Samuel Kvasnicka, BSc

Berechnung des periodisch eingeschwungenen Zustands von nichtlinearen Netzwerken unter Anwendung der Partial Element Equivalent Circuit Methode

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur

Masterstudium Elektrotechnik

eingereicht an der

Technischen Universität Graz

Betreuer

Assoc. Prof. Dr. Thomas Bauernfeind, Dipl.-Ing. Paul Baumgartner

Institut für Grundlagen und Theorie der Elektrotechnik

Graz, September 2020

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

Datum	Unterschrift

Abstract

The object of this master's thesis was to provide numerical methods to compute the periodic steady state solution of nonlinear electrical networks which are supplied with periodic sources.
The modified nodal analysis (MNA) is used for electrical network analysis and it forms the basis for the application of the time domain methods *transient analysis* and *single shooting* as well as the frequency domain method *harmonic balance* to compute the periodic steady state solution.
This master's thesis provides the theory and implementation in MATLAB of the above mentioned methods. The considered methods are verified by the computation of the periodic steady state solution of a simple half and full wave rectifier.
In many applications the quasi-static partial element equivalent circuit (PEEC) method for conductors is an useful method to model an electromagnetic field problem in terms of an equivalent circuit. The application of MNA to analyse such an equivalent circuit of an electromagnetic field problem in combination with an external lumped circuit, that contains linear elements and nonlinear resistive elements, forms the basis of further investigations. On that basis the numerical methods *transient analysis*, *single shooting* and *harmonic balance* can be applied to compute the periodic steady state solution. A near field communication (NFC) system consisting of two inductive air coupled coils in combination with an external lumped circuit containing nonlinear resistive elements served as a test problem. The above mentioned numerical methods are applied and investigated to this test problem to compute the periodic steady state solution.

Kurzfassung

Für elektrische Netzwerke mit nichtlinearen Elementen und periodischer Anregung, werden ausgewählte numerische Methoden vorgestellt, welche die Berechnung des periodisch eingeschwungenen Zustands erlauben. Der Schwerpunkt liegt hierbei im Großsignalverhalten, sowie auf elektrische Netzwerke, welche durch periodische Quellen angeregt werden und durch nicht autonome Differentialgleichungen beschrieben werden können.
Die Analyse eines elektrischen Netzwerkes erfolgt durch das modifizierte Knotenspannungsverfahren, welches die Basis für die Anwendung der Zeitbereichsmethoden *transiente Analyse* und *einfaches Shooting*, sowie der Frequenzbereichsmethode *Harmonic Balance* bildet, um den periodisch eingeschwungenen Zustand zu berechnen.
In dieser Masterarbeit wurde die Theorie der drei besprochenen Methoden erarbeitet und in MATLAB implementiert. Die Übereinstimmung der Berechnungen der betrachteten Verfahren, zur Ermittlung des periodisch eingeschwungenen Zustands, wurden anhand einer einfachen Einweg- und Brückengleichrichterschaltung verifiziert.
In vielen Anwendungen ist es möglich mittels der quasi-stationären Partial Element Equivalent Circuit (PEEC) Methode, die leitfähige Struktur eines elektromagnetischen Feldproblems zu modellieren und als äquivalentes elektrisches Netzwerk darzustellen.
Die Anwendung des modifizierten Knotenspannungsverfahrens, zur Analyse eines solchen äquivalenten elektrischen Netzwerkes, in Kombination mit einem externen elektrischen Netzwerk, welches

lineare Elemente, sowie nichtlineare resistive Elemente enthalten kann, bildet die Grundlage für weiterführende Untersuchungen. Auf dieser Grundlage können die numerischen Methoden *transiente Analyse*, *einfaches Shooting* und *Harmonic Balance* angewendet werden, um den periodisch eingeschwungenen Zustand zu berechnen. Als Testproblem dient ein Near Field Communication (NFC) System, bestehend aus zwei luftgekoppelten Spulen, sowie einer externen Beschaltung mit nichtlinearen resistiven Elementen. Für die Ermittlung des periodisch eingeschwungenen Zustands des Testproblems, werden die besprochenen numerischen Methoden angewendet und untersucht.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

DAG Differentiell-algebraische Gleichung(en)

MKV Modifiziertes Knotenspannungsverfahren

NTA Node Tableau Analysis

ANA Augmented Nodal Analysis

BDF Backward Differentiation Formula

HB Harmonic Balance

DFT Diskrete Fourier Transformation

IDFT Inverse Diskrete Fourier Transformation

FFT Fast Fourier Transformation

PEEC Partial Element Equivalent Circuit

NFC Near Field Communication

Bezeichnungen

$\in$	: Elementrelation
$\subseteq$	: Teilmengenrelation
$\setminus$	: Mengendifferenz
$\to$, $\mapsto$	: Zuordnungssymbole bei einer Funktionsdefinition, z.B. $f : D \to Z : x \mapsto f(x)$ bedeutet, dass D die Definitionsmenge und Z die Zielmenge von f sind und f ordnet jedem $x \in D$, eindeutig das Element $f(x)$ zu
$\emptyset$	: leere Menge
$\mathbb{N}$, $\mathbb{R}$	: Menge der natürlichen und der reellen Zahlen
$\mathbb{R}^m$	: Reeller m- dimensionaler euklidischer Vektorraum
$\mathbb{R}^{m\times n}$	: Reeller Vektorraum der $m \times n$ Matrizen
$\mathbf{I}_{m\times m}$	: Einheits- oder Identitätsmatrix $\mathbf{I}_{m\times m} \in \mathbb{R}^{m\times m}$
$\mathbf{0}_{m\times n}$	: Nullmatrix $\mathbf{0}_{m\times n} \in \mathbb{R}^{m\times n}$
$\mathbf{0}_m$	: Nullvektor $\mathbf{0}_m \in \mathbb{R}^m$
$\mathbf{J_f}(\mathbf{a})$	: Jacobi-Matrix von $\mathbf{f}$ im Punkt $\mathbf{a}$. Siehe S. 3, Def. 1.7
$\mathrm{diag}\big(a_1, ..., a_n\big)$	: $n \times n$ Diagonalmatrix mit den Elementen $a_1, \ldots, a_n$ auf der Hauptdiagonale
$\mathbf{J}_{\mathbf{x};\mathbf{f}}(\mathbf{a}, \mathbf{b})$	: Partielle Jacobi-Matrix von $\mathbf{f}$ bzgl. $\mathbf{x}$, im Punkt $(\mathbf{a}, \mathbf{b})$. Siehe S. 4, Def. 1.10
$\mathrm{rang}(\mathbf{A})$	: Rang der Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, d.h. die Anzahl der linear unabhängigen Spalten- oder Zeilenvektoren von $\mathbf{A}$
$\ker(\mathbf{A})$	: Kern der Matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, d.h. $\ker(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n \,:\, \mathbf{A}\cdot\mathbf{x} = \mathbf{0}_m\}$
$C^1(D, \mathbb{R}^m)$	: Menge der stetig differenzierbaren Funktionen $\mathbf{f} : D \to \mathbb{R}^m$. Siehe S. 3
$P(T, \mathbb{R})$	: Menge der T- periodischen Funktionen, welche auf $[0, T]$ quadratisch integrierbar sind. Siehe S. 46
$[P(T, \mathbb{R})]^m$	: m- Tupel von $P(T, \mathbb{R})$, d.h. für $\mathbf{x} = (x_1, \ldots, x_m) \in [P(T, \mathbb{R})]^m$ gilt $x_i \in P(T, \mathbb{R})$
$\widehat{P}_K(T, \mathbb{R})$	: Approximation mit K Harmonischen für $x \in P(T, \mathbb{R})$. Siehe S. 47
$[\widehat{P}_K(T, \mathbb{R})]^m$	: m- Tupel von $\widehat{P}_K(T, \mathbb{R})$, d.h. für $\mathbf{x} \in [\widehat{P}_K(T, \mathbb{R})]^m$ gilt $x_i \in \widehat{P}_K(T, \mathbb{R})$

Einleitung

In vielen Anwendungsbereichen werden elektrische Netzwerke durch periodische Quellen angeregt. Beispiele hierfür sind Gleichrichterschaltungen, Switched-Capacitor Anwendungen oder schmalbandige Verstärker und Filter (siehe z.B. [16, Chapter 2]). Bei der Analyse der Funktionsweise sind vor allem Kennwerte des Langzeitverhaltens bzw. des periodisch eingeschwungenen Zustands von Interesse. Beispielsweise interessiert man sich für die effektive Strom- oder Leistungsaufnahme oder die harmonische Verzerrung (engl. total harmonic distortion). Besteht das elektrische Netzwerk nur aus linearen Elementen, so ist eine effiziente Berechnung des eingeschwungenen Zustands mit Phasoren möglich. Sind hingegen nichtlineare Elemente vorhanden, so werden zur Bestimmung des Großsignalverhaltens, im Allgemeinen numerische Verfahren zur Ermittlung des eingeschwungenen Zustands benötigt.

Mit dem Ziel, elektrische Netzwerke für einen großen Anwendungsbereich mathematisch beschreiben zu können, erfolgt die Modellierung mit dem modifizierten Knotenspannungsverfahren (MKV). Aufbauend auf [17] und [21] wird das MKV hergeleitet, wobei zeitvariante und nichtlineare resistive, kapazitive und induktive Elemente, sowie gesteuerte und ungesteuerte Quellen berücksichtigt werden. Die Modellierung mit dem MKV ist einfach und automatisierbar, dabei geht man allerdings den Kompromiss ein, dass die mathematische Beschreibung mit differentiell-algebraischen Gleichungen (DAG) erfolgt. Die numerische Lösung von DAG ist dabei im Vergleich zu gewöhnlichen Differentialgleichungssystemen anspruchsvoller. Zur Charakterisierung und Auswahl geeigneter numerischer Verfahren, erweist sich allerdings das Konzept des DAG Index als nützlich.
Um bei elektrischen Netzwerken mit nichtlinearen Elementen, numerisch den eingeschwungenen Zustand zu ermitteln, eignen sich z.B. die Zeitbereichsmethoden *transiente Analyse* und *Shooting*, die Frequenzbereichsmethoden *Harmonic Balance* (HB) und *Volterra Reihe*, sowie das *mixed frequency-time* (MFT) Verfahren (siehe beispielsweise [16] oder [15]).
Die transiente Analyse beruht auf der Lösung eines Anfangswertproblems. Ist das Anfangswertproblem wohlgestellt, dann liegt nach abgeklungenem transienten Vorgang, der periodische eingeschwungene Zustand vor. Hierfür kann die erforderliche Simulationszeit allerdings groß werden, wenn die Periodendauer im Vergleich zu Umladevorgängen sehr klein ist. Mit den Grundlagen der transienten Analyse kann das Shooting Verfahren beschrieben werden, welches unmittelbar den periodisch eingeschwungenen Zustand berechnet. Eine Alternative zu den Zeitbereichsmethoden ist das Harmonic Balance (HB) Verfahren. Dieses Verfahren ermittelt für den periodisch eingeschwungenen Zustand eine approximierte Fourierreihe mit vorgegebener Anzahl an Harmonischen, wobei die Fourierkoeffizienten effizient durch die DFT berechnet werden können. Im Shooting und HB Verfahren wird dabei die Suche nach dem periodisch eingeschwungenen Zustand, auf die Lösung eines Randwertproblems zurückgeführt.
In vielen Anwendungen kann ein elektromagnetisches Feldproblem, unter Anwendung der quasistationären Partial Element Equivalent Circuit Methode, durch ein äquivalentes elektrisches Netzwerk, bestehend aus resistiven, sowie gekoppelten kapazitiven und induktiven Elementen, beschrieben bzw. modelliert werden. Damit ist es möglich, ein externes elektrisches Netzwerk mit dem elektromagnetischen Feldproblem direkt zu koppeln. Für die Berechnung des periodisch eingeschwungenen Zustands für dieses resultierende elektrische Netzwerk, können numerische Methoden verwendet werden, welche für die Analyse elektrischer Netzwerke geeignet sind.

Kapitel 2 enthält die Herleitung eines allgemeinen modifizierten Knotenspannungsverfahrens, welches nichtlineare und zeitvariante Elemente, sowie gesteuerte und unabhängige Quellen berücksichtigt. Ausgehend von diesen Grundlagen, wird in Kapitel 3 ein Modellproblem definiert, welches lineare resistive, induktive und kapazitive Elemente, sowie nichtlineare resistive Elemente und unabhängige periodische Quellen berücksichtigt. Auf dieses Modellproblem werden zur Ermittlung des eingeschwungenen Zustands, die transiente Analyse, das einfache Shooting Verfahren, sowie das HB Verfahren eingeführt und numerische Lösungsstrategien besprochen. Zur Verifikation der einzelnen numerischen Methoden für das Modellproblem, wird eine einfache Einweg- und Brückengleichrichterschaltung mit ohmsch-kapazitiver Last berechnet.
In Kapitel 4 wird beschrieben, unter welchen Voraussetzungen ein elektromagnetisches Feldproblem, unter Anwendung der PEEC Methode, mit einem externen elektrischen Netzwerk gekoppelt werden kann, sodass eine Beschreibung der Anwendung mit dem Modellproblem aus Kapitel 3 möglich wird. Für solche Anwendungen ist es möglich die numerischen Methoden aus Kapitel 3, zur Berechnung des periodisch eingeschwungenen Zustands anzuwenden.
Die ausgewählten numerischen Methoden haben gemeinsam, dass für elektrische Netzwerke mit nichtlinearen Elementen, nichtlineare Gleichungssysteme gelöst werden müssen, wobei hierfür stets das Newton Verfahren verwendet wird. Die mathematischen Grundlagen in Kapitel 1 dienen als Ergänzung, um einerseits Notationen anzugeben und andererseits um ergänzende mathematische Grundlagen zur Verfügung zu stellen, welche dem interessierten Leser helfen sollen, mathematische Zusammenhänge besser verstehen zu können.

1 Mathematische Grundlagen

Dieser Abschnitt dient als mathematische Ergänzung und kann ggf. übersprungen werden.

1.1 Graphentheorie

Dieser Abschnitt enthält einige Definitionen und Aussagen aus der Graphentheorie, welche zur Beschreibung von elektrischen Netzwerken benötigt werden.

Definition 1.1 (Gerichteter (Multi-) Graph, siehe [17, Abschnitt 5.1.1.1])**:**
Seien $V \neq \emptyset$, E endliche Mengen. Die Elemente von V werden als Knoten (engl.: vertices) bezeichnet und die Elemente von E werden als Kanten (engl.: edges) bezeichnet.
Durch eine Funktion $\alpha : E \to V \times V$ sei für jede Kante $e \in E$ festgelegt, in welcher Reihenfolge zwei Knoten durch die Kante e verbunden werden. D.h. ist für $e \in E$

$$\alpha(e) = (\alpha_1(e), \alpha_2(e)) = (v_1, v_2) \in V \times V$$

gegeben, dann ist $v_1 = \alpha_1(e)$ der Startknoten und $v_2 = \alpha_2(e)$ der Endknoten von e. Man sagt dazu auch, dass die Kante e inzident mit den Knoten v_1 und v_2 ist. Ein gerichteter (Multi-) Graph wird dann durch das Tripel (V, E, α) definiert.

Der Begriff Multigraph bezieht sich darauf, dass zwei Knoten durch verschiedene Kanten verbunden werden können. Bei elektrischen Netzwerken entspricht dies einer Parallelschaltung. Im Folgenden wird ein gerichteter Multigraph aber stets als gerichteter Graph bezeichnet. Die Menge der Knoten V entspricht in den elektrischen Netzwerken den Knotenpunkten und die Kantenmenge E entspricht den Zweigen (engl.: branches).

Definition 1.2 (Pfad, geschlossener Pfad, Schleife, siehe [17, Abschnitt 5.1.1.2])**:**
Sei (V, E, α) ein gerichteter Graph. Für $\ell \in \mathbb{N}$ seien $v_0, v_\ell \in V$.

(a) Ein Pfad ist eine endliche Folge $(v_0, e_1, v_1, \ldots, v_{\ell-1}, e_\ell, v_\ell)$, wobei für alle $1 \leq i \leq \ell$ die Kante e_i mit den Knoten v_{i-1} und v_i inzident ist, d.h. es gilt $\alpha(e_i) = (v_{i-1}, v_i)$ oder $\alpha(e_i) = (v_i, v_{i-1})$.

(b) Ein Pfad heißt geschlossen, wenn $v_0 = v_\ell$ gilt.

(c) Eine Schleife (engl.: loop) ist ein geschlossener Pfad mit $\ell \geq 2$, wobei $e_i \neq e_j$ und $v_i \neq v_j$ für alle $1 \leq i < j \leq \ell$ gelte ($\ell \geq 2$, da im Folgenden Eigenschleifen mit $\ell = 1$ ausgeschlossen werden). Im Folgenden genügt es eine Schleife als die Menge der enthaltenen Kanten $\{e_1, \ldots, e_\ell\}$ zu betrachten.

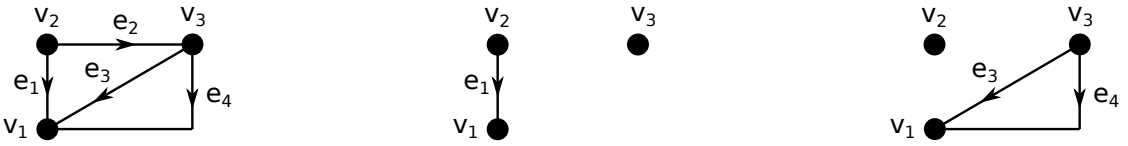
Es sei darauf hingewiesen, dass in Definition 1.2,(a) die gerichtete Kante als ungerichtete Kante betrachtet wird. In Definition 1.3 stimmen die definierten Begriffe somit mit jenen in einem ungerichteten Graphen überein.

Definition 1.3 (Zusammenhängender Graph, Schnittmenge, siehe [17, Abschnitt 5.1.1.2])**:**
Sei (V, E, α) ein gerichteter Graph.

(a) Der gerichtete Graph (V, E, α) heißt zusammenhängend, wenn für alle Knoten $v, w \in V$ mit $v \neq w$ ein Pfad existiert, welcher die Knoten v und w verbindet.

(b) Sei (V, E, α) ein zusammenhängender gerichteter Graph. Eine Teilmenge $K \subseteq E$ der Kantenmenge heißt Schnittmenge (engl.: cutset), wenn der gerichtete Graph $(V, E \setminus K, \alpha)$ nicht mehr zusammenhängend ist und wenn K in dieser Eigenschaft minimal ist. Dies bedeutet, dass für alle echt kleineren Teilmengen $K' \subset K$, der gerichtete Graph $(V, E \setminus K', \alpha)$ zusammenhängend ist.

Abbildung 1.1 zeigt Beispiele für Schnittmengen des gerichteten vollständigen Graphen gemäß Abbildung 1.1a, mit $V = \{v_1, v_2, v_3\}$ und $E = \{e_1, e_2, e_3, e_4\}$. Gemäß Abbildung 1.1b und 1.1c werden die Schnittmengen mit K' und K'' bezeichnet.



(a) Vollständiger Graph (b) $K' = \{e_2, e_3, e_4\}, E \setminus K' = \{e_1\}$ (c) $K'' = \{e_1, e_2\}, E \setminus K'' = \{e_3, e_4\}$

Abbildung 1.1: Beispiele von Schnittmengen eines vollständigen Graphen.

Definition 1.4 (Vollständige und (reduzierte) Inzidenzmatrix, siehe [17, Abschnitt 5.1.1.4])**:**
Seien $n, b \in \mathbb{N}$ mit $b \geq 1$ und $n \geq 2$. Seien $V = \{v_1, \ldots, v_n\}$ eine n-elementige Knotenmenge und $E = \{e_1, \ldots, e_b\}$ eine b-elementige Kantenmenge. Sei (V, E, α) ein gerichteter Graph, welcher keine Eigenschleifen, also Schleifen der Länge 1, enthält.

(a) Die vollständige Inzidenzmatrix (auch Knoten-Kanten-Matrix) $\widetilde{\mathbf{A}} \in \mathbb{R}^{n \times b}$ des gerichteten Graphen (V, E, α), ist eine Matrix welche die Beziehungen (d.h. die Inzidenz) der Knoten und Kanten des gerichteten Graphen speichert. Die Elemente $\left(\widetilde{a}_{i,j}\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,b}}$ der vollständigen Inzidenzmatrix sind folgendermaßen definiert (mit $\alpha = (\alpha_1, \alpha_2)$)

$$\widetilde{a}_{i,j} = \begin{cases} 1 & \text{, wenn } v_i \text{ der Startknoten der Kante } e_j \text{ ist, d.h. } \alpha_1(e_j) = v_i \\ -1 & \text{, wenn } v_i \text{ der Endknoten der Kante } e_j \text{ ist, d.h. } \alpha_2(e_j) = v_i \\ 0 & \text{, wenn Kante } e_j \text{ nicht mit Knoten } v_i \text{ inzident ist, d.h. } \alpha_1(e_j) \neq v_i,\ \alpha_2(e_j) \neq v_i \end{cases}$$

(b) Sei (V, E, α) ein zusammenhängender gerichteter Graph. Wählt man aus der vollständigen Inzidenzmatrix $\widetilde{\mathbf{A}} \in \mathbb{R}^{n \times b}$ eine beliebige Zeile i mit $1 \leq i \leq n$ aus und streicht die Elemente dieser Zeile aus der vollständigen Inzidenzmatrix $\widetilde{\mathbf{A}}$, dann wird die daraus erhaltende Matrix als (reduzierte) Inzidenzmatrix $\mathbf{A} \in \mathbb{R}^{(n-1) \times b}$ bezeichnet.

(c) Sei (V, E, α) ein zusammenhängender gerichteter Graph und sei $\mathbf{A} \in \mathbb{R}^{(n-1) \times b}$ eine beliebige (reduzierte) Inzidenzmatrix. Für $k \in \mathbb{N}$ und $1 \leq k < b$ sei $K = \{e_{j_1}, \ldots, e_{j_k}\} \subseteq E$ eine beliebige k-elementige Teilmenge von E, wobei $1 \leq j_1, \ldots, j_k \leq b$ Indizes sind.
Dann bezeichnet $\mathbf{A}_K \in \mathbb{R}^{(n-1) \times k}$ die Matrix, welche die Spalten $j_1, \ldots, j_k$ von $\mathbf{A}$ enthält.
Und $\mathbf{A}_{E \setminus K} \in \mathbb{R}^{(n-1) \times (b-k)}$ bezeichnet die Matrix, welche durch $\mathbf{A}$ entsteht, wenn die Spalten $j_1, \ldots, j_k$ aus $\mathbf{A}$ entfernt werden.

Das nachfolgende Lemma liefert eine wichtige Aussage über die vollständige und (reduzierte) Inzidenzmatrix eines zusammenhängenden gerichteten Graphen. Im Folgenden wird mit Inzidenzmatrix stets die reduzierte Inzidenzmatrix bezeichnet.

Lemma 1.5 (siehe [17, Abschnitt 5.1.1.4])**:**
Seien $n, b \in \mathbb{N}$ mit $b \geq 1$ und $n \geq 2$. Seien V eine n-elementige Knotenmenge und E eine b-elementige Kantenmenge. Sei (V, E, α) ein zusammenhängender gerichteter Graph, welcher keine Eigenschleifen besitzt. Gemäß Definition 1.4 sei $\widetilde{\mathbf{A}} \in \mathbb{R}^{n \times b}$ die vollständige Inzidenzmatrix und

für ein beliebiges $1 \leq i \leq n$ und der Entfernung der i-ten Zeile aus $\widetilde{\mathbf{A}}$, sei $\mathbf{A} \in \mathbb{R}^{(n-1)\times b}$ die (reduzierte) Inzidenzmatrix. Dann gelten folgende Eigenschaften.

(a) $\operatorname{rang}(\widetilde{\mathbf{A}}) = n-1$. *D.h. $\widetilde{\mathbf{A}}$ besitzt $(n-1)$ linear unabhängige Zeilen und Spalten.*

(b) $\operatorname{rang}(\mathbf{A}) = n-1$. *D.h. unabhängig von der entfernten Zeile aus $\widetilde{\mathbf{A}}$, stimmt der Rang von $\mathbf{A}$ und $\widetilde{\mathbf{A}}$ überein.*

(c) Sei $K \subseteq E$ eine beliebige k-elementige Teilmenge von E mit $1 \leq k < b$ und sei $\mathbf{A}_K \in \mathbb{R}^{(n-1)\times k}$ und $\mathbf{A}_{E\setminus K} \in \mathbb{R}^{(n-1)\times(b-k)}$ gemäß Definition 1.4,(c). Dann gilt

(i) Die Kantenmenge K beinhaltet keine Schleifen genau dann, wenn $\mathbf{A}_K$ vollen Spaltenrang besitzt (d.h. $\operatorname{rang}(\mathbf{A}_K) = k$).

(ii) Die Kantenmenge K beinhaltet keine Schnittmenge genau dann, wenn $\mathbf{A}_{E\setminus K}$ vollen Zeilenrang besitzt (d.h. $\operatorname{rang}(\mathbf{A}_{E\setminus K}) = n-1$).

1.2 Differentialrechnung mehrerer Variabler

Notation 1.6 (Vektoren und Vektoren als Funktionsargumente)**:**
Seien $n, m, p \in \mathbb{N}$. In diesem Dokument sei ein Vektor $\mathbf{x} \in \mathbb{R}^n$ stets ein Spaltenvektor und $\mathbf{x}^\top$ ein Zeilenvektor. Häufig werden Vektoren in Funktionsargumenten benötigt und hierbei wird zur einfacheren Lesbarkeit die Konvention getroffen, dass Vektoren als Funktionsargumente stets Zeilenvektoren sind. Beispielsweise für eine Funktion $\mathbf{f} : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^p : (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y})$ sei das Funktionsargument $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{n+m}$ ein Zeilenvektor (wären stattdessen das Funktionsargument und die Vektoren weiterhin Spaltenvektoren, dann müsste man $(\mathbf{x}^\top, \mathbf{y}^\top)^\top$ schreiben).

Definition 1.7 (Jacobi-Matrix, siehe [27, Abschnitt 3.5])**:**
Seien $m, n \in \mathbb{N}$. Es sei $\emptyset \neq D \subseteq \mathbb{R}^n$ offen und $\mathbf{f} : D \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ differenzierbar im Punkt $\mathbf{a} \in D$. Dann heißt die Matrix

$$\mathbf{J}_\mathbf{f}(\mathbf{a}) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_1}{\partial x_n}(\mathbf{a}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a}) & \dots & \frac{\partial f_m}{\partial x_n}(\mathbf{a}) \end{pmatrix} \in \mathbb{R}^{m\times n} ,$$

die Jacobi-Matrix (oder Funktionalmatrix) von $\mathbf{f}$ im Punkt $\mathbf{a}$.
Wenn $m = 1$ ist, dann gilt zudem $\mathbf{J}_\mathbf{f}(\mathbf{a}) = \operatorname{grad} \mathbf{f}(\mathbf{a})$ (d.h. Gradient als Zeilenvektor).

Satz 1.8 (Stetige Differenzierbarkeit, C^1- Funktion, siehe [27, Abschnitt 3.8, 3.9])**:**
Seien $m, n \in \mathbb{N}$ und sei $\emptyset \neq D \subseteq \mathbb{R}^n$ offen, sowie $\mathbf{f} : D \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$. Wenn für alle $\mathbf{x} \in D$ und für alle $1 \leq i \leq m$ und $1 \leq j \leq n$ die partiellen Ableitungen von $f_i : D \to \mathbb{R}$ nach x_j existieren und wenn die partiellen Ableitungen $\frac{\partial f_i}{\partial x_j} : D \to \mathbb{R}$ stetig sind, so ist $\mathbf{f}$ eine stetig differenzierbare Funktion, d.h. $\mathbf{f} \in C^1(D, \mathbb{R}^m)$. Insbesondere ist $\mathbf{f}$ dann differenzierbar auf D.

Satz 1.9 (Kettenregel, siehe [27, Abschnitt 3.10])**:**
Seien $m, n, p \in \mathbb{N}$. Es sei $\emptyset \neq U \subseteq \mathbb{R}^n$ und $\emptyset \neq V \subseteq \mathbb{R}^m$ offen. Seien $\mathbf{f} : U \to V : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ differenzierbar im Punkt $\mathbf{a} \in U$ und $\mathbf{g} : V \to \mathbb{R}^p : \mathbf{y} \mapsto \mathbf{g}(\mathbf{y})$ differenzierbar im Punkt $\mathbf{b} := \mathbf{f}(\mathbf{a}) \in V$. Dann ist die zusammengesetzte Funktion $\mathbf{h} : U \to \mathbb{R}^p : \mathbf{x} \mapsto \mathbf{h}(\mathbf{x}) = \mathbf{g}(\mathbf{f}(\mathbf{x}))$ differenzierbar in $\mathbf{a} \in U$ und es gilt mit $\mathbf{J}_\mathbf{f}(\mathbf{a}) \in \mathbb{R}^{m\times n}$ und $\mathbf{J}_\mathbf{g}(\mathbf{b}) \in \mathbb{R}^{p\times m}$

$$\mathbf{J}_\mathbf{h}(\mathbf{a}) = \mathbf{J}_\mathbf{g}(\mathbf{f}(\mathbf{a})) \cdot \mathbf{J}_\mathbf{f}(\mathbf{a}) \in \mathbb{R}^{p\times n} .$$

Definition 1.10 (Partielle Jacobi-Matrix, siehe [27, Abschnitt 4.4])**:**
Seien $k, m, n \in \mathbb{N}$. Es sei $\emptyset \neq D \subseteq \mathbb{R}^k \times \mathbb{R}^n = \mathbb{R}^{k+n}$ offen und $\mathbf{f} : D \to \mathbb{R}^m : (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y})$ differenzierbar im Punkt $(\mathbf{a}, \mathbf{b}) \in D$, mit $\mathbf{a}, \mathbf{x} \in \mathbb{R}^k$ und $\mathbf{b}, \mathbf{y} \in \mathbb{R}^n$. Dann heißen die Matrizen

$$\mathbf{J}_{\mathbf{x};\mathbf{f}}(\mathbf{a},\mathbf{b}) := \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{a},\mathbf{b}) & \dots & \frac{\partial f_1}{\partial x_k}(\mathbf{a},\mathbf{b}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{a},\mathbf{b}) & \dots & \frac{\partial f_m}{\partial x_k}(\mathbf{a},\mathbf{b}) \end{pmatrix}, \quad \mathbf{J}_{\mathbf{y};\mathbf{f}}(\mathbf{a},\mathbf{b}) := \begin{pmatrix} \frac{\partial f_1}{\partial y_1}(\mathbf{a},\mathbf{b}) & \dots & \frac{\partial f_1}{\partial y_n}(\mathbf{a},\mathbf{b}) \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial y_1}(\mathbf{a},\mathbf{b}) & \dots & \frac{\partial f_m}{\partial y_n}(\mathbf{a},\mathbf{b}) \end{pmatrix},$$

die partiellen Jacobi-Matrizen von $\mathbf{f}$ im Punkt $(\mathbf{a}, \mathbf{b})$ bezüglich dieser Produktzerlegung, wobei $\mathbf{J}_{\mathbf{x};\mathbf{f}}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{m \times k}$ und $\mathbf{J}_{\mathbf{y};\mathbf{f}}(\mathbf{a}, \mathbf{b}) \in \mathbb{R}^{m \times n}$. Insbesondere gilt $\mathbf{J}_{\mathbf{f}}(\mathbf{a}, \mathbf{b}) = \big(\mathbf{J}_{\mathbf{x};\mathbf{f}}(\mathbf{a}, \mathbf{b}), \mathbf{J}_{\mathbf{y};\mathbf{f}}(\mathbf{a}, \mathbf{b})\big)$. Eine Darstellung für eine Produktzerlegung von mehr als zwei Vektoren erfolgt analog.

Die Notation der partiellen Jacobi-Matrix mag auf den ersten Blick etwas ungewohnt sein. Ein Vorteil dieser Notation ist aber, dass durch die Indizes die berücksichtigte Funktion und die Variablen ersichtlich sind, nach denen differenziert wird. Das **J** deutet zudem darauf hin, dass es sich im Allgemeinen um eine Matrix handelt. Anstatt der Notation der partiellen Jacobi-Matrix im Punkt $(\mathbf{a}, \mathbf{b})$ gemäß Definition 1.10, sind auch folgende Notationen in der Literatur zu finden

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{a},\mathbf{b}), \frac{\partial \mathbf{f}}{\partial \mathbf{y}}(\mathbf{a},\mathbf{b}), \quad \text{oder} \quad \mathbf{f}_{\mathbf{x}}(\mathbf{a},\mathbf{b}), \mathbf{f}_{\mathbf{y}}(\mathbf{a},\mathbf{b}).$$

Satz 1.11 (Hauptsatz über implizite Funktionen, siehe [27, Abschnitt 4.5])**:**
Seien $m, n \in \mathbb{N}$. Es sei $\emptyset \neq D \subseteq \mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{n+m}$ offen und $\mathbf{f} : D \to \mathbb{R}^m : (\mathbf{x}, \mathbf{y}) \mapsto \mathbf{f}(\mathbf{x}, \mathbf{y})$ eine stetig differenzierbare Funktion, d.h. $\mathbf{f} \in C^1(D, \mathbb{R}^m)$, mit $\mathbf{x} \in \mathbb{R}^n$ und $\mathbf{y} \in \mathbb{R}^m$. Für einen Punkt $(\mathbf{a}, \mathbf{b}) \in D$ gelte

$$\mathbf{f}(\mathbf{a},\mathbf{b}) = \mathbf{0} \quad \text{und} \quad \det\big(\mathbf{J}_{\mathbf{y};\mathbf{f}}(\mathbf{a},\mathbf{b})\big) \neq 0,$$

d.h. der Punkt $(\mathbf{a}, \mathbf{b})$ ist eine Nullstelle von $\mathbf{f}$ und $\mathbf{J}_{\mathbf{y};\mathbf{f}}$ ist in $(\mathbf{a}, \mathbf{b})$ invertierbar.
Dann existieren offene Mengen $U \subseteq \mathbb{R}^n$ mit $\mathbf{a} \in U$ und $V \subseteq \mathbb{R}^m$ mit $\mathbf{b} \in V$ und eine stetig differenzierbare Funktion $\mathbf{g} : U \to V : \mathbf{x} \mapsto \mathbf{g}(\mathbf{x})$, sodass für alle $(\mathbf{x}, \mathbf{y}) \in U \times V$ Folgendes gilt

$$\mathbf{f}(\mathbf{x},\mathbf{y}) = \mathbf{0} \iff \mathbf{y} = \mathbf{g}(\mathbf{x}).$$

Insbesondere gilt somit für alle $\mathbf{x} \in U$, dass $\mathbf{f}(\mathbf{x}, \mathbf{g}(\mathbf{x})) = \mathbf{0}$.

1.3 Newton Verfahren

In vielen Anwendungen ist es erforderlich eine Lösung eines nichtlinearen Gleichungssystems zu ermitteln. Ziel dieses Abschnitts ist es, die Idee des lokalen Newton Verfahrens zu vermitteln und auch Algorithmen für eine globalisierte Version vorzustellen, sofern die Jacobi-Matrix der zu untersuchenden Funktion bekannt ist. Es gibt auch viele Anwendungen bei denen die Angabe der Jacobi-Matrix nicht möglich ist oder nur mit großem Aufwand berechnet werden kann. Hierfür stellen z.B. Newton-artige oder Quasi- Newton Verfahren, Alternativen zum klassischen Newton Verfahren dar. Darauf wird im Folgenden aber nicht weiter eingegangen und für weitere Informationen sei z.B. auf die angegebene Literatur verwiesen.

Lokales Newton Verfahren

Für die nachfolgenden Überlegungen sei eine offene Menge $D \subseteq \mathbb{R}^n$, sowie eine stetig differenzierbare Funktion $\mathbf{F} : D \to \mathbb{R}^n : \mathbf{x} \mapsto \mathbf{F}(\mathbf{x})$ vorausgesetzt. Gesucht wird eine Nullstelle $\mathbf{x}^* \in D$, des im Allgemeinen nichtlinearen Gleichungssystems

$$\mathbf{F}(\mathbf{x}^*) = \mathbf{0}. \tag{1.1}$$

Die Idee ist nun, ausgehend von einer Näherungslösung $\mathbf{x}^{(k)} \in D$, die Funktion $\mathbf{F}$, komponentenweise mit Hilfe der mehrdimensionalen Taylorreihe, im Punkt $\mathbf{x}^{(k)}$ zu approximieren. Unter Verwendung der Jacobi-Matrix $\mathbf{J_F}(\mathbf{x}^{(k)})$ (siehe Definition 1.7) wird die folgende affin lineare Approximation verwendet

$$\widetilde{\mathbf{F}} : \mathbb{R}^n \to \mathbb{R}^n : \mathbf{x} \mapsto \widetilde{\mathbf{F}}(\mathbf{x}) = \mathbf{F}(\mathbf{x}^{(k)}) + \mathbf{J_F}(\mathbf{x}^{(k)}) \cdot (\mathbf{x} - \mathbf{x}^{(k)}) \, .$$

Wird vorausgesetzt, dass die Jacobi-Matrix $\mathbf{J_F}(\mathbf{x}^{(k)})$ invertierbar ist, dann kann die Gleichung $\widetilde{\mathbf{F}}(\mathbf{d}^{(k)} + \mathbf{x}^{(k)}) = \mathbf{0}$ eindeutig gelöst werden. Im lokalen Newton Verfahren berechnet sich dann die neue Iterierte durch $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \mathbf{d}^{(k)}$ und $\mathbf{x}^{(k+1)}$ kann als eine neue Näherung einer Nullstelle $\mathbf{x}^*$ in (1.1) interpretiert werden. $\mathbf{d}^{(k)} \in \mathbb{R}^n$ wird auch als Newton-Schritt bezeichnet und zur Berechnung von $\mathbf{d}^{(k)}$ ist eine der beiden äquivalenten Gleichungen zu lösen

$$\mathbf{J_F}(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)}) \, , \tag{1.2a}$$

$$\mathbf{d}^{(k)} = -\left(\mathbf{J_F}(\mathbf{x}^{(k)})\right)^{-1} \cdot \mathbf{F}(\mathbf{x}^{(k)}) \, . \tag{1.2b}$$

Gleichung (1.2a) wird auch als Newton-Gleichung bezeichnet. Beginnend von einem Startwert $\mathbf{x}^{(0)}$ kann diese Vorgehensweise solange wiederholt werden bis eine Iterierte $\mathbf{F}(\mathbf{x}^{(k)}) = \mathbf{0}$ erfüllt, oder wenn sichergestellt ist, dass die Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ gegen eine Nullstelle $\mathbf{x}^*$ konvergiert und man 'nahe genug' bei der Nullstelle das Verfahren abbricht.
Das lokale Newton Verfahren kann in dieser Form einfach implementiert werden. In der Praxis kann dieses Verfahren aber auch häufig fehlschlagen, da eine Konvergenz unter bestimmten Voraussetzungen, vor allem nur in einer kleinen lokalen Umgebung um eine Nullstelle $\mathbf{x}^*$, gewährleistet werden kann. Eine wichtige Voraussetzung für die Konvergenz des Verfahrens ist, dass für die Nullstelle $\mathbf{F}(\mathbf{x}^*) = \mathbf{0}$ die Jacobi-Matrix $\mathbf{J_F}(\mathbf{x}^*)$ invertierbar ist. Dadurch wird auch sichergestellt, dass $\mathbf{J_F}(\mathbf{x})$ in einer Umgebung um $\mathbf{x}^*$ invertierbar ist und dass es in einer kleinen Umgebung um $\mathbf{x}^*$ keine weitere Nullstelle gibt. Ist $\mathbf{J_F}(\mathbf{x}^*)$ singulär, dann wird das Verhalten des Newton Verfahrens sehr kompliziert. Bei skalaren Gleichungen kann es zwar in bestimmten Fällen konvergieren, doch im mehrdimensionalen Fall tritt meist keine Konvergenz auf.
Die Konvergenz des lokalen Newton Verfahrens hängt auch stark von dem Startwert $\mathbf{x}^{(0)} \in D$ ab, da dieser in unmittelbarer 'Nähe' zur Nullstelle $\mathbf{x}^*$ liegen sollte. Dies begründet auch die Bezeichnung lokales Newton Verfahren. Für eine detaillierte Betrachtung des lokalen Newton Verfahrens, inkl. Konvergenzaussagen, wird z.B. auf [22, Abschnitt 5.1] oder [26, Abschnitt 10.1] verwiesen. Wichtig anzumerken ist, dass bei Erfüllung der nötigen Voraussetzungen, im lokalen Newton Verfahren q-quadratische Konvergenz auftritt, d.h. es existiert eine Konstante $C > 0$, sodass

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \leq C \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^*\|^2 \, , \qquad \text{für alle } k \geq 0$$

gilt. Um die Abhängigkeit durch einen geeigneten Startwert $\mathbf{x}^{(0)}$ abzuschwächen, können globalisierte Newton Verfahren verwendet werden. Damit ist unter bestimmten Voraussetzungen eine Konvergenz zur Nullstelle $\mathbf{x}^*$ gewährleistet, auch wenn $\mathbf{x}^{(0)}$ nicht 'nahe' bei $\mathbf{x}^*$ liegt. In [9, Kapitel 3] kann z.B. eine ausführliche Behandlung globaler Newton Verfahren nachgelesen werden.

