for the convergence theory it is necessary to assume that the unperturbed DAE has a global unique solution. Furthermore we needed to prove

the global unique solvability of our considered coupled systems to mathematically justify their coupling approach. We provided sufficient criteria for the global unique solvability of differential-algebraic equations with an arbitrary index, see Theorem 5.13, and applied this theoretical result to our circuit application, see Theorem 5.20.
Furthermore, we dealt with challenges of the applicability, the stability and the convergence of numerical methods. It is known that standard ODE methods like the implicit Euler methods, the BDF methods or the Radau IIA methods may loose their convergence if applied to DAEs, cf. [GP83, LMT13]. We identified the source of these instabilities and provided sufficient convergence criteria for the standard ODE methods, see Theorem 6.6. Then we introduced a class of methods which overcomes these instability problems and proved their convergence, see Theorem 6.15.
In the last chapter we investigated half-explicit methods applied to DAEs. Since it is no longer possible to accelerate CPUs as it were in the past, parallelizing algorithms becomes more and more important. Hence explicit methods are focus even more nowadays because they can be parallelized very efficiently. We introduce a new class of of half-explicit multistep methods for index 3 DAEs and prove their convergence, see Theorem 7.6.

The coupled systems, which were considered in this thesis, can be embedded in a more general network approach, cf. [JT14]. This general network approach includes various network applications such as water or gas transportation networks, blood flow networks or electric circuits. For all of these networks it is important to be able to simulate the behavior of the respective network in advance. In comparison to blood flow networks or electric circuits the flow rate in a water transportation network is slow relative to distance the water has to cover. Therefore it is necessary to anticipate changes in the water demand and know how to react to those changes before handed, since it may take hours up to days till the water arrives at the needed locations. In the case of a blood flow network it is of great interest to know how the quantities of the network react if a new substance, for example a medicine, is insert into the network. With the help of simulations the effects of new drugs can be tested without putting test subjects in potential danger. In particular if new drugs for children are developed simulations can be of great importance. The next objective is to apply the results of this thesis to all the other kinds of network besides electric circuits.

Bibliography

[ABG04] G. Ali, A. Bartel, and M. Günther. Parabolic differential-algebraic models in electrical network design. *SIAM MMS*, 2004.

[ABGT03] G. Ali, A. Bartel, M. Günther, and C. Tischendorf. Elliptic partial differential algebraic multiphysics models in electrical network design. *Math. Models Meth. Appl. Sci.*, 13(9):1261–1278, 2003.

[Ada75] R.A. Adams. *Sobolev Spaces.* Academic Press, New York, 1975.

[Ami92] F.M.L. Amirouche. *Computional Methods in Multibody Dynamics.* Prentice Hall, New Jersey, 1992.

[Arn98] M. Arnold. Half-explicit Runge-Kutta methods with explicit stages for differential-algebraic systems of index 2. *BIT Numerical Mathematics*, 38(3):415–438, 1998.

[ASW93] M. Arnold, K. Strehmel, and R. Weiner. Half-explicit Runge-Kutta methods for semi-explicit differential-algebraic equations of index 1. *Numerische Mathematik*, 64(1):409–431, 1993.

[Aul04] B. Aulbach. *Gewöhnliche Differenzialgleichungen.* Spektrum Akademischer Verlag, 2004.

[Bar04] Andreas Bartel. *Partial Differential-Algebraic Models in Chip Design - Thermal and Semiconductor Problems. Fortschritt-Berichte VDI.* Number 391 in 20. VDI-Verlag, Düsseldorf, 2004.

[Bau12] S. Baumanns. *Coupled Electromagnetic Field/Circuit Simulation: Modeling and Numerical Analysis.* PhD thesis, Universität zu Köln, 2012.

[BBS11] A. Bartel, S. Baumanns, and S. Schöps. Structural Analysis of Electrical Circuits Including Magnetoquasistatic Devices. *Applied Numerical Mathematics*, 61(12):1257–1270, 2011.

[BCP96] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical solution of initial-value problems in differential-algebraic equations*, volume 14. SIAM, 1996.

[BH93] V. Brasey and E. Hairer. Half-explicit Runge-Kutta methods for differential-algebraic systems of index 2. *SIAM J. Numer. Anal.*, 30(2):538–552, April 1993.

[Bod07] M. Bodestedt. *Perturbation Analysis of Refined Models in Circuit Simulation.* PhD thesis, Technical University of Berlin, 2007.