Gedämpftes Newton Verfahren

Mit den Algorithmen 1 und 2 werden nun zwei globalisierte Newton Verfahren vorgestellt.
Um die Konvergenz von globalen Newton Verfahren gewährleisten zu können, muss die Funktion $\mathbf{F} : D \to \mathbb{R}^n$ und der Startwert $\mathbf{x}^{(0)} \in D$ einige Voraussetzungen erfüllen. Eine Konvergenzanalyse inkl. Voraussetzungen der vorgestellten Algorithmen, kann in [9] und [3] nachgelesen werden. Für die nachfolgenden Algorithmen kann vereinfachend aber angenommen werden, dass $\mathbf{F} : D \to \mathbb{R}^n$ auf der offenen Menge $D \subseteq \mathbb{R}^n$ stetig differenzierbar ist, die Jabobi-Matrix $\mathbf{J_F}(\mathbf{x})$ für alle $\mathbf{x} \in \mathbb{R}^n$

invertierbar ist und das $\mathbf{x}^* \in D$ die einzige Nullstelle in D ist.

Algorithmus 1 : Einfaches gedämpftes Newton Verfahren

Eingabe : $\mathbf{F} : D \to \mathbb{R}^n$, $\mathbf{J_F} : D \to \mathbb{R}^{n \times n}$, $\mathbf{x}^{(0)} \in D \subseteq \mathbb{R}^n$
Ausgabe : Bei Terminierung ist das aktuelle $\mathbf{x}^{(k)}$ eine Näherung für $\mathbf{x}^*$
Voraussetzung : Streng monoton fallende Funktion $t_{Damp} : \mathbb{N} \to (0, 1]$ mit $t_{Damp}(0) = 1$

$k \leftarrow 0$
while $\mathbf{F}(\mathbf{x}^{(k)}) \neq \mathbf{0}$ *ODER Abbruchkriterium nicht erfüllt* **do**
  /* Berechne Newton-Schritt $\mathbf{d}^{(k)}$ durch Lösen der Newton- Gleichung */
  $\mathbf{J_F}(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)})$
  $\ell \leftarrow 0$
  while $\|\mathbf{F}(\mathbf{x}^{(k)} + t_{Damp}(\ell) \cdot \mathbf{d}^{(k)})\| \geq \|\mathbf{F}(\mathbf{x}^{(k)})\|$ **do**
    $\ell \leftarrow \ell + 1$
  /* Berechnung des Dämpfungsfaktors t_k und der neuen Iterierten $\mathbf{x}^{(k+1)}$ */
  $t_k \leftarrow t_{Damp}(\ell)$; $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + t_k \cdot \mathbf{d}^{(k)}$
  $k \leftarrow k + 1$

Algorithmus 2 : Gedämpftes Newton Verfahren nach Bank-Rose [3, Abschnitt 3]

Eingabe : $\mathbf{F} : D \to \mathbb{R}^n$, $\mathbf{J_F} : D \to \mathbb{R}^{n \times n}$, $\mathbf{x}^{(0)} \in D \subseteq \mathbb{R}^n$, $\omega \in (0, 1)$
Ausgabe : Bei Terminierung ist das aktuelle $\mathbf{x}^{(k)}$ eine Näherung für $\mathbf{x}^*$

$K \leftarrow 0$, $k \leftarrow 0$
$\mathbf{F}^{(k)} \leftarrow \mathbf{F}(\mathbf{x}^{(k)})$; $F^{(k)}_{Norm} \leftarrow \|\mathbf{F}(\mathbf{x}^{(k)})\|$ // k == 0
while $F^{(k)}_{Norm} \neq 0$ *ODER Abbruchkriterium nicht erfüllt* **do**
  /* Berechne Newton-Schritt $\mathbf{d}^{(k)}$ durch Lösen der Newton- Gleichung */
  $\mathbf{J_F}(\mathbf{x}^{(k)}) \cdot \mathbf{d}^{(k)} = -\mathbf{F}(\mathbf{x}^{(k)})$
  /* Ermittlung des Dämpfungsfaktors t_k */
  $Status_{DF} \leftarrow 0$ // $Status_{DF} == 0$: t_k nicht akzeptiert; $Status_{DF} == 1$: t_k akzeptiert
  while $Status_{DF} == 0$ **do**
    $t_k \leftarrow \left(1 + K \cdot F^{(k)}_{Norm}\right)^{-1}$ // Dämpfungsfaktor t_k berechnen
    /* Berechnung der neuen Iterierten $\mathbf{x}^{(k+1)}$ und Auswertung der Funktion $\mathbf{F}$ */
    $\mathbf{x}^{(k+1)} \leftarrow \mathbf{x}^{(k)} + t_k \cdot \mathbf{d}^{(k)}$; $\mathbf{F}^{(k+1)} \leftarrow \mathbf{F}(\mathbf{x}^{(k+1)})$; $F^{(k+1)}_{Norm} \leftarrow \|\mathbf{F}(\mathbf{x}^{(k+1)})\|$
    if $\left(1 - \frac{F^{(k+1)}_{Norm}}{F^{(k)}_{Norm}}\right) \cdot t_k^{-1} < \omega$ **then**
      if $K == 0$ **then**
        $K \leftarrow 1$
      else
        $K \leftarrow 10 \cdot K$
    else
      $K \leftarrow \frac{K}{10}$; $k \leftarrow k + 1$
      $Status_{DF} \leftarrow 1$ // Dämpfungsfaktor t_k wird akzeptiert

Algorithmus 1 zeigt ein einfaches gedämpftes Newton Verfahren. Entlang der Richtung, welche durch den Newton-Schritt $\mathbf{d}^{(k)}$ vorgegeben wird, wird für jeden Schritt ein skalarer Dämpfungsfaktor t_k gesucht, bei dem es zu einer Abnahme des Residuums kommt. Zur Existenz eines solchen Dämpfungsfaktors und für einen Beweis der Konvergenz von Algorithmus 1 sei auf [9, Abschnitt 3.1.3 und 3.2] verwiesen. In den Algorithmen 1 und 2 ist zudem das Ziel, dass der Dämpfungsfaktor t_k gegen 1 konvergiert und sich somit in einer Umgebung um der Nullstelle $\mathbf{x}^*$ wie das lokale New-

ton Verfahren verhält und am Ende q-quadratisch konvergiert (Algorithmus 1 ohne *while* Schleife in Zeile 5 und 6 und mit $t_k = 1$ entspricht dem lokalen Newton Verfahren).
In Algorithmus 1 wird ein passender Dämpfungsfaktor t_k nach dem *Trial and Error* Prinzip ermittelt. Beispiele für $t_{Damp} : \mathbb{N} \to (0, 1]$ sind gemäß [6, Abschnitt 2.6.2], z.B.

$$t_{Damp}(\ell) = 2^{-\ell} , \quad \text{oder} \quad t_{Damp}(\ell) = 2^{-\frac{\ell\cdot(\ell+1)}{2}} , \quad \text{für } \ell = 0, 1, 2, \dots .$$

Algorithmus 2 wurde aus [3, Abschnitt 3] entnommen. Hierbei werden bei der Berechnung des Dämpfungsfaktors t_k auch Informationen aus dem Residuum berücksichtigt. Die notwendigen Voraussetzungen und Konvergenznachweise können in [3, Abschnitt 2] nachgelesen werden. Gemäß [3, Abschnitt 1] ist der Algorithmus auch für approximierte Newton Verfahren geeignet. Weshalb auch eine Konvergenz gewährleistet werden kann, wenn z.B. eine Approximation der Jacobi-Matrix $\mathbf{J_F}(\mathbf{x}^{(k)})$ verwendet wird.
Es folgt nun noch eine Diskussion über Abbruchkriterien und worauf bei der Berechnung der Newton-Gleichung (1.2a) zu achten ist.

Abbruchkriterien

In diesem Abschnitt soll eine kleine Übersicht über Abbruchkriterien im Newton Verfahren gegeben werden. Die Grundlagen hierzu wurden vorwiegend aus [10] entnommen.
Für die theoretische Analyse von Newton Verfahren erfolgt ein Abbruch der Iteration wenn für eine Iterierte $\mathbf{x}^{(k)} = \mathbf{x}^*$, bzw. $\mathbf{F}(\mathbf{x}^{(k)}) = \mathbf{0}$ gilt. In diesem Fall endet der Algorithmus nach endlich vielen Schritten. Ansonsten wird eine unendliche Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ erzeugt, welche im Erfolgsfall gegen $\mathbf{x}^*$ konvergiert. Bei der Implementierung ist ein Algorithmus aber immer nach endlich vielen Iterationen abzubrechen und hierbei muss entschieden werden, unter welchen Kriterien eine Iterierte $\mathbf{x}^{(k)}$ als ausreichende Näherung von $\mathbf{x}^*$ gilt, bzw. ob die Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ überhaupt konvergiert.
Es werden nun einige Kriterien vorgestellt, wobei jedes Vor- und Nachteile besitzt und daher kann auch empfohlen werden, verschiedene Kriterien zu kombinieren. Zusätzlich soll eine maximale Iterationszahl vorgegeben werden. Wird diese erreicht, könnten z.B. andere Abbruchkriterien zu streng sein oder die Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ besitzt keinen Grenzwert.
Im Folgenden seien Parameter $\mathrm{Tol}_{abs}, \mathrm{Tol}_{rel}, \mathrm{Tol}_{Fkt} > 0$ gegeben, welche den Toleranzen für den absoluten und relativen Fehler bzw. die Toleranz bei der Nullstellenbewertung bzw. des Residuums entsprechen. Wäre die Nullstelle $\mathbf{x}^*$ bekannt und ist $\widetilde{\mathbf{x}}$ eine Näherungslösung, dann könnte eines der folgenden Abbruchkriterien verwendet werden

$$\|\widetilde{\mathbf{x}} - \mathbf{x}^*\| \le \mathrm{Tol}_{abs} , \tag{1.3a}$$
$$\|\widetilde{\mathbf{x}} - \mathbf{x}^*\| \le \mathrm{Tol}_{rel} \cdot \|\mathbf{x}^*\| , \tag{1.3b}$$
$$\|\mathbf{F}(\widetilde{\mathbf{x}})\| \le \mathrm{Tol}_{Fkt} . \tag{1.3c}$$

Nachdem die Lösung $\mathbf{x}^*$ gewöhnlich nicht bekannt ist, werden anstatt (1.3a), (1.3b) auch folgende Kriterien als Schätzungen des absoluten und relativen Fehlers verwendet

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \mathrm{Tol}_{abs} , \tag{1.4a}$$
$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \mathrm{Tol}_{rel} \cdot \|\mathbf{x}^{(k+1)}\| . \tag{1.4b}$$

Zur Vermeidung von Schwierigkeiten bei $\|\mathbf{x}^{(k+1)}\| \approx 0$, kann auch die Kombination

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \le \mathrm{Tol}_{abs} + \mathrm{Tol}_{rel} \cdot \|\mathbf{x}^{(k+1)}\| \tag{1.4c}$$

verwendet werden.
Ob (1.4a) bis (1.4c) zufriedenstellende Abbruchkriterien darstellen, ist u.a. von der Konvergenzgeschwindigkeit der Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ abhängig. Angenommen die Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ konvergiert schnell

gegen $\mathbf{x}^*$, d.h. es gelte $\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \ll \|\mathbf{x}^{(k)} - \mathbf{x}^*\|$. Wegen der Dreieicksungleichung

$$\left| \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| - \|\mathbf{x}^* - \mathbf{x}^{(k)}\| \right| \leq \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \leq \|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| + \|\mathbf{x}^* - \mathbf{x}^{(k)}\| \approx \|\mathbf{x}^* - \mathbf{x}^{(k)}\| \,,$$

folgt $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \approx \|\mathbf{x}^{(k)} - \mathbf{x}^*\|$. Aus $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \leq \mathrm{Tol}_{abs}$ folgt somit $\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \ll \mathrm{Tol}_{abs}$. Wird allerdings angenommen, dass die Folge $\left(\mathbf{x}^{(k)}\right)_{k\in\mathbb{N}}$ langsam gegen $\mathbf{x}^*$ konvergiert, dann gilt $\|\mathbf{x}^{(k+1)} - \mathbf{x}^*\| \approx \|\mathbf{x}^{(k)} - \mathbf{x}^*\|$ und es kann $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\|$ sehr klein sein, obwohl die Iterierten noch nicht nahe genug bei $\mathbf{x}^*$ liegen. Es könnte daher zu einem zu frühen Abbruch der Iterationen kommen. Ob eine zu langsame Konvergenz vorhanden ist, könnte z.B. mit dem Kriterium (eps entspricht der Maschinengenauigkeit)

$$\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \leq \mathrm{eps} \cdot \|\mathbf{x}^{(k)}\| \tag{1.4d}$$

abgefragt werden und zur Terminierung des Algorithmus führen. Bei einem Abbruch des Algorithmus bei zu langsamer Konvergenz, sollte allerdings auch eine Meldung ausgegeben werden, weil eine langsame Konvergenz auftreten könnte, wenn das Problem schlecht konditioniert ist oder die Toleranzparameter zu klein gewählt wurden oder aufgrund von Rundungsfehlern die Iterierten in einer Umgebung um $\mathbf{x}^*$ oszillieren.
Als Abbruchkriterium kann auch das Residuum (1.3c) verwendet werden, wobei empfohlen wird, dieses in Kombination mit den Kriterien (1.4a) bis (1.4c) zu kombinieren. Angenommen für die Norm der Jacobi-Matrix in $\widetilde{\mathbf{x}}$ gelte $\|\mathbf{J}_\mathbf{F}(\widetilde{\mathbf{x}})\| \gg 1$ (z.B. wenn partielle Ableitungen groß sind), dann könnte der Fehler $\|\widetilde{\mathbf{x}} - \mathbf{x}^*\|$ bereits klein sein, auch wenn $\|\mathbf{F}(\widetilde{\mathbf{x}})\| > \mathrm{Tol}_{Fkt}$ gilt.

Berechnung der Newton-Gleichung

Die Methoden welche in diesem Unterkapitel vorgestellt wurden, haben gemeinsam, dass für die neue Iterierte $\mathbf{x}^{(k+1)}$, die Berechnung des Newton-Schritts $\mathbf{d}^{(k)}$ gemäß Gleichung (1.2a) oder (1.2b) erforderlich ist. Gleichung (1.2b) ist gut für das Verständnis, sollte aber bei Implementierungen vermieden werden, da der Aufwand für eine Matrixinversion sehr groß und ineffizient ist. Für die Implementierung wird hingegen das Lösen der Newton-Gleichung (1.2a) empfohlen, da es viele effiziente Möglichkeiten gibt, um ein lineares Gleichungssystem zu lösen. Unterschieden wird hierbei zwischen iterativen und direkten Lösern. Iterative Löser können das lineare Gleichungssystem im Allgemeinen nicht exakt lösen, dafür kann aber meist in wenigen Iterationen bereits eine gute Approximation der Lösung gefunden werden. Bei großen Gleichungssystemen, z.B. Dimensionen größer als 10000, werden zunehmend iterative Löser bevorzugt. Abgesehen von Rundungsfehlern, können direkte Löser ein Gleichungssystem exakt lösen. Für eine invertierbare Matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ kann zum direkten Lösen des linearen Gleichungssystems $\mathbf{A}\cdot\mathbf{x} = \mathbf{b}$, z.B. die LR Zerlegung (links-rechts Zerlegung, engl.: lower-upper bzw. LU-Zerlegung) für $n \leq 10000$ empfohlen werden. Dabei existiert eine invertierbare untere Dreiecksmatrix $\mathbf{L} \in \mathbb{R}^{n\times n}$ und eine invertierbare obere Dreiecksmatrix $\mathbf{R} \in \mathbb{R}^{n\times n}$, sodass $\mathbf{A} = \mathbf{L}\cdot\mathbf{R}$ gilt. Der Rechenaufwand für die arithmetischen Operationen liegt bei $\mathcal{O}\left(\frac{2}{3}\cdot n^3\right)$. Das lineare Gleichungssystem lässt sich aufgrund der Dreiecksstruktur von $\mathbf{L}, \mathbf{R}$, durch Vorwärts- und Rückwärtssubstitutionen, folgendermaßen effizienter lösen

$$\mathbf{L}\cdot\widetilde{\mathbf{b}} = \mathbf{b}\,,$$
$$\mathbf{R}\cdot\mathbf{x} = \widetilde{\mathbf{b}}\,.$$

Der Rechenaufwand der arithmetischen Operationen für die Vorwärts- und Rückwärtssubstitutionen beträgt $\mathcal{O}(n^2)$ und kann gegebenüber dem Aufwand für die LR Zerlegung vernachlässigt werden.

1.4 Differentiell-algebraische Gleichung(en) (DAG)

Ziel dieses Abschnitts ist es, die grundlegenden Ideen und Begriffe von differentiell-algebraischen Gleichungen (engl.: differential-algebraic equation(s), Abk.: DAE) zu vermitteln und welche numerischen Methoden darauf angewendet werden können. Dieses Themengebiet ist sehr umfangreich und für eine weiterführende detailliertere Behandlung, u.a. mit Schwerpunkt auf elektrische Netzwerke, sei beispielsweise auf die angegebene Literatur, sowie auf [12] verwiesen.

1.4.1 Einleitung

Definition von differentiell-algebraischen Gleichungen und Motivation

Die Grundlagen dieses Abschnitts wurden aus [4, Abschnitt 1.1] und [17, Abschnitt 1] entnommen. Ausgangspunkt ist die Darstellung eines impliziten gewöhnlichen Gleichungssystems, welches Funktionen und Ableitungen 1. Ordnung enthält. Hierfür seien $m \in \mathbb{N}$ und eine offenen Menge $D \subseteq \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} = \mathbb{R}^{2\cdot m+1}$, sowie eine (hinreichend oft) stetig differenzierbare Funktion $\mathbf{F} : D \to \mathbb{R}^m : (\mathbf{v}, \mathbf{w}, t) \mapsto \mathbf{F}(\mathbf{v}, \mathbf{w}, t)$ gegeben. D.h. abhängig von der Anwendung, erfüllen die Komponenten von $\mathbf{F}$, Eigenschaften wie Stetigkeit oder seien ausreichend oft stetig differenzierbar. Existiert für ein geeignetes Zeitintervall $\mathcal{T} \subseteq \mathbb{R}$ eine stetig (differenzierbare) Funktion $\mathbf{x} : \mathcal{T} \to \mathbb{R}^m$, sodass für alle $t \in \mathcal{T}$

$$\mathbf{F}\Big(\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}, \mathbf{x}(t), t\Big) = \mathbf{0} \tag{1.6}$$

erfüllt ist, so wird die Funktion $\mathbf{x}(t)$ eine klassische Lösung des impliziten Gleichungssystems genannt. Klassisch bezieht sich in diesem Fall auf die stetige Differenzierbarkeit von allen Komponenten von $\mathbf{x}(t)$. Diese Voraussetzung kann abhängig von der Anwendung auch abgeschwächt werden, wenn z.B. nicht alle Komponenten von $\mathbf{x}(t)$ differenziert werden müssen. Einen wichtigen Spezialfall bilden explizite gewöhnliche Differentialgleichungssysteme 1. Ordnung, d.h. wenn es möglich ist Gleichung (1.6) folgendermaßen zu schreiben

$$\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t) \ . \tag{1.7}$$

Es sei darauf hingewiesen, dass für Gleichung (1.6) zumindest eine lokale Darstellung gemäß (1.7) existiert, wenn die nötigen Voraussetzungen von Satz 1.11 erfüllt sind, bzw. wenn speziell $\mathbf{F}(\mathbf{v}_0, \mathbf{w}_0, t_0) = \mathbf{0}$ und $\det(\mathbf{J}_{\mathbf{v};\mathbf{F}}(\mathbf{v}_0, \mathbf{w}_0, t_0)) \neq 0$ für einen Punkt $(\mathbf{v}_0, \mathbf{w}_0, t_0) \in D$ gilt. In diesem Fall bezeichnet man Gleichung (1.6) als implizites gewöhnliches Differentialgleichungssystem 1. Ordnung. Wenn in einer Differentialgleichung höhere Ableitungen enthalten sind, so ist es aber stets möglich diese Differentialgleichung als Differentialgleichungssystem 1. Ordnung zu schreiben. Im Folgenden sind nun implizite Gleichungssysteme gemäß (1.6) interessant, bei denen eine Darstellung gemäß (1.7) weder lokal noch global möglich oder weniger wünschenswert ist. Ein System von DAG zeichnet sich durch algebraische Nebenbedingungen einzelner Variablen aus (d.h. Gleichungen ohne Ableitungen). Die Nebenbedingungen können beispielsweise auftreten, wenn gemäß (1.6) die partielle Jacobi-Matrix (siehe Definition 1.10) $\mathbf{J}_{\mathbf{v};\mathbf{F}}(\mathbf{v}, \mathbf{w}, t)$ singulär ist oder wenn das implizite Gleichungssystem (1.6) durch eine Funktion $\mathbf{F}_1(\mathbf{v}_1, \mathbf{w}_1, \mathbf{w}_2, t)$ mit invertierender partieller Jacobi-Matrix $\mathbf{J}_{\mathbf{v}_1;\mathbf{F}_1}(\mathbf{v}_1, \mathbf{w}_1, \mathbf{w}_2, t)$ und einer Funktion $\mathbf{F}_2(\mathbf{w}_1, \mathbf{w}_2, t)$ für die Nebenbedingungen, gemäß

$$\mathbf{F}_1\Big(\frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t}, \mathbf{y}(t), \mathbf{z}(t), t\Big) = \mathbf{0} \ , \tag{1.8a}$$

$$\mathbf{F}_2(\mathbf{y}(t), \mathbf{z}(t), t) = \mathbf{0} \ , \tag{1.8b}$$

dargestellt werden kann, wobei $\mathbf{x}(t) = (\mathbf{y}(t)^\top, \mathbf{z}(t)^\top)^\top$ mit $0 < p < m$, $\mathbf{y}(t) \in \mathbb{R}^p$ und $\mathbf{z}(t) \in \mathbb{R}^{m-p}$ gelte.
Differentiell-algebraische Gleichungen treten vor allem bei der Simulation physikalischer Probleme auf (z.B. modifiziertes Knotenspannungsverfahren oder Mehrkörpersysteme). Ein großer Vorteil ist hierbei, dass die Modellierung des physikalischen Problems automatisiert werden kann und die DAG- Variablen für gewöhnlich auch eine physikalische Bedeutung haben. Zudem ist es teilweise nur mit großem Aufwand möglich, physikalische Probleme durch ein explizites Differentialgleichungssystem gemäß (1.7) zu modellieren und dieses Vorgehen kann im Allgemeinen auch nicht automatisiert werden. DAG sind somit für viele physikalische Probleme einfach zu erhalten, allerdings ist die numerische Berechnung für viele DAG nicht trivial. Dies liegt u.a. daran, dass einige Komponenten der Lösung $\mathbf{x}(t)$ redundant sind oder es auch versteckte Nebenbedingungen gibt, weshalb die Anfangszustände meist nicht beliebig wählbar sind (siehe Anmerkung zu Definition 1.16). Vor allem für die Numerik von DAG hilft eine Kennzahl, welche Index genannt wird und im nachfolgenden Abschnitt erläutert wird.

Index-Konzept für DAG

Um unterschiedliche Typen von differentiell-algebraischen Gleichungen charakterisieren zu können, ist der ganzzahlige DAG Index ≥ 0, eine wichtige Kennzahl. Der Index ist u.a. ein Maß dafür, wie schwierig es ist, die numerische Lösung der dazugehörigen DAG zu bestimmen. Zunächst sei darauf hingewiesen, dass es einige verschiedene Definitionen für den DAG-Index gibt, z.B. die englischen Bezeichnungen *Differentiation-Index*, *Tractability-Index*, *Geometric-Index*, *Perturbation-Index* (eine kurze Übersicht über diese Indizes können in [17] nachgelesen werden). Dies liegt u.a. daran, dass abhängig von dem Typ der DAG, Kriterien zur Ermittlung eines speziellen DAG-Index einfacher zu überprüfen sind, als für einen anderen Index. Die Vielfalt an verschiedenen Indexdefinitionen soll aber nicht als Nachteil gesehen werden, da sich für eine bestimmte Anwendung, die Werte der unterschiedlichen Indizes meist nur um Werte ± 1 oder gar nicht unterscheiden.
Zunächst wird der globale Index definiert, welcher vor allem die Idee des Index-Konzepts bei bestimmten DAG-Typen am besten erklärt. und die Bedeutung in den nachfolgenden Beispielen zeigt.

Definition 1.12 (Globaler Index, siehe [4, Abschnitt 2.2, Definition 2.2.2])**:**
Der globale Index $\nu \in \mathbb{N}$ ist die minimale Anzahl an Zeitableitungen, einzelner Komponenten oder aller Gleichungen von (1.6)*, bis $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}$ als stetige Funktion, welche nur von $\mathbf{x}(t)$ und t abhängt, dargestellt werden kann (d.h. wenn es eine explizite Darstellung gemäß* (1.7) *gibt).*

Neben Definition 1.12 haben auch andere Indexdefinitionen gemeinsam, dass die Gleichung (1.6) einen DAG-Index 0 besitzt, wenn (1.6) durch eine explizite Darstellung gemäß (1.7) ersetzt werden kann. Das nachfolgende Beispiel einer semi-expliziten DAG entspricht der Struktur von (1.8a), (1.8b) und soll die Ermittlung des globalen Index zeigen. Durch dieses Beispiel lässt sich nachfolgend auch eine Index-Abhängigkeit, bestimmter DAG-Eigenschaften nachvollziehen.

Beispiel 1.13 (Semi-explizite DAG, siehe [4, Abschnitt 2.2])**:**
Seien $m, n \in \mathbb{N}$, $D \subseteq \mathbb{R}^m \times \mathbb{R}^n \times \mathbb{R}$ eine offene Menge, sowie (hinreichend oft) stetig differenzierbare Funktionen $\mathbf{f}_1 : D \to \mathbb{R}^m : (\mathbf{y}, \mathbf{z}, t) \mapsto \mathbf{f}_1(\mathbf{y}, \mathbf{z}, t)$, $\mathbf{f}_2 : D \to \mathbb{R}^n : (\mathbf{y}, \mathbf{z}, t) \mapsto \mathbf{f}_2(\mathbf{y}, \mathbf{z}, t)$ gegeben. Eine semi-explizite DAG besitzt folgende Darstellung

$$\frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t} = \mathbf{f}_1(\mathbf{y}(t), \mathbf{z}(t), t) \ , \tag{1.9a}$$

$$\mathbf{0} = \mathbf{f}_2(\mathbf{y}(t), \mathbf{z}(t), t) \ . \tag{1.9b}$$

Wird (1.9b) *nach* t *abgeleitet und setzt man* (1.9a) *ein, so führt dies mit Hilfe der Kettenregel* (*gemäß Satz 1.9*) *und den partiellen Jacobi-Matrizen* (*gemäß Definition 1.10*)*, zu folgender Gleichung*

$$\begin{aligned}\mathbf{0} &= \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{f}_2(\mathbf{y}(t),\mathbf{z}(t),t)\\ &= \mathbf{J}_{\mathbf{y};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)\cdot\frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t} + \mathbf{J}_{\mathbf{z};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)\cdot\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} + \mathbf{J}_{t;\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)\cdot 1\\ &= \mathbf{J}_{\mathbf{y};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)\cdot\mathbf{f}_1(\mathbf{y}(t),\mathbf{z}(t),t) + \mathbf{J}_{\mathbf{z};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)\cdot\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t} + \mathbf{J}_{t;\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t) \qquad (1.9c)\end{aligned}$$

Ist die partielle Jacobi-Matrix $\mathbf{J}_{\mathbf{z};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)$ *invertierbar, dann kann Gleichung* (1.9c) *von links mit* $\mathbf{J}_{\mathbf{z};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)^{-1}$ *multipliziert werden und nach* $\frac{\mathrm{d}\mathbf{z}(t)}{\mathrm{d}t}$ *umgeformt werden. Diese umgeformte Gleichung* (1.9c) *entspricht der expliziten Darstellung gemäß* (1.7)*. In diesem Fall besitzt die semi-explizite DAG nach Definition 1.12 den globalen Index 1. Gilt hingegen* $\det(\mathbf{J}_{\mathbf{z};\,\mathbf{f}_2}(\mathbf{y}(t),\mathbf{z}(t),t)) = 0$ (*d.h. die partielle Jacobi-Matrix ist nicht invertierbar*)*, dann wird angenommen, dass durch algebraische Umformungen und Koordinatentransformationen Gleichung* (1.9c) *in eine Darstellung gemäß* (1.9a), (1.9b) *gebracht werden kann, allerdings mit neuen Funktionen* $\widetilde{\mathbf{y}}(t)$, $\widetilde{\mathbf{z}}(t)$*. Ist es dann möglich, wie bei* (1.9c) *eine explizite Darstellung gemäß* (1.7) *zu erhalten, dann besitzt die DAG* (1.9a), (1.9b) *den globalen Index 2. Ist es nicht möglich, dann kann das Vorgehen weiter wiederholt werden und die Anzahl der benötigten zeitlichen Ableitungen ergibt den globalen Index.*

Die ggf. in Beispiel 1.13 benötigten algebraischen Umformungen und Koordinatentransformationenen sollen im nachfolgenden Beispiel erläutert werden.

Beispiel 1.14 (Quasilineare DAG, siehe [13, Abschnitt 1])**:**
Seien $m \geq 2$*, eine Matrix* $\mathbf{M} \in \mathbb{R}^{m\times m}$*, eine offene Menge* $D \subseteq \mathbb{R}^m \times \mathbb{R}$*, sowie eine* (*hinreichend oft*) *stetig differenzierbare Funktion* $\mathbf{f} : D \to \mathbb{R}^m : (\mathbf{x},t) \mapsto \mathbf{f}(\mathbf{x},t)$*, gegeben, welche die folgende quasilineare-DAG beschreiben*

$$\mathbf{M}\cdot\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t),t)\ . \qquad (1.10a)$$

Wenn $\mathbf{M}$ *invertierbar ist, dann besitzt* (1.10a) *einen globalen Index 0. Ist* $\mathbf{M}$ *singulär, dann wird angenommen, dass* $\operatorname{rang}(\mathbf{M}) = r$ *mit* $0 < r < m$*. Durch das Gaußsche Eliminationsverfahren ist es möglich,* $\mathbf{M}$ *derart umzuformen, dass lediglich Diagonalelemente den Wert* 1 *besitzen und alle restlichen Elemente den Wert* 0*. Diese Umformung kann durch zwei invertierbare Matrizen* $\mathbf{S},\mathbf{T} \in \mathbb{R}^{m\times m}$*, sowie der Einheitsmatrix* $\mathbf{I}_r \in \mathbb{R}^{r\times r}$*, mit* $n = (m-r)$*, folgendermaßen geschrieben werden*

$$\mathbf{M} = \mathbf{S}\cdot\begin{pmatrix}\mathbf{I}_r & \mathbf{0}_{n\times n}\\ \mathbf{0}_{n\times n} & \mathbf{0}_{n\times n}\end{pmatrix}\cdot\mathbf{T}$$

Mit der Transformation

$$\begin{pmatrix}\mathbf{y}(t)\\ \mathbf{z}(t)\end{pmatrix} := \mathbf{T}\cdot\mathbf{x}(t)\ , \qquad \textit{bzw.} \qquad \mathbf{x}(t) = \mathbf{T}^{-1}\cdot\begin{pmatrix}\mathbf{y}(t)\\ \mathbf{z}(t)\end{pmatrix}\ ,$$

wobei $\mathbf{y}(t) \in \mathbb{R}^r$ *und* $\mathbf{z}(t) \in \mathbb{R}^n$ *gelte, lässt sich* (1.10a) *nach Multiplikation mit* $\mathbf{S}^{-1}$ *von links und Funktionen* $\widetilde{\mathbf{f}},\widetilde{\mathbf{f}}_1,\widetilde{\mathbf{f}}_2$*, dann folgendermaßen als semi-explizite DAG schreiben*

$$\begin{pmatrix}\frac{\mathrm{d}\mathbf{y}(t)}{\mathrm{d}t}\\ \mathbf{0}_n\end{pmatrix} = \mathbf{S}^{-1}\cdot\mathbf{f}\left(\mathbf{T}^{-1}\cdot\begin{pmatrix}\mathbf{y}(t)\\ \mathbf{z}(t)\end{pmatrix},t\right)$$

$$= \widetilde{\mathbf{f}}(\mathbf{y}(t), \mathbf{z}(t), t)$$
$$= \begin{pmatrix} \widetilde{\mathbf{f}}_1(\mathbf{y}(t), \mathbf{z}(t), t) \\ \widetilde{\mathbf{f}}_2(\mathbf{y}(t), \mathbf{z}(t), t) \end{pmatrix} . \tag{1.10b}$$

Um den globalen Index von (1.10a) *zu ermitteln, kann nun ausgehend von* (1.10b)*, wie in Beispiel 1.13 beschrieben wurde, vorgegangen werden.*

Wird die Idee des globalen Index im Hinblick auf nichtlineare DAG verallgemeinert, so führt dies zur Definition des *Differential-Index* (im Folgenden auch als Index bezeichnet).

Definition 1.15 (Differentiation-Index, siehe [4, Abschnitt 2.5.1, Definition 2.5.1])**:**
Ausgehend von Gleichung (1.6)*, sei der Differentiation-Index $k \in \mathbb{N}$, das kleinste k bei dem das Gleichungssystem*

$$\begin{aligned} \mathbf{F}\Big(\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}, \mathbf{x}(t), t\Big) &= \mathbf{0} , \\ \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{F}\Big(\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}, \mathbf{x}(t), t\Big) &= \mathbf{0} , \\ &\vdots \\ \frac{\mathrm{d}^k}{\mathrm{d}t^k}\,\mathbf{F}\Big(\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}, \mathbf{x}(t), t\Big) &= \mathbf{0} , \end{aligned}$$

die Funktion $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}$ eindeutig als stetige Funktion, abhängig von $\mathbf{x}(t)$ und t, liefert und somit eindeutig durch eine stetige Funktion $\mathbf{f}$ wie in (1.7) *dargestellt werden kann.*

Zur Motivation des *Differentiation-Index* sei auch auf [17, Abschnitt 3.7] verwiesen (hier lässt sich auch nachlesen, dass eine semi-explizite DAG (gemäß Beispiel 1.13) mit globalen Index 1, auch einen *Differentiation-Index* 1 besitzt).
Im Folgenden werden zwei Spezialfälle für DAG angegeben, welche man u.a. bei der Modellierung eines elektrischen Netzwerkes, durch das MKV erhält (siehe Unterkapitel 2.2).

Lineare DAG

Für $m \geq 1$ seien konstante Matrizen $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times m}$, sowie auf einem offenen Zeitintervall $\mathcal{T} \subseteq \mathbb{R}$ eine Funktion $\mathbf{q} : \mathcal{T} \to \mathbb{R}^m : t \mapsto \mathbf{q}(t)$ gegeben. Eine zeitinvariante lineare DAG besitzt folgende Darstellung (im Allgemeinen ist $\mathbf{A}$ singulär)

$$\mathbf{A}\cdot\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} + \mathbf{B}\cdot\mathbf{x}(t) = \mathbf{q}(t) . \tag{1.11}$$

Gleichung (1.11) tritt beispielsweise bei linearen elektrischen Netzwerken auf, wenn dieses mit dem MKV modelliert wird. Eine zeitvariante lineare DAG entspricht der Form (1.11), allerdings mit zeitabhängigen Matrizen $\mathbf{A}(t), \mathbf{B}(t)$. Es gibt eine umfangreiche Theorie zu linearen DAG und hierfür sei beispielsweise auf [4, Abschnitt 2.3] oder [17, Kapitel 2] verwiesen.

Quasilineare DAG

Seien $m \geq 1$, eine offene Menge $D \subseteq \mathbb{R}^m \times \mathbb{R}$, sowie eine (hinreichend oft) stetig differenzierbare Funktion $\mathbf{f} : D \to \mathbb{R}^m : (\mathbf{x}, t) \mapsto \mathbf{f}(\mathbf{x}, t)$ und eine Matrix-Funktion $\mathbf{A} : D \to \mathbb{R}^{m\times m} : (\mathbf{x}, t) \mapsto \mathbf{A}(\mathbf{x}, t)$ mit (hinreichend oft) stetig differenzierbaren Komponenten gegeben. Dann wird (nach [17, Abschnitt 1.4.3])

$$\mathbf{A}(\mathbf{x}(t), t)\cdot\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} = \mathbf{f}(\mathbf{x}(t), t) . \tag{1.12}$$

als quasilineare DAG bezeichnet. Diese allgemeine Form tritt beispielsweise auf, wenn ein elektrisches Netzwerk, welches nichtlineare oder zeitvariante Elemente besitzt, mit dem MKV modelliert wird. Die Zeitvarianz bei energiespeichernden Elementen, wird in der Matrix $\mathbf{A}(\mathbf{x}, t)$ durch die Komponente t berücksichtigt. In einem zeitinvarianten elektrischen Netzwerk entfällt diese Komponente und Nichtlinearitäten in Spulen und Kondensatoren werden durch eine Matrix $\mathbf{A}(\mathbf{x})$ berücksichtigt. Eine quasilineare DAG gemäß Beispiel 1.14, tritt beispielsweise in einem elektrischen Netzwerk auf, welches lineare Spulen und Kondensatoren besitzt, sowie resistive Elemente mit linearen Verhalten oder mit algebraischen Nichtlinearitäten (z.B. Dioden).

Konsistente Anfangsbedingungen

Zunächst wird der Begriff *konsistente Anfangsbedingung* definiert, wobei von den Bezeichnungen für Gleichung (1.6) ausgegangen wird.

Definition 1.16 (Konsistente Anfangsbedingung, siehe [20, Abschnitt 2, Definition 2.1])**:**
Ein Vektor $\mathbf{x}_0 \in \mathbb{R}^m$ *wird konsistente Anfangsbedingung für Gleichung* (1.6) *genannt, wenn eine Lösung* $\mathbf{x}(t)$ *für* (1.6) *existiert, welche* $\mathbf{x}(t_0) = \mathbf{x}_0$ *erfüllt.*

Ein Vorteil von expliziten Differentialgleichungen der Darstellung (1.7) ist, dass eine geschlossene Lösungstheorie existiert, vor allem Aussagen über die Existenz und Eindeutigkeit von Anfangswertproblemen (siehe Sätze von Peano oder Picard-Lindelöf). Erfüllt z.B. die rechte Seite einfach überprüfbare Stetigkeitskriterien, so ist sichergestellt, dass für jeden beliebigen Anfangswert eine Lösung existiert und diese ggf. auch eindeutig ist (diese Eigenschaften sind auch wesentlich für die Numerik von expliziten Differentialgleichungen). Bei DAG ist hingegen die Wahl der Anfangsbedingung im Allgemeinen nicht beliebig. Einerseits muss die Lösung $\mathbf{x}(t)$ von DAG auch algebraische Bedingungen erfüllen und zusätzlich können DAG auch versteckte Nebenbedingungen beinhalten (diese treten auf, wenn die algebraischen Bedinungen nach der Zeit abgeleitet werden). Eine versteckte Nebenbedingung entspricht z.B. Gleichung 1.9c einer semi-explizite DAG mit Index 1, gemäß Beispiel 1.13, d.h. eine konsistente Anfangsbedingung muss (1.9b) und (1.9c) erfüllen. Aus diesem Beispiel wird auch ersichtlich, dass die Anzahl der versteckten Nebenbedingungen mit dem Index zunimmt. Dies hat z.B. auch eine Auswirkungen bei der numerischen Lösung von MKV Modellen für elektrische Netzwerke. So kann z.B. bei Index 0 oder 1 DAG, eine *DC-Analyse* konsistente Anfangsbedinungen liefern, allerdings kann dies bei einem größerem Index nicht mehr der Fall sein (siehe z.B. [19], [20]).

1.4.2 Numerische Methoden für DAG

Wie bereits erwähnt wurde, ist der Index eine wichtige Kennzahl für die Auswahl von numerischen Methoden, zur numerischen Berechnung bestimmter DAG-Typen. Zuverlässige numerische Methoden gibt es nur für eine breite Klasse von DAG bis zu einem Index von 2. Es gibt DAG mit einem Index größer 2, welche eine bestimmte Struktur besitzen und bei denen eine numerische Berechnung möglich ist. Bei allgemeinen DAG mit einem Index größer 2 kann aber eine numerische Berechnung nicht sichergestellt werden oder es kommt z.B. zu Konvergenzproblemen bei kleinen Schrittweiten (siehe z.B. [1, Abschnitt 10.1.4]). Es gilt daher, je kleiner der Index einer DAG ist, umso besser kann die DAG auch numerisch berechnet werden. In Simulationsprogrammen kommen daher u.a. auch Algorithmen zur Indexreduzierung zum Einsatz, um für eine breite Anwendung die Konvergenz sicherzustellen (Details zur Indexreduzierung können z.B. in [19, Abschnitt 2.4] oder [17, Kapitel 3] nachgelesen werden). Die Idee der Indexreduzierung kann anhand von Beispiel 1.13 gut nachvollzogen werden. So besitzt die DAG, welche neben (1.9a) und (1.9b), auch die versteckte Nebenbedingung (1.9c) enthält, einen Index der gegenüber einer DAG bestehend aus (1.9a),

(1.9b), um 1 niedriger ist (d.h. durch zeitliche Ableitung der algebraischen Nebenbedingungen konnte der Index um 1 reduziert werden).

Zur Beschreibung der Konvergenzeigenschaften der nachfolgenden numerischen Verfahren für DAG, wird zunächst der Begriff Konvergenzordnung erklärt.

$\mathcal{O}$-Notation und Konvergenzordnung

Im Folgenden wird das Landau Symbol $\mathcal{O}$ (*Big-O*) zur Beschreibung verschiedener Fehler verwendet (siehe [1, Kapitel 3]). In numerischen Verfahren ist eine wesentliche Eigenschaft, wie sich ein Fehler in Abhängigkeit einer Schrittweit $h > 0$ verhält, d.h. mit welchem Verhalten wird der Fehler kleiner, wenn h kleiner wird. Für $m \geq 1$ und einem Vektor der von h abhängt, d.h. $\mathbf{d} = \mathbf{d}(h) \in \mathbb{R}^m$, bedeutet die Notation

$$\|\mathbf{d}\| = \mathcal{O}(h^p) \ ,$$

dass zwei Konstanten $p > 0$ und $C > 0$ existieren, sodass für alle $h > 0$ und h klein genug,

$$\|\mathbf{d}\| \leq C \cdot h^p \ ,$$

gilt. C darf dabei nicht von h abhängen und $\|\cdot\|$ bezeichnet eine beliebige Vektornorm im $\mathbb{R}^m$, z.B. die euklidische Norm. Aufgrund der Normäquivalenz in endlich dimensionalen normierten Vektorräumen, sind zwei Normen im $\mathbb{R}^m$ mit einer Konstante abschätzbar, weshalb die obige Definition unabhängig von der verwendeten Norm ist.
Mit Hilfe der $\mathcal{O}$-Notation wird nun der Begriff der Konvergenzordnung erklärt. Hierfür sei ein Zeitintervall $[a, b] \subseteq \mathbb{R}$ gegeben, sowie $N + 1$ Zeitschritte $a = t_0 < t_1 < \cdots < t_N = b$ und die maximale Schrittweite

$$h := h_{max} = \max_{1 \leq n \leq N} h_n = \max_{1 \leq n \leq N} (t_n - t_{n-1}) \ .$$

Sei $\mathbf{x}(t)$ die Lösung einer DAG gemäß (1.6) und $\mathbf{x}_n$ sei der durch ein numerisches Verfahren ermittelte approximierte Lösungsvektor zum Zeitpunkt t_n. Das numerische Verfahren heißt dann konvergent mit der Ordnung p, wenn der globale Fehler $\mathbf{e}_n = \mathbf{x}_n - \mathbf{x}(t_n)$ mit $\mathbf{e}_0 = \mathbf{0}$, für alle $n \in \{1, \ldots, N\}$ die Eigenschaft $\|\mathbf{e}_n\| = \mathcal{O}(h^p)$ besitzt.

BDF Verfahren

Das erste numerische Verfahren, welches zur Lösung von impliziten DAG der Form (1.6) betrachtet wird, ist das Backward Differentiation Formula (BDF) Verfahren, welches bei konstanter Schrittweite $h > 0$ erläutert wird. Dieses Verfahren funktioniert nur für eine Index 1 DAG zuverlässig, bzw. lässt sich im Allgemeinen die Konvergenz bei einem Index größer 1 nicht gewährleistet. Das BDF Verfahren ist ein lineares Mehrschrittverfahren und dies bedeutet, dass bei der Berechnung eines neuen Zeitschritts, auch mehrere Vorgängerwerte berücksichtigt werden können. Sei mit $\mathbf{x}(t)$ die Lösung von (1.6) bezeichnet, dann werden für $k < 7$, bei einem k-Schritt BDF (oder BDFk) Verfahren, k bekannte Werte $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-k+1}$ vorausgesetzt, welche die Lösungen $\mathbf{x}(t_n), \mathbf{x}(t_{n-1}), \ldots \mathbf{x}(t_{n-k+1})$ approximieren. Die Idee des BDFk Verfahrens ist nun, ein Interpolationspolynom zu erstellen, welches in den Zeitpunkten $t_n, t_{n-1}, \ldots, t_{n-k+1}$ (wobei $h = t_i - t_{i-1}$), mit $\mathbf{x}_n, \mathbf{x}_{n-1}, \ldots, \mathbf{x}_{n-k+1}$ übereinstimmt (d.h. ein Polynom je Komponente). Weiters erfülle dieses Interpolationspolynom die Eigenschaft, dass die Zeitableitung zum Zeitpunkt t_{n+1}, näherungsweise mit $\frac{\mathrm{d}\mathbf{x}(t_{n+1})}{\mathrm{d}t}$ übereinstimmt. In Tabelle 1.1 sind die Koeffizienten des Mehrschrittverfahrens BDFk angegeben, welche sich nach diesem Ansatz ergeben (siehe [1, Abschnitt 5.1.2, Table 5.3]). BDF1

Schritte k	β_0	α_0	α_1	α_2	α_3	α_4	α_5	α_6
1	1	1	-1					
2	$\frac{2}{3}$	1	$-\frac{4}{3}$	$\frac{1}{3}$				
3	$\frac{6}{11}$	1	$-\frac{18}{11}$	$\frac{9}{11}$	$-\frac{2}{11}$			
4	$\frac{12}{25}$	1	$-\frac{48}{25}$	$\frac{36}{25}$	$-\frac{16}{25}$	$\frac{3}{25}$		
5	$\frac{60}{137}$	1	$-\frac{300}{137}$	$\frac{300}{137}$	$-\frac{200}{137}$	$\frac{75}{137}$	$-\frac{12}{137}$	
6	$\frac{60}{147}$	1	$-\frac{360}{147}$	$\frac{450}{147}$	$-\frac{400}{147}$	$\frac{225}{147}$	$-\frac{72}{147}$	$\frac{10}{147}$

Tabelle 1.1: Koeffizienten für das k-Schritt BDF Verfahren.

stimmt zudem mit dem *Backward Euler* Verfahren überein. Ausgehend von der Darstellung einer impliziten DAG der Form (1.6), ist für das BDFk Verfahren (gemäß Tabelle 1.1), die Gleichung

$$\mathbf{F}\left(\frac{1}{\beta_0 \cdot h} \cdot \sum_{i=0}^{k} \left(\alpha_i \cdot \mathbf{x}_{n+1-i}\right),\, \mathbf{x}_{n+1},\, t_{n+1}\right) = \mathbf{0}\,, \tag{1.13}$$

für $\mathbf{x}_{n+1}$ zu erfüllen (siehe [1, Abschnitt 10.1.2]). $\mathbf{x}_{n+1}$ kann z.B. mit dem Newton Verfahren gemäß Unterkapitel 1.3 ermittelt werden und entspricht der Approximation von $\mathbf{x}(t_{n+1})$. Es sei zudem angemerkt, dass aufgrund (1.6), der erste Term der Funktion $\mathbf{F}$ in (1.13), einer Approximation von $\frac{\mathrm{d}\mathbf{x}(t_{n+1})}{\mathrm{d}t}$ entspricht.
Gemäß [1, Abschnitt 10.1.2] bzw. [4, Abschnitt 3.2.2, Theorem 3.2.1] gilt folgendes Konvergenzverhalten. Für eine implizite DAG mit Index 0 oder 1 und für $k < 7$, konvergiert das BDFk Verfahren gemäß (1.13) mit $\mathcal{O}(h^k)$, besitzt also die Konvergenzordnung k, wenn die k Startwerte korrekt mit $\mathcal{O}(h^k)$ sind (d.h. für $0 \leq n \leq k-1$ gilt $\|\mathbf{x}_n - \mathbf{x}(t_n)\| = \mathcal{O}(h^k)$) und wenn Gleichung (1.13) in jedem Schritt mit einer Genauigkeit von $\mathcal{O}(h^{k+1})$ berechnet wurde. Wenn also das Newton Verfahren, Gleichung (1.13) nur näherungsweise löst, soll sich der Fehler mit $\mathcal{O}(h^{k+1})$ verhalten.
Für ein BDFk Verfahren kann gemäß [1, Abschnitt 5.1.3] bzw. [5, Abschnitt 8.2], folgende Startstrategie empfohlen werden. Ausgehend von einem Startwert $\mathbf{x}_0 = \mathbf{x}(t_0)$ werden der Reihe nach die Verfahren BDF1, ..., BDF$(k-1)$ angewendet um die erforderlichen k Startwerte für das BDFk Verfahren zu erhalten. In den weiteren Schritten wird dann stets das BDFk Verfahren verwendet.

Radau IIA Verfahren

Das zweite numerische Verfahren welches zur Lösung von impliziten DAG der Form (1.6) betrachtet wird, ist das Radau IIA Verfahren, mit konstanter Schrittweite $h > 0$. Dieses ist ein implizites Runge-Kutta Einschrittverfahren und kann zur Lösung von semi-expliziten DAG, gemäß der Struktur in Beispiel 1.13, bis zu einem Index 2 zuverlässig verwendet werden. Zudem sei angemerkt, dass implizite Runge-Kutta Verfahren auch für quasilineare DAG der Form (1.12) angewendet werden können, da diese invariant gegenüber Transformationen sind, welche in Beispiel 1.14 vorgestellt wurden (siehe Abschnitt *Classes of implicit Runge-Kutta methods* in [13, Kapitel 2]).
Bevor das Verfahren vorgestellt wird, soll zunächst die Idee eines Einschrittverfahren (inkl. Runge-Kutta Verfahren) vermittelt werden (siehe [1, Abschnitt 4.1]). Ausgangspunkt ist bei $h = t_n - t_{n-1}$ die Gleichung

$$\mathbf{x}(t_n) - \mathbf{x}(t_{n-1}) = \int_{t_{n-1}}^{t_n} \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}\, \mathrm{d}t$$

Die Methode des Einschrittverfahrens wird nun durch das Integrationsverfahren bestimmt, welches zur Lösung des Integrals verwendet wird. Hierbei ist das Ziel, dass das Integral für Polynome

mit möglichst großem Polynomgrad (als Integrand), exakt berechnet wird. Hierbei gilt als obere Schranke, dass s-stufige Integrationsverfahren, Polynome mit maximalen Polynomgrad $2{\cdot}s$, exakt berechnen können. Ein Integrationsverfahren zeichnet sich dabei durch die Wahl der Stützstellen und der Wahl des Interpolationspolynoms aus. Das Radau IIA Verfahren leitet sich dabei aus den Radau Integrationsformeln ab.
Es ist üblich, die Koeffizienten eines s-stufigen Einschrittverfahrens in einem *Butcher-Tableau* anzugeben (siehe Tabelle 1.2a). In den Tabellen 1.2b bis 1.2d sind zudem für $s \in \{1, 2, 3\}$ die Koeffizienten für das s-stufige Radau IIA($2{\cdot}s-1$) Verfahren angegeben, wobei die Zahl in der Klammer, der maximalen Ordnung entspricht (siehe [14, Abschnitt IV.5, Table 5.5,5.6] oder [13, Kapitel 2, Table 2.1,2.2]). Ausgehend von der Darstellung einer impliziten DAG der Form (1.6),

c_1	a_{11}	$\dots$	a_{1s}
c_2	a_{21}	$\dots$	a_{2s}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
c_s	a_{s1}	$\dots$	a_{ss}
	b_1	$\dots$	b_s

(a) Butcher Tableau

1	1
	1

(b) Radau IIA(1)

$\frac{1}{3}$	$\frac{5}{12}$	$-\frac{1}{12}$
1	$\frac{3}{4}$	$\frac{1}{4}$
	$\frac{3}{4}$	$\frac{1}{4}$

(c) Radau IIA(3)

$\frac{4-\sqrt{6}}{10}$	$\frac{88-7\cdot\sqrt{6}}{360}$	$\frac{296-169\cdot\sqrt{6}}{1800}$	$\frac{-2+3\cdot\sqrt{6}}{225}$
$\frac{4+\sqrt{6}}{10}$	$\frac{296+169\cdot\sqrt{6}}{1800}$	$\frac{88+7\cdot\sqrt{6}}{360}$	$\frac{-2-3\cdot\sqrt{6}}{225}$
1	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$
	$\frac{16-\sqrt{6}}{36}$	$\frac{16+\sqrt{6}}{36}$	$\frac{1}{9}$

(d) Radau IIA(5)

Tabelle 1.2: Butcher Tableau und Koeffizienten des Radau IIA($2{\cdot}s-1$) Verfahrens mit $s \in \{1, 2, 3\}$.

sind für das s-stufige Radau IIA Verfahren (gemäß Tabelle 1.2), die Gleichungen

$$\mathbf{F}\left(\mathbf{k}_i,\, \mathbf{x}_n + h \cdot \sum_{j=1}^{s} \left(a_{ij} \cdot \mathbf{k}_j\right),\, t_n + c_i \cdot h\right) = \mathbf{0}\,, \qquad \text{für alle } i = 1, 2, \dots, s\,, \tag{1.14a}$$

für $\mathbf{k}_1, \dots, \mathbf{k}_s$ zu erfüllen (siehe [1, Abschnitt 10.1.3]). $\mathbf{k}_1, \dots, \mathbf{k}_s$ können z.B. mit dem Newton Verfahren gemäß Unterkapitel 1.3 ermittelt werden und entsprechen den approximierten zeitlichen Ableitungen von $\mathbf{x}(t)$, zum Zeitpunkt $t_n + c_i \cdot h$. Der neue Zeitschritt $\mathbf{x}_{n+1}$, d.h. die Approximation von $\mathbf{x}(t_{n+1})$, ergibt sich dann durch

$$\mathbf{x}_{n+1} = \mathbf{x}_n + h \cdot \sum_{i=1}^{s} \left(b_i \cdot \mathbf{k}_i\right)\,. \tag{1.14b}$$

Gemäß [1, Abschnitt 10.1.3], [13, Kapitel 2, Table 2.3] bzw. [14, Kapitel VI,VII], gilt folgendes Konvergenzverhalten. Zunächst wird vorausgesetzt, dass eine konsistente Anfangsbedingung $\mathbf{x}_0$ (gemäß Definition 1.16) vorliegt und die Gleichungen (1.14a) für $\mathbf{k}_i$ hinreichend genau gelöst wurden. Für eine implizite DAG mit Index 0 oder 1 konvergiert das s-stufige Radau IIA($2{\cdot}s-1$) Verfahren mit $\mathcal{O}(h^{(2\cdot s-1)})$ Bei einer Index 2 semi-expliziten DAG gemäß Beispiel 1.13, konvergiert die differentielle Komponente $\mathbf{y}_{n+1}$ mit $\mathcal{O}(h^{(2\cdot s-1)})$ und die algebraische Komponente $\mathbf{z}_{n+1}$ mit $\mathcal{O}(h^s)$. Diese Eigenschaft gilt auch für die differentielle und algebraische Komponente bei quasilinearen Index 2 DAG.
Ein Radau IIA Verfahren hat im Vergleich mit dem BDF Verfahren den Vorteil, dass eine größere Konvergenzordnung erzielt werden kann und das damit auch Index 2 Gleichungen gelöst werden können. Das BDF Verfahren ist hingegen bei DAG einfacher zu implementieren und das nichtlineare Gleichungssystem, welches in jedem Zeitschritt gelöst werden muss, ist kleiner (vergleiche (1.13) mit (1.14a)). Bei einem Radau IIA Verfahren wird zudem neben dem Startwert $\mathbf{x}_0 = \mathbf{x}(t_0)$ auch die erste Ableitung $\frac{\mathrm{d}\mathbf{x}(t_0)}{\mathrm{d}t}$ benötigt, da zur Lösung des Gleichungssystems (1.14a) z.B. bei Radau IIA(3), bei der ersten Newton Iteration der Startwert $\mathbf{k}_1 = \mathbf{k}_2 = \frac{\mathrm{d}\mathbf{x}(t_0)}{\mathrm{d}t}$ empfohlen wird (siehe [5, Abschnitt 8.3, Gl. (8.45)]). Eine Approximation von $\frac{\mathrm{d}\mathbf{x}(t_0)}{\mathrm{d}t}$ könnte z.B. durch das Radau IIA(1) oder das BDF1 Verfahren ermittelt werden (beides entspricht dem impliziten Euler Verfahren).

2 Knotenspannungsverfahren

In diesem Kapitel werden die Grundlagen der Knotenspannungsverfahren (engl.: nodal analysis) bereitgestellt, mit dem Ziel, durch das Modell des modifizierten Knotenspannungsverfahrens (MKV) (engl.: modified nodal analysis, Abk.: MNA), eine sehr allgemeine Beschreibung elektrischer Netzwerke zu erhalten.

2.1 Grundlegende Aspekte für elektrische Netzwerke

Dieses Unterkapitel stellt die wesentlichen Grundlagen und Voraussetzungen zur Verfügung, welche für Knotenspannungsverfahren benötigt werden. Die Grundlagen dieses Unterkapitels wurden aus [17, Abschnitt 5.1] und [21] entnommen.

2.1.1 Anwendung der Graphentheorie auf elektrische Netzwerke

In diesem Abschnitt wird ein Zusammanhang zwischen elektrischen Netzwerken und der Graphentheorie gegeben. Dadurch ist es möglich strukturelle Eigenschaften der Graphentheorie auf elektrische Netzwerke anzuwenden.

Allgemeine Voraussetzungen

Die grundlegenden Voraussetzungen sind, dass sich ein elektrisches Netzwerk durch Zweipole bzw. Eintore aufgebauen lässt und es in diesem elektrischen Netzwerk nur ein Bezugspotential bzw. eine Bezugsmasse gibt. Dies stellt aber keine allzu große Einschränkung dar, da ein 'isoliertes', elektrisches Netzwerk eine 'virtuelle' Masse besitzen kann, welche über einen hohen Isolationswiderstand mit dem Bezugspotential verbunden ist. Beispielsweise könnte bei einem Einphasentransformator, die Primär- und Sekundärseite über einen Isolationswiderstand verbunden werden.
In einem elektrischen Netzwerk wird nun jedem Zweig, welche je einem Zweipolelement entspricht, eine beliebige (Zählpfeil-) Richtung zugeordnet. Die Richtung des Zweiges legt nun die positive Bezugsrichtung des Strom- und Spannungszählpfeils des Zweiges fest. Diese Konvention gilt nicht nur für passive Bauteile, sondern auch für Strom- und Spannungsquellen. Sei beispielsweise eine Spannungsquelle in einem elektrischen Netzwerk gegeben, bei dem die positive Richtung des Strom- und des Spannungszählpfeils vom Plus- zum Minuspol der Spannungsquelle zeigt. Gibt die Spannungsquelle Leistung ab, so wird aufgrund der Kirchhoffschen Gesetze gewährleistet, dass die Berechnung ein negatives Vorzeichen für den Strom der Spannungsquelle liefert. Die Zählpfeile legen daher lediglich die positive Bezugsrichtung fest.
Mit diesen Voraussetzungen ist es nun möglich das elektrische Netzwerk als zusammenhängenden gerichteten Graphen (V, E, α), gemäß Definition 1.1 zu betrachten. Hierbei enthält die Menge V die n Knoten und die Menge E die b Zweige des zu modellierenden elektrischen Netzwerkes. Nachdem nichttriviale elektrische Netzwerke mindestens 2 Knoten und 2 Zweige besitzen, wird zudem vorausgesetzt, dass $b, n \geq 2$ gilt. Zusätzlich darf das elektrische Netzwerk keine Eigenschleifen besitzen, weshalb die beiden Anschlüsse eines Zweiges nicht mit demselben Knoten inzident bzw. verbunden sein dürfen. In der Funktion α sei zudem die (Zählpfeil-) Richtung der Zweige gespeichert.

Konvention zur Nummerierung der Knoten und Zweige

Es wird nun folgende Konvention bzgl. der Nummerierung der Knoten und der Zweige getroffen. Aus den n Knoten kann nun ein beliebiger Knoten als Bezugsknoten bestimmt werden, wodurch das Bezugspotential definiert ist. Dieser Knoten wird mit k_0 bezeichnet. Die restlichen Knoten werden beliebig mit $k_1, \ldots, k_{n-1}$ bezeichnet. Die b Zweige werden nun in 5 Gruppen eingeteilt, nämlich in b_R resistive Zweige, b_C kapazitive Zweige, b_L induktive Zweige, b_U Zweige mit gesteuerten und ungesteuerten Spannungsquellen und b_I Zweige mit gesteuerten und ungesteuerten Stromquellen, sodass $b = b_R + b_C + b_L + b_U + b_I$ gilt. Für $1 \leq j \leq b$ wird ein Zweig mit z_j bezeichnet. Die Reihenfolge der Nummerierung der Zweige erfolgt so, dass resistive Zweige, kapazitive Zweige, induktive Zweige, Zweige mit Spannungsquellen und Zweige mit Stromquellen in dieser Reihenfolge aufsteigend nummeriert werden. Damit entsprechen $z_1, \ldots, z_{b_R}$ den resistiven Zweigen, $z_{b_R+1}, \ldots, z_{b_R+b_C}$ den kapazitiven Zweigen und bei Fortsetzung dieser Reihenfolge $z_{b_R+b_C+b_L+b_U+1}, \ldots, z_b$ den Zweigen der Stromquellen. Wenn $b_\ell = 0$ für $\ell \in \{R, C, L, U, I\}$ gilt, dann wird diese Gruppe bei der Nummerierung übersprungen. Ist beispielsweise $b_L = 0$, dann gibt es keinen induktiven Zweig im gerichteten Graphen. Es sei bemerkt, dass die gewählte Konvention der Gruppenreihenfolge und Nummerierung der Elemente im elektrischen Netzwerk, lediglich der einheitlichen Darstellung in der weiteren Betrachtung dient, im Allgemeinen aber beliebig wählbar ist.

Inzidenzmatrix

Durch die gewählte Nummerierung der Zweige und Knoten ist es nun möglich die vollständige Inzidenzmatrix $\widetilde{\mathbf{A}} = \left(\widetilde{a}_{i,j}\right)_{\substack{i=0,\ldots,n-1 \\ j=1,\ldots,b}} \in \mathbb{R}^{n\times b}$, gemäß Definition 1.4,(a) anzugeben (Knoten werden anstatt mit v_i mit k_i bezeichnet und Kanten bzw. Zweige anstatt mit e_j durch z_j). Gemäß Lemma 1.5,(a) sind die Zeilen der vollständigen Inzidenzmatrix $\widetilde{\mathbf{A}}$ linear abhängig. Die Wahl eines Bezugsknotens enspricht nun der Entfernung der Zeile in $\widetilde{\mathbf{A}}$, welche dem Bezugsknoten k_0 entspricht. Dies führt gemäß Definition 1.4,(b) zur (reduzierten) Inzidenzmatrix $\mathbf{A} = \left(a_{i,j}\right)_{\substack{i=1,\ldots,n-1 \\ j=1,\ldots,b}} \in \mathbb{R}^{(n-1)\times b}$, mit den Koeffizienten

$$a_{i,j} = \begin{cases} 1 & \text{, wenn } k_i \text{ der Startknoten des gerichteten Zweiges } z_j \text{ ist} \\ -1 & \text{, wenn } k_i \text{ der Endknoten des gerichteten Zweiges } z_j \text{ ist} \\ 0 & \text{, wenn der Zweig } z_j \text{ nicht mit Knoten } k_i \text{ inzident bzw. verbunden ist} \end{cases} .$$

Um dies an einem Beispiel zu erläutern, werden die Koeffizienten für den Zweig z_r in Abbildung 2.1 angegeben. Hierbei ergibt sich $a_{i,r} = -1$ und $a_{j,r} = 1$, sowie $a_{\ell,r} = 0$ für $\ell \notin \{i, j\}$.

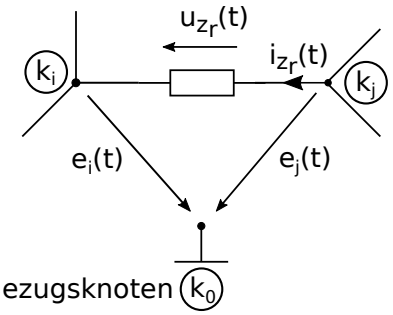


Abbildung 2.1: Zusammenhang zwischen Zweig- und Knotenspannungen.

Die (reduzierte) Inzidenzmatrix $\mathbf{A}$ wird ab nun nur mehr als Inzidenzmatrix bezeichnet, da die vollständige Inzidenzmatrix im Folgenden nicht mehr benötigt wird. Gemäß Lemma 1.5,(b) besitzt die Inzidenzmatrix $\mathbf{A}$ den Zeilenrang $(n-1)$, und daher sind alle Zeilen von $\mathbf{A}$ linear unabhängig.

Durch die gewählte Reihenfolge der gruppierten Elemente des elektrischen Netzwerkes, ist es nun möglich die Inzidenzmatrix folgendermaßen zu schreiben

$$\mathbf{A} := \left(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_L, \mathbf{A}_U, \mathbf{A}_I\right) \text{ , mit } \mathbf{A}_\ell \in \mathbb{R}^{(n-1)\times b_\ell} \text{ für } \ell \in \{R, C, L, U, I\} \text{ ,} \tag{2.1}$$

wobei für $\ell \in \{R, C, L, U, I\}$, die Matrix $\mathbf{A}_\ell$ nur dann in $\mathbf{A}$ enthalten ist, wenn $b_\ell \geq 1$ gilt. Ist beispielsweise $b_L = 0$, dann fehlt die Matrix $\mathbf{A}_L$ in $\mathbf{A}$.

UC-Schleifen und IL- Schnittmengen

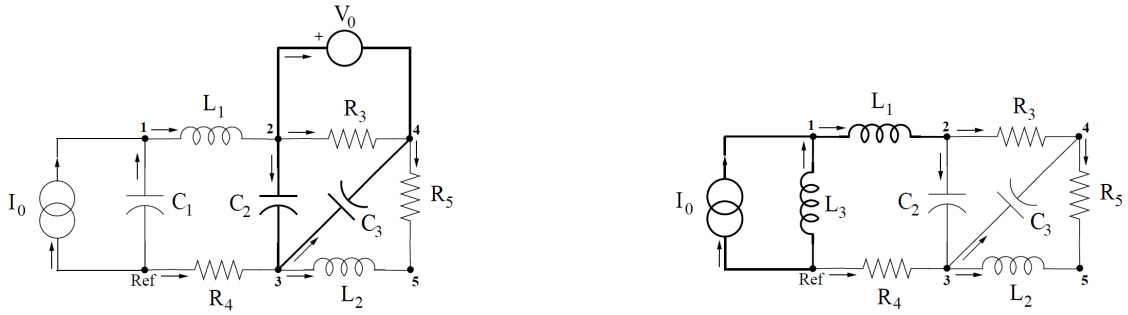
Im Folgenden werden bestimmte Schleifen bzw. Maschen, sowie Schnittmengen definiert, welche u.a. bei der Analyse des MKV für numerische Methoden einen wesentlichen Beitrag liefern.

Definition 2.1 (UC-Schleife, IL-Schnittmenge, siehe [21, Seite 6])**:**
Es sei ein elektrisches Netzwerk gegeben, welches die bisher beschriebenen Voraussetzungen aus Unterkapitel 2.1.1 erfüllt und (V, E, α) sei der zusammenhängende gerichtete Graph, welcher die Struktur des elektrischen Netzwerkes abbildet.

(a) Eine UC-Schleife, ist gemäß Definition 1.2,(c) eine Schleife bzw. Masche, welche nur Zweige von Spannungsquellen und kapazitive Zweige besitzt.

(b) Eine IL-Schnittmenge, ist gemäß Definition 1.3,(b) eine Schnittmenge, welche nur Zweige von Stromquellen und/oder induktive Zweige besitzt.

Abbildung 2.2 zeigt Beispiele für eine *UC-Schleife* und eine *IL-Schnittmenge*.



(a) *UC-Schleife*[1] (b) *IL-Schnittmenge*[2]

Abbildung 2.2: Beispiele für eine *UC-Schleife* und eine *IL-Schnittmenge*.

Gemäß Definition 2.1,(a) sind reine *U-Schleifen* oder *C-Schleifen*, also Schleifen, welche nur aus Zweigen mit Spannungsquellen oder nur aus kapazitiven Zweigen bestehen, keine *UC-Schleifen*. Andererseits sind reine *I-Schnittmengen* oder *L-Schnittmengen*, also Schnittmengen, welche nur aus Zweigen mit Stromquellen oder nur aus induktiven Zweigen bestehen, gemäß Definition 2.1,(b) *IL-Schnittmengen*.
Im Folgenden wird nun zusätzlich vorausgesetzt, dass das elektrische Netzwerk keine reinen *U-Schleifen* und keine reinen *I-Schnittmengen* besitzt. Ansonsten könnten beispielsweise bei unabhängigen Quellen die Kirchhoffschen Gesetze nicht erfüllt sein. Diese Voraussetzung wurde von [21, Theorem 1.1] bzw. [17, Abschnitt 5.1.2.1, Seite 206, well-posed circuit] übernommen.
Es soll nun eine Möglichkeit angegeben werden, mit der die vorhergehenden Eigenschaften, mit Hilfe der Inzidenzmatrix $\mathbf{A}$ überprüft werden können. Gemäß Lemma 1.5,(c),(i) besitzt das elektrische Netzwerk genau dann keine *UC-Schleifen*, wenn die Matrix $(\mathbf{A}_C, \mathbf{A}_U)$ vollen Spaltenrang

[1] Die Grafik von Abbildung 2.2a wurde aus [17, Abschnitt 5.4.1, Example 2, Fig. 5.2, Seite 229] entnommen.
[2] Die Grafik von Abbildung 2.2b wurde aus [17, Abschnitt 5.4.1, Example 3, Fig. 5.3, Seite 231] entnommen.

besitzt, d.h. wenn $\operatorname{rang}(\mathbf{A}_C, \mathbf{A}_U) = b_C + b_U$ gilt. Weiters gibt es genau dann keine *U-Schleifen*, wenn die Matrix $\mathbf{A}_U$ vollen Spaltenrang besitzt, d.h. wenn $\operatorname{rang}(\mathbf{A}_U) = b_U$ gilt. Gemäß Lemma 1.5,(c),(ii) besitzt das elektrische Netzwerk genau dann keine *IL-Schnittmenge*, wenn die Matrix $(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_U)$ vollen Zeilenrang besitzt, d.h. wenn $\operatorname{rang}(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_U) = n - 1$ gilt. Weiters gibt es genau dann keine *I-Schnittmenge*, wenn die Matrix $(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_L, \mathbf{A}_U)$ vollen Zeilenrang besitzt, d.h. wenn $\operatorname{rang}(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_L, \mathbf{A}_U) = n - 1$ gilt.

Kirchhoffsche Gesetze

Gemäß der Struktur der Inzidenzmatrix $\mathbf{A}$ in (2.1) werden nun zeitabhängige Spaltenvektoren definiert, welche die Zweigströme und die Zweigspannungen des elektrischen Netzwerkes enthalten

$$\mathbf{i}(t) := (\mathbf{i}_R(t)^\top, \mathbf{i}_C(t)^\top, \mathbf{i}_L(t)^\top, \mathbf{i}_U(t)^\top, \mathbf{i}_I(t)^\top)^\top \in \mathbb{R}^{b_R+b_C+b_L+b_U+b_I} = \mathbb{R}^b \,, \tag{2.2}$$

$$\mathbf{u}(t) := (\mathbf{u}_R(t)^\top, \mathbf{u}_C(t)^\top, \mathbf{u}_L(t)^\top, \mathbf{u}_U(t)^\top, \mathbf{u}_I(t)^\top)^\top \in \mathbb{R}^{b_R+b_C+b_L+b_U+b_I} = \mathbb{R}^b \,, \tag{2.3}$$

wobei für $\ell \in \{R, C, L, U, I\}$ die Vektoren $\mathbf{i}_\ell(t)$ und $\mathbf{u}_\ell(t)$ nur dann in $\mathbf{i}(t)$ und $\mathbf{u}(t)$ enthalten sind, wenn $b_\ell \geq 1$ gilt. Ist beispielsweise $b_L = 0$, dann sind die Vektoren $\mathbf{i}_L(t)$ und $\mathbf{u}_L(t)$ in $\mathbf{i}(t)$ und $\mathbf{u}(t)$ nicht enthalten.

Mit dieser Konvention der Zweigströme und Zweigspannungen können nun die Kirchhoffschen Gesetze formuliert werden, welche die Grundlage zur Analyse elektrischer Netzwerke liefert. Es handelt sich dabei um die Knotenregel und die Maschenregel. Diese Gesetzmäßigkeiten ergeben sich bei stationären oder quasistationären Vorgängen aus der Voraussetzung, dass das elektrische Feld wirbelfrei ist und aus dem Kontinuitätsgesetz für elektrische Strömungsfelder.
Die Kirchhoffsche Knotenregel liefert eine Aussage über die Ströme der Zweige, welche in einem Knoten zusammengeschaltet sind. Es gilt, dass die Summe der von einem Knoten abfließenden Ströme in jedem Zeitpunkt gleich Null ist, wobei zufließende Ströme negativ einzusetzen sind. Die Zählweise ob ein Strom zu- oder abfließend ist, ergibt sich aus der gewählten Richtung des Zweiges. Mit der Inzidenzmatrix $\mathbf{A}$ gemäß (2.1) und dem Stromvektor $\mathbf{i}(t)$ gemäß (2.2), kann die Kirchhoffsche Knotenregel auch folgendermaßen angegeben werden

$$\mathbf{A} \cdot \mathbf{i}(t) = \mathbf{0} \,. \tag{2.4}$$

Gemäß Lemma 1.5,(b) erhält man hiermit $(n-1)$ linear unabhängige Gleichungen. Die Kirchhoffsche Maschenregel bezieht sich auf die Zweigspannungen einer Schleife bzw. Masche. Es gilt, dass bei einem elektrischen Netzwerk die Summe der Zweigspannungen in einer Schleife, in jedem Zeitpunkt gleich Null ist, wobei die Zweigspannung negativ einzusetzen ist, falls die Zählpfeilrichtung der Zweigspannung nicht mit der Durchlaufrichtung der Schleife übereinstimmt. Für $1 \leq i \leq (n-1)$ sei $e_i(t)$ die zeitabhängige Knotenspannung des Knotens k_i, bezogen auf das Bezugspotential am Bezugsknoten k_0, wobei der Knotenspannungszählpfeil von $e_i(t)$ von k_i nach k_0 zeigt (siehe Abbildung 2.1 auf Seite 18). Es werden nun alle Knotenspannungen $e_i(t)$ der Knoten k_i in einem Knotenspannungsvektor $\mathbf{e}(t) \in \mathbb{R}^{n-1}$ zusammengefasst. Nachdem das elektrische Netzwerk als zusammenhängend vorausgesetzt wurde, ist es nun möglich, jede Zweigspannung im elektrischen Netzwerk, mit Hilfe der Kirchhoffschen Maschenregel zu beschreiben. Die Durchlaufrichtung der Schleife, zur Beschreibung einer Zweigspannung, stimmt hierbei mit der gewählten (Zählpfeil-) Richtung des Zweiges überein. Mit der Inzidenzmatrix $\mathbf{A}$ gemäß (2.1) und dem Spannungsvektor $\mathbf{u}(t)$ gemäß (2.3), gilt folgender Zusammenhang zwischen Zweigspannung und Knotenspannung

$$\mathbf{u}(t) - \mathbf{A}^\top \cdot \mathbf{e}(t) = \mathbf{0} \,. \tag{2.5}$$

2.1.2 Elemente im elektrischen Netzwerk

In diesem Abschnitt wird ein Überblick über die Elemente des elektrischen Netzwerkes für das Knotenspannungsverfahren gegeben. Hierbei soll mit der Berücksichtigung von nichtlinearen und zeitvarianten Elemente, sowie unabhängigen und gesteuerten Quellen ein großer Anwendungsbereich abgedeckt werden. Die Beschreibung von resistiven, kapazitiven und induktiven Elementen muss allerdings mit geeigneten explizite Gleichungen möglich sein. Dabei werden u.a. Funktionen vorausgesetzt, welche global definiert sind und z.B. Eigenschaften wie stetige Differenzierbarkeit besitzen. Dabei sei angemerkt, dass stetige Differenzierbarkeit auf einer Definitionsmenge, auch Stetigkeit impliziert. Diese Voraussetzungen gewährleisten dann auch den Einsatz geeigneter numerischer Verfahren zur Lösung des Knotenspannungsverfahrens. Bei der numerischen Lösung liegt in weiterer Folge allerdings der Fokus auf dem modifizierten Knotenspannungsverfahren.
Es sei aber darauf hingewiesen, dass in den Anwendungen auch Modelle interessant sein können, welche diese globalen Eigenschaften nicht erfüllen, wie beispielsweise Elemente dessen Kennlinien stückweise definiert sind oder Elemente bei denen nur lokale Beschreibungen existieren. Hierauf wird im Folgenden nicht weiter eingegangen und im wesentlichen müssten diese Spezialfälle einzeln analysiert werden, speziell auch hinsichtlich der Auswahl geeigneter numerischer Methoden.

Kapazitive Elemente

Die elektromagnetische Beziehung von kapazitiven Elementen, ist durch folgende Differentialgleichung gegeben

$$\mathbf{i}_C(t) = \frac{\mathrm{d}\mathbf{q}_C(t)}{\mathrm{d}t} \,, \tag{2.6}$$

wobei $\mathbf{q}_C(t)$ den Ladungen und $\mathbf{i}_C(t)$ den Strömen in den Zweigen der kapazitiven (konzentrierten) Elementen entspricht. Im Allgemeinen könnte das Verhalten kapazitiver Elemente über eine stetig differenzierbare Funktion

$$\mathbf{g}_C : \mathbb{R}^{b_C} \times \mathbb{R}^{b_C} \times \mathbb{R} \to \mathbb{R}^{b_C} : (\mathbf{q}, \mathbf{u}, t) \mapsto \mathbf{g}_C(\mathbf{q}, \mathbf{u}, t) \,,$$

d.h. $\mathbf{g}_C \in C^1(\mathbb{R}^{b_C} \times \mathbb{R}^{b_C} \times \mathbb{R}, \mathbb{R}^{b_C})$ (siehe Satz 1.8) und die folgende implizite Gleichung

$$\mathbf{g}_C(\mathbf{q}_C(t), \mathbf{u}_C(t), t) = \mathbf{0} \,, \tag{2.7}$$

beschrieben werden. Im Folgenden wird nun aber vorausgesetzt, dass die Ladung der kapazitiven Zweige, durch eine Funktion dargestellt werden kann, welche abhängig von den kapazitiven Zweigspannungen und der Zeit ist. D.h. es existiere eine stetig differenzierbare Funktion $\boldsymbol{\gamma}_C \in C^1(\mathbb{R}^{b_C} \times \mathbb{R}, \mathbb{R}^{b_C})$, mit

$$\boldsymbol{\gamma}_C : \mathbb{R}^{b_C} \times \mathbb{R} \to \mathbb{R}^{b_C} : (\mathbf{u}, t) \mapsto \boldsymbol{\gamma}_C(\mathbf{u}, t) \,,$$

und es gilt

$$\mathbf{q}_C(t) = \boldsymbol{\gamma}_C(\mathbf{u}_C(t), t) \,. \tag{2.8}$$

Für Knotenspannungsverfahren wird eine explizite Darstellung gemäß Gleichung (2.8) benötigt. Falls eine implizite Darstellung gemäß Gleichung (2.7) existiert, so sei darauf hingewiesen, dass bei Erfüllung der Voraussetzungen von Satz 1.11, d.h. für einen Punkt $(\mathbf{q}_0^\top, \mathbf{u}_0^\top, t_0)^\top \in \mathbb{R}^{b_C} \times \mathbb{R}^{b_C} \times \mathbb{R}$ gelte insbesondere $\mathbf{g}_C(\mathbf{q}_0, \mathbf{u}_0, t_0) = \mathbf{0}$ und $\det(\mathbf{J}_{\mathbf{q};\mathbf{g}_C}(\mathbf{q}_0, \mathbf{u}_0, t_0)) \neq 0$, zumindest eine lokale Darstellung $\mathbf{q} = \boldsymbol{\gamma}_{C,loc}(\mathbf{u}, t)$ um den Punkt $(\mathbf{q}_0^\top, \mathbf{u}_0^\top, t_0)^\top$ existiert.
Wird nun Gleichung (2.8) in Gleichung (2.6) eingesetzt und definiert man $\mathbf{f}(t) = (\mathbf{u}_C(t)^\top, t)^\top$,

dann führt die Anwendung der Kettenregel (gemäß Satz 1.9) auf $\boldsymbol{\gamma}_C(\mathbf{f}(t))$ und die Verwendung der partiellen Jacobi-Matrizen (gemäß Definition 1.10), zu folgender Gleichung

$$\begin{aligned}\mathbf{i}_C(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}_C(\mathbf{u}_C(t),t)\\ &= \mathbf{J}_{\boldsymbol{\gamma}_C}(\mathbf{f}(t))\cdot\mathbf{J}_{\mathbf{f}}(t)\\ &= \mathbf{J}_{\mathbf{u},\boldsymbol{\gamma}_C}(\mathbf{u}_C(t),t)\cdot\frac{\mathrm{d}\mathbf{u}_C(t)}{\mathrm{d}t}+\mathbf{J}_{t,\boldsymbol{\gamma}_C}(\mathbf{u}_C(t),t)\cdot 1\\ &= \mathbf{C}(\mathbf{u}_C(t),t)\cdot\frac{\mathrm{d}\mathbf{u}_C(t)}{\mathrm{d}t}+\mathbf{J}_{t,\boldsymbol{\gamma}_C}(\mathbf{u}_C(t),t)\ .\end{aligned} \tag{2.9}$$

Hierbei kann $\mathbf{J}_{t,\boldsymbol{\gamma}_C}(\mathbf{u}_C(t),t)\neq\mathbf{0}$ nur auftreten, wenn es zeitvariante kapazitive Elemente im elektrischen Netzwerk gibt. Weiters wird die Matrix $\mathbf{C}(\mathbf{u},t):=\mathbf{J}_{\mathbf{u},\boldsymbol{\gamma}_C}(\mathbf{u},t)$ in Gleichung (2.9) als Kapazitätsmatrix bezeichnet und es gilt

$$\mathbf{C}(\mathbf{u},t):=\begin{pmatrix}\frac{\partial\gamma_{C,1}}{\partial u_1}(\mathbf{u},t) & \dots & \frac{\partial\gamma_{C,1}}{\partial u_{b_C}}(\mathbf{u},t)\\ \vdots & & \vdots\\ \frac{\partial\gamma_{C,b_C}}{\partial u_1}(\mathbf{u},t) & \dots & \frac{\partial\gamma_{C,b_C}}{\partial u_{b_C}}(\mathbf{u},t)\end{pmatrix}\in\mathbb{R}^{b_C\times b_C}\ . \tag{2.10}$$

Wenn die kapazitiven Elemente im elektrischen Netzwerk ungekoppelt sind, dann ist die Kapazitätsmatrix $\mathbf{C}(\mathbf{u},t)$ eine Diagonalmatrix. Falls die Beschreibung der kapazitiven Elemente gemäß Gleichung (2.8) linear und zeitinvariant ist und somit $\boldsymbol{\gamma}_C(\mathbf{u},t)=\boldsymbol{\gamma}_C(\mathbf{u})=\mathbf{C}\cdot\mathbf{u}$ gilt, dann ist die Kapazitätsmatrix $\mathbf{C}(\mathbf{u},t)=\mathbf{C}\in\mathbb{R}^{b_C\times b_C}$ konstant.