[BST10] S. Baumanns, M. Selva Soto, and C. Tischendorf. Consistent initialization for coupled circuit-device simulation. In L.R.J. Costa J. Roos, editor, *Mathematics in Industry – Scientific Computing in Electrical Engineering (SCEE 2008).* Springer Verlag, 2010.

[Cam87] S.L. Campbell. A general form for solvable linear time varying singular systems of differential equations. *SIAM journal on Mathematical Analysis*, 18(4):1101–1115, 1987.

[CC07] V.F. Chistyakov and E.V. Chistyakova. Nonlocal theorems on existence of solutions of differential-algebraic equations of index 1. *Russian Mathematics (Iz Vuz)*, 51(1):71–76, 2007.

[CDK87] L.O. Chua, C.A. Desoer, and E.S. Kuh. *Linear and nonlinear circuits.* McGraw-Hill, Singapore, 1987.

[CG95a] S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM J. Sci. Comput.*, 16:257–270, 1995.

[CG95b] S.L. Campbell and C. W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[Chu11] L.O. Chua. Resistance switching memories are memristors. *Applied Physics A: Materials Science & Processing*, 102(4):765–783, 2011.

[CL75] L.O. Chua and P.M. Lin. *Computer-aided analysis of electronic circuits: algorithms and computational techniques.* Prentice-Hall series in electrical and computer engineering. Prentice-Hall, 1975.

[CW01] M. Clemens and T. Weiland. Discrete Electromagnetism with the Finite Integration Technique. *Progress In Electromagnetics Research*, 32:65–87, 2001.

[DK84] C.A. Desoer and E.S. Kuh. *Basic Circuit Theory.* International student edition. McGraw-Hill, 1984.

[ESF98] E. Eich-Soellner and C. Führer. *Numerical Methods in Multibody Systems.* Lecture Notes in Mathematics. Teubner, Stuttgart, 1998.

[Est00] D. Estévez Schwarz. *Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation.* Dissertation, Humboldt-Universität zu Berlin, 2000.

[ET00] D. Estévez Schwarz and C. Tischendorf. Structural analysis of electric circuits and consequences for MNA. *International journal of Circuit Theory and Applications*, 28(2):131–162, 2000.

[Eva10] L.C. Evans. *Partial Differential Equations.* Graduate Studies in Mathematics. American Mathematical Society, 2010.

[For96] O. Forster. *Analysis 3.* Vieweg-Verlag, 1996.

[Gaj93] H. Gajewski. Analysis und Numerik von Ladungstransport in Halbleitern. Technical Report 6, Weierstrass Institut für Andgewandte Analysis und Stochastik, 1993.

[Gaj94] H. Gajewski. In *On uniqueness of solutions to the drift-diffusion-model of semiconductor devices*, volume 117, pages 171–183, 1994.

[GJ09] L. Grüne and O. Junge. *Gewöhnliche Differentialgleichungen: Eine Einführung aus der Perspektive der dynamischen Systeme.* Teubner, 2009.

[GJH$^+$14] S. Grundel, L. Jansen, N. Hornung, T. Clees, C. Tischendorf, and P. Benner. Model order reduction of differential algebraic equations arising from the simulation of gas transport networks. *Progress in Differential-Algebraic Equations - Deskriptor 2013*, pages 183–206, 2014.

[GM86] E. Griepentrog and R. März. *Differential-Algebraic Equations and Their Numerical Treatment.* Teubner, Leipzig, 1986.

[GP83] C.W. Gear and L.R. Petzold. Differential/algebraic systems and matrix pencils. In B. Kågström and A. Ruhe, editors, *Matrix Pencils*, volume 973 of *Lecture Notes in Mathematics*, pages 75–89. Springer, Berlin/Heidelberg, 1983.

[GS00] T. Grasser and S. Selberherr. Mixed-mode device simulation. *Microelectronics Journal*, 31(11-12):873–881, 2000.

[Hau89] E.J. Haug. *Computer Aided Kinematics and Dynamics of Mechanical Systems, volume 1: Basic Methods.* Allyn and Bacon, 1989.

[HLR89] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods.* Springer, New York, 1989.

[HM76] A.Y. Hannalla and D.C. MacDonald. Numerical analysis of transient field problems in electrical machines. *Proceedings of the Institution of Electrical Engineers*, 123(9):893–898, 1976.

[HM04] I. Higueras and R. März. Differential algebraic equations with properly stated leading terms. *Computers & Mathematics with Applications*, 48(1-2):215–235, 2004.

[HMM98] M. Hanke, E.I. Macana, and R. März. On asymptotics in case of linear index-2 DAE's. *SIAM J. Numer. Anal.*, 35(4):1326–1346, 1998.