Induktive Elemente

Abgesehen von unterschiedlichen Indizes und Variablenbezeichnungen stimmen die mathematischen Eigenschaften und Überlegungen von kapazitiven und induktiven Elementen überein. Insbesondere gilt dies auch für die Beschreibung von induktiven Elementen durch implizite Gleichungen. Im Folgenden werden allerdings nur jene Eigenschaften von induktiven Elementen angegeben, welche für Knotenspannungsverfahren benötigt werden. Im Anschluss werden als Anwendungsbeispiele, die Gleichungen von zwei gekoppelten Spulen angegeben, sowie die Approximation eines idealen Übertragers gezeigt, da für Knotenspannungsverfahren die Implementierung eines idealen Übertragers mit Kopplungsfaktor 1 nicht möglich ist.

Die elektromagnetische Beziehung von induktiven Elementen, ist durch folgende Differentialgleichung gegeben

$$\mathbf{u}_L(t)=\frac{\mathrm{d}\boldsymbol{\phi}_L(t)}{\mathrm{d}t}\ , \tag{2.11}$$

wobei $\boldsymbol{\phi}_L(t)$ dem magnetischen Fluss und $\mathbf{u}_L(t)$ den Spannungen in den Zweigen der induktiven (konzentrierten) Elementen entspricht. Im Folgenden wird nun vorausgesetzt, dass der magnetische Fluss der induktiven Zweige, durch eine Funktion dargestellt werden kann, welche abhängig von den induktiven Zweigströmen und der Zeit ist. D.h. es existiere eine stetig differenzierbare Funktion $\boldsymbol{\gamma}_L\in C^1(\mathbb{R}^{b_L}\times\mathbb{R},\mathbb{R}^{b_L})$, mit

$$\boldsymbol{\gamma}_L:\mathbb{R}^{b_L}\times\mathbb{R}\to\mathbb{R}^{b_L}:(\mathbf{i},t)\mapsto\boldsymbol{\gamma}_L(\mathbf{i},t)\ ,$$

und es gilt

$$\boldsymbol{\phi}_L(t)=\boldsymbol{\gamma}_L(\mathbf{i}_L(t),t)\ . \tag{2.12}$$

Wird nun Gleichung (2.12) in Gleichung (2.11) eingesetzt, so erhält man

$$\begin{aligned}\mathbf{u}_L(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}_L(\mathbf{i}_L(t),t)\\ &= \mathbf{J}_{\mathbf{i},\boldsymbol{\gamma}_L}(\mathbf{i}_L(t),t)\cdot\frac{\mathrm{d}\mathbf{i}_L(t)}{\mathrm{d}t} + \mathbf{J}_{t,\boldsymbol{\gamma}_L}(\mathbf{i}_L(t),t)\cdot 1\\ &= \mathbf{L}(\mathbf{i}_L(t),t)\cdot\frac{\mathrm{d}\mathbf{i}_L(t)}{\mathrm{d}t} + \mathbf{J}_{t,\boldsymbol{\gamma}_L}(\mathbf{i}_L(t),t)\,. \end{aligned} \tag{2.13}$$

Hierbei kann $\mathbf{J}_{t,\boldsymbol{\gamma}_L}(\mathbf{i}_L(t),t) \neq \mathbf{0}$ nur auftreten, wenn es zeitvariante induktive Elemente im elektrischen Netzwerk gibt. Weiters wird die Matrix $\mathbf{L}(\mathbf{i},t) := \mathbf{J}_{\mathbf{i},\boldsymbol{\gamma}_L}(\mathbf{i},t)$ in Gleichung (2.13) als Induktivitätsmatrix bezeichnet und es gilt

$$\mathbf{L}(\mathbf{i},t) := \begin{pmatrix} \frac{\partial\gamma_{L,1}}{\partial i_1}(\mathbf{i},t) & \dots & \frac{\partial\gamma_{L,1}}{\partial i_{b_L}}(\mathbf{i},t)\\ \vdots & & \vdots\\ \frac{\partial\gamma_{L,b_L}}{\partial i_1}(\mathbf{i},t) & \dots & \frac{\partial\gamma_{L,b_L}}{\partial i_{b_L}}(\mathbf{i},t)\end{pmatrix} \in \mathbb{R}^{b_L\times b_L}\,. \tag{2.14}$$

Wenn die induktiven Elemente im elektrischen Netzwerk ungekoppelt sind, dann ist die Induktivitätsmatrix $\mathbf{L}(\mathbf{i},t)$ eine Diagonalmatrix. Falls die Beschreibung der induktiven Elemente gemäß Gleichung (2.12) linear und zeitinvariant ist, und somit $\boldsymbol{\gamma}_L(\mathbf{i},t) = \boldsymbol{\gamma}_L(\mathbf{i}) = \mathbf{L}\cdot\mathbf{i}$ gilt, dann ist die Induktivitätsmatrix $\mathbf{L}(\mathbf{i},t) = \mathbf{L} \in \mathbb{R}^{b_L\times b_L}$ konstant.

Linearer Übertrager:

Für gekoppelte Induktivitäten ist folgende Beziehung zwischen Strom und Spannung bekannt (für eine physikalische Begründung der Bauteilbeziehungen sei z.B. auf [23, Abschnitt 7.2] verwiesen):

$$u_{L_1}(t) = L_1\cdot\frac{\mathrm{d}i_{L_1}(t)}{\mathrm{d}t} + M\cdot\frac{\mathrm{d}i_{L_2}(t)}{\mathrm{d}t}\,, \tag{2.15a}$$

$$u_{L_2}(t) = M\cdot\frac{\mathrm{d}i_{L_1}(t)}{\mathrm{d}t} + L_2\cdot\frac{\mathrm{d}i_{L_2}(t)}{\mathrm{d}t}\,. \tag{2.15b}$$

Hierbei entsprechen L_1, L_2 den Selbstinduktivitäten der Spulen 1 und 2 und M der Gegeninduktivität des Spulenpaares. Die Gleichungen gelten unter den Zählpfeilen, welche in Abbildung 2.3 angegeben sind. Die Punkte bei den Spulen in Abbildung 2.3 geben eine Information über den gemeinsamen Bezugssinn der Verkettungsflüsse, welcher vom Wicklungssinn der Spulen abhängt. In den Gleichungen (2.15a) und (2.15b) gilt $M>0$ bei gleichem Wicklungssinn 2.3a und $M<0$ bei entgegengesetztem Wicklungssinn 2.3b. Sollte bei einer Schaltung in Abbildung 2.3 ein Zählpfeil umgedreht werden, so muss bei dieser Größe in den Gleichungen (2.15a) und (2.15b) das Vorzeichen geändert werden.

(a) Gleicher Wicklungssinn (b) Entgegengesetzter Wicklungssinn

Abbildung 2.3: Schaltzeichen und Zählpfeile zweier gekoppelter Induktivitäten.

Wie in [23, Abschnitt 3.1.3.1] erläutert, gelten für die Selbstinduktivitäten $L_1, L_2 \geq 0$ und für die Gegeninduktivität $M^2 \leq L_1 \cdot L_2$. Gemäß [23, Abschnitt 7.2.4.1] wird der Kopplungsfaktor k durch die Gleichung $M = k \cdot \sqrt{L_1 \cdot L_2}$ definiert und es gilt $|k| \leq 1$. Wird in den Gleichungen (2.15a) und (2.15b) die Beziehung (2.11) eingesetzt und nach der Variable t integriert (mit Integrationskonstante Null), dann erhält man folgende Darstellung gemäß der Beziehung (2.12)

$$\phi_{L_1}(t) = L_1 \cdot i_{L_1}(t) + M \cdot i_{L_2}(t) \ , \tag{2.16a}$$
$$\phi_{L_2}(t) = M \cdot i_{L_1}(t) + L_2 \cdot i_{L_2}(t) \ . \tag{2.16b}$$

Approximation des idealen Übertragers:

Gemäß [23, Abschnitt 3.1.3] bzw. [23, Abschnitt 7.2] können aus den Gleichungen (2.15a) und (2.15b), die nachfolgenden Gleichungen des idealen Übertragers hergeleitet werden

$$u_{L_1}(t) = \ddot{u} \cdot u_{L_2}(t) \ , \tag{2.17a}$$
$$i_{L_1}(t) = -\frac{1}{\ddot{u}} \cdot i_{L_2}(t) \ , \tag{2.17b}$$

wobei $\ddot{u} = \pm \sqrt{\frac{L_1}{L_2}}$ dem Übertragungsverhältnis entspricht, mit dem selben Vorzeichen wie M.
Die algebraischen Gleichungen (2.17a) und (2.17b), entsprechen gemäß Gleichung (2.18), einem resistiven Element in impliziter Darstellung. Nachdem bei Knotenspannungsverfahren, für resistive Elemente, eine Beschreibung durch explizite spannungsgesteuerte algebraische Gleichungen benötigt wird, kann der ideale Übertrager nicht als resistives Element implementiert werden.
Ein idealer Übertrager kann aber approximiert werden, wenn $M^2 = L_1 \cdot L_2$ gilt und wenn bei gleichbleibenden Übertragungsverhältnis $\ddot{u}$, $\sqrt{L_1 \cdot L_2} \gg 1$ gilt. Gemäß [23, Abschnitt 3.1.3.1] würde unter diesen Bedingungen, der ideale Übertrager folgen, wenn der Grenzwert $\sqrt{L_1 \cdot L_2} \to \infty$ gebildet wird.

Resistive Elemente

Als resistive Elemente werden jene Elemente im elektrischen Netzwerk bezeichnet, welche eine *algebraische* Beschreibung zwischen resistiven Zweigspannungen und Zweigströmen besitzen (algebraisch bedeutet insbesondere, dass die Beschreibung keine Differentialquotienten enthält). Im Allgemeinen könnte das Verhalten resistiver Elemente über eine stetig differenzierbare Funktion $\mathbf{g}_R \in C^1(\mathbb{R}^{b_R} \times \mathbb{R}^{b_R} \times \mathbb{R}, \mathbb{R}^{b_R})$ und die folgende implizite Gleichung

$$\mathbf{g}_R(\mathbf{u}_R(t), \mathbf{i}_R(t), t) = \mathbf{0} \ , \tag{2.18}$$

beschrieben werden. In Knotenspannungsverfahren wird hingegen eine Beschreibung der resistiven Elemente vorausgesetzt, bei der die resistiven Zweigströme von den resistiven Zweigspannungen und der Zeit abhängen. D.h. es existiere eine stetig differenzierbare Funktion $\boldsymbol{\gamma}_R \in C^1(\mathbb{R}^{b_R} \times \mathbb{R}, \mathbb{R}^{b_R})$ mit

$$\boldsymbol{\gamma}_R : \mathbb{R}^{b_R} \times \mathbb{R} \to \mathbb{R}^{b_R} : (\mathbf{u}, t) \mapsto \boldsymbol{\gamma}_R(\mathbf{u}, t) \ ,$$

und es gilt

$$\mathbf{i}_R(t) = \boldsymbol{\gamma}_R(\mathbf{u}_R(t), t) \ . \tag{2.19}$$

Durch eine explizite Darstellung der resistiven Elemente gemäß Gleichung (2.19), können nun beispielsweise lineare ohmsche Widerstände, ein Gyrator mit Gyratorkonstante $g > 0$ der Form

$$i_{Gyr_1}(t) = g \cdot u_{Gyr_2}(t) \ ,$$

$$i_{Gyr_2}(t) = -\,g \cdot u_{Gyr_1}(t)\ ,$$

oder Dioden, welche durch die Shockley Gleichung mit reellen Konstanten I_0, U_0

$$i_D(t) = I_0 \cdot \left(\exp\left(\frac{u_D(t)}{U_0}\right) - 1\right)$$

definiert sind, berücksichtigt werden.
Bei der Anwendung von numerischen Methoden auf das modifizierte Knotenspannungsverfahren zeigt sich, dass die partielle Jacobi-Matrix $\mathbf{J}_{\mathbf{u},\boldsymbol{\gamma}_R}(\mathbf{u},t)$ benötigt wird (siehe Kapitel 3). Dies motiviert zur Definition der Leitwertsmatrix $\mathbf{G}(\mathbf{u},t) := \mathbf{J}_{\mathbf{u},\boldsymbol{\gamma}_R}(\mathbf{u},t)$, mit der Darstellung

$$\mathbf{G}(\mathbf{u},t) := \begin{pmatrix} \frac{\partial \gamma_{R,1}}{\partial u_1}(\mathbf{u},t) & \dots & \frac{\partial \gamma_{R,1}}{\partial u_{b_R}}(\mathbf{u},t) \\ \vdots & & \vdots \\ \frac{\partial \gamma_{R,b_R}}{\partial u_1}(\mathbf{u},t) & \dots & \frac{\partial \gamma_{R,b_R}}{\partial u_{b_R}}(\mathbf{u},t) \end{pmatrix} \in \mathbb{R}^{b_R \times b_R}\ . \tag{2.20}$$

Wenn die Beschreibung der resistiven Elemente gemäß Gleichung (2.19) linear ist, dann ist die Leitwertsmatrix $\mathbf{G}(\mathbf{u},t) = \mathbf{G} \in \mathbb{R}^{b_R \times b_R}$ konstant. Die Leitwertsmatrix $\mathbf{G}(\mathbf{u},t)$ ist zudem eine Diagonalmatrix, wenn die Leitwerte im elektrischen Netzwerk ungekoppelt sind.

Strom- und Spannungsquellen

In Knotenspannungsverfahren können unabhängige oder gesteuerte Strom- und Spannungsquellen berücksichtigt werden. Für den allgemeinen Fall der gesteuerten Quellen, könnten diese im Allgemeinen von jedem Zweigstrom- oder Zweigspannung abhängig sein. Die Abhängigkeit der Strom- und Spannungsquellen sei hierbei durch die Funktionen

$$\begin{aligned} \mathbf{u}_Q &: \mathbb{R}^b \times \mathbb{R}^{b_C} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \times \mathbb{R} \to \mathbb{R}^{b_U} : (\mathbf{u}, \mathbf{i}_C, \mathbf{i}_L, \mathbf{i}_U, t) \mapsto \mathbf{u}_Q(\mathbf{u}, \mathbf{i}_C, \mathbf{i}_L, \mathbf{i}_U, t)\ , \\ \mathbf{i}_Q &: \mathbb{R}^b \times \mathbb{R}^{b_C} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \times \mathbb{R} \to \mathbb{R}^{b_I} : (\mathbf{u}, \mathbf{i}_C, \mathbf{i}_L, \mathbf{i}_U, t) \mapsto \mathbf{i}_Q(\mathbf{u}, \mathbf{i}_C, \mathbf{i}_L, \mathbf{i}_U, t) \end{aligned}$$

beschrieben. Es sei darauf hingewiesen, dass eine Abhängigkeit von Strömen aus resistiven Zweigen oder von Zweigen mit Stromquellen, durch die angegebenen Variablen, auch berücksichtigt wird. Sollten in einem elektrischen Netzwerk keine gesteuerten Quellen vorhanden sein, dann kann die Quellenbeschreibung vereinfacht durch zeitabhängige Funktionen $\mathbf{u}_Q : \mathbb{R} \to \mathbb{R}^{b_U} : t \mapsto \mathbf{u}_Q(t)$ und $\mathbf{i}_Q : \mathbb{R} \to \mathbb{R}^{b_I} : t \mapsto \mathbf{i}_Q(t)$ dargestellt werden.
Mit der Beschreibung der Quellen folgt dann im allgemeinen Fall

$$\mathbf{u}_U(t) = \mathbf{u}_Q(\mathbf{u}(t), \mathbf{i}_C(t), \mathbf{i}_L(t), \mathbf{i}_U(t), t)\ , \tag{2.21}$$

$$\mathbf{i}_I(t) = \mathbf{i}_Q(\mathbf{u}(t), \mathbf{i}_C(t), \mathbf{i}_L(t), \mathbf{i}_U(t), t)\ . \tag{2.22}$$

Gewöhnlich hängt eine gesteuerte Strom- oder Spannungsquelle nur von einer Variable ab. In diesem Fall unterscheidet man zwischen spannungsgesteuerten Spannungsquellen (engl.: voltage-controlled voltage sources, Abk.: VCVS), stromgesteuerten Spannungsquellen (engl.: current-controlled voltage sources, Abk.: CCVS), spannungsgesteuerten Stromquellen (engl.: voltage-controlled current sources, Abk.: VCCS) und stromgesteuerten Stromquellen (engl.: current-controlled current sources, Abk.: CCCS).

Passivität

Die nachfolgende Definition definiert den Begriff der positiven Definitheit für quadratische Matrizen. Dabei sei hingewiesen, dass in dieser Definition, die Matrix nicht symmetrisch sein muss.

Definition 2.2 (positive Definitheit)**:**
Seien $m, n \in \mathbb{N}$.

(a) Eine Matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$ *heißt genau dann positiv definit, wenn für alle* $\mathbf{v} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$

$$\mathbf{v}^\top \cdot \mathbf{M} \cdot \mathbf{v} > 0$$

gilt.

(b) Sei $D \subseteq \mathbb{R}^n$ *und* $\mathbf{M} : D \to \mathbb{R}^{m\times m} : \mathbf{x} \mapsto \mathbf{M}(\mathbf{x})$ *eine von dem Parameter* $\mathbf{x} \in D$ *abhängige Matrix* $\mathbf{M}(\mathbf{x})$. $\mathbf{M}$ *ist genau dann positiv definit, wenn für alle* $\mathbf{x} \in D$ *und für alle* $\mathbf{v} \in \mathbb{R}^m \setminus \{\mathbf{0}\}$

$$\mathbf{v}^\top \cdot \mathbf{M}(\mathbf{x}) \cdot \mathbf{v} > 0$$

gilt (d.h. wenn $\mathbf{M}(\mathbf{x})$ *punktweise positiv definit ist).*

Die kapazitiven, induktiven oder resistiven Elemente im elektrischen Netzwerk werden *strikt lokal passiv* genannt, wenn die den Elementen entsprechende Kapazitiäts-, Induktivitäts- oder Leitwertsmatrix gemäß Definition 2.2,(b) positiv definit sind. Die physikalischen Eigenschaften, welche sich aus der Definitheit oder anderen Voraussetzungen ergeben (wie z.B. Reziprozität), können in [7] nachgelesen werden. Es sei auch darauf hingewiesen, dass aus der Eigenschaft der positiven Definitheit einer Matrix, auch stets die Invertierbarkeit der Matrix folgt. Für eine positive definite Matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$, gilt nämlich für dessen Kern, $\ker(\mathbf{M}) = \{\mathbf{x} \in \mathbb{R}^m : \mathbf{M} \cdot \mathbf{x} = \mathbf{0}_m\} = \{\mathbf{0}_m\}$.

2.2 Modifiziertes Knotenspannungsverfahren (MKV)

Mit den Grundlagen aus Unterkapitel 2.1 ist es nun möglich das MKV zu formulieren. Das MKV zeichnet sich vorallem dadurch aus, dass eine große Klasse elektrischer Netzwerke schnell und automatisiert modelliert werden kann. Ein Hauptanwendungsbereich ist beispielsweise die Simulation integrierter Schaltungen. Den Kompromiss den man dabei eingeht ist, dass das MKV differentiell-algebraische Gleichungen (DAG) liefert und hierfür geeignete numerische Verfahren benötigt werden. Alternativ wäre es auch möglich viele Anwendungen elektrischer Netzwerke, mit Hilfe der Graphentheorie durch explizite Differentialgleichungssysteme zu modellieren. Explizite Differentialgleichungssysteme sind numerisch einfacher zu lösen, allerdings kann diese Modellierung des elektrischen Netzwerkes, bei Vorhandensein von nichtlinearen Elemente, im Allgemeinen nicht automatisiert werden.

2.2.1 Herleitung

Die Grundlagen dieses Unterkapitels wurden aus [17, Abschnitt 5.2] und [21] entnommen. In [17] werden zudem auch weitere Knotenspannungsverfahren analysiert, wie z.B. die *Node Tableau Analysis* und die *Augmented Nodal Analysis*.

Zusammenfassung der Voraussetzungen und Notationen zur Erstellung des MKV

Zunächst werden die wesentlichen Voraussetzungen und Notationen aus Unterkapitel 2.1.1 zusammengefasst. Gemäß Abschnitt '*Allgemeine Voraussetzungen*' sei ein zusammenhängendes elektrisches Netzwerk, mit b Zweigen $\{z_1, \ldots, z_b\}$ ohne Eigenschleifen, und n Knoten $\{k_0, k_1, \ldots, k_{n-1}\}$ gegeben, wobei k_0 der Bezugsknoten sei. Hierbei gelten zudem die Nummerierungen und die Bezeichnungen gemäß Abschnitt '*Konvention zur Nummerierung der Knoten und Zweige*', d.h. insbesondere besitze das elektrische Netzwerk b_R resistive Zweige, b_C kapazitive Zweige, b_L induktive

Zweige, b_U Zweige mit gesteuerten und ungesteuerten Spannungsquellen und b_I Zweige mit gesteuerten und ungesteuerten Stromquellen, wobei $b = b_R + b_C + b_L + b_U + b_I$ gelte. Gemäß dem Abschnitt '*Inzidenzmatrix*' erhält die Inzidenzmatrix somit folgende Darstellung (siehe (2.1))

$$\mathbf{A} := \big(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_L, \mathbf{A}_U, \mathbf{A}_I\big) \text{ , mit } \mathbf{A}_\ell \in \mathbb{R}^{(n-1)\times b_\ell} \text{ für } \ell \in \{R, C, L, U, I\} \text{ .}$$

Weiters sei gemäß Abschnitt '*UC-Schleifen und IL- Schnittmengen*' vorausgesetzt, dass das elektrische Netzwerk keine reinen *U-Schleifen* und keine reinen *I-Schnittmengen* besitzt.
Gemäß dem Abschnitt '*Kirchhoffsche Gesetze*' kann nun mit den Vektoren (siehe (2.2) und (2.3))

$$\begin{aligned}\mathbf{i}(t) &:= (\mathbf{i}_R(t)^\top, \mathbf{i}_C(t)^\top, \mathbf{i}_L(t)^\top, \mathbf{i}_U(t)^\top, \mathbf{i}_I(t)^\top)^\top \in \mathbb{R}^{b_R+b_C+b_L+b_U+b_I} = \mathbb{R}^b \text{ ,}\\ \mathbf{u}(t) &:= (\mathbf{u}_R(t)^\top, \mathbf{u}_C(t)^\top, \mathbf{u}_L(t)^\top, \mathbf{u}_U(t)^\top, \mathbf{u}_I(t)^\top)^\top \in \mathbb{R}^{b_R+b_C+b_L+b_U+b_I} = \mathbb{R}^b \text{ ,}\end{aligned}$$

die Knotenregel mit $\mathbf{A}\cdot\mathbf{i}(t) = \mathbf{0}$ formuliert werden und der Zusammenhang zwischen Zweigspannungsvektor $\mathbf{u}(t)$ und Knotenspannungsvektor $\mathbf{e}(t)$ ist durch $\mathbf{u}(t) - \mathbf{A}^\top\cdot\mathbf{e}(t) = \mathbf{0}$ gegeben.
Es folgen nun noch die wesentlichen Voraussetzungen aus Unterkapitel 2.1.2. Für $\ell \in \{R, C, L\}$ seien stetig differenzierbare Funktionen $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell}\times\mathbb{R} \to \mathbb{R}^{b_\ell}$ gegeben, sodass $\mathbf{q}_C(t) = \boldsymbol{\gamma}_C(\mathbf{u}_C(t), t)$, $\boldsymbol{\phi}_L(t) = \boldsymbol{\gamma}_L(\mathbf{i}_L(t), t)$ und $\mathbf{i}_R(t) = \boldsymbol{\gamma}_R(\mathbf{u}_R(t), t)$ gelte. Weiters beschreiben zudem Funktionen $\mathbf{u}_Q$ und $\mathbf{i}_Q$ das Verhalten der Zweige mit unabhängigen und gesteuerten Spannungs- und Stromquellen.

NTA, ANA und grundlegende Gleichungen des MKV

Der Ausgangspunkt bei der Herleitung des MKV bilden die Grundgleichungen der *Node Tableau Analysis* (NTA) (2.23), wobei die Begründung der Gleichungen im Anschluss erfolgt.

$$\mathbf{0}_{b_C} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{u}_C(t), t)\big) - \mathbf{i}_C(t) \text{ ,} \tag{2.23a}$$
$$\mathbf{0}_{b_L} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_L(\mathbf{i}_L(t), t)\big) - \mathbf{u}_L(t) \text{ ,} \tag{2.23b}$$
$$\mathbf{0}_{b_R} = \mathbf{i}_R(t) - \gamma_R(\mathbf{u}_R(t), t) \text{ ,} \tag{2.23c}$$
$$\mathbf{0}_{b_U} = \mathbf{u}_U(t) - \mathbf{u}_Q\big(\mathbf{A}^\top\cdot\mathbf{e}(t), \mathbf{i}_C(t), \mathbf{i}_L(t), \mathbf{i}_U(t), t\big) \text{ ,} \tag{2.23d}$$
$$\mathbf{0}_{b_I} = \mathbf{i}_I(t) - \mathbf{i}_Q(\mathbf{A}^\top\cdot\mathbf{e}(t), \mathbf{i}_C(t), \mathbf{i}_L(t), \mathbf{i}_U(t), t) \text{ ,} \tag{2.23e}$$
$$\mathbf{0}_{n-1} = \mathbf{A}_R\cdot\mathbf{i}_R(t) + \mathbf{A}_L\cdot\mathbf{i}_L(t) + \mathbf{A}_C\cdot\mathbf{i}_C(t) + \mathbf{A}_U\cdot\mathbf{i}_U(t) + \mathbf{A}_I\cdot\mathbf{i}_I(t) \text{ ,} \tag{2.23f}$$
$$\mathbf{0}_{b_R} = \mathbf{u}_R(t) - \mathbf{A}_R^\top\cdot\mathbf{e}(t) \text{ ,} \tag{2.23g}$$
$$\mathbf{0}_{b_C} = \mathbf{u}_C(t) - \mathbf{A}_C^\top\cdot\mathbf{e}(t) \text{ ,} \tag{2.23h}$$
$$\mathbf{0}_{b_L} = \mathbf{u}_L(t) - \mathbf{A}_L^\top\cdot\mathbf{e}(t) \text{ ,} \tag{2.23i}$$
$$\mathbf{0}_{b_U} = \mathbf{u}_U(t) - \mathbf{A}_U^\top\cdot\mathbf{e}(t) \text{ ,} \tag{2.23j}$$
$$\mathbf{0}_{b_C} = \mathbf{u}_I(t) - \mathbf{A}_I^\top\cdot\mathbf{e}(t) \text{ .} \tag{2.23k}$$

Gleichung (2.23a) entspricht den Gleichungen der kapazitiven Elemente gemäß (2.6) und (2.8). Gleichung (2.23b) entspricht den Gleichungen der induktiven Elemente gemäß (2.11) und (2.12). Die resistiven Elemente werden in Gleichung (2.23c) gemäß (2.19) berücksichtigt. Die Gleichungen der Quellen (2.23d) und (2.23e) entsprechen (2.21) und (2.22). Gleichung (2.23f) entspricht den Kirchhoffschen Knotengleichungen gemäß (2.4) und die Gleichungen (2.23g) bis (2.23k) entsprechen dem Zusammenhang zwischen Zweig- und Knotenspannungen gemäß (2.5).
Es können nun einige Variablen des Gleichungssystems (2.23) eliminiert werden, ohne die Struktur des Modells wesentlich zu ändern. Dies führt dann zu den Grundgleichungen der *Augmented Nodal Analysis* (ANA) (2.24), wobei die Begründung der Gleichungen im Anschluss erfolgt.

$$\mathbf{0}_{b_C} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{u}_C(t), t)\big) - \mathbf{i}_C(t) \text{ ,} \tag{2.24a}$$

$$\mathbf{0}_{b_L} = \frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_L(\mathbf{i}_L(t),t)\big) - \mathbf{A}_L^\top\cdot\mathbf{e}(t)\ , \tag{2.24b}$$

$$\begin{aligned}\mathbf{0}_{n-1} = {}& \mathbf{A}_R\cdot\gamma_R(\mathbf{A}_R^\top\cdot\mathbf{e}(t),t) + \mathbf{A}_L\cdot\mathbf{i}_L(t) + \mathbf{A}_C\cdot\mathbf{i}_C(t) + \mathbf{A}_U\cdot\mathbf{i}_U(t) + \\ & + \mathbf{A}_I\cdot\mathbf{i}_Q(\mathbf{A}^\top\cdot\mathbf{e}(t),\mathbf{i}_C(t),\mathbf{i}_L(t),\mathbf{i}_U(t),t)\ ,\end{aligned} \tag{2.24c}$$

$$\mathbf{0}_{b_C} = \mathbf{u}_C(t) - \mathbf{A}_C^\top\cdot\mathbf{e}(t)\ , \tag{2.24d}$$

$$\mathbf{0}_{b_U} = \mathbf{A}_U^\top\cdot\mathbf{e}(t) - \mathbf{u}_Q\big(\mathbf{A}^\top\cdot\mathbf{e}(t),\mathbf{i}_C(t),\mathbf{i}_L(t),\mathbf{i}_U(t),t\big)\ . \tag{2.24e}$$

Wird in Gleichung (2.23b) $\mathbf{u}_L(t)$ durch die Beziehung (2.23i) ersetzt, so führt dies zu Gleichung (2.24b). Werden die Beziehungen von $\mathbf{i}_R(t)$ und $\mathbf{i}_I(t)$ gemäß (2.23c) und (2.23e) in Gleichung (2.23f) eingesetzt und wird zudem $\mathbf{u}_R(t)$ durch die Beziehung (2.23c) ersetzt, so führt dies zu Gleichung (2.24c). Wird in Gleichung (2.23d) $\mathbf{u}_U(t)$ durch die Beziehung (2.23j) ersetzt, so führt dies zu Gleichung (2.24e). Gleichung (2.23k) wird nicht benötigt, da $\mathbf{u}_I(t)$ ggf. nur für gesteuerte Quellen benötigt wird und dies wird in den Funktionen $\mathbf{i}_Q$ bzw. $\mathbf{u}_Q$ durch das Argument $\mathbf{A}^\top\cdot\mathbf{e}(t)$ berücksichtigt.

Ausgehend von den Gleichungen (2.24) können nun folgende Grundgleichungen des modifizierten Knotenspannungsverfahrens abgeleitet werden. Hierfür wird zunächst in Gleichung (2.24a) $\mathbf{u}_C(t)$ durch die Beziehung (2.24d) ersetzt. Diese Gleichung wird nach $\mathbf{i}_C(t)$ umgeformt und in die Gleichungen (2.24c) und (2.24e) eingesetzt. Dies führt zu den MKV Grundgleichungen (2.25), mit den Unbekannten $(\mathbf{e}(t)^\top,\mathbf{i}_L(t)^\top,\mathbf{i}_U(t)^\top)^\top\in\mathbb{R}^{n-1+b_L+b_U}$.

$$\begin{aligned}\mathbf{0}_{n-1} = {}& \mathbf{A}_C\cdot\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\big) + \mathbf{A}_R\cdot\gamma_R(\mathbf{A}_R^\top\cdot\mathbf{e}(t),t) + \mathbf{A}_L\cdot\mathbf{i}_L(t) + \mathbf{A}_U\cdot\mathbf{i}_U(t) + \\ & + \mathbf{A}_I\cdot\mathbf{i}_Q\Big(\mathbf{A}^\top\cdot\mathbf{e}(t),\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\big),\mathbf{i}_L(t),\mathbf{i}_U(t),t\Big)\ , \\ \mathbf{0}_{b_L} = {}& \frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_L(\mathbf{i}_L(t),t)\big) - \mathbf{A}_L^\top\cdot\mathbf{e}(t)\ , \\ \mathbf{0}_{b_U} = {}& \mathbf{A}_U^\top\cdot\mathbf{e}(t) - \mathbf{u}_Q\Big(\mathbf{A}^\top\cdot\mathbf{e}(t),\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\big),\mathbf{i}_L(t),\mathbf{i}_U(t),t\Big)\ .\end{aligned} \tag{2.25}$$

Ausgehend von den Gleichungen (2.25) können nun die Gleichungen des konventionellen MKV und des ladungsorientierten MKV angegeben werden. Hierbei wird in den Anwendungen bei nichtlinearen kapazitiven oder induktiven Elementen meist das ladungsorientierte MKV bevorzugt, da bei diesem Modell u.a. das Prinzip der Ladungserhaltung besser erfüllbar ist (dies ist beispielsweise bei Switched-Capacitor Anwendungen oder bei Ladungspumpen erforderlich). Weitere Details können in [12, Kapitel I, Abschnitt 3-4] nachgelesen werden.

Konventionelles modifiziertes Knotenspannungsverfahren

Wird in den Gleichungen (2.25) die zeitliche Ableitung gemäß (2.9) und (2.13) ausgeführt, dann führt dies zum konventionellen MKV (2.26), mit den Unbekannten $(\mathbf{e}(t)^\top,\mathbf{i}_L(t)^\top,\mathbf{i}_U(t)^\top)^\top\in\mathbb{R}^{n-1+b_L+b_U}$.

$$\begin{aligned}\mathbf{0}_{n-1} = {}& \mathbf{A}_C\cdot\mathbf{C}(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\cdot\mathbf{A}_C^\top\cdot\frac{\mathrm{d}\mathbf{e}(t)}{\mathrm{d}t} + \mathbf{A}_C\cdot\mathbf{J}_{t;\gamma_C}(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t) + \mathbf{A}_R\cdot\gamma_R(\mathbf{A}_R^\top\cdot\mathbf{e}(t),t) + \\ & + \mathbf{A}_L\cdot\mathbf{i}_L(t) + \mathbf{A}_U\cdot\mathbf{i}_U(t) + \mathbf{A}_I\cdot\mathbf{i}_Q\Big(\mathbf{A}^\top\cdot\mathbf{e}(t),\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\big),\mathbf{i}_L(t),\mathbf{i}_U(t),t\Big)\ , \\ \mathbf{0}_{b_L} = {}& \mathbf{L}(\mathbf{i}_L(t),t)\cdot\frac{\mathrm{d}\mathbf{i}_L(t)}{\mathrm{d}t} + \mathbf{J}_{t;\gamma_L}(\mathbf{i}_L(t),t) - \mathbf{A}_L^\top\cdot\mathbf{e}(t)\ , \\ \mathbf{0}_{b_U} = {}& \mathbf{A}_U^\top\cdot\mathbf{e}(t) - \mathbf{u}_Q\Big(\mathbf{A}^\top\cdot\mathbf{e}(t),\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}(t),t)\big),\mathbf{i}_L(t),\mathbf{i}_U(t),t\Big)\ .\end{aligned} \tag{2.26}$$

Dabei ist gemäß (2.10) $\mathbf{C}(\mathbf{A}_C^\top \cdot \mathbf{e}(t), t) \in \mathbb{R}^{b_C \times b_C}$ die Kapazitätsmatrix, ausgewertet im Punkt $(\mathbf{A}_C^\top \cdot \mathbf{e}(t), t) \in \mathbb{R}^{b_C} \times \mathbb{R}$, sowie gemäß (2.14) $\mathbf{L}(\mathbf{i}_L(t), t) \in \mathbb{R}^{b_L \times b_L}$ die Induktivitätsmatrix, ausgewertet im Punkt $(\mathbf{i}_L(t), t) \in \mathbb{R}^{b_L} \times \mathbb{R}$. Weiters treten die Vektoren $\mathbf{J}_{t;\gamma_C}(\mathbf{A}_C^\top \cdot \mathbf{e}(t), t) \in \mathbb{R}^{b_C}$ oder $\mathbf{J}_{t;\gamma_L}(\mathbf{i}_L(t), t) \in \mathbb{R}^{b_L}$ nur auf, wenn zeitvariante Kondensatoren oder Induktivitäten im elektrischen Netzwerk vorhanden sind.
In (2.26) werden in den Funktionen $\mathbf{i}_Q$ und $\mathbf{u}_Q$, die Abhängigkeit der kapazitiven Ströme, durch $\frac{\mathrm{d}}{\mathrm{d}t}\big(\gamma_C(\mathbf{A}_C^\top \cdot \mathbf{e}(t), t)\big)$ berücksichtigt. Es sei angemerkt, dass die Abhängigkeit der kapazitiven Ströme auch durch $\mathbf{A}_C^\top \cdot \frac{\mathrm{d}\mathbf{e}(t)}{\mathrm{d}t}$ berücksichtigt werden kann, sofern man die Funktionen $\mathbf{i}_Q$ und $\mathbf{u}_Q$, gemäß den Beziehungen (2.9) und $\mathbf{u}_C(t) = \mathbf{A}_C^\top \cdot \mathbf{e}(t)$ anpasst.

Ladungsorientiertes modifiziertes Knotenspannungsverfahren

Wird das Gleichungssystem (2.25) mit den Gleichungen (2.8) und (2.12) erweitert, dann führt dies zum ladungsorientierten MKV (2.27), mit den Unbekannten $(\mathbf{e}(t)^\top, \mathbf{i}_L(t)^\top, \mathbf{i}_U(t)^\top, \mathbf{q}_C(t)^\top, \boldsymbol{\phi}_L(t)^\top)^\top \in \mathbb{R}^{n-1+b_L+b_U+b_C+b_L}$.

$$
\begin{aligned}
\mathbf{0}_{n-1} &= \mathbf{A}_C \cdot \frac{\mathrm{d}\mathbf{q}_C(t)}{\mathrm{d}t} + \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}(t), t) + \mathbf{A}_L \cdot \mathbf{i}_L(t) + \mathbf{A}_U \cdot \mathbf{i}_U(t) + \\
&\quad + \mathbf{A}_I \cdot \mathbf{i}_Q\Big(\mathbf{A}^\top \cdot \mathbf{e}(t), \frac{\mathrm{d}\mathbf{q}_C(t)}{\mathrm{d}t}, \mathbf{i}_L(t), \mathbf{i}_U(t), t\Big) , \\
\mathbf{0}_{b_L} &= \frac{\mathrm{d}\boldsymbol{\phi}_L(t)}{\mathrm{d}t} - \mathbf{A}_L^\top \cdot \mathbf{e}(t) , \\
\mathbf{0}_{b_U} &= \mathbf{A}_U^\top \cdot \mathbf{e}(t) - \mathbf{u}_Q\Big(\mathbf{A}^\top \cdot \mathbf{e}(t), \frac{\mathrm{d}\mathbf{q}_C(t)}{\mathrm{d}t}, \mathbf{i}_L(t), \mathbf{i}_U(t), t\Big) , \\
\mathbf{0}_{b_C} &= \mathbf{q}_C(t) - \gamma_C(\mathbf{A}_C^\top \cdot \mathbf{e}(t), t) , \\
\mathbf{0}_{b_L} &= \boldsymbol{\phi}_L(t) - \gamma_L(\mathbf{i}_L(t), t) .
\end{aligned}
\tag{2.27}
$$

Das ladungsorientierten MKV wird bei der numerischen Berechnung vor allem dann bevorzugt, wenn das elektrische Netzwerk nichtlineare kapazitive oder induktive Elemente besitzt. Mathematisch sind die Gleichungssysteme (2.26) und (2.27) äquivalent, es zeigt sich aber beispielsweise bei der Verwendung numerischer Methoden, dass das in vielen Anwendungen benötigte Prinzip der Ladungserhaltung, mit (2.27) besser erfüllt werden kann.

Struktur des MKV bei zeitinvarianten Elementen und unabhängigen Quellen

Die Grundgleichungen der MKV (2.25) sind in einer sehr allgemeinen Form angegeben, welche u.a. nichtlineare und zeitvariante resistive, kapazitive und induktive Elemente, sowie gesteuerte Quellen berücksichtigen. Häufig sind aber elektrische Netzwerke mit zeitinvarianten resistiven, kapazitiven und induktiven Elementen von Interesse, d.h. für $\ell \in \{R, C, L\}$ werden die Elemente durch Funktionen $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \to \mathbb{R}^{b_\ell}$ beschrieben (anstatt mit $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \times \mathbb{R} \to \mathbb{R}^{b_\ell}$). Weiters wird angenommen, dass lediglich unabhängige Quellen im elektrischen Netzwerk vorhanden sind. D.h. die Funktionen für die Quellen $\mathbf{u}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_U}$ und $\mathbf{i}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_I}$ sind lediglich zeitabhängig.
Wird mit $\mathbf{x}(t) := (\mathbf{e}(t)^\top, \mathbf{i}_L(t)^\top, \mathbf{i}_U(t)^\top)^\top \in \mathbb{R}^{n-1+b_L+b_U}$ der Vektor der Unbekannten bezeichnet, dann besitzen die Grundgleichungen der MKV (2.25) folgende kompakte Darstellung

$$
\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\mathbf{x}(t)) + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0} , \tag{2.28}
$$

mit den gegebenen Funktionen

$$
\mathbf{q} : \mathbb{R}^{n-1} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \to \mathbb{R}^{n-1+b_L+b_U} : (\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) \mapsto \mathbf{q}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \begin{pmatrix} \mathbf{A}_C \cdot \gamma_C(\mathbf{A}_C^\top \cdot \mathbf{e}) \\ \gamma_L(\mathbf{i}_L) \\ \mathbf{0}_{b_U} \end{pmatrix} ,
$$

$$\mathbf{f}:\mathbb{R}^{n-1}\times\mathbb{R}^{b_L}\times\mathbb{R}^{b_U}\to\mathbb{R}^{n-1+b_L+b_U}:(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)\mapsto\mathbf{f}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)=\begin{pmatrix}\mathbf{A}_R\cdot\gamma_R(\mathbf{A}_R^\top\cdot\mathbf{e})+\mathbf{A}_L\cdot\mathbf{i}_L+\mathbf{A}_U\cdot\mathbf{i}_U\\ -\mathbf{A}_L^\top\cdot\mathbf{e}\\ \mathbf{A}_U^\top\cdot\mathbf{e}\end{pmatrix},$$

$$\mathbf{b}:\mathbb{R}\to\mathbb{R}^{n-1+b_L+b_U}:t\mapsto\mathbf{b}(t)=\begin{pmatrix}\mathbf{A}_I\cdot\mathbf{i}_Q(t)\\ \mathbf{0}_{b_L}\\ -\mathbf{u}_Q(t)\end{pmatrix}.$$

Wird in (2.28) die Zeitableitung durchgeführt, so erhält man mit Hilfe der Kettenregel (gemäß Satz 1.9), folgende kompakte Darstellung des konventionellen MKV (2.26)

$$\mathbf{M}(\mathbf{x}(t))\cdot\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}+\mathbf{f}(\mathbf{x}(t))+\mathbf{b}(t)=\mathbf{0}\,, \tag{2.29}$$

wobei $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}=\frac{\mathrm{d}}{\mathrm{d}t}\big((\mathbf{e}(t)^\top,\mathbf{i}_L(t)^\top,\mathbf{i}_U(t)^\top)^\top\big)$ gelte und sich $\mathbf{M}(\mathbf{x})\in\mathbb{R}^{(n-1+b_L+b_U)\times(n-1+b_L+b_U)}$ mit Hilfe der partiellen Jacobi-Matrix (gemäß Definition 1.10) folgendermaßen berechnen lässt

$$\begin{aligned}\mathbf{M}(\mathbf{x})=\mathbf{M}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)=\mathbf{J}_{\mathbf{q}}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)&=\big(\mathbf{J}_{\mathbf{e};\mathbf{q}}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U),\mathbf{J}_{\mathbf{i}_L;\mathbf{q}}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U),\mathbf{J}_{\mathbf{i}_U;\mathbf{q}}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)\big)\\&=\begin{pmatrix}\mathbf{A}_C\cdot\mathbf{C}(\mathbf{A}_C^\top\cdot\mathbf{e})\cdot\mathbf{A}_C^\top & \mathbf{0}_{(n-1)\times b_L} & \mathbf{0}_{(n-1)\times b_U}\\ \mathbf{0}_{b_L\times(n-1)} & \mathbf{L}(\mathbf{i}_L) & \mathbf{0}_{b_L\times b_U}\\ \mathbf{0}_{b_U\times(n-1)} & \mathbf{0}_{b_U\times b_L} & \mathbf{0}_{b_U\times b_U}\end{pmatrix}.\end{aligned}$$

Die Gleichungsstruktur von (2.29) entspricht somit jener, einer quasilinearen DAG gemäß (1.12).

2.2.2 DAG Index für MKV Gleichungen

Der nachfolgende Satz liefert eine Aussage darüber, unter welchen Voraussetzungen das konventionelle MKV und das ladungsorientierte MKV einen (*Differentiation-*)*Index* ≤ 2 garantieren (siehe Definition 1.15). In [21, Abschnitt 2] wird auch darauf hingewiesen, dass für den Satz der *Tractability-Index* mit dem *Differentiation-Index* übereinstimmt.

Satz 2.3 (Index von MKV, siehe [17, Abschnitt 5.4.2, Theorem 5.2] und [21, Abschnitt 2, Theoreme 2.1 und 2.2])**:**
Es sei ein elektrisches Netzwerk gegeben, welches die Voraussetzungen aus Unterkapitel 2.1 erfüllt (siehe auch Abschnitt 'Zusammenfassung der Voraussetzungen und Notationen zur Erstellung des MKV' in Unterkapitel 2.2.1). Dieses elektrische Netzwerk bestehe aus kapazitiven, induktiven und resistiven Elementen, sowie ungesteuerten Quellen (d.h. anstatt (2.21) *und* (2.22) *gilt* $\mathbf{u}_U(t)=\mathbf{u}_Q(t)$ *und* $\mathbf{i}_I(t)=\mathbf{i}_Q(t)$*). Die Kapazitätsmatrix* $\mathbf{C}(\mathbf{u},t)$ *(gemäß* (2.10)*), Induktivitätsmatrix* $\mathbf{L}(\mathbf{i},t)$ *(gemäß* (2.14)*) und Leitwertsmatrix* $\mathbf{G}(\mathbf{u},t)$ *(gemäß* (2.20)*) seien positiv definit (gemäß Definition 2.2). Es gelten dann folgende Eigenschaften für den (Differentiation-)Index des konventionellen MKV* (2.26) *und des ladungsorientierten MKV* (2.27) *(siehe Definition 1.15).*

(a) Das konventionelle MKV- System (2.26) *besitzt den (Differentiation-) Index 0 genau dann, wenn*
  (i) keine Spannungsquellen existieren, und
  (ii) kein kapazitiver Baum im elektrischen Netzwerk existiert, d.h. es existiert im elektrischen Netzwerk kein zusammenhängender Teilgraph ohne Schleifen, welcher nur aus kapazitiven Zweigen besteht.

Angenommen, es sei eine der Bedingungen (a),(i) oder (a),(ii) nicht erfüllt.

(b) Die MKV- Systeme (2.26) *und* (2.27) *besitzen den (Differentiation-)Index 1 genau dann, wenn das elektrische Netzwerk weder IL-Schnittmengen noch UC-Schleifen besitzt (siehe Definition 2.1).*

(c) Die MKV- Systeme (2.26) *und* (2.27) *besitzen den (Differentiation-)Index 2 genau dann, wenn das elektrische Netzwerk IL-Schnittmengen und/oder UC-Schleifen besitzt.*

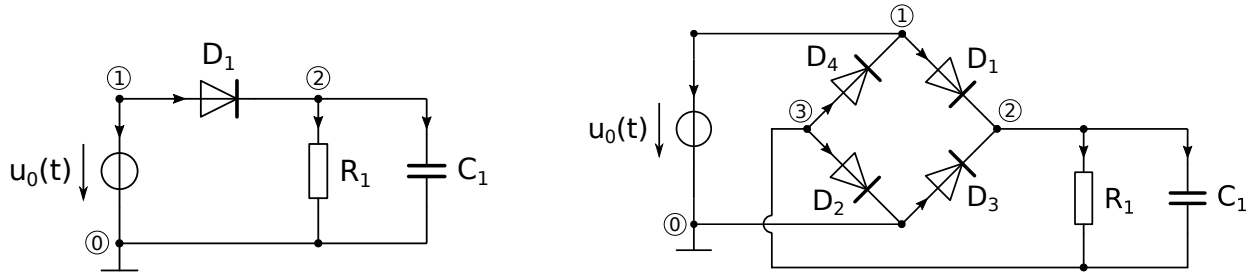
Satz 2.3 wurde für ungesteuerte Quellen formuliert, er behält aber auch für eine große Klasse von gesteuerten Quellen seine Gültigkeit, wobei diese bestimmte Eigenschaften erfüllen müssen. Für eine umfangreiche Formulierung des Satzes mit gesteuerten Quellen, sei auf [21, Abschnitt 2, Theoreme 2.1 und 2.2, Corollary 2.3] verwiesen. Dass MKV Systeme mit einem höheren Index nicht ausgeschlossen werden können, zeigen z.B. Beispiele mit Index 3 in [12, Kapitel I, Abschnitt 7-8]. Satz 2.3 und dessen umfangreiche Erweiterungen für gesteuerte Quellen können daher auch als Designregeln für elektrische Netzwerke gesehen werden, um die Konvergenz bei geeigneten numerischen Verfahren zu gewährleisten. Der Index ist hierbei eine wichtige Kennzahl bei der Anwendung numerischer Zeitschrittverfahren für DAG. Bei einem Index ≤ 2 gibt es zuverlässige numerische Verfahren, bei einem Index > 2 kann eine Konvergenz numerischer Verfahren im Allgemeinen nicht gewährleistet werden (siehe Unterkapitel 1.4.2). Es sei auch darauf hingewiesen, dass es bei einem Index ≥ 2 zudem schwieriger wird, konsistente Anfangsbedingungen anzugeben (siehe Unterkapitel 1.4.1). In einem Simulationsprogramm sind neben den Integrationsverfahren und Schrittweitensteuerungen, vor allem Konzepte zur Indexreduktion und die Ermittlung von konsistenten Anfangsbedingungen wesentliche Bestandteile Eine ausführlichere Übersicht hierzu, wird z.B. in [12] gegeben.

2.2.3 Beispiele

In diesem Abschnitt wird für den Einweg- und den Brückengleichrichter gemäß Abbildung 2.4, das konventionelle MKV angewendet. Die Pfeile in den Schaltungen entsprechen der gewählten (Zählpfeil-) Richtung der Zweige und in der Inzidenzmatrix entspricht Zeile i dem Knoten i. Die Gleichrichterschaltungen zeichnen sich dadurch aus, dass die Elemente zeitinvariant sind und der Kondensator, sowie der Widerstand linear sind, d.h. es gelte $q_{C_1} = C_1 \cdot u_{C_1}$ und $i_{R_1} = g_1 \cdot u_{R_1}$ (bei konstanter Kapazität C_1 und konstantem Leitwert g_1). Durch die Dioden treten allerdings algebraische Nichtlinearitäten auf. Für die Modellierung der Diode wird die Shockley Gleichung

$$i_D(u_D) = I_S \cdot \left(\exp\left(\frac{u_D}{n \cdot U_T} \right) - 1 \right)$$

verwendet. Hierbei entspricht I_S dem Sättigungssperrstrom der Diode, n dem Emissionskoeffizienten und U_T der Temperaturspannung. Mit der Boltzmannkonstante k_B, der absoluten Temperatur T und der Elementarladung e gilt zudem $U_T = \frac{k_B \cdot T}{e}$.



(a) Einweggleichrichter (b) Brückengleichrichter

Abbildung 2.4: Elektrisches Netzwerk eines Einweg- und Brückengleichrichter.

Für die beiden Gleichrichterschaltungen wird nun das konventionelle MKV gemäß (2.26) erstellt. Hierbei seien die wesentlichen Bezeichnungen dem Abschnitt '*Zusammenfassung der Voraussetzungen und Notationen zur Erstellung des MKV*' in Unterkapitel 2.2.1 zu entnehmen.

Nachdem in den Beispielen zeitinvariante Bauteile vorausgesetzt werden und es nur eine unabhängige Spannungsquelle $u_0(t)$ gibt, kann das konventionelle MKV gemäß (2.29) dargestellt werden. In den nachfolgenden Beispielen sind die Unbekannten die Knotenspannungen $e_i(t)$ und der Strom der Spannungsquelle $i_0(t)$. Nachdem in den betrachteten elektrischen Netzwerken keine Induktivitäten vorhanden sind, enthält der Vektor der Unbekannten auch keine Induktivitätsströme $\mathbf{i}_L(t)$.

Beispiel 2.4 (Eingweggleichrichter gemäß Abbildung 2.4a)**:**
Es gilt $n = 3$, $b_C = 1$, $b_L = 0$, $b_R = 2$, $b_U = 1$, $b_I = 0$. *Der Vektor der Unbekannten sei* $(\mathbf{e}(t)^\top, i_0(t))^\top = (e_1(t), e_2(t), i_0(t))^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^3$ *und der Vektor der Zweigspannungen sei* $\mathbf{u}(t) = (u_{D_1}(t), u_{R_1}(t), u_{C_1}(t), u_0(t))^\top \in \mathbb{R}^4$.

$$\mathbf{A}_R = \overbrace{\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}}^{D_1 \mid R_1}, \quad \mathbf{A}_C = \overbrace{\begin{pmatrix} 0 \\ 1 \end{pmatrix}}^{C_1}, \quad \mathbf{A}_U = \overbrace{\begin{pmatrix} 1 \\ 0 \end{pmatrix}}^{u_0}, \quad \mathbf{A} = \big(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_U\big), \quad \mathbf{u}(t) = \mathbf{A}^\top \cdot \mathbf{e}(t),$$
$$\boldsymbol{\gamma}_R(\mathbf{u}) = \boldsymbol{\gamma}_R(u_1, u_2) = (i_{D_1}(u_1), g_1 \cdot u_2)^\top, \quad \mathbf{C} = C_1, \quad \mathbf{u}_Q(t) = u_0(t), \quad \mathbf{i}_U(t) = i_0(t).$$

Das konventionelle MKV (2.26) *des Eingweggleichrichters führt zur Darstellung*

$$\begin{aligned} \mathbf{A}_C \cdot \mathbf{C} \cdot \mathbf{A}_C^\top \cdot \frac{\mathrm{d}\mathbf{e}(t)}{\mathrm{d}t} + \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}(t)) + \mathbf{A}_U \cdot i_0(t) &= \mathbf{0}_2, \\ \mathbf{A}_U^\top \cdot \mathbf{e}(t) - u_0(t) &= 0. \end{aligned} \tag{2.30}$$

Die Gleichungen (2.30) *können nun gemäß* (2.29) *auch als quasilineare DAG dargestellt werden.*

$$\underbrace{\begin{pmatrix} \mathbf{A}_C \cdot \mathbf{C} \cdot \mathbf{A}_C^\top & 0 \\ \mathbf{0}_2^\top & 0 \end{pmatrix}}_{:=\mathbf{M}} \cdot \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \mathbf{e}(t) \\ i_0(t) \end{pmatrix} + \underbrace{\begin{pmatrix} \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}(t)) + \mathbf{A}_U \cdot i_0(t) \\ \mathbf{A}_U^\top \cdot \mathbf{e}(t) \end{pmatrix}}_{:=\mathbf{f}(\mathbf{e}(t), i_0(t))} + \underbrace{\begin{pmatrix} \mathbf{0}_2 \\ -u_0(t) \end{pmatrix}}_{:=\mathbf{b}(t)} = \begin{pmatrix} \mathbf{0}_2 \\ 0 \end{pmatrix} \tag{2.31}$$

Die Anwendung von Satz 2.3 aus Unterkapitel 2.2.2 liefert für (2.31) *den DAG Index 1.*

Beispiel 2.5 (Brückengleichrichter gemäß Abbildung 2.4b)**:**
Es gilt $n = 4$, $b_C = 1$, $b_L = 0$, $b_R = 5$, $b_U = 1$, $b_I = 0$. *Der Vektor der Unbekannten sei* $(\mathbf{e}(t)^\top, i_0(t))^\top = (e_1(t), e_2(t), e_3(t), i_0(t))^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^4$ *und der Vektor der Zweigspannungen sei* $\mathbf{u}(t) = (u_{D_1}(t), u_{D_2}(t), u_{D_3}(t), u_{D_4}(t), u_{R_1}(t), u_{C_1}(t), u_0(t))^\top \in \mathbb{R}^7$.

$$\mathbf{A}_R = \overbrace{\begin{pmatrix} 1 & 0 & 0 & -1 & 0 \\ -1 & 0 & -1 & 0 & 1 \\ 0 & 1 & 0 & 1 & -1 \end{pmatrix}}^{D_1 \mid D_2 \mid D_3 \mid D_4 \mid R_1}, \quad \mathbf{A}_C = \overbrace{\begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}}^{C_1}, \quad \mathbf{A}_U = \overbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}}^{u_0}, \quad \mathbf{A} = \big(\mathbf{A}_R, \mathbf{A}_C, \mathbf{A}_U\big),$$
$$\boldsymbol{\gamma}_R(\mathbf{u}) = \boldsymbol{\gamma}_R(u_1, u_2, u_3, u_4, u_5) = (i_{D_1}(u_1), i_{D_2}(u_2), i_{D_3}(u_3), i_{D_4}(u_4), g_1 \cdot u_5)^\top, \quad \mathbf{C} = C_1,$$
$$\mathbf{u}_Q(t) = u_0(t), \quad \mathbf{i}_U(t) = i_0(t), \quad \mathbf{u}(t) = \mathbf{A}^\top \cdot \mathbf{e}(t).$$

Das konventionelle MKV (2.26) *des Brückengleichrichters führt zur Darstellung*

$$\begin{aligned} \mathbf{A}_C \cdot \mathbf{C} \cdot \mathbf{A}_C^\top \cdot \frac{\mathrm{d}\mathbf{e}(t)}{\mathrm{d}t} + \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}(t)) + \mathbf{A}_U \cdot i_0(t) &= \mathbf{0}_3, \\ \mathbf{A}_U^\top \cdot \mathbf{e}(t) - u_0(t) &= 0. \end{aligned} \tag{2.32}$$

Die Gleichungen (2.32) *können nun gemäß* (2.29) *auch als quasilineare DAG dargestellt werden.*

$$\underbrace{\begin{pmatrix} \mathbf{A}_C \cdot \mathbf{C} \cdot \mathbf{A}_C^\top & 0 \\ \mathbf{0}_3^\top & 0 \end{pmatrix}}_{:=\mathbf{M}} \cdot \frac{\mathrm{d}}{\mathrm{d}t} \begin{pmatrix} \mathbf{e}(t) \\ i_0(t) \end{pmatrix} + \underbrace{\begin{pmatrix} \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}(t)) + \mathbf{A}_U \cdot i_0(t) \\ \mathbf{A}_U^\top \cdot \mathbf{e}(t) \end{pmatrix}}_{:=\mathbf{f}(\mathbf{e}(t), i_0(t))} + \underbrace{\begin{pmatrix} \mathbf{0}_3 \\ -u_0(t) \end{pmatrix}}_{:=\mathbf{b}(t)} = \begin{pmatrix} \mathbf{0}_3 \\ 0 \end{pmatrix} \tag{2.33}$$

Die Anwendung von Satz 2.3 aus Unterkapitel 2.2.2 liefert für (2.33) *den DAG Index 1.*

3 Ausgewählte Methoden zur Berechnung des eingeschwungenen Zustands

In diesem Kapitel werden drei ausgewählte numerische Verfahren vorgestellt, welche zur Ermittlung des periodisch eingeschwungenen Zustands verwendet werden können. Die Implementierung soll zudem für ein Modellproblem beschrieben werden. Als Grundlage dieses Kapitels dienten [16] und [2].

Periodisch eingeschwungener Zustand

Ziel ist es, bei periodischen Vorgängen den periodisch eingeschwungenen Zustand zu bestimmen, wobei vorausgesetzt wird, dass ein eingeschwungener Zustand existiert, also sämtliche transiente Vorgänge abgeklungen sind. Als periodisch eingeschwungener Zustand sei eine Lösung zu verstehen, dessen asymptotisches Verhalten (d.h. für $t \to \infty$), eine Linearkombination zwischen Gleichanteil, auch DC-Anteil genannt, und einer Fourierreihe entspricht. Der periodisch eingeschwungene Zustand kann z.B. eine Eigenschwingung eines Oszillators sein oder aufgrund periodischer Eingänge entstehen. Wird ein elektrisches Netzwerk durch eine Differentialgleichung beschrieben, so ist das asymptotische Verhalten im Allgemeinen, bei Vorhandensein von nichtlinearen Elementen, von der Anfangsbedingung abhängig. Zudem kann es abhängig von der Anfangsbedingung keinen, einen oder mehrere eingeschwungene Zustände geben. Im Folgenden wird aber angenommen, dass das elektrische Netzwerk '*hinreichend stabil*' ist, d.h. jede Anfangsbedingung der Differentialgleichung, soll bei periodischen Eingängen zu einem periodisch eingeschwungenen Zustand führen. Dies ist allerdings eine der wesentlichen Vorgaben, bei der Entwicklung eines elektrischen Netzwerkes, für eine bestimmte Anwendung.
Neben periodisch eingeschwungenen Zuständen sind in den Anwendungen auch quasi-periodische Vorgänge interessant. Es treten daher im elektrischen Netzwerk Frequenzen auf, welche in keinem ganzzahligen Vielfachen zu einer Grundschwingung stehen. Dies ist u.a. bei Mixern relevant. Quasi-periodische Vorgänge werden in diesem Kapitel aber nicht betrachtet. Eine ausführlichere Übersicht zum Thema eingeschwungener Zustand (engl.: steady state), kann z.B. in [16, Chapter 1, Abschnitt 2] nachgelesen werden.

Einteilung der Verfahren zur Ermittlung des eingeschwungenen Zustands

Verfahren zur Ermittlung des eingeschwungenen Zustands, können in Zeit- und Frequenzbereichsverfahren eingeteilt werden (siehe z.B. [16, Chapter 1, Abschnitt 4] oder [15]). Transiente Analyse (siehe Unterkapitel 3.1) ist ein Verfahren im Zeitbereich und man erhält den eingeschwungenen Zustand nach ausreichend großer Simulationsdauer. Treten im elektrischen Netzwerk aber sehr lange transiente Vorgänge auf (d.h. mit großen Zeitkonstanten) oder sind z.B. Eingangssignale mit weit auseinanderliegenden Frequenzen vorhanden, dann kann die Simulationszeit sehr groß werden und das Verfahren somit ineffizient. Dieser Nachteil der transienten Analyse kann durch das Zeitverfahren Shooting (siehe Unterkapitel 3.2) verbessert werden. Aufgrund der Periodizität stimmt der Anfangszustand mit dem Wert nach einer Periodendauer überein und daher wird

anstatt eines Anfangswertproblems ein Randwertproblem betrachtet. Im Shooting Verfahren werden dann die Anfangsbedingungen solange variiert, bis der Endzustand nach einer Periodendauer mit dem Anfangszustand übereinstimmt. Ein weiteres Zeitbereichsverfahren wäre beispielsweise, wenn das Randwertproblem durch finite Differenzen berechnet wird und hierbei die Periodizität berücksichtigt wird. Es sei angemerkt, dass im Shooting Verfahren oder auch bei der Lösung des Randwertproblems durch finite Differenzen, bei großer zeitlicher Auflösung, bereits für kleinere elektrische Netzwerke große lineare Gleichungssysteme zu berechnen sind. Für eine ausführlichere Betrachtung von Zeitbereichsmethoden sei z.B. auf [16, Chapter 4]) verwiesen.
Eine Alternative zu Zeitbereichsverfahren sind Verfahren im Frequenzbereich. Hierbei arbeitet man mit den Koeffizienten der Fourierreihe, der zu suchenden Lösung. Im Allgemeinen besitzt eine Fourierreihe unendlich viele Summanden. Daher wird für praktische Berechnungen lediglich eine Approximation mit einer bestimmten Anzahl an Harmonischen berechnet. Ein Vorteil von Frequenzbereichsverfahren ist, dass durch eine trigonometrische Reihe, ein periodischer Vorgang viel besser beschrieben werden kann, als durch ein Zeitschrittverfahren. Grund dafür ist, dass Zeitschrittverfahren dafür optimiert sind, polynomiale Lösungen möglichst exakt zu berechnen. Sinus und Kosinus können aber durch Polynome nur schlecht approximiert werden, wodurch kleine Schrittweiten erforderlich werden). In einem linearen elektrischen Netzwerk kann u.a. eine periodische Lösung mit Frequenzbereichsverfahren sehr schnell und effizient berechnet werden. Aufwendiger wird die Berechnung allerdings, wenn das elektrische Netzwerk nichtlineare Elemente besitzt. Einerseits gilt hierbei nicht mehr das Prinzip der Superposition und andererseits gibt es keinen bekannten Weg, wie bei einem nichtlinearen elektrischen Netzwerk, die Koffizienten der Lösung aus den Koeffizienten der Eingangssignale ermittelt werden könnten. Aufgrund von nichtlinearen Elementen treten in der Lösung auch sehr viele Mischfrequenzen auf, welche u.a. Summen- und Differenzfrequenzen der Frequenzen der Eingangssignale sind. Wenn man daher bei der Berechnung der Lösung zu wenige Harmonische berücksichtigt, kann der Fehler zwischen der berechneten Approximation und der wahren Lösung, im Allgemeinen nicht mehr vernachlässigt werden. Ein Beispiel einer Frequenzbereichsmethode ist z.B. die *Volterra Reihe*, welche im Frequenzbereich die Nichtlinearitäten berücksichtigt und bei schwach nichtlinearen Systemen verwendet werden kann. Als ein weiteres Beispiel für Frequenzbereichsverfahren kann Harmonic Balance genannt werden (siehe Unterkapitel 3.3 oder [16, Chapter 5,6], [2, Kapitel 3]). Bei Harmonic Balance werden die Nichtlinearitäten zwar im Zeitbereich berücksichtigt, allerdings findet die Berechnung ausschließlich mit Fourierkoeffizienten statt. Je nach Definition könnte man Harmonic Balance daher auch als Vertreter eines gemischten Zeit-Frequenz-Verfahrens ansehen. Weiters gibt es auch gemischte Zeit-Frequenz-Verfahren. Hierfür sei beispielsweise auf das *mixed frequency-time* Verfahren für quasi-periodische Lösungen in [16, Chapter 7] verwiesen, wobei gemäß [16, Chapter 3, Abschnitt 1, Seite 29] ein Zusammenhang zwischen periodischen und quasi-periodischen Funktionen besteht. Weiters werden in [16] zudem Methoden für periodische und quasi-periodische eingeschwungene Zustände vorgestellt, welche die Anwendung nicht nur auf konzentrierte elektrische Elemente beschreiben, sondern auch auf verteilte Systeme, wie z.B. Leitungen.

Modellproblem

Modellproblem 3.1 (Spezialfall des MKV)**:**
Es sei ein elektrisches Netzwerk mit folgenden Eigenschaften gegeben.

(i) Es gelten die grundlegenden Voraussetzungen und Notationen gemäß Abschnitt 'Zusammenfassung der Voraussetzungen und Notationen zur Erstellung des MKV' in Unterkapitel 2.2.1.

(ii) Die resistiven, kapazitiven und induktiven Elemente sind zeitinvariant, d.h. für $\ell \in \{R, C, L\}$ werden die Elemente durch Funktionen $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \to \mathbb{R}^{b_\ell}$ beschrieben.

(iii) Die induktiven und kapazitiven Elemente sind linear, d.h. es existieren Matrizen $\mathbf{C} \in \mathbb{R}^{b_C \times b_C}$ *und* $\mathbf{L} \in \mathbb{R}^{b_L \times b_L}$*, sodass* $\boldsymbol{\gamma}_C(\mathbf{u}_C) = \mathbf{C} \cdot \mathbf{u}_C$ *und* $\boldsymbol{\gamma}_L(\mathbf{i}_L) = \mathbf{L} \cdot \mathbf{i}_L$ *gelte.*

(iv) Es sind nur unabhängige periodische Quellen mit Periode $T > 0$ *im elektrischen Netzwerk vorhanden, d.h. die Funktionen für die Quellen* $\mathbf{u}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_U}$ *und* $\mathbf{i}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_I}$ *sind lediglich zeitabhängig und es gilt* $\mathbf{u}_Q(t) = \mathbf{u}_Q(t+T)$*,* $\mathbf{i}_Q(t) = \mathbf{i}_Q(t+T)$ *für alle* $t \in \mathbb{R}$*.*

Gemäß Abschnitt 'Struktur des MKV bei zeitinvarianten Elementen und unabhängigen Quellen' in Unterkapitel 2.2.1, wird mit $m := n - 1 + b_L + b_U$ *und dem Vektor der Unbekannten* $\mathbf{x}(t) := (\mathbf{e}(t)^\top, \mathbf{i}_L(t)^\top, \mathbf{i}_U(t)^\top)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m$*, das elektrische Netzwerk durch die Gleichung*

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\mathbf{x}(t)) + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m\ , \tag{3.1a}$$

beschrieben, wobei das MKV System als nicht autonom vorausgesetzt wird und somit $\mathbf{b}$ *nicht der Nullvektorfunktion entspricht, bzw.* $\mathbf{b} \neq \mathbf{0}$ *gelte. Die Funktionen* $\mathbf{q}, \mathbf{f}$ *und* $\mathbf{b}$ *sind dabei durch*

$$\mathbf{q} : \mathbb{R}^{n-1} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \to \mathbb{R}^m : (\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) \mapsto \mathbf{q}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \begin{pmatrix} \mathbf{A}_C \cdot \gamma_C(\mathbf{A}_C^\top \cdot \mathbf{e}) \\ \gamma_L(\mathbf{i}_L) \\ \mathbf{0}_{b_U} \end{pmatrix}\ ,$$

$$\mathbf{f} : \mathbb{R}^{n-1} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \to \mathbb{R}^m : (\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) \mapsto \mathbf{f}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \begin{pmatrix} \mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}) + \mathbf{A}_L \cdot \mathbf{i}_L + \mathbf{A}_U \cdot \mathbf{i}_U \\ -\mathbf{A}_L^\top \cdot \mathbf{e} \\ \mathbf{A}_U^\top \cdot \mathbf{e} \end{pmatrix}\ ,$$

$$\mathbf{b} : \mathbb{R} \to \mathbb{R}^m : t \mapsto \mathbf{b}(t) = \begin{pmatrix} \mathbf{A}_I \cdot \mathbf{i}_Q(t) \\ \mathbf{0}_{b_L} \\ -\mathbf{u}_Q(t) \end{pmatrix}\ .$$

gegeben. Wird in (3.1a) *die Kettenregel angewendet, so führt dies zu der äquivalenten Gleichung*

$$\mathbf{M} \cdot \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m\ , \tag{3.1b}$$

wobei die Matrix $\mathbf{M}$ *die folgende Darstellung besitzt*

$$\mathbf{M} = \begin{pmatrix} \mathbf{A}_C \cdot \mathbf{C} \cdot \mathbf{A}_C^\top & \mathbf{0}_{(n-1)\times b_L} & \mathbf{0}_{(n-1)\times b_U} \\ \mathbf{0}_{b_L \times (n-1)} & \mathbf{L} & \mathbf{0}_{b_L \times b_U} \\ \mathbf{0}_{b_U \times (n-1)} & \mathbf{0}_{b_U \times b_L} & \mathbf{0}_{b_U \times b_U} \end{pmatrix} \in \mathbb{R}^{m \times m}\ .$$

Mit $\mathbf{x} := (\mathbf{e}^\top, \mathbf{i}_L^\top, \mathbf{i}_U^\top)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m$ *gilt zudem für die Jacobi-Matrix* $\mathbf{J}_\mathbf{f}(\mathbf{x}) \in \mathbb{R}^{m \times m}$*,*

$$\mathbf{J}_\mathbf{f}(\mathbf{x}) = \mathbf{J}_\mathbf{f}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \left(\mathbf{J}_{\mathbf{e};\mathbf{f}}, \mathbf{J}_{\mathbf{i}_L;\mathbf{f}}, \mathbf{J}_{\mathbf{i}_U;\mathbf{f}}\right) = \begin{pmatrix} \mathbf{A}_R \cdot \mathbf{J}_{\gamma_R}(\mathbf{A}_R^\top \cdot \mathbf{e}) \cdot \mathbf{A}_R^\top & \mathbf{A}_L & \mathbf{A}_U \\ -\mathbf{A}_L^\top & \mathbf{0}_{b_L \times b_L} & \mathbf{0}_{b_L \times b_U} \\ \mathbf{A}_U^\top & \mathbf{0}_{b_U \times b_L} & \mathbf{0}_{b_U \times b_U} \end{pmatrix}\ . \tag{3.2}$$

Mit der Leitwertsmatrix $\mathbf{G}$ *gemäß* (2.20) *gilt zudem* $\mathbf{G}(\mathbf{A}_R^\top \cdot \mathbf{e}) = \mathbf{J}_{\gamma_R}(\mathbf{A}_R^\top \cdot \mathbf{e})$*.*

Wie dem Modellproblem 3.1 zu entnehmen ist, werden in diesem Kapitel lediglich eingeschwungene Zustände berücksichtigt, welche durch periodische Eingänge entstehen. Nachdem $\mathbf{b} \neq \mathbf{0}$ vorausgesetzt wird, bedeutet dies, dass nur elektrische Netzwerke betrachtet werden, welche durch nicht autonome Differentialgleichungen beschrieben werden können. Insbesondere werden hierbei keine Oszillatoren berücksichtigt, da diese durch autonome Differentialgleichungen beschrieben werden. Bei Oszillatoren ist zudem die numerische Berechnung im Allgemeinen schwieriger, vor allem auch deswegen, weil meist die Eigenfrequenz nicht bekannt ist. Voraussetzung (iii) des Modellproblems 3.1 ist keine Einschränkung der numerischen Methoden, welche in diesem Kapitel

beschrieben werden, sondern führt lediglich zu einer einfacheren Berechnung bei den vorgestellten numerischen Methoden. Bei nichtlinearen induktiven oder kapazitiven Zweigen sei aber eine Darstellung gemäß (2.27) oder (2.28), dem Modell (3.1b) zu bevorzugen. Bei (3.1b) könnte es aufgrund von numerischen Fehlern zu einer Verletzung der Ladungserhaltung kommen, bzw. könnte beim Lösen von nichtlinearen Gleichungssystemen die Jacobi-Matrix für das Newton Verfahren schwieriger zu berechnen sein. Ein weiterer wichtiger Vorteil der Darstellung gemäß (2.27) oder (2.28) ist, dass durch die Beschreibung von nichtlinearen kapazitiven Elementen durch Ladungen und Spannungen bzw. die Beschreibung von nichtlinearen induktiven Elementen durch Flüsse und Ströme, die Nichtlinearitäten algebraisch auftreten. Folglich würde deren Beschreibung keine Differentiale oder Integrale beinhalten, weshalb diese Elemente auch kein 'Gedächtnis' besitzen würden. Dies ist u.a. für viele numerische Verfahren wesentlich, da sich dadurch die Berechnungen vereinfachen. Zudem sind die vorgestellten Methoden auch nicht nur auf unabhängige Quellen beschränkt, d.h. es ist auch der Einsatz von gesteuerten Quellen möglich. Eine allgemeine Behandlung würde allerdings auf Kosten der Übersichtlichkeit gehen, welche zudem keine neue Informationen liefert, da sich bei praxisrelevanten Anwendungen, die Vorgehensweise zwischen unabhängigen und gesteuerten Quellen nicht unterscheidet. Im Modellproblem 3.1 würden im Fall von gesteuerten Quellen z.B. $\mathbf{b}(t)$ nur unabhängige Quellen beinhalten und gesteuerte Quellen würden in $\mathbf{f}(\mathbf{x})$ berücksichtigt werden. Es sei aber darauf hingewiesen, dass das Modellproblem 3.1 das elektrische Netzwerk nicht mit einer Differentialgleichung modelliert, sondern durch eine DAG. Bei allen Methoden kann daher nur empfohlen werden, den DAG Index so klein wie möglich zu halten. Durch den DAG Index wird nämlich nicht nur der Einsatz von Zeitschrittverfahren beschränkt, sondern es wird auch bei einem DAG Index ≥ 2 schwieriger konsistente Anfangsbedingungen zu ermitteln (siehe auch Unterkapitel 1.4). In Unterkapitel 2.2.2 liefert z.B. Satz 2.3 eine Einschätzung für den DAG Index des Modellproblems 3.1.
Es sei darauf hingewiesen, dass Voraussetzung (v) des Modellproblems 3.1 auch Eingangssignale mit unterschiedlicher Frequenz erlaubt, sofern die Frequenzen in einem ganzzahligen Vielfachen zueinander stehen.

3.1 Transiente Analyse

Dieser Abschnitt gibt einen kurzen Überblick über die transiente Analyse, sowie notwendige Berechnungen zur Implementierung des BDF und des Radau IIA Verfahrens, angewendet auf das Modellproblem 3.1.

Überblick

Gemäß Unterkapitel 1.4.2 werden nun Algorithmen vorgestellt, mit denen das Modellproblem 3.1 bei konstanter Schrittweite berechnet werden kann. Mit den Bezeichnungen aus dem Modellproblem 3.1, sei eine Funktion $\widetilde{\mathbf{F}}(\mathbf{v}, \mathbf{w}, t)$ folgendermaßen gegeben

$$\widetilde{\mathbf{F}} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m : (\mathbf{v}, \mathbf{w}, t) \mapsto \widetilde{\mathbf{F}}(\mathbf{v}, \mathbf{w}, t) := \mathbf{M} \cdot \mathbf{v} + \mathbf{f}(\mathbf{w}) + \mathbf{b}(t) \ , \tag{3.3}$$

sodass $\widetilde{\mathbf{F}}(\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}, \mathbf{x}(t), t) = \mathbf{0}_m$ mit Gleichung (3.1a) übereinstimmt. Diese Funktion wird u.a. in den nachfolgenden Algorithmen und Berechnungen benötigt.
Die nachfolgenden Algorithmen sollen vor allem einen Einblick in zwei Lösungsverfahren geben, mit denen im Allgemeinen DAG numerisch gelöst werden können. In professionellen Softwarepaketen sind die Integrationsmethoden zur Lösung der DAG aber nur ein kleiner Bestandteil und weitere wichtige Aufgaben eines professionellen Softwarepakts sind z.B. Indexreduktion, Schrittweitensteuerung und die Ermittlung von konsistenten Anfangszuständen. Ein Ablauf für ein professionelles Softwareprogramm sei z.B. Abbildung 3.1 zu entnehmen. Ein professionelles Softwarepaket

auf Basis des BDF Verfahrens ist z.B. DASSL (siehe [4, Kapitel 5]) und auf Basis des Radau IIA Verfahrens beispielsweise RADAU5 (siehe [13, Kapitel 10]).

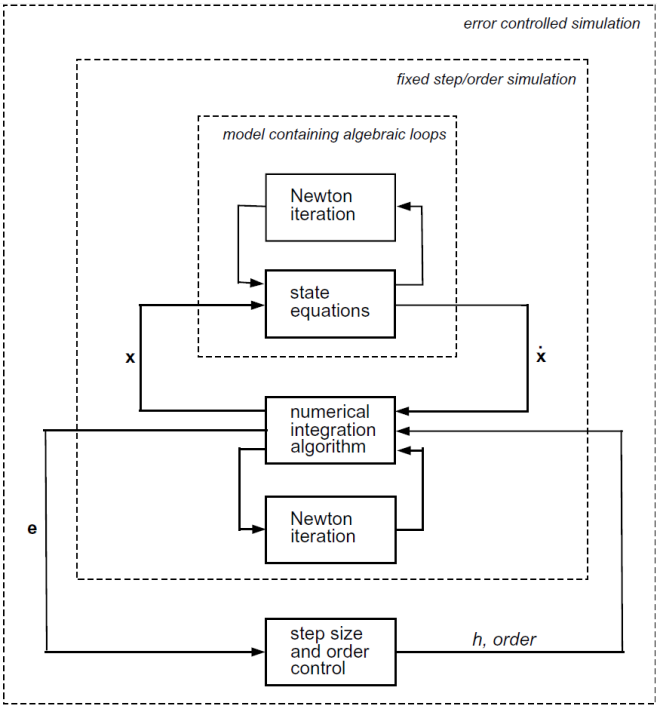


Abbildung 3.1: Ablaufdiagramm einer transienten Analyse[3].

Algorithmus zur Lösung des Modellproblems durch das BDF Verfahren

Wie bereits in Unterkapitel 1.4.2 erwähnt wurde, kann eine zuverlässige Anwendung des BDF Verfahrens nur gewährleistet werden, wenn der DAG Index des Modellproblems 3.1 höchstens 1 ist. Im Folgenden wird der Algorithmus für das BDFk Verfahren angegeben, wobei $1 \leq k \leq 6$ gelte und die Koeffizienten des BDFk Verfahrens Tabelle 1.1 zu entnehmen sind.