[HMT03] I. Higueras, R. März, and C. Tischendorf. Stability Preserving Integeration of Index-2 DAEs. *Applied Numerical Mathematics*, 45(2-3):201–229, 2003.

[HNW02] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems.* Springer series in Computational Mathematics. Springer, Berlin, 2 edition, 2002.

[Hor76] E.H. Horneber. *Analyse nichtlinearer RLCÜ-Netzwerke mit Hilfe der gemischten Potentialfunktion mit einer systematischen Darstellung der Analyse nichtlinearer dynamischer Netzwerke.* Dissertation, Universität Kaiserslautern, 1976.

[HWL06] Ernst Hairer, Gerhard Wanner, and Christian Lubich. *Geometric Numerical Integration. Structure-Preserving Algorithms for Ordinary Differential Equations.* Springer Verlag, Berlin Heidelberg, 2006.

[JMT13] L. Jansen, M. Matthes, and C. Tischendorf. Global Unique Solvability of Nonlinear Index-1 DAEs with Monotonicity Properties. *Int. J. Circ. Theor. Appl.*, 2013.

[JP13] L. Jansen and Jonas Pade. Global Unique Solvability for a Quasi-Stationary Water Network Model. *Preprint, Humboldt-Universität zu Berlin*, 2013.

[JT14] L. Jansen and C. Tischendorf. Multiphysical Modeling and Numerical Simulation of Flow Networks. *Progress in Differential-Algebraic Equations - Deskriptor 2013*, 2014.

[KM06] P. Kunkel and V. Mehrmann. *Differential-algebraic equations: analysis and numerical solution.* EMS textbooks in mathematics. European Mathematical Society, 2006.

[KM07] P. Kunkel and V. Mehrmann. Stability properties of differential-algebraic equations and spin-stabilized discretizations. *Electronic Transactions on Numerical Analysis*, 26:385–420, 2007.

[KMST93] A. Konrad, G. Meunier, J.C. Sabonnadière, and I.A. Tsukerman. Coupled field-circuit problems: trends and accomplishments. *IEEE Transactions on Magnetics*, 29(2):1701–1704, 1993.

[KRS92] W. Kampowski, P. Rentrop, and W. Schmidt. Classification and numerical simulation of electric circuits. *Surveys on Mathematics for Industry*, 2(1):23–65, 1992.

[Lam07] R. Lamour. A projector based representation of the Strangeness Index Concept. *Preprint, Humboldt-UniversitÃ¤t zu Berlin*, 2007.

[LM14] Vu Hoang Linh and Volker Mehrmann. Efficient integration of strangeness-free non-stiff differential-algebraic equations by half-explicit methods. *J. Comput. Appl. Math.*, 262:346–360, May 2014.

[LMT13] R. Lamour, R. März, and C. Tischendorf. *Differential Algebraic Equations: A Projector Based Analysis.* Springer, 2013.

[LRS86] M. M. Lavrentev, V. G. Romanov, and S. P. Shishatskii. *Ill-Posed Problems of Mathematical Physics and Analysis.* American Mathematical Society, Berlin, 1986.

[Mar86] P. A. Markowich. *The stationary semiconductor device equations.* Springer, 1986.

[Mär98] R. März. Characterizing differential algebraic equations without the use of derivative arrays. *Computers Math. Appl.*, 50:1141–1156, 1998.

[Mär02] R. März. The index of linear differential algebraic equations with properly stated leading terms. *Results in Mathematics*, 42(3-4):308–338, 2002.

[Mat13] M. Matthes. *Numerical Analysis of Nonlinear Partial Differential Algebraic Equations: A Coupled and an Abstract Systems Approach.* PhD thesis, Universität zu Köln, 2013.

[Meh12] V. Mehrmann. Index concepts for differential-algebraic equations. *Preprint, TU Berlin*, 2012.

[MG05] J. ter Maten M. Günther, U. Feldmann. *Modelling and discretization of circuit problems*, volume 13 of *Handbook of numerical analysis.* Elsevier, 2005.

[Moc83] M. S. Mock. *Analysis of Mathematical Models of Semiconductor Devices.* Boole Press, 1983.

[MSW99] J. Miller, W. Schilders, and S. Wang. Application of finite element methods to the simulation of semiconductor devices. *Rep. Prog. Phys.*, 62:277–353, 1999.

[Mur97] A. Murua. Partitioned half-explicit Runge-Kutta methods for differential-algebraic systems of index 2. *Computing*, 59(1):43–61, 1997.

[OR70] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Academic Press, 1970.