Algorithmus 3 : BDFk Verfahren zur Lösung des Modellproblems 3.1

Eingabe : $\widetilde{\mathbf{F}} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ gemäß (3.3), Startzeitpunkt t_0, Schrittweite $h > 0$, Schrittanzahl $N \in \mathbb{N}$ $(N \geq k)$, Startwerte $\mathbf{x}_0, \ldots \mathbf{x}_{k-1} \in \mathbb{R}^m$, wobei $\mathbf{x}_n = \mathbf{x}(t_0 + n \cdot h) = \mathbf{x}(t_n)$

Ausgabe : Zeitschritte $\mathbf{x}_k, \ldots \mathbf{x}_N$

1 **for** $n \leftarrow k-1, \ldots, N-1$ **do**

`/* Löse gemäß Tabelle 1.1 die Gleichung (1.13) nach` $\mathbf{x}_{n+1}$ `auf */`

2 $$\widetilde{\mathbf{F}}\left(\frac{1}{\beta_0 \cdot h} \cdot \sum_{i=0}^{k} \left(\alpha_i \cdot \mathbf{x}_{n+1-i}\right),\ \mathbf{x}_{n+1},\ t_{n+1}\right) = \mathbf{0}_m$$

Die nötigen Startwerte in Algorithmus 3 könnten beispielsweise mit einem Verfahren höherer Ordnung bestimmt werden, wie z.B. mit einem Radau IIA Verfahren. Alternativ könnte auch die Startstrategie gemäß Unterkapitel 1.4.2 verwendet werden, welche zu Algorithmus 4 führt. Dabei ist nur ein Startwert erforderlich und bei einem BDFk Verfahren werden in den ersten $k-1$ Schritten, in aufsteigender Reihenfolge die Verfahren BDF1 bis BDF$(k-1)$ verwendet und ab dem k-ten Schritt das BDFk Verfahren. Für die erforderliche konsistente Anfangsbedingung $\mathbf{x}_0$

[3]Die Grafik von Abbildung 3.1 wurde aus [5, Abschnitt 8.1, FIGURE 8.1., Seite 320] entnommen.

wären z.B. die Werte einer DC-Analyse naheliegend, wobei darauf hingewiesen wird, dass dies bei einem größeren DAG Index, keine Garantie für eine konsistente Anfangsbedingung sein muss (siehe hierfür z.B. Unterkapitel 1.4.1 oder [19], [20]).

Algorithmus 4 : Aufsteigendes BDFk Verfahren zur Lösung des Modellproblems 3.1

Eingabe : $\widetilde{\mathbf{F}} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ gemäß (3.3), Startzeitpunkt t_0, Schrittweite $h > 0$, Schrittanzahl $N \in \mathbb{N}$ $(N \geq k)$, Startwert $\mathbf{x}_0 = \mathbf{x}(t_0) \in \mathbb{R}^m$

Ausgabe : Zeitschritte $\mathbf{x}_1, \ldots \mathbf{x}_N$

1 **for** $n \leftarrow 0, \ldots, N-1$ **do**

/* Benutze in Zeile 6 das BDF$\widetilde{k}$ Verfahren gemäß Tabelle 1.1 (d.h. beginne in aufsteigender Reihenfolge mit BDF1 bis BDFk) */

2 **if** $(n+1) < k$ **then**

3 $\widetilde{k} \leftarrow n+1$ // Benutze nachfolgend in aufsteigender Reihenfolge BDF1 bis BDF$(k-1)$

4 **else**

5 $\widetilde{k} \leftarrow k$ // Benutze nachfolgend das BDFk Verfahren

/* Löse gemäß Tabelle 1.1 die Gleichung (1.13) nach $\mathbf{x}_{n+1}$ auf */

6 $\widetilde{\mathbf{F}}\left(\frac{1}{\beta_0 \cdot h} \cdot \sum_{i=0}^{\widetilde{k}} \left(\alpha_i \cdot \mathbf{x}_{n+1-i}\right),\, \mathbf{x}_{n+1},\, t_{n+1}\right) = \mathbf{0}_m$

Berechnungen für die Algorithmen 3 und 4, bei Verwendung des Newton Verfahrens

Nachfolgend werden die wesentlichen Berechnungen für das lokale Newton Verfahren angegeben (gemäß Unterkapitel 1.3), welche in Zeile 2 von Algorithmus 3, bzw. Zeile 6 von Algorithmus 4 erforderlich sind, um $\mathbf{x}_{n+1}$ bzw. eine neue Iterierte $\mathbf{x}_{n+1}^{(\ell+1)}$ zu berechnen. Diese Berechnungen sind ebenfalls die Grundlagen für die Anwendung eines gedämpften Newton Verfahrens. Ausgehend von Algorithmus 3 werden für ein $k \in \{1, \ldots, 6\}$ die folgenden Funktionen definiert

$$\varphi : \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} : \mathbf{x} \mapsto \varphi(\mathbf{x}) = \left(\frac{\alpha_0}{\beta_0 \cdot h} \cdot \mathbf{x}^\top + \frac{1}{\beta_0 \cdot h} \cdot \sum_{i=1}^{k} \left(\alpha_i \cdot \mathbf{x}_{n+1-i}^\top\right),\, \mathbf{x}^\top,\, t_{n+1}\right)^\top$$

$$\mathbf{F} : \mathbb{R}^m \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{F}(\mathbf{x}) = \widetilde{\mathbf{F}}(\varphi(\mathbf{x})) \,.$$

Ausgehend von den bekannten Vektoren $\mathbf{x}_{n-(k-1)}, \ldots, \mathbf{x}_n$, wird nun $\mathbf{x}_{n+1}$ gesucht, welches die Gleichung $\mathbf{F}(\mathbf{x}_{n+1}) = \mathbf{0}_m$ erfüllt. Unter Verwendung der Kettenregel (siehe Satz 1.9) und der partiellen Jacobi-Matrix (siehe Definition 1.10) wird folgende Jabobi-Matrix von $\mathbf{F}$ berechnet

$$\mathbf{J}_{\mathbf{F}}(\mathbf{x}) = \mathbf{J}_{\mathbf{v};\widetilde{\mathbf{F}}}(\varphi(\mathbf{x})) \cdot \frac{\alpha_0}{\beta_0 \cdot h} \cdot \mathbf{I}_{m\times m} + \mathbf{J}_{\mathbf{w};\widetilde{\mathbf{F}}}(\varphi(\mathbf{x})) \cdot \mathbf{I}_{m\times m} + \mathbf{J}_{t;\widetilde{\mathbf{F}}}(\varphi(\mathbf{x})) \cdot \mathbf{0}_m^\top = \frac{\alpha_0}{\beta_0 \cdot h} \cdot \mathbf{M} + \mathbf{J}_{\mathbf{f}}(\mathbf{x}) \,,$$

wobei die Matrizen $\mathbf{M}$, $\mathbf{J}_{\mathbf{f}}(\mathbf{x})$ und die Funktion $\mathbf{f}$ gemäß Modellproblem 3.1 gegeben sind.
Die neue Iterierte $\mathbf{x}_{n+1}^{(\ell+1)}$ im lokalen Newton Verfahren berechnet sich dann folgendermaßen

$$\mathbf{x}_{n+1}^{(\ell+1)} = \mathbf{x}_{n+1}^{(\ell)} - \left(\mathbf{J}_{\mathbf{F}}(\mathbf{x}_{n+1}^{(\ell)})\right)^{-1} \cdot \mathbf{F}(\mathbf{x}_{n+1}^{(\ell)}) \,. \tag{3.4}$$

Als Startwert für die erste Newton Iteration kann z.B. der Wert des letzten Zeitschritts gewählt werden, d.h. $\mathbf{x}_{n+1}^{(0)} = \mathbf{x}_n$.

Algorithmus zur Lösung des Modellproblems durch das Radau IIA Verfahren

Wie bereits in Unterkapitel 1.4.2 erwähnt wurde, kann eine zuverlässige Anwendung des Radau IIA Verfahrens nur gewährleistet werden, wenn der DAG Index des Modellproblems 3.1 höchstens 2 ist. Im Folgenden wird für $s \in \{1,2,3\}$ der Algorithmus für das s-stufige Radau IIA$(2{\cdot}s-1)$ Verfahren angegeben, wobei die Koeffizienten gemäß den Butcher Tableaus von Tabelle 1.2 zu entnehmen sind.

Algorithmus 5 : s-stufiges Radau IIA$(2{\cdot}s-1)$ Verfahren zur Lösung des Modellproblems 3.1

Eingabe : $\widetilde{\mathbf{F}} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ gemäß (3.3), Startzeitpunkt t_0, Schrittweite $h > 0$, Schrittanzahl $N \in \mathbb{N}$ $(N \geq 1)$, Startwerte $\mathbf{x}_0 \in \mathbb{R}^m$, wobei $\mathbf{x}_n = \mathbf{x}(t_0 + n{\cdot}\, h) = \mathbf{x}(t_n)$

Ausgabe : Zeitschritte $\mathbf{x}_1, \ldots \mathbf{x}_N$

1 **for** $n \leftarrow 0, \ldots, N-1$ **do**

/* Löse gemäß Tabelle 1.2 die Gleichungen (1.14a) nach $\mathbf{k}_1, \ldots, \mathbf{k}_s$ auf und bestimme $\mathbf{x}_{n+1}$ gemäß Gleichung (1.14b) */

2 $$\widetilde{\mathbf{F}}\left(\mathbf{k}_i,\, \mathbf{x}_n + h \cdot \sum_{j=1}^{s} \left(a_{ij} \cdot \mathbf{k}_j\right),\, t_n + c_i \cdot h\right) = \mathbf{0}_m \;, \qquad \text{für alle } i = 1, 2, \ldots, s$$

3 $$\mathbf{x}_{n+1} \leftarrow \mathbf{x}_n + h \cdot \sum_{i=1}^{s} \left(b_i \cdot \mathbf{k}_i\right)$$

Berechnungen für Algorithmus 5, bei Verwendung des Newton Verfahrens

Nachfolgend werden die wesentlichen Berechnungen für das lokale Newton Verfahren angegeben (gemäß Unterkapitel 1.3), welche in Zeile 2 von Algorithmus 5 erforderlich sind, um $\mathbf{k}_1, \ldots, \mathbf{k}_s$ bzw. die neuen Iterierten $\mathbf{k}_1^{(\ell+1)}, \ldots, \mathbf{k}_s^{(\ell+1)}$ zu berechnen, welche ebenfalls die Grundlagen für die Anwendung eines gedämpften Newton Verfahrens sind. Ausgehend von Algorithmus 5 werden für $1 \leq i \leq s$ die folgenden Funktionen definiert

$$\varphi_i : \underbrace{\mathbb{R}^m \times \ldots \times \mathbb{R}^m}_{=\mathbb{R}^{s\cdot m}} \to \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} : (\mathbf{k}_1, \ldots, \mathbf{k}_s) \mapsto \left(\mathbf{k}_i^\top,\, \mathbf{x}_n^\top + h \cdot \sum_{j=1}^{s} \left(a_{ij} \cdot \mathbf{k}_j^\top\right),\, t_n + c_i \cdot h\right)^\top$$

$$\mathbf{F} : \underbrace{\mathbb{R}^m \times \ldots \times \mathbb{R}^m}_{=\mathbb{R}^{s\cdot m}} \to \mathbb{R}^{s\cdot m} : (\mathbf{k}_1, \ldots, \mathbf{k}_s) \mapsto \mathbf{F}(\mathbf{k}_1, \ldots, \mathbf{k}_s) = \begin{pmatrix} \widetilde{\mathbf{F}}(\varphi_1(\mathbf{k}_1, \ldots, \mathbf{k}_s)) \\ \vdots \\ \widetilde{\mathbf{F}}(\varphi_s(\mathbf{k}_1, \ldots, \mathbf{k}_s)) \end{pmatrix} .$$

Gesucht wird nun $\mathbf{k} := (\mathbf{k}_1^\top, \ldots, \mathbf{k}_s^\top)^\top \in \mathbb{R}^{s\cdot m}$, sodass $\mathbf{F}(\mathbf{k}) = \mathbf{F}(\mathbf{k}_1, \ldots, \mathbf{k}_s) = \mathbf{0}_{s\cdot m}$ gilt. Unter Verwendung der Kettenregel (siehe Satz 1.9) und der partiellen Jacobi-Matrix (siehe Definition 1.10), ergibt sich folgende Jabobi-Matrix von $\mathbf{F}$ in Blockdarstellung

$$\mathbf{J}_{\mathbf{F}}(\mathbf{k}) = \mathbf{J}_{\mathbf{F}}(\mathbf{k}_1, \ldots, \mathbf{k}_s) = \begin{pmatrix} [\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{11} & \ldots & [\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{1s} \\ \vdots & \ddots & \vdots \\ [\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{s1} & \ldots & [\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{ss} \end{pmatrix} \in \mathbb{R}^{s\cdot m \times s\cdot m} ,$$

wobei für $1 \leq i, \ell \leq s$, die Blöcke $[\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{i\ell} = [\mathbf{J}_{\mathbf{F}}(\mathbf{k}_1, \ldots, \mathbf{k}_s)]_{i\ell} \in \mathbb{R}^{m\times m}$ die Darstellung

$$[\mathbf{J}_{\mathbf{F}}(\mathbf{k})]_{i\ell} = \begin{cases} \mathbf{J}_{\mathbf{v};\widetilde{\mathbf{F}}}(\varphi_i(\mathbf{k})) + \mathbf{J}_{\mathbf{w};\widetilde{\mathbf{F}}}(\varphi_i(\mathbf{k})) \cdot h{\cdot}a_{ii} = \mathbf{M} + h{\cdot}a_{ii} \cdot \mathbf{J}_{\mathbf{f}}\left(\mathbf{x}_n + h \cdot \sum_{j=1}^{s} (a_{ij} \cdot \mathbf{k}_j)\right) & \text{, für } i = \ell \\ \mathbf{J}_{\mathbf{w};\widetilde{\mathbf{F}}}(\varphi_i(\mathbf{k})) \cdot h{\cdot}a_{i\ell} = h{\cdot}a_{i\ell} \cdot \mathbf{J}_{\mathbf{f}}\left(\mathbf{x}_n + h \cdot \sum_{j=1}^{s} (a_{ij} \cdot \mathbf{k}_j)\right) & \text{, für } i \neq \ell \end{cases}$$

besitzen, mit den Matrizen $\mathbf{M}$, $\mathbf{J_f}(\mathbf{x})$ und der Funktion $\mathbf{f}$ gemäß Modellproblem 3.1.
Die neue Iterierte $\mathbf{k}^{(\ell+1)} \in \mathbb{R}^{s\cdot m}$ im lokalen Newton Verfahren berechnet sich dann folgendermaßen

$$\mathbf{k}^{(\ell+1)} = \mathbf{k}^{(\ell)} - \left(\mathbf{J_F}(\mathbf{k}^{(\ell)})\right)^{-1} \cdot \mathbf{F}(\mathbf{k}^{(\ell)}) \; . \tag{3.5}$$

In Algorithmus 5 entsprechen die Vektoren $\mathbf{k}_i$ Approximationen von $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}$, an Stützstellen im Intervall $[t_n, t_{n+1}]$. D.h. insbesondere, dass in der ersten Newton Iteration bereits eine hinreichend gute Näherung von $\frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t}$ an diesen Stützstellen benötigt wird. Beispielsweise wird in [5, Abschnitt 8.3, Gl. (8.45)], für das 2-stufige Radau IIA(3) Verfahren empfohlen, dass bei der ersten Newton Iteration der Startwert $\mathbf{k}_1^{(0)} = \mathbf{k}_2^{(0)} = \frac{\mathrm{d}\mathbf{x}(t_0)}{\mathrm{d}t}$ gewählt wird. Eine Approximation von $\frac{\mathrm{d}\mathbf{x}(t_0)}{\mathrm{d}t}$ könnte z.B. durch das Radau IIA(1) oder das BDF1 Verfahren ermittelt werden (beides entspricht dem impliziten Euler Verfahren).

3.2 Einfaches Shooting

Ziel dieses Unterkapitels ist es die Idee des einfachen Shooting Verfahrens zu vermitteln und eine mögliche Implementierung für das Modellproblem 3.1 anzugeben. Hierfür wurden die Grundlagen aus [16, Kapitel 4, Abschnitt 2] entnommen.

Idee

Der Ausgangspunkt ist Gleichung (3.1b) zusammen mit einer konsistenten Anfangsbedingung $\mathbf{x}(t_0) = \mathbf{x}_0$ (siehe Definition 1.16). Es wird nun vorausgesetzt, dass für jede konsistente Anfangsbedingung $\mathbf{x}_0 \in \mathbb{R}^m$, eine eindeutige Lösung $\mathbf{x}(t)$ existiert, dessen asymptotisches Verhalten periodisch mit Periodendauer $T > 0$ ist. Folglich werden sämtliche transiente Vorgänge als abgeklungen angenommen und für t groß genug gelte $\mathbf{x}(t) = \mathbf{x}(t+T)$. Wird nun eine Anfangsbedingung gewählt bei dem der transiente Vorgang bereits abgeklungen ist, so führt dies unmittelbar zur periodischen Lösung und es gilt $\mathbf{x}(t) = \mathbf{x}(t+T)$ für alle $t \in \mathbb{R}$. Das Ziel des Shooting Verfahrens ist es nun, eine geeignete Anfangsbedingung $\mathbf{x}_0$ zu finden, die zu einer periodischen Lösung führt. Mathematisch bedeutet dies, dass das Randwertproblem

$$\mathbf{M} \cdot \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m \; , \quad \mathbf{x}(t_0) = \mathbf{x}_0, \quad \mathbf{x}(t_0 + T) = \mathbf{x}_0 \tag{3.6}$$

zu lösen ist. Nachdem für die periodische Lösung, $\mathbf{x}(t) = \mathbf{x}(t+T)$ für alle t gelten muss, kann im Randwertproblem (3.6) $t_0 \in \mathbb{R}$ beliebig gewählt werden. Dies rechtfertigt im Folgenden die Wahl $t_0 = 0$.
Die Abhängigkeit der Lösung $\mathbf{x}(t)$ von dem Anfangszustand $\mathbf{x}(0)$ soll nun mit Hilfe der Transitionsfunktion $\mathbf{\Phi}$ ausgedrückt werden, sodass $\mathbf{x}(t) = \mathbf{\Phi}(\mathbf{x}_0, t)$ gelte. Das Randwertproblem (3.6) kann nun auch für konsistente Anfangsbedingungen $\mathbf{x}$, als Fixpunktgleichung für $\mathbf{\Phi}(\mathbf{x}, t)$ interpretiert werden. Hierbei wird der Fixpunkt $\mathbf{x}_0$ gesucht, welcher

$$\mathbf{\Phi}(\mathbf{x}_0, T) - \mathbf{x}_0 = \mathbf{0}_m$$

erfüllt. Ausgehend von einer konsistenten Anfangsbedingung $\mathbf{x}_0^{(0)}$ kann der gesuchte Fixpunkt durch die Fixpunktiteration $\mathbf{x}_0^{(\ell+1)} = \mathbf{\Phi}(\mathbf{x}_0^{(\ell)}, T)$ ermittelt werden. Dieses Vorgehen ist zwar intuitiv, allerdings in den meisten Anwendungen langsam. Daher versucht man die Suche nach dem Fixpunkt zu beschleunigen. Gemäß [16, Kapitel 4, Abschnitt 2.1] kann dieses Verfahren beispielsweise mittels Extrapolation beschleunigt werden, wobei für Details auf die angegebene Literatur

verwiesen wird. Gemäß [16, Kapitel 4, Abschnitt 2.2] soll nun allerdings ein beschleunigtes Verfahren unter Verwendung des lokalen Newton Verfahrens gemäß Unterkapitel 1.3 erläutert werden. Für praktische Implementierungen sei allerdings z.B. ein gedämpftes Newton Verfahren gemäß Unterkapitel 1.3 zu empfehlen. Es sei angemerkt, dass ein gedämpftes Newton Verfahren die nachfolgende Vorgehensweise nicht beeinflusst und die Implementierung auf den nachfolgenden Berechnungen aufbaut.
Um das klassische Newton Verfahren anwenden zu können, wird nun vorausgesetzt, dass $\mathbf{\Phi}(\mathbf{x}, T)$ als Funktion in $\mathbf{x}$, stetig differenzierbar ist, d.h. die Komponenten der partiellen Jacobi-Matrix $\mathbf{J}_{\mathbf{x};\,\mathbf{\Phi}}(\mathbf{x}, T)$ sind stetig. Es wird nun $\mathbf{S}(\mathbf{x}) := \mathbf{J}_{\mathbf{x};\,\mathbf{\Phi}}(\mathbf{x}, T)$ definiert, wobei $\mathbf{S}(\mathbf{x}) \in \mathbb{R}^{m\times m}$ als Sensitivitätsmatrix bezeichnet wird. Wird nun die Funktion $\mathbf{F}(\mathbf{x}) := \mathbf{\Phi}(\mathbf{x}, T) - \mathbf{x}$ definiert, so wird die Lösung $\mathbf{F}(\mathbf{x}_0) = \mathbf{0}_m$ gesucht. Bei Vorgabe einer konsistenten Anfangsbedingung $\mathbf{x}_0^{(0)}$, ergibt sich gemäß dem lokalen Newton Verfahren, die neue Iterierte $\mathbf{x}_0^{(\ell+1)}$ durch

$$\mathbf{x}_0^{(\ell+1)} = \mathbf{x}_0^{(\ell)} - \left(\mathbf{J}_{\mathbf{F}}(\mathbf{x}_0^{(\ell)})\right)^{-1} \cdot \mathbf{F}(\mathbf{x}_0^{(\ell)}) \,,$$

wobei $\mathbf{F}(\mathbf{x}_0^{(\ell)}) = \mathbf{\Phi}(\mathbf{x}_0^{(\ell)}, T) - \mathbf{x}_0^{(\ell)}$ und $\mathbf{J}_{\mathbf{F}}(\mathbf{x}_0^{(\ell)}) = \mathbf{J}_{\mathbf{x};\,\mathbf{\Phi}}(\mathbf{x}_0^{(\ell)}, T) - \mathbf{I}_{m\times m} = \mathbf{S}(\mathbf{x}_0^{(\ell)}) - \mathbf{I}_{m\times m}$ gilt.
Die bisherige Herleitung geht stillschweigend davon aus, dass für konsistente Anfangsbedingungen $\mathbf{x}_0^{(\ell)}$, die Transitionsfunktion $\mathbf{\Phi}(\mathbf{x}_0^{(\ell)}, t)$, sowie die Sensitivitätsmatrix $\mathbf{S}(\mathbf{x}_0^{(\ell)})$ bekannt sind. In praktischen Anwendungen ist aber im Allgemeinen beides unbekannt, weshalb für praktische Berechnungen die Transitionsfunktion und dessen Sensitivitätsmatrix approximiert werden müssen. Eine ausreichend gute Approximation $\widehat{\mathbf{\Phi}}$ und $\widehat{\mathbf{S}}(\mathbf{x})$ der Transitionsfunktion $\mathbf{\Phi}$ und der Sensitivitätsmatrix $\mathbf{S}(\mathbf{x})$ erhält man unter Verwendung von Zeitschrittverfahren, welche u.a. bei der transienten Analyse eingesetzt werden. Im Folgenden wird eine Approximation basierend auf dem BDF Verfahren vorgestellt (siehe Unterkapitel 3.1).

Herleitung basierend auf dem BDF Verfahren

Ausgangspunkt der nachfolgenden Herleitung sei Algorithmus 3. Für eine konstante Schrittweite $h > 0$, ein $k \in \{1, \ldots, 6\}$ und bei gegebener konsistenter Anfangsbedingung $\mathbf{x}_0$, sei für ein $N \in \mathbb{N}$ und für alle $n \in \{0, \ldots, N\}$, die approximierte Transitionsfunktion $\widehat{\mathbf{\Phi}}$ durch

$$\widehat{\mathbf{\Phi}}(\mathbf{x}_0, t_n) := \widehat{\mathbf{\Phi}}(\mathbf{x}_0, n{\cdot}h) = \mathbf{x}_n \tag{3.7}$$

gegeben, wobei $t_N = T$ gelte und $\mathbf{x}_n$ gemäß Algorithmus 3 berechnet wurde. D.h. $\widehat{\mathbf{\Phi}}$ ist durch die berechneten Zeitschritte aus der transienten Analyse definiert. Gemäß dem BDFk Verfahren in Algorithmus 3 und bei Verwendung der approximierten Transitionsfunktion $\widehat{\mathbf{\Phi}}$, wird die Zeitableitung folgendermaßen approximiert

$$\frac{\mathrm{d}\mathbf{x}(t_{n+1})}{\mathrm{d}t} \approx \frac{1}{\beta_0{\cdot}\,h} \cdot \sum_{i=0}^{k} \left(\alpha_i \cdot \mathbf{x}_{n+1-i}\right) = \frac{1}{\beta_0{\cdot}\,h} \cdot \sum_{i=0}^{k} \left(\alpha_i \cdot \widehat{\mathbf{\Phi}}(\mathbf{x}_0, t_{n+1-i})\right) .$$

Damit lässt sich das Randwertproblem (3.6) für $n = k-1, \ldots, N-1$ folgendermaßen diskretisieren

$$\mathbf{M} \cdot \frac{1}{\beta_0{\cdot}\,h} \cdot \sum_{i=0}^{k} \Big(\alpha_i \cdot \underbrace{\widehat{\mathbf{\Phi}}(\mathbf{x}_0, t_{n+1-i})}_{=\mathbf{x}_{n+1-i}}\Big) + \mathbf{f}\Big(\underbrace{\widehat{\mathbf{\Phi}}(\mathbf{x}_0, t_{n+1})}_{=\mathbf{x}_{n+1}}\Big) + \mathbf{b}(t_{n+1}) = \mathbf{0}_m \,. \tag{3.8}$$

In Gleichung (3.8) ist nun jeder Zeitschritt $\mathbf{x}_j$ gemäß (3.7) von der Anfangsbedingung $\mathbf{x}_0$ abhängig und diese Abhängigkeit wird durch die Transitionsfunktion $\widehat{\mathbf{\Phi}}$ berücksichtigt. Wird nun auf beiden Seiten der Gleichung (3.8) die partielle Jacobi-Matrix bzgl. der Variable $\mathbf{x}_0$ berechnet, so erhält

man die Sensitivität der Zeitschritte $\mathbf{x}_j$, in Abhängigkeit von der Anfangsbedingung $\mathbf{x}_0$. Mit der Definition von $\widehat{\mathbf{S}}_n(\mathbf{x}_0) := \mathbf{J}_{\mathbf{x}_0;\,\widehat{\mathbf{\Phi}}}(\mathbf{x}_0, t_n)$ für alle $n \in \{0, \ldots, N\}$, führt dies zu

$$\mathbf{M} \cdot \frac{1}{\beta_0 \cdot h} \cdot \underbrace{\sum_{i=0}^{k} \left(\alpha_i \cdot \mathbf{J}_{\mathbf{x}_0;\,\widehat{\mathbf{\Phi}}}(\mathbf{x}_0, t_{n+1-i})\right)}_{=\alpha_0 \cdot \widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0) + \sum_{i=1}^{k} \alpha_i \cdot \widehat{\mathbf{S}}_{n+1-i}(\mathbf{x}_0)} + \mathbf{J}_{\mathbf{f}}(\mathbf{x}_{n+1}) \cdot \underbrace{\mathbf{J}_{\mathbf{x}_0;\,\widehat{\mathbf{\Phi}}}(\mathbf{x}_0, t_{n+1})}_{=\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0)} = \mathbf{0}_{m\times m} \,. \tag{3.9}$$

Hierbei ist die Matrix $\mathbf{J}_{\mathbf{f}}(\mathbf{x}_{n+1})$ gemäß Modellproblem 3.1 gegeben und $\mathbf{J}_{\mathbf{f}}(\mathbf{x}_{n+1})$ ist zudem bekannt, wenn Zeile 2 in Algorithmus 3 mit dem Newton Verfahren gelöst wird.
Wird nun (3.9) nach $\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0)$ umgeformt, dann führt dies zu

$$\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0) = -\Big(\frac{\alpha_0}{\beta_0 \cdot h} \cdot \mathbf{M} + \mathbf{J}_{\mathbf{f}}(\mathbf{x}_{n+1})\Big)^{-1} \cdot \mathbf{M} \cdot \frac{1}{\beta_0 \cdot h} \cdot \sum_{i=1}^{k} \left(\alpha_i \cdot \widehat{\mathbf{S}}_{n+1-i}(\mathbf{x}_0)\right) \,. \tag{3.10}$$

Nachdem $\widehat{\mathbf{\Phi}}(\mathbf{x}_0, 0) = \mathbf{x}_0$ gilt, folgt $\widehat{\mathbf{S}}_0(\mathbf{x}_0) = \mathbf{I}_{m\times m}$. Ist nun $k > 1$, so wird bei der Implementierung des einfachen Shooting Verfahrens empfohlen, in den ersten $k-1$ Schritten, in aufsteigender Reihenfolge, die Verfahren BDF1 bis BDF$(k-1)$ zu verwenden, wodurch ab dem k-ten Schritt die Matrizen $\widehat{\mathbf{S}}_{n+1-k}(\mathbf{x}_0), \ldots, \widehat{\mathbf{S}}_n(\mathbf{x}_0)$ zur Verfügung stehen und somit ab dem k-ten Schritt das BDFk Verfahren verwendet werden kann. Somit kann $\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0)$ gemäß (3.10) durch eine Matrizenmultiplikation berechnet werden. In einer Implementierung sollte allerdings die Berechnung einer inversen Matrix vermieden werden, weshalb empfohlen wird, dass zu (3.10) äquivalente lineare Gleichungssystem zu lösen.
Wird nun $\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0)$ für alle $n \in \{0, \ldots, N-1\}$ gemäß (3.10) berechnet, so gilt $\widehat{\mathbf{S}}(\mathbf{x}_0) = \widehat{\mathbf{S}}_N(\mathbf{x}_0)$ und somit erhält man eine Approximation der Sensitivitätsmatrix $\mathbf{S}(\mathbf{x}_0)$. Wird nun die Funktion $\widehat{\mathbf{F}}(\mathbf{x}) := \widehat{\mathbf{\Phi}}(\mathbf{x}, T) - \mathbf{x}$ definiert, so wird die Lösung $\widehat{\mathbf{F}}(\mathbf{x}_0) = \mathbf{0}_m$ gesucht. Bei Vorgabe einer konsistenten Anfangsbedingung $\mathbf{x}_0^{(0)}$, ergibt sich gemäß dem lokalen Newton Verfahren, die neue Iterierte $\mathbf{x}_0^{(\ell+1)}$ durch

$$\mathbf{x}_0^{(\ell+1)} = \mathbf{x}_0^{(\ell)} - \left(\mathbf{J}_{\widehat{\mathbf{F}}}(\mathbf{x}_0^{(\ell)})\right)^{-1} \cdot \widehat{\mathbf{F}}(\mathbf{x}_0^{(\ell)}) \,, \tag{3.11}$$

wobei $\widehat{\mathbf{F}}(\mathbf{x}_0^{(\ell)}) = \widehat{\mathbf{\Phi}}(\mathbf{x}_0^{(\ell)}, T) - \mathbf{x}_0^{(\ell)} = \mathbf{x}_N^{(\ell)} - \mathbf{x}_0^{(\ell)}$ und $\mathbf{J}_{\widehat{\mathbf{F}}}(\mathbf{x}_0^{(\ell)}) = \widehat{\mathbf{S}}(\mathbf{x}_0^{(\ell)}) - \mathbf{I}_{m\times m}$.
Es sei darauf hingewiesen, dass die Berechnungen von (3.10) und (3.11) bei größeren elektrischen Netzwerken, schon bei kleiner Schrittanzahl N aufwendig sind (u.a. ist im Allgemeinen $\widehat{\mathbf{S}}(\mathbf{x}_0^{(\ell)})$ eine vollbesetzte Matrix). Zu bedenken sei zudem, dass bei einer Sinus ähnlichen Lösung, N nicht zu klein gewählt werden darf, damit die approximierte Lösung immer noch eine ausreichend gute Näherung darstellt. Bei größeren elektrischen Netzwerken kann das einfache Shooting Verfahren daher sehr aufwendig bzw. ineffizient werden. Zu einer Verbesserung kann hierbei z.B. das mehrfache Shooting Verfahren führen, bei dem Abläufe parallelisiert werden (siehe z.B. [16, Chapter 4, Abschnitt 2.5]).

Algorithmus basierend auf dem BDF Verfahren

Mit Hilfe der vorhergehenden Herleitung ist es nun möglich den Algorithmus für das einfache Shooting, unter Verwendung des lokalen Newton Verfahrens anzugeben. Algorithmus 6 enthält allerdings auch die wesentlichen Schritte, welche z.B. bei einem gedämpften Newton Verfahren benötigt werden (siehe Unterkapitel 1.3).

Algorithmus 6 : Einfaches Shooting mit BDFk, zur Lösung des Modellproblems 3.1

Eingabe : $\widetilde{\mathbf{F}} : \mathbb{R}^m \times \mathbb{R}^m \times \mathbb{R} \to \mathbb{R}^m$ gemäß (3.3), $\mathbf{J_f} : \mathbb{R}^m \to \mathbb{R}^{m\times m}$ gemäß (4.10), konsistente Anfangsbedingung $\mathbf{x}_0^{(0)}$, absoluter Fehler $\mathrm{Tol}_{abs} > 0$, Schrittweite $h > 0$, Schrittanzahl $N \in \mathbb{N}$ ($N \geq k$), sodass $\mathbf{x}_N = \mathbf{x}(N \cdot h) = \mathbf{x}(T)$

Ausgabe : Zeitschritte $\mathbf{x}_0^{(\ell)}, \ldots, \mathbf{x}_N^{(\ell)}$

1 Berechne $\mathbf{x}_0^{(0)}, \ldots, \mathbf{x}_N^{(0)}$ gemäß Algorithmus 3
2 $\ell \leftarrow 0$ // Iterationsindex für das lokale Newton Verfahren
3 **while** $\|\mathbf{x}_N^{(\ell)} - \mathbf{x}_0^{(\ell)}\| \geq \mathrm{Tol}_{abs}$ **do**
4 $\widehat{\mathbf{S}}_0(\mathbf{x}_0^{(\ell)}) \leftarrow \mathbf{I}_{m\times m}$
5 **for** $n \leftarrow 0, \ldots, N-1$ **do**
  /* Benutze in Zeile 10 das BDF$\widetilde{k}$ Verfahren gemäß Tabelle 1.1 (d.h. beginne in aufsteigender Reihenfolge mit BDF1 bis BDFk) */
6 **if** $(n+1) < k$ **then**
7 $\widetilde{k} \leftarrow n+1$ // Benutze nachfolgend in aufsteigender Reihenfolge BDF1 bis BDF$(k-1)$
8 **else**
9 $\widetilde{k} \leftarrow k$ // Benutze nachfolgend das BDFk Verfahren
10 $\widehat{\mathbf{S}}_{n+1}(\mathbf{x}_0) \leftarrow -\left(\frac{\alpha_0}{\beta_0 \cdot h} \cdot \mathbf{M} + \mathbf{J_f}(\mathbf{x}_{n+1})\right)^{-1} \cdot \mathbf{M} \cdot \frac{1}{\beta_0 \cdot h} \cdot \sum_{i=1}^{\widetilde{k}} \left(\alpha_i \cdot \widehat{\mathbf{S}}_{n+1-i}(\mathbf{x}_0)\right)$ // gemäß (3.10)
11 $\widehat{\mathbf{S}}(\mathbf{x}_0^{(\ell)}) \leftarrow \widehat{\mathbf{S}}_N(\mathbf{x}_0^{(\ell)})$ // approximierte Sensitivitätsmatrix
12 $\mathbf{x}_0^{(\ell+1)} \leftarrow \mathbf{x}_0^{(\ell)} - \left(\widehat{\mathbf{S}}(\mathbf{x}_0^{(\ell)}) - \mathbf{I}_{m\times m}\right)^{-1} \cdot \left(\mathbf{x}_N^{(\ell)} - \mathbf{x}_0^{(\ell)}\right)$ // neue Newton Iterierte gemäß (3.11)
13 Berechne $\mathbf{x}_0^{(\ell+1)}, \ldots, \mathbf{x}_N^{(\ell+1)}$ gemäß Algorithmus 3
14 $\ell \leftarrow \ell + 1$

3.3 Harmonic Balance Verfahren

Ziel dieses Unterkapitels ist es, ausgehend von Gleichung (3.1a) des Modellproblems 3.1, das Randwertproblem

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{q}(\mathbf{x}(t)) + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m\,, \quad \mathbf{x}(0) = \mathbf{x}_0, \quad \mathbf{x}(T) = \mathbf{x}_0\,. \tag{3.12}$$

mit dem Harmonic Balance (HB) Verfahren zu lösen. Zunächst werden hierfür die Grundlagen der reellen Fourierreihe und der diskreten Fourier Transformation (DFT) angegeben. Darauf aufbauend werden die wesentlichen Herleitungsschritte für das HB Verfahren angegeben, welche zur Lösung von (3.12) benötigt werden. Damit lässt sich zum Abschluss das Verfahren in einem Algorithmus zusammenfassen.

Reelle Fourierreihe

Der Raum der reellwertigen quadratisch integrierbaren Funktionen wird durch

$$L^2([0,T]) := \left\{ f : [0,T] \to \mathbb{R} \;\middle|\; f \text{ messbar}, \int_0^T |f(t)|^2 \,\mathrm{d}t < \infty \right\}$$

definiert. Physikalisch kann dies als Raum der Signale mit endlicher Energie im Intervall $[0,T]$ betrachtet werden. $L^2([0,T])$ wird zu einem unendlich dimensionalen Hilbert-Raum, wenn für

$f, g \in L^2([0,T])$ das Skalarprodukt $\langle \cdot, \cdot \rangle_{L^2([0,T])}$ mit der daraus induzierten Norm $\| \cdot \|_{L^2([0,T])}$ gemäß

$$\langle f, g \rangle_{L^2([0,T])} := \int_0^T f(t) \cdot g(t) \, \mathrm{d}t \, , \quad \|f\|^2_{L^2([0,T])} := \int_0^T |f(t)|^2 \, \mathrm{d}t \, ,$$

sowie Konvergenz im quadratischen Mittel verwendet wird.
Wird für eine Funktion $f : \mathbb{R} \to \mathbb{R}$ die Einschränkung von f auf $[0,T]$ durch $f|_{[0,T]} : [0,T] \to \mathbb{R} : t \mapsto f(t)$ definiert, so bezeichnet

$$P(T, \mathbb{R}) := \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; f(t) = f(t+T) \text{ für alle } t \in \mathbb{R}, \; f|_{[0,T]} \in L^2([0,T]) \right\} \tag{3.13}$$

die Menge der T- periodischen Funktionen, welche auf $[0,T]$ quadratisch integrierbar sind. Für eine Funktion $x \in P(T, \mathbb{R})$ mit Kreisfrequenz $\omega_0 = \frac{2 \cdot \pi}{T}$ gilt nun die Fourierreihendarstellung (siehe z.B. [9, Abschnitt 7.3.3])

$$x(t) = X(0) + \sum_{k=1}^{\infty} X^c(k) \cdot \cos(k \cdot \omega_0 \cdot t) + X^s(k) \cdot \sin(k \cdot \omega_0 \cdot t) \, , \tag{3.14}$$

mit den reellen Koeffizienten

$$X^c(k) := \frac{2}{T} \cdot \int_0^T x(t) \cdot \cos(k \cdot \omega_0 \cdot t) \, \mathrm{d}t \, , \; X^s(k) := \frac{2}{T} \cdot \int_0^T x(t) \cdot \sin(k \cdot \omega_0 \cdot t) \, \mathrm{d}t \, , \; X(0) := \frac{X^c(0)}{2} \, , \tag{3.15}$$

sowie die Parsevalsche Gleichung (siehe z.B. [27, Abschnitt 10.15])

$$\frac{2}{T} \cdot \int_0^T |x(t)|^2 \, \mathrm{d}t = 2 \cdot |X(0)|^2 + \sum_{k=1}^{\infty} \left(|X^c(k)|^2 + |X^s(k)|^2 \right) < \infty \, .$$

Damit folgt auch, dass $\lim\limits_{k \to \infty} X^c(k) = 0$ und $\lim\limits_{k \to \infty} X^s(k) = 0$ gilt.
Nachdem die Fourierreihe analytisch ist, ist sie auch differenzierbar und es gilt

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = \sum_{k=1}^{\infty} (-k \cdot \omega_0 \cdot X^c(k)) \cdot \sin(k \cdot \omega_0 \cdot t) + (k \cdot \omega_0 \cdot X^s(k)) \cdot \cos(k \cdot \omega_0 \cdot t) \, , \tag{3.16}$$

Approximation der Fourierreihe und diskrete Fourier Transformation (DFT)

Für numerische Berechnungen ist es im Allgemeinen nicht möglich für $x \in P(T, \mathbb{R})$ die Fourierreihe mit unendlich vielen Koeffizienten zu bestimmen. Im ersten Schritt wird daher für ein $K \in \mathbb{N}$ die Funktion $x(t)$ durch $\widetilde{x}(t)$ approximiert. Hierfür wählt man die reelle trigonometrische Interpolation bzw. Fourier-Galerkin Approximation (siehe z.B. [8, Abschnitt 8.7.2] oder [9, Abschnitt 7.3.3])

$$x(t) \approx \widetilde{x}(t) = X(0) + \sum_{k=1}^{K} X^c(k) \cdot \cos(k \cdot \omega_0 \cdot t) + X^s(k) \cdot \sin(k \cdot \omega_0 \cdot t) \, ,$$

wobei die Koeffizienten weiterhin durch (3.15) gegeben sind. Zur numerischen Berechnung der Koeffizienten, ist es allerdings auch nötig die Integrale in (3.15) näherungsweise zu bestimmen. Wird nun das Intervall $[0,T]$ in $(2 \cdot K + 1)$ äquidistante Intervalle der Länge $h = \frac{T}{2 \cdot K + 1}$ unterteilt,

mit den Stützstellen $t_n = h{\cdot}n$ für $0 \le n \le 2{\cdot}K{+}1$, sowie die Koeffizienten in (3.15) mit der Trapezregel näherungsweise berechnet, dann erhält man

$$\widehat{X}^c(k) := \frac{2}{2{\cdot}K{+}1} \cdot \sum_{n=0}^{2\cdot K} x(t_n) \cdot \cos(k{\cdot}\omega_0{\cdot}t_n)\,,\ \widehat{X}^s(k) := \frac{2}{2{\cdot}K{+}1} \cdot \sum_{n=0}^{2\cdot K} x(t_n) \cdot \sin(k{\cdot}\omega_0{\cdot}t_n)\,,\ \widehat{X}(0) := \frac{\widehat{X}^c(0)}{2}\,. \tag{3.17}$$

Die approximierten Koeffizienten in (3.17) werden im Folgenden als DFT Koeffizienten bezeichnet und dies führt zu folgender Approximation der Fourierreihe (3.14) (siehe z.B. [8, Abschnitt 8.7.2] oder [9, Abschnitt 7.3.3])

$$\widehat{x}(t) = \widehat{X}(0) + \sum_{k=1}^{K} \widehat{X}^c(k) \cdot \cos(k{\cdot}\omega_0{\cdot}t) + \widehat{X}^s(k) \cdot \sin(k{\cdot}\omega_0{\cdot}t)\,. \tag{3.18}$$

Ausgehend von der Darstellung (3.18), wird nun mit

$$\begin{aligned}\widehat{P}_K(T,\mathbb{R}) := \Big\{ \widehat{x} : \mathbb{R} \to \mathbb{R} \;\Big|\; & \widehat{\mathbf{X}} = (\widehat{X}(0), \widehat{X}^c(1), \widehat{X}^s(1), \ldots, \widehat{X}^c(K), \widehat{X}^s(K))^\top \in \mathbb{R}^{2\cdot K+1}, \\ & \widehat{x}(t) = \widehat{X}(0) + \sum_{k=1}^{K} \widehat{X}^c(k) \cdot \cos(k{\cdot}\omega_0{\cdot}t) + \widehat{X}^s(k) \cdot \sin(k{\cdot}\omega_0{\cdot}t) \Big\}\,,\end{aligned} \tag{3.19}$$

die Menge der T- periodischen Funktionen mit K Harmonischen definiert.
Für $\widehat{x} \in \widehat{P}_K(T,\mathbb{R})$ kann nun folgender Zusammenhang zwischen den DFT Koeffizienten und den Funktionswerten $\widehat{x}(t_0), \ldots, \widehat{x}(t_{2\cdot K})$ hergestellt werden

$$\underbrace{\begin{pmatrix} \widehat{x}(t_0) \\ \widehat{x}(t_1) \\ \vdots \\ \widehat{x}(t_{2\cdot K}) \end{pmatrix}}_{:=\widehat{\mathbf{x}}} = \underbrace{\begin{pmatrix} 1 & \cos(\omega_0{\cdot}t_0) & \sin(\omega_0{\cdot}t_0) & \ldots & \cos(K{\cdot}\omega_0{\cdot}t_0) & \sin(K{\cdot}\omega_0{\cdot}t_0) \\ 1 & \cos(\omega_0{\cdot}t_1) & \sin(\omega_0{\cdot}t_1) & \ldots & \cos(K{\cdot}\omega_0{\cdot}t_1) & \sin(K{\cdot}\omega_0{\cdot}t_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos(\omega_0{\cdot}t_{2\cdot K}) & \sin(\omega_0{\cdot}t_{2\cdot K}) & \ldots & \cos(K{\cdot}\omega_0{\cdot}t_{2\cdot K}) & \sin(K{\cdot}\omega_0{\cdot}t_{2\cdot K}) \end{pmatrix}}_{:=\boldsymbol{\Gamma}^{-1}} \cdot \underbrace{\begin{pmatrix} \widehat{X}(0) \\ \widehat{X}^c(1) \\ \widehat{X}^s(1) \\ \vdots \\ \widehat{X}^c(K) \\ \widehat{X}^s(K) \end{pmatrix}}_{:=\widehat{\mathbf{X}}}\,. \tag{3.20}$$

Dabei kann die Matrix $\boldsymbol{\Gamma}^{-1} \in \mathbb{R}^{(2\cdot K+1)\times(2\cdot K+1)}$ zur Berechnung der inversen diskreten Fourier Transformation (IDFT) (3.20) dienen, sofern die DFT Koeffizienten $\widehat{\mathbf{X}}$ bekannt sind. Die inverse Matrix $\boldsymbol{\Gamma}$ von $\boldsymbol{\Gamma}^{-1}$, welche $\boldsymbol{\Gamma}{\cdot}\,\boldsymbol{\Gamma}^{-1} = \mathbf{I}$ erfüllt, ergibt sich dann gemäß (3.17) zu

$$\boldsymbol{\Gamma} := \frac{2}{2{\cdot}K{+}1} \cdot \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & \ldots & \frac{1}{2} \\ \cos(\omega_0{\cdot}t_0) & \cos(\omega_0{\cdot}t_1) & \ldots & \cos(\omega_0{\cdot}t_{2\cdot K}) \\ \sin(\omega_0{\cdot}t_0) & \sin(\omega_0{\cdot}t_1) & \ldots & \sin(\omega_0{\cdot}t_{2\cdot K}) \\ \vdots & \vdots & \ddots & \vdots \\ \cos(K{\cdot}\omega_0{\cdot}t_0) & \cos(K{\cdot}\omega_0{\cdot}t_1) & \ldots & \cos(K{\cdot}\omega_0{\cdot}t_{2\cdot K}) \\ \sin(K{\cdot}\omega_0{\cdot}t_0) & \sin(K{\cdot}\omega_0{\cdot}t_1) & \ldots & \sin(K{\cdot}\omega_0{\cdot}t_{2\cdot K}) \end{pmatrix} \in \mathbb{R}^{(2\cdot K+1)\times(2\cdot K+1)}\,. \tag{3.21}$$

Damit lässt sich bei bekannten Funktionswerten $\widehat{\mathbf{x}}$, die diskrete Fourier Transformation (DFT) (3.22) berechnen

$$\widehat{\mathbf{X}} = \boldsymbol{\Gamma}{\cdot}\,\widehat{\mathbf{x}}\,. \tag{3.22}$$

Idee des HB Verfahrens

In den nachfolgenden Abschnitten gelte für sämtliche zeitabhängige vektorwertige Funktionen $\mathbf{x}(t)$, dass gemäß (3.13) $\mathbf{x} \in [P(T,\mathbb{R})]^m$. Für numerische Berechnungen kann die exakte Darstellung (3.14) im Allgemeinen nicht verwendet werden. Stattdessen wird für ein beliebiges aber fest gewähltes $K \in \mathbb{N}$ die approximierte trigonometrische Interpolation $\widehat{\mathbf{x}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ gemäß (3.19) verwendet. Damit können erforderliche Berechnungen im Folgenden effizient mit Hilfe der DFT durchgeführt werden. Für eine gegebene Funktion $\mathbf{b} \in [P(T,\mathbb{R})]^m$ wird nun eine approximierte Fourierreihe $\widehat{\mathbf{x}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ gesucht, welche gemäß (3.23)

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\widehat{\mathbf{x}}(t)) + \mathbf{f}(\widehat{\mathbf{x}}(t)) + \mathbf{b}(t) \approx \mathbf{0}_m\,, \quad \widehat{\mathbf{x}}(0) = \mathbf{x}_0, \quad \widehat{\mathbf{x}}(T) = \mathbf{x}_0\,, \tag{3.23}$$

das Randwertproblem (3.12) näherungsweise erfüllt. Aufgrund der T- Periodizität von $\widehat{\mathbf{x}}(t)$ sind die Randwertbedingungen stets erfüllt und die Randwertaufgabe reduziert sich auf die Suche nach den DFT Koeffizienten $\widehat{\mathbf{X}}$ von $\widehat{\mathbf{x}}(t)$.
Seien $\widehat{\mathbf{B}}$ die DFT Koeffizienten von $\widehat{\mathbf{b}} \in [\widehat{P}_K(T,\mathbb{R})]^m$, wobei $\widehat{\mathbf{b}}(t)$ die approximierte Fourierreihe von $\mathbf{b} \in [P(T,\mathbb{R})]^m$ darstellt. Dann kann (3.23) folgendermaßen in symbolischer Schreibweise, in Abhängigkeit von den DFT Koeffizienten $\widehat{\mathbf{X}}$ formuliert werden

$$\mathrm{DFT}\Big(\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}\big(\mathrm{IDFT}(\widehat{\mathbf{X}})\big)\Big) + \mathrm{DFT}\Big(\mathbf{f}\big(\mathrm{IDFT}(\widehat{\mathbf{X}})\big)\Big) + \widehat{\mathbf{B}} = \mathbf{0}\,, \tag{3.24}$$

wobei der folgende symbolische Zusammenhang gelte

$$\mathrm{IDFT}\Big(\mathrm{DFT}\Big(\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}\big(\mathrm{IDFT}(\widehat{\mathbf{X}})\big)\Big) + \mathrm{DFT}\Big(\mathbf{f}\big(\mathrm{IDFT}(\widehat{\mathbf{X}})\big)\Big) + \widehat{\mathbf{B}}\Big) \approx \begin{pmatrix} \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\widehat{\mathbf{x}}(t_0)) + \mathbf{f}(\widehat{\mathbf{x}}(t_0)) + \mathbf{b}(t_0) \\ \vdots \\ \frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\widehat{\mathbf{x}}(t_{2\cdot K})) + \mathbf{f}(\widehat{\mathbf{x}}(t_{2\cdot K})) + \mathbf{b}(t_{2\cdot K}) \end{pmatrix}.$$

Diese symbolische Schreibweise von (3.24) wird nun im Folgenden begründet. Zunächst wird aufgrund der T- Periodizität, der Zusammenhang zwischen den DFT Koeffizienten $\widehat{\mathbf{X}}$ und $\widehat{\mathbf{x}}(t)$ verwendet. Nachdem die Funktionen $\mathbf{f}(\mathbf{x})$ und $\mathbf{q}(\mathbf{x})$ lediglich algebraische Nichtlinearitäten darstellen, ist sichergestellt, dass $\mathbf{f}(\widehat{\mathbf{x}}(t))$ und $\mathbf{q}(\widehat{\mathbf{x}}(t))$ ebenfalls T- periodisch sind. D.h. insbesondere, dass $\widehat{\mathbf{f}}$, $\widehat{\mathbf{q}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ existieren, welche $\mathbf{f}(\widehat{\mathbf{x}}(t))$ und $\mathbf{q}(\widehat{\mathbf{x}}(t))$ approximieren, wobei die entsprechenden DFT Koeffizienten mit $\widehat{\mathbf{F}}$ und $\widehat{\mathbf{Q}}$ bezeichnet werden. Wird nun mit $\widehat{\mathbf{h}}(t) := \frac{\mathrm{d}\widehat{\mathbf{q}}(t)}{\mathrm{d}t}$ die Zeitableitung von $\widehat{\mathbf{q}}(t)$ bezeichnet, so besteht gemäß (3.16) ein Zusammenhang zwischen den DFT Koeffizienten $\widehat{\mathbf{H}}$ und $\widehat{\mathbf{Q}}$. Dieser Zusammenhang kann durch eine Matrix $\mathbf{\Omega}$ und die Beziehung $\widehat{\mathbf{H}} = \mathbf{\Omega}\cdot\widehat{\mathbf{Q}}$ berücksichtigt werden.
Mit diesen Erläuterungen lässt sich Gleichung (3.23) und die Suche nach geeigneten DFT Koeffizienten $\widehat{\mathbf{X}}$, auf die Lösung eines algebraischen nichtlinearen Gleichungssystems zurückführen

$$\mathbf{\Omega}\cdot\widehat{\mathbf{Q}}(\widehat{\mathbf{X}}) + \widehat{\mathbf{F}}(\widehat{\mathbf{X}}) + \widehat{\mathbf{B}} = \mathbf{0}\,.$$

Zusammenfassend lässt sich feststellen, dass die DFT Koeffizienten $\widehat{\mathbf{X}}$ die Veränderlichen dieses Gleichungssystems sind. Zudem sei festgestellt, dass das algebraische nichtlineare Verhalten von $\mathbf{f}(\mathbf{x})$ und $\mathbf{q}(\mathbf{x})$, die wesentliche Voraussetzung für eine effiziente Implementierung des HB Verfahrens ist. Weiters ist zu beobachten, dass die algebraischen Nichtlinearitäten im Zeitbereich behandelt werden und mit Hilfe der DFT bzw. IDFT zwischen Frequenz- und Zeitbereich gewechselt wird.

Herleitung des HB Verfahrens

Aufbauend auf der vorhergehenden Idee, erfolgt nun eine detaillierte Herleitung, welche insbesondere auch für die Implementierung benötigt wird.
Gemäß (3.18) wurde der Zusammenhang zwischen DFT und der approximierten Fourierreihe für eine skalare Funktion $\widehat{x} \in \widehat{P}_K(T,\mathbb{R})$ dargestellt. Nachdem aber $\mathbf{x} \in [P(T,\mathbb{R})]^m$ in (3.12), bzw. dessen Approximierte $\widehat{\mathbf{x}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ in (3.23) vektorwertige periodische Funktionen sind, muss zunächst geklärt werden, wie die DFT auf $\widehat{\mathbf{x}}(t)$ angewendet werden kann. Für die Funktionen $\widehat{x}_1(t),\dots,\widehat{x}_m(t)$ von $\widehat{\mathbf{x}}(t)$ werden nun für $1 \le i \le m$, durch

$$\widehat{\mathbf{x}}_i := (\widehat{x}_i(t_0),\dots,\widehat{x}_i(t_{2\cdot K}))^\top \in \mathbb{R}^{2\cdot K+1} \tag{3.25a}$$
$$\widehat{\mathbf{X}}_i := (\widehat{X}_i(0),\widehat{X}_i^c(1),\widehat{X}_i^s(1),\dots,\widehat{X}_i^c(K),\widehat{X}_i^s(K))^\top \in \mathbb{R}^{2\cdot K+1} \tag{3.25b}$$

die Vektoren der Funktionswerte $\widehat{\mathbf{x}}_i$ und der DFT Koeffizienten $\widehat{\mathbf{X}}_i$ von $\widehat{x}_i(t)$ definiert. Mit

$$\widehat{\mathbf{x}} := (\widehat{\mathbf{x}}_1^\top,\dots,\widehat{\mathbf{x}}_m^\top)^\top \in \mathbb{R}^{m\cdot(2\cdot K+1)} \tag{3.26a}$$
$$\widehat{\mathbf{X}} := (\widehat{\mathbf{X}}_1^\top,\dots,\widehat{\mathbf{X}}_m^\top)^\top \in \mathbb{R}^{m\cdot(2\cdot K+1)} \tag{3.26b}$$

werden die Funktionswerte $\widehat{\mathbf{x}}$ und die DFT Koeffizienten $\widehat{\mathbf{X}}$ von $\widehat{x}_1(t),\dots,\widehat{x}_m(t)$ zusammengefasst. Es soll darauf hingewiesen werden, dass die gewählte Komponentenreihenfolge in (3.25) und (3.26) beliebig ist, allerdings wurde diese so gewählt, dass mit Hilfe der Matrizen $\mathbf{\Gamma}$ und $\mathbf{\Gamma}^{-1}$ gemäß (3.20) und (3.21), einfach zwischen Zeit- und Frequenzbereich gewechselt werden kann. Hierfür werden nun folgendermaßen Matrizen $\mathbf{\Gamma}_{\mathrm{DFT}}, \mathbf{\Gamma}_{\mathrm{DFT}}^{-1} \in \mathbb{R}^{m\cdot(2\cdot K+1)\times m\cdot(2\cdot K+1)}$ in Blockdarstellung definiert

$$\mathbf{\Gamma}_{\mathrm{DFT}} := \begin{pmatrix} [\mathbf{\Gamma}_{\mathrm{DFT}}]_{11} & \dots & [\mathbf{\Gamma}_{\mathrm{DFT}}]_{1m} \\ \vdots & \ddots & \vdots \\ [\mathbf{\Gamma}_{\mathrm{DFT}}]_{m1} & \dots & [\mathbf{\Gamma}_{\mathrm{DFT}}]_{mm} \end{pmatrix}, \quad [\mathbf{\Gamma}_{\mathrm{DFT}}]_{ij} := \begin{cases} \mathbf{\Gamma} & \text{, für } i=j \\ \mathbf{0}_{(2\cdot K+1)\times(2\cdot K+1)} & \text{, für } i\neq j \end{cases},$$
$$\mathbf{\Gamma}_{\mathrm{DFT}}^{-1} := \begin{pmatrix} [\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}]_{11} & \dots & [\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}]_{1m} \\ \vdots & \ddots & \vdots \\ [\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}]_{m1} & \dots & [\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}]_{mm} \end{pmatrix}, \quad [\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}]_{ij} := \begin{cases} \mathbf{\Gamma}^{-1} & \text{, für } i=j \\ \mathbf{0}_{(2\cdot K+1)\times(2\cdot K+1)} & \text{, für } i\neq j \end{cases}.$$

Damit kann nun die DFT mit $\widehat{\mathbf{X}} = \mathbf{\Gamma}_{\mathrm{DFT}}\cdot\widehat{\mathbf{x}}$, sowie die IDFT durch $\widehat{\mathbf{x}} = \mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}$ berechnet werden.

Ein großer Vorteil von Fourierreihen ist, dass die Ableitung gemäß (3.16) einfach berechnet werden kann. Dieser Zusammenhang gilt zudem auch für die approximierte Fourierreihe gemäß (3.18). Sei nun $\widehat{\mathbf{z}}(t) := \frac{\mathrm{d}\widehat{\mathbf{x}}(t)}{\mathrm{d}t}$, dann gilt gemäß (3.16) $\widehat{\mathbf{Z}} = \mathbf{\Omega}\cdot\widehat{\mathbf{X}}$, wobei $\mathbf{\Omega} \in \mathbb{R}^{m\cdot(2\cdot K+1)\times m\cdot(2\cdot K+1)}$ folgendermaßen in Blockdarstellung gegeben ist

$$\mathbf{\Omega} := \begin{pmatrix} [\mathbf{\Omega}]_{11} & \dots & [\mathbf{\Omega}]_{1m} \\ \vdots & \ddots & \vdots \\ [\mathbf{\Omega}]_{m1} & \dots & [\mathbf{\Omega}]_{mm} \end{pmatrix}, \quad [\mathbf{\Omega}]_{ij} := \begin{cases} \mathbf{\Omega}_{\omega_0} & \text{, für } i=j \\ \mathbf{0}_{(2\cdot K+1)\times(2\cdot K+1)} & \text{, für } i\neq j \end{cases}.$$

Hierbei ist $\mathbf{\Omega}_{\omega_0} \in \mathbb{R}^{(2\cdot K+1)\times(2\cdot K+1)}$, für $1 \le i,j \le (2\cdot K+1)$ und $k \in \{1,\dots,K\}$, durch folgende Koeffizienten definiert

$$(\mathbf{\Omega}_{\omega_0})_{ij} := \begin{cases} k\cdot\omega_0 & \text{, für } i=2\cdot k,\ j=2\cdot k+1 \\ -k\cdot\omega_0 & \text{, für } i=2\cdot k+1,\ j=2\cdot k \\ 0 & \text{, sonst} \end{cases},$$

d.h. $\mathbf{\Omega}_{\omega_0}$ besitzt folgende Darstellung

$$\mathbf{\Omega}_{\omega_0} := \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \omega_0 & 0 & 0 & \dots & 0 & 0 \\ 0 & -\omega_0 & 0 & 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 2{\cdot}\omega_0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & -2{\cdot}\omega_0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots \\ 0 & 0 & 0 & 0 & 0 & \dots & 0 & K{\cdot}\omega_0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & -K{\cdot}\omega_0 & 0 \end{pmatrix} .$$

Damit in weiterer Folge für die Funktion $\mathbf{f}(\widehat{\mathbf{x}}(t))$, die approximierte Fourierreihe $\widehat{\mathbf{f}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ in Abhängigkeit von den DFT Koeffizienten $\widehat{\mathbf{X}}$ angegeben werden kann, wird im nächsten Schritt die Funktion $\widetilde{\mathbf{f}}$ definiert. Nachdem die algebraische Nichtlinearität von $\mathbf{f}(\mathbf{x})$ im Zeitbereich berücksichtigt wurde, müssen die DFT Koeffizienten $\widehat{\mathbf{F}}$ ermittelt werden. Um $\widehat{\mathbf{F}}$ effizient mit $\mathbf{\Gamma}_{\mathrm{DFT}}$ berechnen zu können, muss bei der Definition von $\widetilde{\mathbf{f}}$, insbesondere die Reihenfolge der Zeitschritte gemäß (3.26a) berücksichtigt werden. Für eine übersichtlichere Schreibweise wird zunächst folgende Indexfunktion definiert

$$\alpha : \mathbb{N} \to \mathbb{N} : i \mapsto \alpha(i) = (i-1) \cdot (2{\cdot}K{+}1)$$

Mit $\mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{f}(\mathbf{x})$ gemäß dem Modellproblem 3.1, ergibt sich folgende Definition

$$\widetilde{\mathbf{f}} : \mathbb{R}^{m\cdot(2\cdot K+1)} \to \mathbb{R}^{m\cdot(2\cdot K+1)} : (y_1,\dots,y_{m\cdot(2\cdot K+1)}) \mapsto \begin{pmatrix} f_1(y_{1+\alpha(1)}, y_{1+\alpha(2)}, \dots, y_{1+\alpha(m)}) \\ f_1(y_{2+\alpha(1)}, y_{2+\alpha(2)}, \dots, y_{2+\alpha(m)}) \\ \vdots \\ f_1(y_{2\cdot K+1+\alpha(1)}, y_{2\cdot K+1+\alpha(2)}, \dots, y_{2\cdot K+1+\alpha(m)}) \\ \vdots \\ f_m(y_{1+\alpha(1)}, y_{1+\alpha(2)}, \dots, y_{1+\alpha(m)}) \\ \vdots \\ f_m(y_{2\cdot K+1+\alpha(1)}, y_{2\cdot K+1+\alpha(2)}, \dots, y_{2\cdot K+1+\alpha(m)}) \end{pmatrix} .$$

Die im Folgenden benötigte Jabobi-Matrix $\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y}) = \mathbf{J}_{\widetilde{\mathbf{f}}}(y_1,\dots,y_{m\cdot(2\cdot K+1)}) \in \mathbb{R}^{m\cdot(2\cdot K+1)\times m\cdot(2\cdot K+1)}$ der Funktion $\widetilde{\mathbf{f}}(\mathbf{y})$, ist für $1 \le i,j \le m$ und $1 \le \ell,p \le (2{\cdot}K{+}1)$ folgendermaßen durch Diagonalmatrizen $[\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{ij} \in \mathbb{R}^{(2\cdot K+1)\times(2\cdot K+1)}$ in Blockdarstellung gegeben

$$\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y}) := \begin{pmatrix} [\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{11} & \dots & [\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{1m} \\ \vdots & \ddots & \vdots \\ [\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{m1} & \dots & [\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{mm} \end{pmatrix} , \quad \left([\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{ij}\right)_{\ell p} := \begin{cases} \frac{\partial f_i}{\partial x_j}(y_{\ell+\alpha(1)}, \dots, y_{\ell+\alpha(m)}) & , \text{für } \ell = p \\ 0 & , \text{für } \ell \neq p \end{cases} ,$$

$$[\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{ij} := \begin{pmatrix} \frac{\partial f_i}{\partial x_j}(y_{1+\alpha(1)}, \dots, y_{1+\alpha(m)}) & & \\ & \ddots & \mathbf{0} \\ \mathbf{0} & & \frac{\partial f_i}{\partial x_j}(y_{2\cdot K+1+\alpha(1)}, \dots, y_{2\cdot K+1+\alpha(m)}) \end{pmatrix} . \tag{3.27}$$

Die Differentialquotienten $\frac{\partial f_i}{\partial x_j}(\mathbf{x})$ der Matrix $[\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})]_{ij}$ sind dabei die Koeffizienten von $\mathbf{J}_{\mathbf{f}}(\mathbf{x})$, gemäß (4.10) des Modellproblems 3.1.
Die vorhergehenden Überlegungen gelten auch in äquivalenter Weise für die Funktion $\mathbf{q}(\widehat{\mathbf{x}}(t))$. Mit

$\mathbf{q} : \mathbb{R}^m \to \mathbb{R}^m : \mathbf{x} \mapsto \mathbf{q}(\mathbf{x})$, gemäß dem Modellproblem 3.1, kann analog zur Funktion $\widetilde{\mathbf{f}}(\mathbf{y})$ die Funktion $\widetilde{\mathbf{q}} : \mathbb{R}^{m\cdot(2\cdot K+1)} \to \mathbb{R}^{m\cdot(2\cdot K+1)} : \mathbf{y} \mapsto \widetilde{\mathbf{q}}(\mathbf{y})$ definiert werden und damit ergibt sich analog zur Jacobi-Matrix $\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})$ die im Folgenden benötigte Jacobi-Matrix $\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y}) = \mathbf{J}_{\widetilde{\mathbf{q}}}(y_1, \ldots, y_{m\cdot(2\cdot K+1)}) \in \mathbb{R}^{m\cdot(2\cdot K+1)\times m\cdot(2\cdot K+1)}$, wobei lediglich die Funktionen f_i durch die Funktionen q_i ersetzt werden müssen. Nachdem die Matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$ des Modellproblems 3.1 konstant ist, gilt für die Funktion $\mathbf{q}(\mathbf{x})$ im Modellproblem 3.1, auch die äquivalente Darstellung $\mathbf{q}(\mathbf{x}) = \mathbf{M}\cdot\mathbf{x}$. Nachdem $\mathbf{J}_{\mathbf{q}}(\mathbf{x}) = \mathbf{M}$ gilt, erhält man für die Jacobi-Matrix $\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})$, die folgende vereinfachte Darstellung

$$\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y}) := \begin{pmatrix} [\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})]_{11} & \ldots & [\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})]_{1m} \\ \vdots & \ddots & \vdots \\ [\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})]_{m1} & \ldots & [\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})]_{mm} \end{pmatrix}, \quad \left([\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})]_{ij}\right)_{\ell p} := \begin{cases} \frac{\partial q_i}{\partial x_j}(y_{\ell+\alpha(1)}, \ldots, y_{\ell+\alpha(m)}) = (\mathbf{M})_{ij}\ , \\ \qquad\qquad \text{für } \ell = p \\ 0 \qquad\qquad , \text{ für } \ell \neq p \end{cases}, \tag{3.28}$$

wobei $1 \leq i, j \leq m$ und $1 \leq \ell, p \leq (2\cdot K+1)$ und $(\mathbf{M})_{ij}$ die Koeffizienten von $\mathbf{M}$ bezeichnet.
Mit $\widehat{\mathbf{b}} \in [\widehat{P}_K(T, \mathbb{R})]^m$ wird nun noch die approximierte Fourierreihe von $\mathbf{b} \in [P(T, \mathbb{R})]^m$ bezeichnet, welche die DFT Koeffizienten $\widehat{\mathbf{B}}$ besitze.

Es sind nun alle Grundlagen vorhanden um das Randwertproblem (3.12) näherungsweise in Abhängigkeit von den DFT Koeffizienten $\widehat{\mathbf{X}}$ zu formulieren. Um im Folgenden eine kompakte Schreibweise zu ermöglichen, wird die Funktion $\mathbf{F} : \mathbb{R}^{m\cdot(2\cdot K+1)} \to \mathbb{R}^{m\cdot(2\cdot K+1)} : \widehat{\mathbf{X}} \mapsto \mathbf{F}(\widehat{\mathbf{X}})$ definiert, wobei

$$\mathbf{F}(\widehat{\mathbf{X}}) := \boldsymbol{\Omega}\cdot\boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{q}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{f}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \widehat{\mathbf{B}}\ . \tag{3.29}$$

Damit kann das Randwertproblem (3.12) folgendermaßen näherungsweise im Frequenzbereich formuliert werden

$$\mathbf{F}(\widehat{\mathbf{X}}) = \boldsymbol{\Omega}\cdot\boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{q}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{f}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \widehat{\mathbf{B}} = \mathbf{0}_{m\cdot(2\cdot K+1)}\ . \tag{3.30}$$

Algorithmus des HB Verfahrens

Aufgrund der vorhergehenden Herleitung kann das HB Verfahren in kompakter Schreibweise durch Algorithmus 7 angegeben werden.

Algorithmus 7 : HB Verfahren zur Lösung des Modellproblems 3.1

Eingabe : $\mathbf{F} : \mathbb{R}^{m\cdot(2\cdot K+1)} \to \mathbb{R}^{m\cdot(2\cdot K+1)}$ gemäß (3.29), Periodendauer $T > 0$, Anzahl der Harmonischen $K \in \mathbb{N}$
Ausgabe : $\widehat{\mathbf{x}} \in [\widehat{P}_K(T, \mathbb{R})]^m$

1 $\omega_0 \leftarrow \frac{2\cdot\pi}{T}$ // Grundschwingungskreisfrequenz
/* Löse gemäß Gleichung (3.30), $\mathbf{F}(\widehat{\mathbf{X}}) = \mathbf{0}_{m\cdot(2\cdot K+1)}$ nach $\widehat{\mathbf{X}}$ auf */
2 $\boldsymbol{\Omega}\cdot\boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{q}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \boldsymbol{\Gamma}_{\mathrm{DFT}}\cdot\widetilde{\mathbf{f}}\left(\boldsymbol{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\right) + \widehat{\mathbf{B}} = \mathbf{0}_{m\cdot(2\cdot K+1)}$
/* Berechne $\widehat{x}_i(t)$ gemäß (3.18), mit $\widehat{\mathbf{X}} := (\widehat{\mathbf{X}}_1^\top, \ldots, \widehat{\mathbf{X}}_m^\top)^\top \in \mathbb{R}^{m\cdot(2\cdot K+1)}$ gemäß (3.26b) */
3 **for** $i \leftarrow 1, \ldots, m$ **do**
4 $\quad \widehat{x}_i(t) \leftarrow \widehat{X}_i(0) + \sum_{k=1}^{K} \widehat{X}_i^c(k)\cdot\cos(k\cdot\omega_0\cdot t) + \widehat{X}_i^s(k)\cdot\sin(k\cdot\omega_0\cdot t)$

Gemäß Algorithmus 7 reduziert sich das HB Verfahren auf die Lösung eines im Allgemeinen nichtlinearen Gleichungssystems. Zeile 2 kann auf verschiedene Möglichkeiten gelöst werden. Im

Folgenden wird auf die Lösung mit Hilfe des Newton Verfahrens näher eingegangen. In [16, Chapter 5, Abschnitt 5 und 6] werden auch alternative Lösungsmethoden mit Vor- und Nachteilen besprochen.

Berechnungen für Algorithmus 7, bei Verwendung des Newton Verfahrens

Nachfolgend werden die wesentlichen Berechnungen für das lokale Newton Verfahren angegeben (gemäß Unterkapitel 1.3), welche in Zeile 2 von Algorithmus 7 erforderlich sind, um $\widehat{\mathbf{X}}$ bzw. eine neue Iterierte $\widehat{\mathbf{X}}^{(\ell+1)}$ zu berechnen. Diese Berechnungen sind ebenfalls die Grundlagen für die Anwendung eines gedämpften Newton Verfahrens.
Gesucht wird nun $\widehat{\mathbf{X}}$, sodass gemäß (3.30) $\mathbf{F}(\widehat{\mathbf{X}}) = \mathbf{0}_{m\cdot(2\cdot K+1)}$ gilt. Unter Verwendung der Kettenregel (siehe Satz 1.9) wird folgende Jabobi-Matrix von $\mathbf{F}$ berechnet

$$\mathbf{J}_{\mathbf{F}}(\widehat{\mathbf{X}}) = \mathbf{\Omega}\cdot\mathbf{\Gamma}_{\mathrm{DFT}}\cdot\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\cdot\mathbf{\Gamma}_{\mathrm{DFT}}^{-1} + \mathbf{\Gamma}_{\mathrm{DFT}}\cdot\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\cdot\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\ ,$$

wobei die Jacobi-Matrizen $\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})$ und $\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})$ gemäß (3.27) und (3.28) gegeben sind.
Aufgrund der Blockdarstellung der Matrizen $\mathbf{J}_{\widetilde{\mathbf{f}}}(\mathbf{y})$ und $\mathbf{J}_{\widetilde{\mathbf{q}}}(\mathbf{y})$, sowie der Blockdiagonaldarstellung von $\mathbf{\Gamma}_{\mathrm{DFT}}$ und $\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}$, gilt folgende Darstellung

$$\mathbf{\Gamma}_{\mathrm{DFT}}\cdot\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\cdot\mathbf{\Gamma}_{\mathrm{DFT}}^{-1} = \begin{pmatrix} \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{11}\cdot\mathbf{\Gamma}^{-1} & \dots & \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{1m}\cdot\mathbf{\Gamma}^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{m1}\cdot\mathbf{\Gamma}^{-1} & \dots & \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{f}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{mm}\cdot\mathbf{\Gamma}^{-1} \end{pmatrix}\ ,$$

$$\mathbf{\Gamma}_{\mathrm{DFT}}\cdot\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\cdot\mathbf{\Gamma}_{\mathrm{DFT}}^{-1} = \begin{pmatrix} \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{11}\cdot\mathbf{\Gamma}^{-1} & \dots & \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{1m}\cdot\mathbf{\Gamma}^{-1} \\ \vdots & \ddots & \vdots \\ \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{m1}\cdot\mathbf{\Gamma}^{-1} & \dots & \mathbf{\Gamma}\cdot\big[\mathbf{J}_{\widetilde{\mathbf{q}}}\big(\mathbf{\Gamma}_{\mathrm{DFT}}^{-1}\cdot\widehat{\mathbf{X}}\big)\big]_{mm}\cdot\mathbf{\Gamma}^{-1} \end{pmatrix}\ .$$

Die neue Iterierte $\widehat{\mathbf{X}}^{(\ell+1)}$ im lokalen Newton Verfahren berechnet sich dann folgendermaßen

$$\widehat{\mathbf{X}}^{(\ell+1)} = \widehat{\mathbf{X}}^{(\ell)} - \big(\mathbf{J}_{\mathbf{F}}(\widehat{\mathbf{X}}^{(\ell)})\big)^{-1}\cdot\mathbf{F}(\widehat{\mathbf{X}}^{(\ell)})\ . \tag{3.31}$$

Gemäß [2, Abschnitt 3.1.5] kann als Startwert für die erste Newton Iteration z.B. $\widehat{\mathbf{X}}^{(0)} = \widehat{\mathbf{X}}_{\mathrm{DC}}$ gewählt werden, wobei $\widehat{\mathbf{X}}_{\mathrm{DC}}$ die DFT Koeffizienten des DC-Arbeitspunktes im elektrischen Netzwerk sind, d.h. $\widehat{\mathbf{X}}_{\mathrm{DC}}$ sind die DFT Koeffizienten von $\widehat{\mathbf{x}}_{\mathrm{DC}} \in [\widehat{P}_0(T,\mathbb{R})]^m$.

Diskussion und Fehlerursachen im HB Verfahren

Zum Abschluss wird nun auf drei Ursachen hingewiesen, welche zu einem Fehler $\|\mathbf{x}-\widehat{\mathbf{x}}\|_{L^2([0,T])} > 0$ führen, wobei $\mathbf{x}(t)$ die Lösung des Randwertproblems (3.12) ist und $\widehat{\mathbf{x}}(t)$ einer approximierten Lösung gemäß (3.23) entspricht, welche durch Gleichung (3.30) bestimmt wurde.
Die erste Fehlerquelle resultiert, wenn Gleichung (3.30) nicht exakt gelöst wird. Dieser Fehler lässt sich aber z.B. mit dem Newton Verfahren beliebig klein machen. Die beiden weiteren Fehlerquellen haben ihren Ursprung darin, dass die Approximation in $[\widehat{P}_K(T,\mathbb{R})]^m$ liegt, d.h. nur durch endlich viele Harmonische dargestellt wird. Hierbei sei erwähnt, dass für eine Funktion $\mathbf{x} \in [\widehat{P}(T,\mathbb{R})]^m$, bei hinreichend großem K, die Funktion $\mathbf{x}(t)$ durch $\widehat{\mathbf{x}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ beliebig genau approximiert werden kann. Nachdem im HB Verfahren die Komponentenfunktionen von $\widehat{\mathbf{x}}(t)$, durch die DFT Koeffizienten von $\widehat{\mathbf{X}}$ vorgegeben werden, resultiert die zweite Fehlerquelle, von der Abweichung zwischen $\widehat{\mathbf{x}}(t)$ und $\mathbf{x}(t)$. Die dritte und dominante Fehlerquelle im HB Verfahren lässt sich durch folgende Überlegung nachvollziehen. Aufgrund der algebraischen Nichtlinearitäten von $\mathbf{f}(\mathbf{x})$ und $\mathbf{q}(\mathbf{x})$, kann zwar für $\widehat{\mathbf{x}} \in [\widehat{P}_K(T,\mathbb{R})]^m$ gewährleistet werden, dass die Funktion

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\widehat{\mathbf{x}}(t)) + \mathbf{f}(\widehat{\mathbf{x}}(t)) + \mathbf{b}(t) \tag{3.32}$$

periodisch ist, allerdings könnten für die approximierende Fourierreihe der Funktion (3.32), mehr als K Harmonische erforderlich sein, um eine gute Näherung zu liefern. Nachdem $\widehat{\mathbf{X}}$ gemäß (3.23) bestimmt wurde, definieren die DFT Koeffizienten $\mathbf{F}(\widehat{\mathbf{X}})$ gemäß (3.29), eine Funktion in $[\widehat{P}_K(T,\mathbb{R})]^m$, welche eine Näherung von (3.32) darstellt. Ob diese Näherung mit K Harmonischen allerdings schon ausreichend ist, ist noch nicht sichergestellt. Dieser Fehler ist vor allem deswegen dominant, da durch die algebraischen Nichtlinearitäten, Terme aus Summen- und Differenzfrequenzen entstehen, welche im Allgemeinen nicht ausreichend durch eine Funktion in $[\widehat{P}_K(T,\mathbb{R})]^m$ berücksichtigt werden können. Zudem sei auch zu bemerken, dass bei allgemeinen algebraischen Nichtlinearitäten, eine kleine Ursache eine große Wirkung erzeugen kann. Weshalb man im Allgemeinen nicht davon ausgehen kann, dass 'kleine' Amplituden bei höheren Frequenzkomponenten keine Auswirkung haben. Dass das HB Verfahren allerdings für eine ausreichend große Anzahl an Harmonischen eine gute Näherung für die Lösung des Randwertproblems (3.12) liefert, kann einerseits durch die Stetigkeit der Funktionen $\mathbf{f}(\mathbf{x})$ und $\mathbf{q}(\mathbf{x})$ gewährleistet werden, sowie der Eigenschaft, dass für eine Funktion aus $P(T,\mathbb{R})$ (gemäß (3.13)), die entsprechenden Fourierkoeffizienten eine Nullfolge bilden. So ist z.B. durch die Stetigkeit von $\mathbf{f}(\mathbf{x})$ gewährleistet, dass für jeden Punkt $\mathbf{x}_0 \in \mathbb{R}^m$, für alle $\varepsilon > 0$ ein von $\mathbf{x}_0$ abhängiges $\delta(\mathbf{x}_0) > 0$ existiert, sodass $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{x}_0)\| < \varepsilon$ für alle $\|\mathbf{x} - \mathbf{x}_0\| < \delta(\mathbf{x}_0)$ gilt. D.h. bei hinreichend kleiner Aussteuerung, ist auch der Abstand zwischen den Funktionswerte klein. Ohne zusätzliche Voraussetzungen an eine stetige Funktion, kann man allerdings im Allgemeinen nicht angeben, wie klein eine Aussteuerung sein muss, damit sich die Funktionswerte nur unwesentlich ändern. Falls die Nichtlinearitäten bestimmte Eigenschaften besitzen, ist es allerdings möglich den Fehler zu schätzen und hierdurch kann z.B. eine minimale erforderliche Anzahl von Harmonischen $K_{\min}$ angegeben werden, womit eine Fehlergenauigkeit garantiert wird. Für eine genauere Fehleranalyse sei z.B. auf [16, Chapter 5, Abschnitt 2.2, 3 und 8] und [9, Abschnitt 7.3.3] verwiesen.
Für Anmerkungen wie eine Implementierung des HB Verfahrens optimiert bzw. beschleunigt werden kann, sei z.B. auf [16, Chapter 6] verwiesen. In [16, Chapter 6, Abschnitt 1] wird auch begründet, weshalb eine Implementierung des HB Verfahrens mit der FFT, im Vergleich zur Implementierung mit der DFT, nur kaum zu einer schnelleren Berechnung führt. Der Hauptgrund ist, dass in der Berechnung, die Anwendung der DFT nur einen kleinen Anteil am Rechenaufwand hat, da z.B. Funktionsauswertungen und die Lösung eines nichtlinearen Gleichungssystems, mehr Ressourcen benötigt. Eine FFT ist zudem auch nur effizient, wenn die Anzahl der Frequenzenkomponenten z.B. einer Zweierpotenz entspricht. Hingegen kann die DFT für eine beliebige Anzahl von Frequenzenkomponenten implementiert werden.