[Ost93] Alexander Ostermann. A class of half-explicit Runge-Kutta methods for differential-algebraic systems of index 3. *Applied Numerical Mathematics*, 13(1-3):165 – 179, 1993.

[Pet82] L.R. Petzold. *Description of DASSL: A differential/algebraic system solver.* Sep 1982.

[Pul12] R. Pulch. Stochastic Collocation and Stochastic Galerkin Methods for Linear Differential Algebraic Equations. *Preprint, Bergische Universität Wuppertal*, 2012.

[Rei91] S. Reich. On an existence and uniqueness theory for nonlinear differential-algebraic equations. *Circuits, Systems and Signal Processing*, 10(3):343–359, 1991.

[Ria11] R. Riaza. Dynamical properties of electrical circuits with fully nonlinear memristors. *Nonlinear Analysis: Real World Applications*, 12(6):3674–3686, 2011.

[RK04] A.G. Rutkas and I.G. Khudoshin. Global solvability of one degenerate semilinear differential operator equation. *Nonlinear Oscillations*, 7:403–417, 2004.

[RR88] R. Schwertassek R.E. Roberson. *Dynamics of multibody systems.* Springer-Verlag, Berlin, Germany, 1988.

[RT11] R. Riaza and C. Tischendorf. Semistate models of electrical circuits including memristors. *International journal of Circuit Theory and Applications*, 39(6):607–627, 2011.

[SBST14] L. Jansen S. Baumanns, M. Selva Soto, and C. Tischendorf. Analysis of semidiscretized differential algebraic equation from coupled circuit device simulation. *Computational and Applied Mathematics*, 2014.

[Sch11] S. Schöps. *Multiscale Modeling and Multirate Time-Integration of Field/Circuit Coupled Problems.* Dissertation, Bergischen Universität Wuppertal, 2011.

[Sel84] S. Selberherr. *Analysis and simulation of semiconductor devices.* Springer-Verlag, 1984.

[Sim95] B. Simeon. Numerical integration software for constrained mechanical motion. *Surveys on Mathematics for Industry*, 5(3):169–202, 1995.

[Sot06] M. Selva Soto. An index analysis from coupled circuit and device simulation. *Mathematics in Industry – Scientific Computing in Electrical Engineering (SCEE 2004)*, pages 121–128, 2006.

[SSSW08] D.B. Strukov, G.S. Snider, D.R. Stewart, and R.S. Williams. The missing memristor found. *Nature*, 453(7191):80–83, May 2008.

[ST05] M. Selva Soto and C. Tischendorf. Numerical Analysis of DAEs from coupled circuit and semiconductor simulation. *Applied Numerical Mathematics*, 53(2-4):471–488, 2005.

[Ste06] A. Steinbrecher. *Numerical Solution of Quasi-Linear Differential-Algebraic Equations and Industrial Simulation of Multibody Systems.* Dissertation, Technische Universität Berlin, 2006.

[Str14] Christian Strohm. Lösbarkeit gekoppelter partieller Differentialgleichungen für die Simulation elektronischer Schaltungen. Master's thesis, University of Cologne, 2014.

[Tis99] C. Tischendorf. Topological index calculation of DAEs in circuit simulation. *Surveys on Mathematics for Industry*, 8(3-4):187–199, 1999.

[Tis03] C. Tischendorf. *Coupled Systems of Differential Algebraic and Partial Differential Equations in Circuit and Device Simulation.* Habilitation Thesis. Humboldt University of Berlin, 2003.

[TW96] P. Thoma and T. Weiland. A consistent subgridding scheme for the finite difference time domain method. *International journal of Numerical Modelling: Electronic Networks, Devices and Fields*, 9(5):359–374, 1996.

[Voi06] S. Voigtmann. *General linear methods for integrated circuit design.* Dissertation, Humboldt-Universität zu Berlin, 2006.

[Wei77] T. Weiland. A discretization model for the solution of Maxwell's equations for six-component fields. *Archiv Elektronik und Übertragungstechnik*, 31(3):116–120, 1977.

[Whi59] E.T. Whittaker. *A Treatise on the Analytical Dynamics of Particles and Rigid Bodies.* Cambridge University Press, Cambridge, 1959.

[Yee66] K.S. Yee. Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media. *IEEE Transactions on Antennas and Propagation*, 14(3):302–307, 1966.

[Zei86] E. Zeidler. *Nonlinear Functional Analysis and its Applications I. Fixed Point Theorems.* Springer-Verlag, New York, 1986.

[Zei90] E. Zeidler. *Nonlinear Functional Analysis and its Applications II/B. Nonlinear Monotone Operators.* Springer Verlag, New York, 1990.