3.4 Anwendungsbeispiele

In diesem Unterkapitel werden die in Kapitel 3 vorgestellten Methoden, auf die Einweg- und Brückengleichrichterschaltung, gemäß der Beispiele 2.4 und 2.5 aus Unterkapitel 2.2.3, angewendet. Für beide elektrische Netzwerke wurde folgende Konfiguration gewählt

$$f_0 = 3\,kHz\ , \quad \omega_0 = 2{\cdot}\pi{\cdot}f_0\ , \quad u_0(t) = 1\,V \cdot \sin(\omega_0 \cdot t)\ , \quad R_1 = 100\,\Omega\ , \quad C_1 = \begin{cases} 0\,F \\ 1\,\mu F \end{cases}\ ,$$

wobei die Gleichrichterschaltungen mit $C_1 = 0\,F$ und $C_1 = 1\,\mu F$ berechnet wurden.
Die Dioden $D_1, \ldots, D_4$ wurden gemäß der Shockley Gleichung

$$i_D(u_D) = I_S \cdot \left(\exp\left(\frac{u_D}{n \cdot U_T} \right) - 1 \right)$$

modelliert und als Schottky Dioden BAR43S implementiert. Die gewählten Parameter sind

$$k_B = 1.380649{\cdot}10^{-23}\,J/K\ , \quad T = 300\,K\ , \quad e = 1.602176634{\cdot}10^{-19}\,C\ , \quad U_T = \frac{k_B \cdot T}{e}\ ,$$
$$n = 1.4622\ , \quad I_S = 0.4345{\cdot}10^{-6}\,A\ ,$$

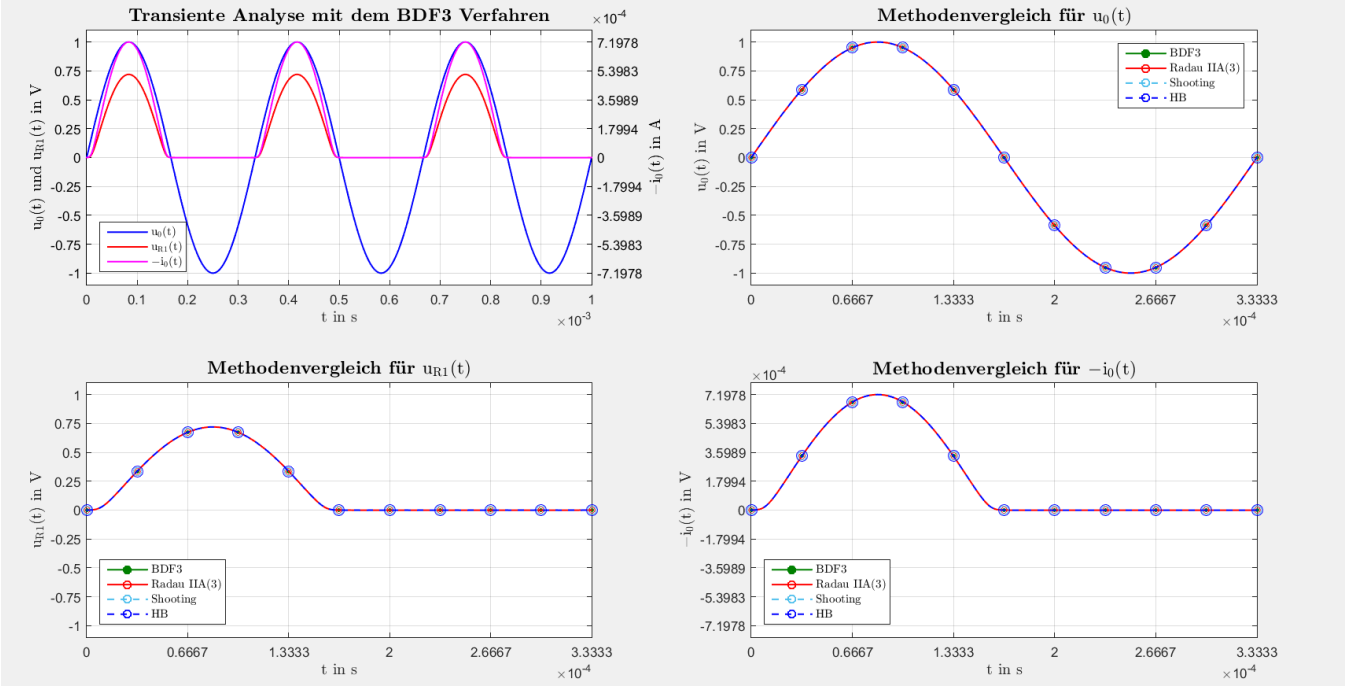
welche gemäß Typ BAR43SFILM, der PSPICE Bibliothek [24] entnommen wurden. Bei der Implementierung der Methoden aus Kapitel 3, war es erforderlich nichtlineare Gleichungssysteme zu lösen. Diese wurden stets mit dem gedämpften Newton Verfahren nach Bank-Rose, gemäß Algorithmus 2 aus Unterkapitel 1.3, gelöst. Der Algorithmus wurde dabei mit folgenden Parametern implementiert

$$\mathrm{Tol}_{abs} = 10^{-7}\ , \quad \mathrm{Tol}_{rel} = 10^{-7}\ , \quad \mathrm{Tol}_{Fkt} = 10^{-7}\ , \quad \omega = 0.9\ .$$

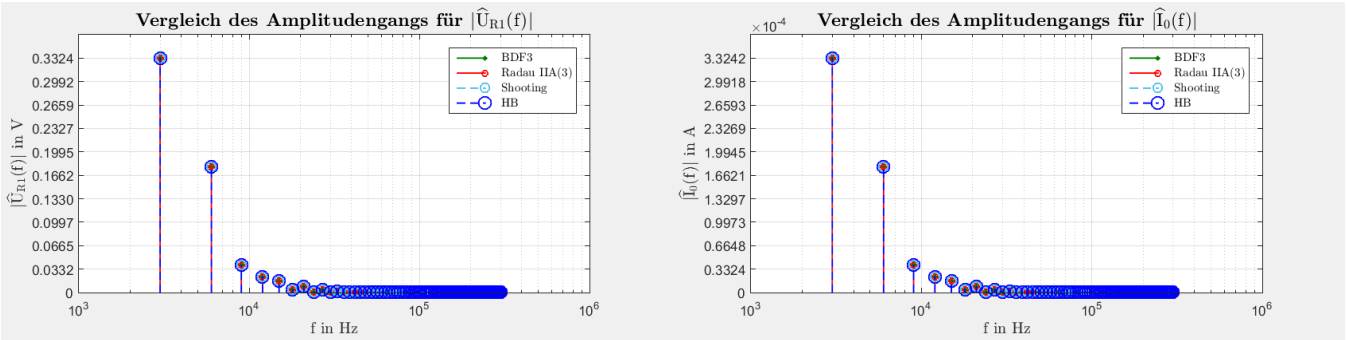
Um die einzelnen Methoden vergleichen zu können, wurde mit $N_{S/T} = 201$, die Anzahl der äquidistanten Stützstellen je Periodendauer vorgegeben. Nachdem bei den implementierten Methoden vor allem der eingeschwungene Zustand von Interesse ist und die transiente Analyse bei Verwendung der Zeitschrittverfahren BDF und Radau IIA das transiente Verhalten ermittelt, wurde für den Vergleich der Methoden im Zeit- und Frequenzbereich, lediglich die letzte Periode analysiert, wobei hierbei bereits der transiente Vorgang abgeklungen war. Bei der Ermittlung des Amplitudengangs, wurden zudem für alle verwendeten Zeitbereichsmethoden, mit Hilfe der $N_{S/T}$ Zeitwerte pro Periode, die Frequenzkomponenten durch die reelle DFT berechnet. Die Vorgabe von $N_{S/T}$ Stützstellen je Periode, sind für Zeitschrittverfahren intuitiv und werden für das HB Verfahren folgendermaßen berücksichtigt. Im HB Verfahren wird die Anzahl der Harmonischen mit $K \in \mathbb{N}$ vorgegeben und aus den daraus ermittelten DFT Koeffizienten, erhält man für jede Komponente der gesuchten Lösung $\mathbf{x}(t)$, eine approximierte Fourierreihe gemäß (3.18). Diese approximierte Fourierreihe wird dann in den $N_{S/T}$ Stützstellen ausgewertet.

Zusammenfassung der Ergebnisse des Einweggleichrichters

In Abbildung 3.2 ist die Auswertung des Beispiels 2.4, bei $C_1 = 0\,F$ zu sehen.
Abbildung 3.2a zeigt für die Größen $u_0(t)$, $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich im Zeitbereich, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand. Zudem ist auch die transiente Analyse enthalten, welche mit dem BDF3 Verfahren berechnet wurde.
Abbildung 3.2b zeigt für die Größen $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich des Amplitudengangs, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand.
Für Abbildung 3.2 wurde dabei das HB Verfahren mit $K = 30$ berechnet.



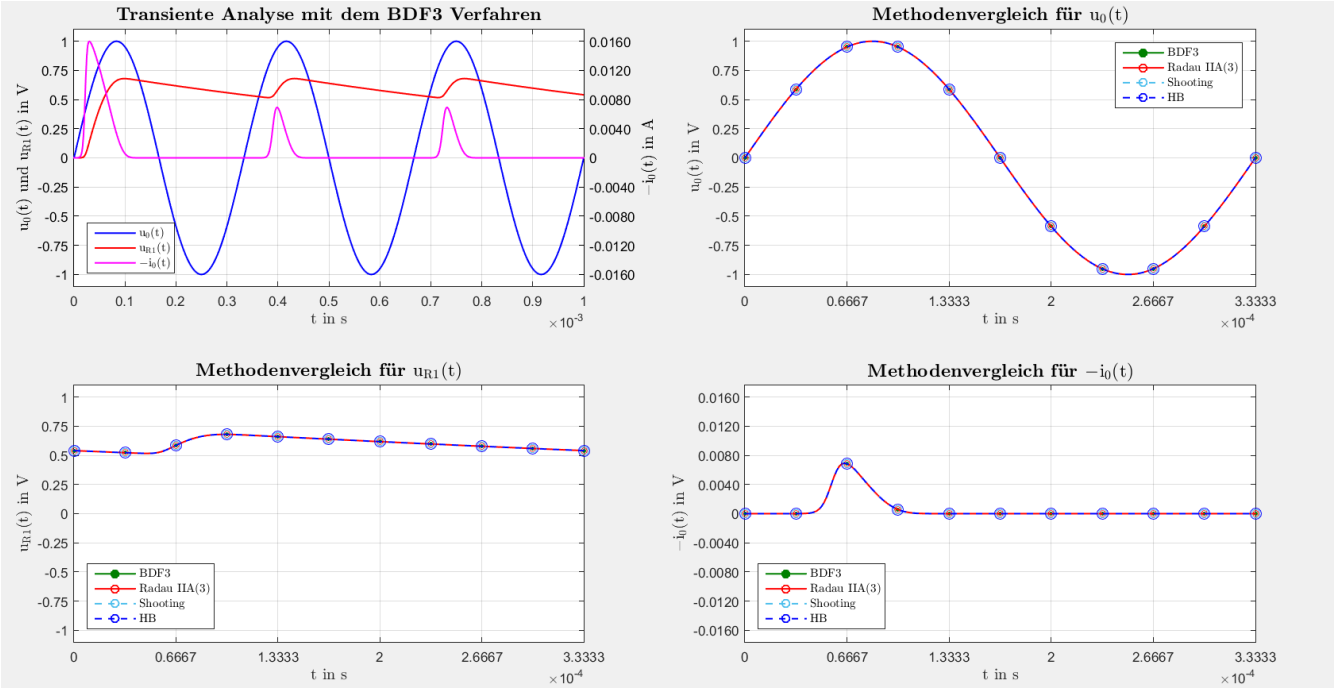
(a) Transiente Analyse und Vergleich der Methoden im Zeitbereich



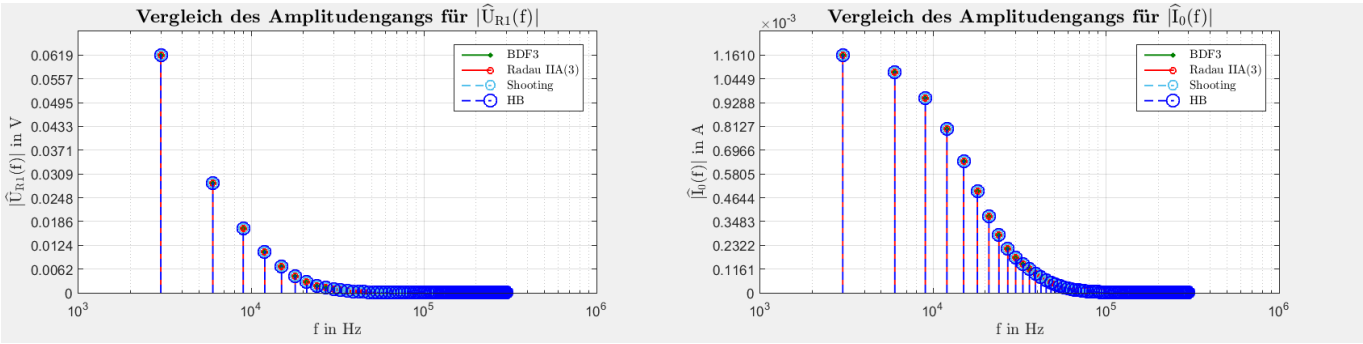
(b) Vergleich des Amplitudengangs

Abbildung 3.2: Vergleich der Methoden im Zeit und Frequenzbereich.

In Abbildung 3.3 ist die Auswertung des Beispiels 2.4, bei $C_1 = 1\,F$ zu sehen.
Abbildung 3.3a zeigt für die Größen $u_0(t)$, $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich im Zeitbereich, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand. Zudem ist auch die transiente Analyse enthalten, welche mit dem BDF3 Verfahren berechnet wurde.
Abbildung 3.3b zeigt für die Größen $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich des Amplitudengangs, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand.
Für Abbildung 3.3 wurde dabei das HB Verfahren mit $K = 30$ berechnet.



(a) Transiente Analyse und Vergleich der Methoden im Zeitbereich



(b) Vergleich des Amplitudengangs

Abbildung 3.3: Vergleich der Methoden im Zeit und Frequenzbereich.

Gemäß Beispiel 2.4 mit $C_1 = 0\,F$, ist in Abbildung 3.4 das Ergebnis des HB Verfahrens bei $K \in \{5, 10\}$ zu sehen. Dass das HB Verfahren bereits für $K = 10$ eine gute Näherung liefert, liegt vor allem am Amplitudengang von $u_{R_1}(t)$ und $-i_0(t)$ gemäß Abbildung 3.2b, da k.te Harmonische mit $k > 10$ nur mehr einen geringen Beitrag leisten.

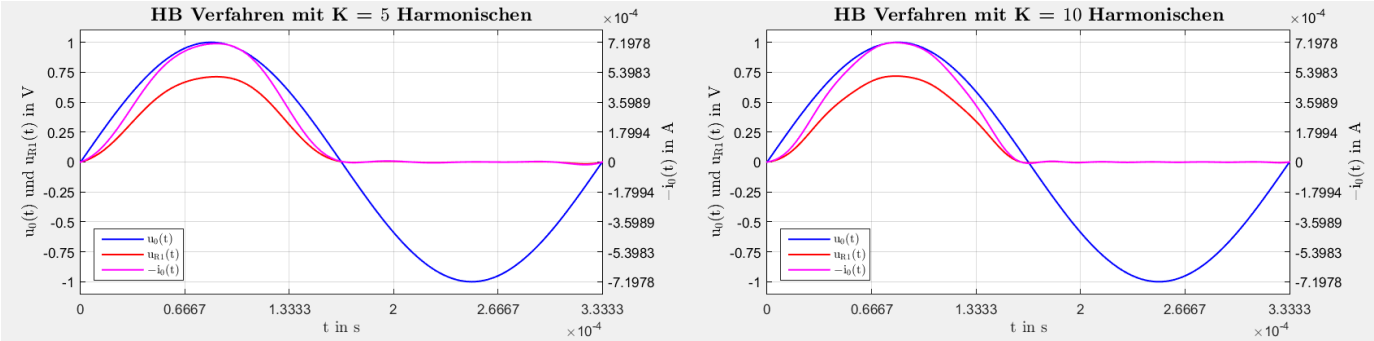


Abbildung 3.4: Vergleich des HB Verfahrens bei $C_1 = 0\,F$ und $K \in \{5, 10\}$.

Gemäß Beispiel 2.4 mit $C_1 = 1\,F$, ist in Abbildung 3.5 das Ergebnis des HB Verfahrens bei $K \in \{5, 10\}$ zu sehen. Der Amplitudengang von $u_{R_1}(t)$ gemäß Abbildung 3.3b, begründet weshalb bereits für $K = 10$ eine gute Näherung von $u_{R_1}(t)$ mit dem HB Verfahren gefunden werden kann. Anderseits wird durch den Amplitudengang von $-i_0(t)$ ersichtlich, weshalb für eine gute Näherung auch k.te Harmonische mit $k > 10$ berechnet werden müssen. Zudem erkennt man anhand des Stroms $-i_0(t)$ in Abbildung 3.5, dass ein nicht physikalisches Verhalten des Sperrstroms durch die Diode D_1 zu beobachten ist, da der Sperrstrom vernachlässigbar klein sein sollte. Dieses Verhalten verbessert sich allerdings, wenn K größer gewählt wird, wie z.B. das Ergebnis in Abbildung 3.3 bei $K = 30$ zeigt.

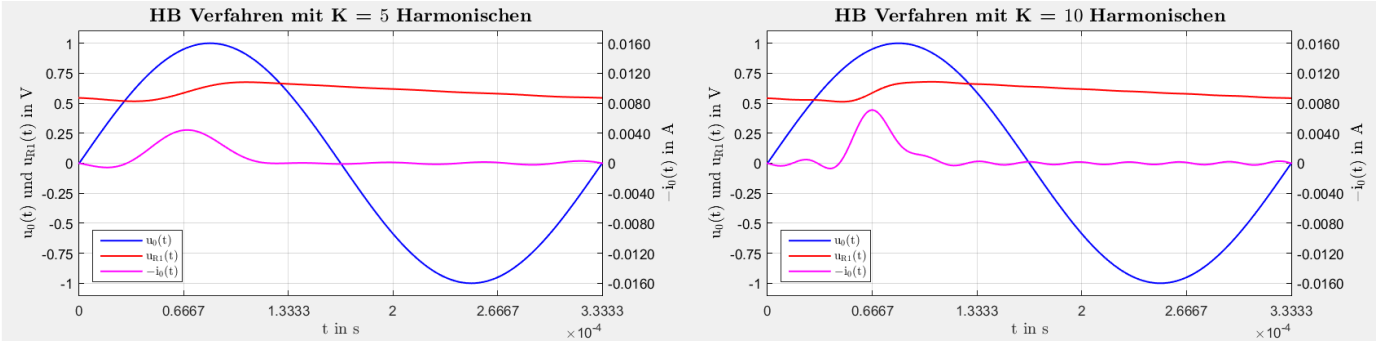
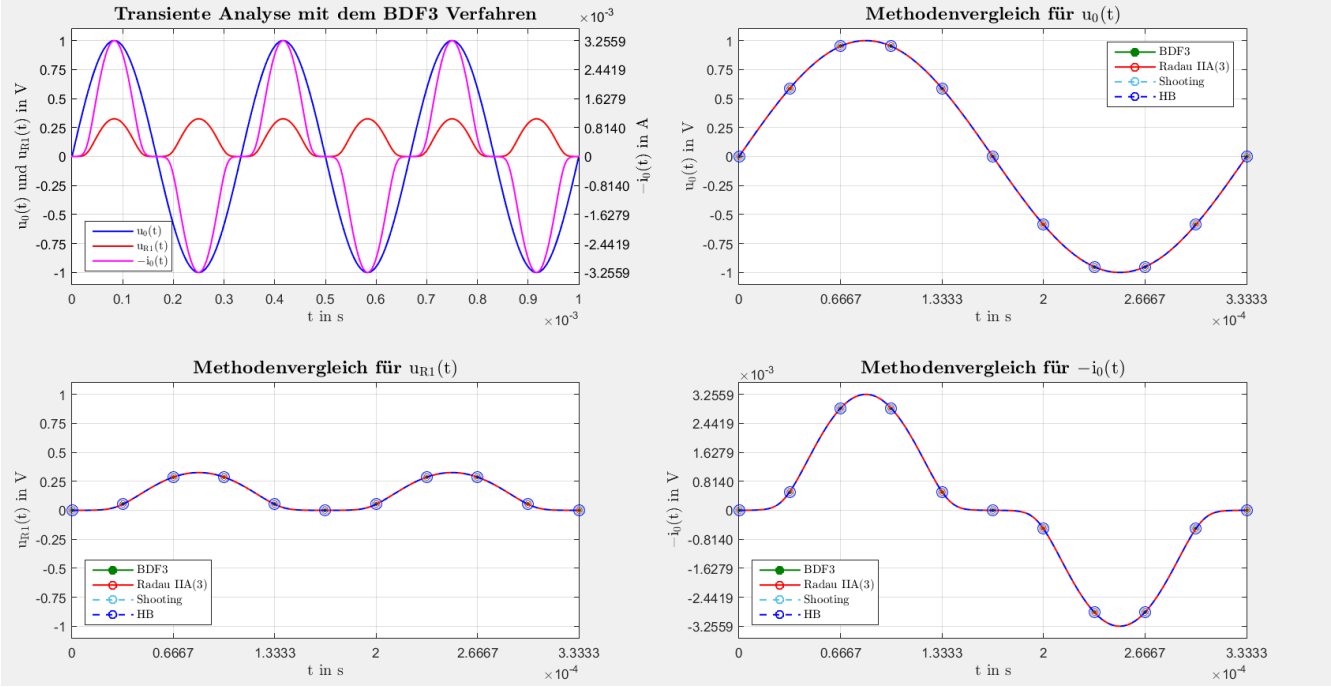


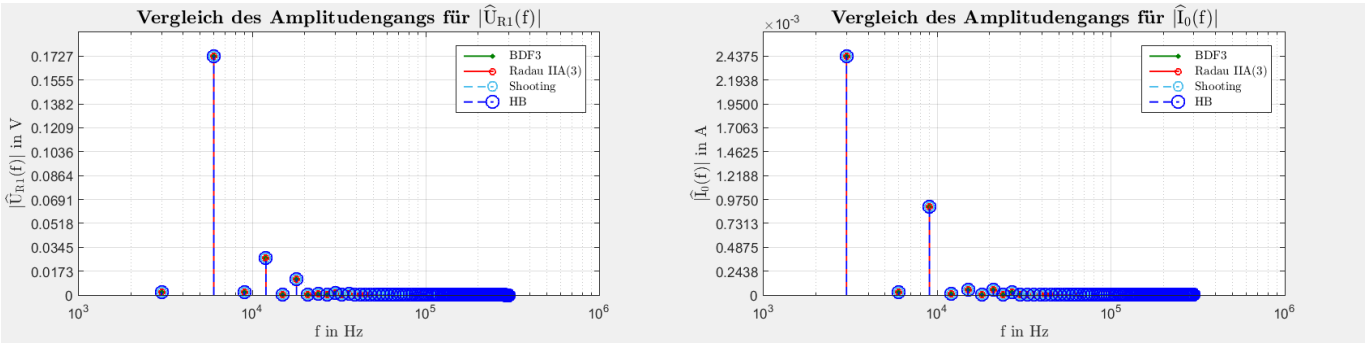
Abbildung 3.5: Vergleich des HB Verfahrens bei $C_1 = 0\,F$ und $K \in \{5, 10\}$.

Zusammenfassung der Ergebnisse des Brückengleichrichters

In Abbildung 3.6 ist die Auswertung des Beispiels 2.5, bei $C_1 = 0\,F$ zu sehen.
Abbildung 3.6a zeigt für die Größen $u_0(t)$, $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich im Zeitbereich, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand. Zudem ist auch die transiente Analyse enthalten, welche mit dem BDF3 Verfahren berechnet wurde.
Abbildung 3.6b zeigt für die Größen $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich des Amplitudengangs, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand.
Für Abbildung 3.6 wurde dabei das HB Verfahren mit $K = 30$ berechnet.



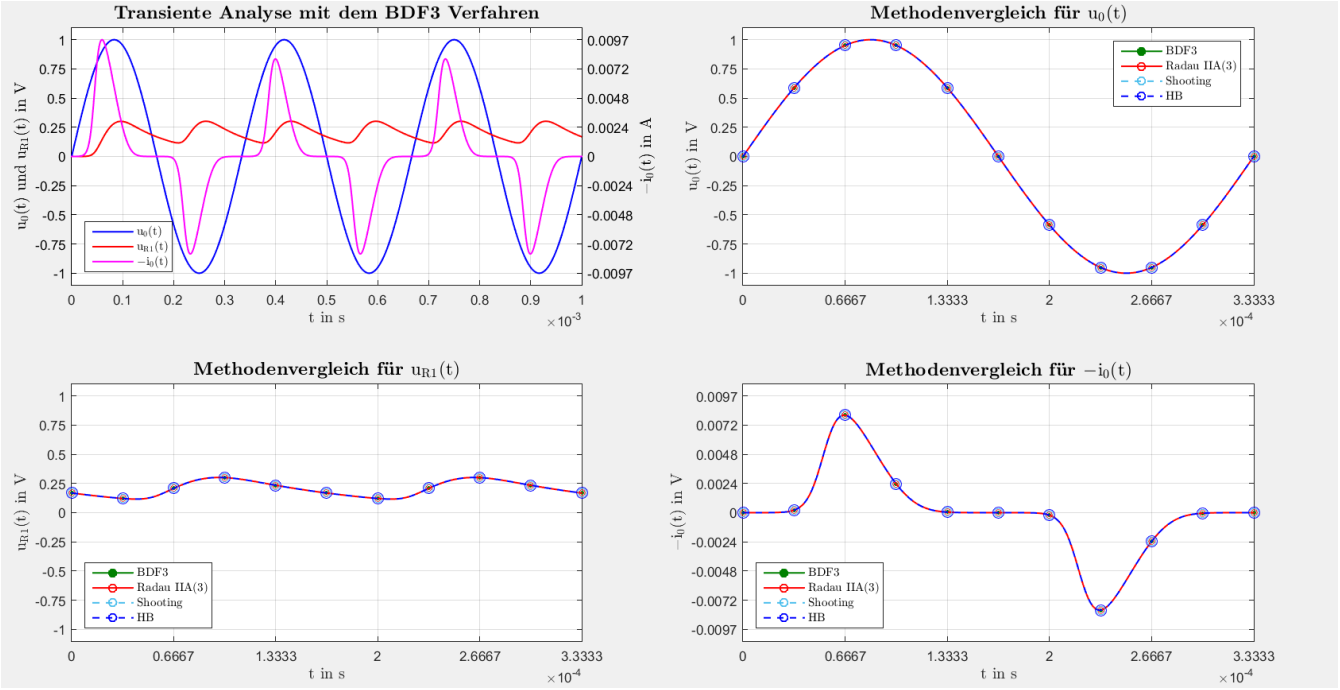
(a) Transiente Analyse und Vergleich der Methoden im Zeitbereich



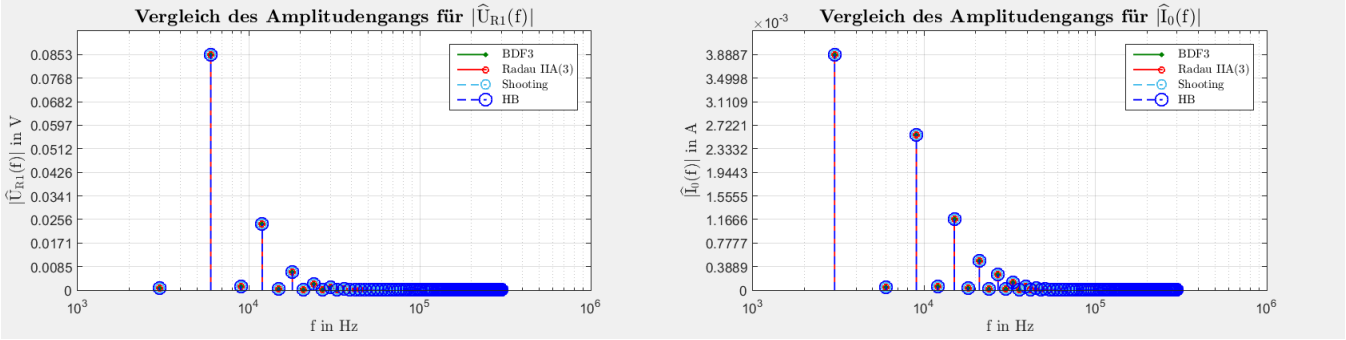
(b) Vergleich des Amplitudengangs

Abbildung 3.6: Vergleich der Methoden im Zeit und Frequenzbereich.

In Abbildung 3.7 ist die Auswertung des Beispiels 2.5, bei $C_1 = 1\,F$ zu sehen.
Abbildung 3.7a zeigt für die Größen $u_0(t)$, $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich im Zeitbereich, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand. Zudem ist auch die transiente Analyse enthalten, welche mit dem BDF3 Verfahren berechnet wurde.
Abbildung 3.7b zeigt für die Größen $u_{R_1}(t)$ und $-i_0(t)$, den Vergleich des Amplitudengangs, der unterschiedlichen Methoden im periodisch eingeschwungenen Zustand.
Für Abbildung 3.7 wurde dabei das HB Verfahren mit $K = 30$ berechnet.



(a) Transiente Analyse und Vergleich der Methoden im Zeitbereich



(b) Vergleich des Amplitudengangs

Abbildung 3.7: Vergleich der Methoden im Zeit und Frequenzbereich.

Gemäß Beispiel 2.5 mit $C_1 = 0\,F$, ist in Abbildung 3.8 das Ergebnis des HB Verfahrens bei $K \in \{5, 10\}$ zu sehen. Dass das HB Verfahren bereits für $K = 10$ eine gute Näherung liefert, liegt vor allem am Amplitudengang von $u_{R_1}(t)$ und $-i_0(t)$ gemäß Abbildung 3.6b, da k.te Harmonische mit $k > 10$ nur mehr einen geringen Beitrag leisten.

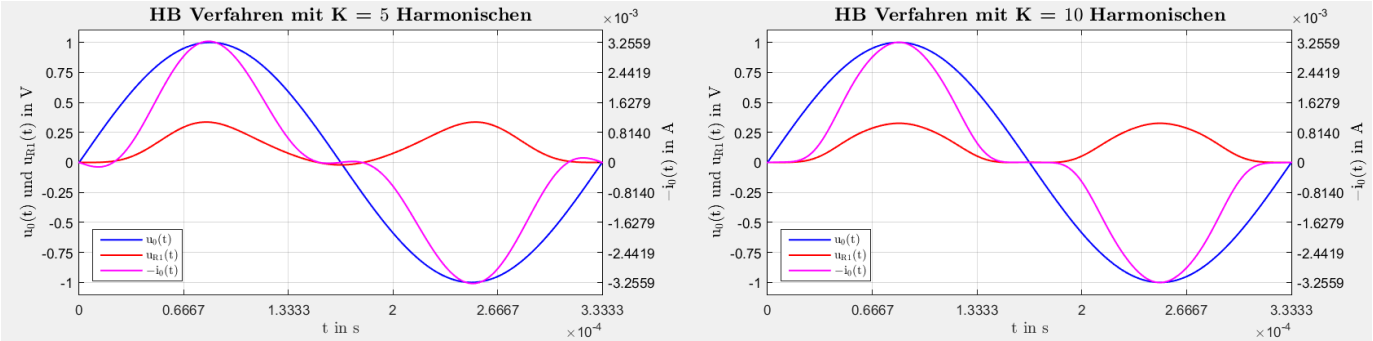


Abbildung 3.8: Vergleich des HB Verfahrens bei $C_1 = 0\,F$ und $K \in \{5, 10\}$.

Gemäß Beispiel 2.5 mit $C_1 = 1\,F$, ist in Abbildung 3.9 das Ergebnis des HB Verfahrens bei $K \in \{5, 10\}$ zu sehen. Der Amplitudengang von $u_{R_1}(t)$ gemäß Abbildung 3.7b, begründet weshalb bereits für $K = 10$ eine gute Näherung von $u_{R_1}(t)$ mit dem HB Verfahren gefunden werden kann. Anderseits wird durch den Amplitudengang von $-i_0(t)$ ersichtlich, weshalb für eine gute Näherung auch k.te Harmonische mit $k > 10$ berechnet werden müssen. Zudem erkennt man anhand des Stroms $-i_0(t)$ in Abbildung 3.9, dass ein nicht physikalisches Verhalten des Sperrstroms durch die Dioden $D_1, \ldots, D_4$ zu beobachten ist, da der Sperrstrom vernachlässigbar klein sein sollte. Dieses Verhalten verbessert sich allerdings, wenn K größer gewählt wird, wie z.B. das Ergebnis in Abbildung 3.7 bei $K = 30$ zeigt.

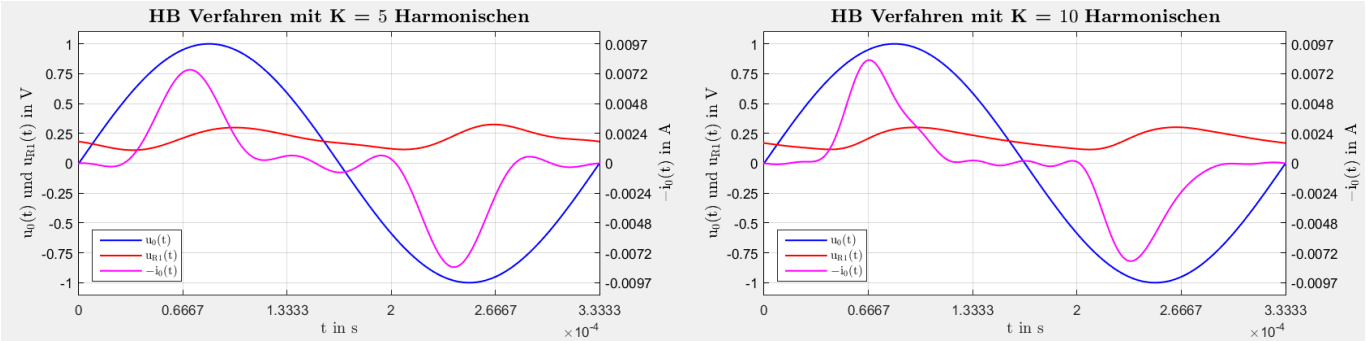


Abbildung 3.9: Vergleich des HB Verfahrens bei $C_1 = 0\,F$ und $K \in \{5, 10\}$.

4 Berechnung des periodisch eingeschwungenen Zustands eines NFC Systems

Unter Verwendung der quasi-stationären Partial Element Equivalent Circuit (PEEC) Methode wird ein gegebenes NFC Spulenpaar durch ein äquivalentes elektrisches Netzwerk modelliert. Dieses wird mit einem externen elektrischen Netzwerk verbunden, welches lineare Elemente, sowie Dioden enthält. Die mathematische Beschreibung des vollständigen elektrischen Netzwerkes, erfolgt durch differentiell algebraische Gleichungen, welche durch die Anwendung des modifizierten Knotenspannungsverfahrens entstehen. Für die Berechnung des periodisch eingeschwungenen Zustands, werden die numerischen Methoden aus Kapitel 3 zur Lösung der DAG angewendet und untersucht.

4.1 Modellierung unter Anwendung der Partial Element Equivalent Circuit Methode

In vielen Anwendungen kann ein elektromagnetisches Feldproblem, unter Anwendung der PEEC Methode, durch ein äquivalentes elektrisches Netzwerk, bestehend aus resistiven, kapazitiven und induktiven Elementen, beschrieben bzw. modelliert werden. Damit ist es möglich, ein externes elektrisches Netzwerk mit dem elektromagnetischen Feldproblem direkt zu koppeln. Für die Berechnung des periodisch eingeschwungenen Zustands für dieses resultierende elektrische Netzwerk, können numerische Methoden verwendet werden, welche für die Analyse elektrischer Netzwerke geeignet sind.
Ausgehend von den Maxwell Gleichungen, können mit Hilfe der Potentialtheorie, Integralgleichungen hergeleitet werden, dessen Diskretisierung und Interpretation zur PEEC Methode führen (siehe z.B. [11, Chapter 3]). Es existieren dabei unterschiedliche Ansätze, welche unter anderem Wellenausbreitung, sowie leitfähige, magnetische und dielektrische Materialien berücksichtigen (siehe z.B. [18] oder [11]).
Der Ausgangspunkt dieses Abschnitts, ist die Modellierung eines geeigneten elektromagnetischen Feldproblems durch ein äquivalentes elektrisches Netzwerk, welches durch die Anwendung der quasi-stationären PEEC Methode für leitfähige Strukturen angegeben werden kann und somit mit einem externen elektrischen Netzwerk verbunden werden kann. Die mathematische Beschreibung des vollständigen elektrischen Netzwerkes, erfolgt durch differentiell algebraische Gleichungen, welche gemäß dem Modellproblems 3.1, durch die Anwendung des modifizierten Knotenspannungsverfahrens entstehen.

Quasi-stationäre 1D PEEC Methode für leitfähige Strukturen

In diesem Abschnitt wurden die Grundlagen zur PEEC Methode aus [11] entnommen.
Eine quasi-stationäre PEEC Methode für leitfähige Strukturen vernachlässigt die endliche Ausbreitungsgeschwindigkeit bzw. den Verschiebestrom und berücksichtigt keine magnetischen oder

dielektrischen Materialen. Die geometrische leitfähige Struktur wird dabei in zylindrische Elemente diskretisiert, welche auch als Sticks bezeichnet werden. Jeder Stick wird dabei durch eine PEEC-Zelle gemäß Abbildung 4.1 modelliert (siehe [11, Abschnitt 3.2.2]).

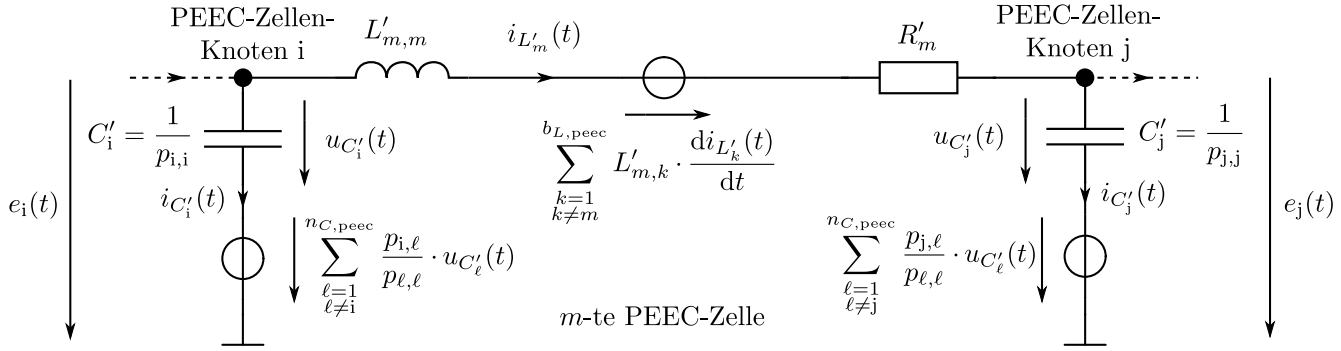


Abbildung 4.1: m-te PEEC-Zelle einer leitfähigen Struktur, zwischen den PEEC-Knoten i, j, bestehend aus $b_{L,\mathrm{peec}}$ PEEC-Zellen und $n_{C,\mathrm{peec}}$ PEEC-Knoten.

Besteht eine leitfähige Struktur aus $b_{L,\mathrm{peec}}$ PEEC-Zellen und $n_{C,\mathrm{peec}}$ PEEC-Knoten, so wird das elektromagnetische Feldproblem gemäß Abbildung 4.1 durch eine symmetrische und invertierbare partielle Potential-Matrix $\mathbf{P} = \big(p_{i,j}\big)_{1\leq i,j\leq n_{C,\mathrm{peec}}}$, eine partielle Induktivitäts-Matrix $\mathbf{L}' = \big(L'_{i,j}\big)_{1\leq i,j\leq b_{L,\mathrm{peec}}}$ und eine partielle Widerstands-Diagonalmatrix $\mathbf{R}' = \mathrm{diag}\big(R'_1, \ldots, R'_{b_{L,\mathrm{peec}}}\big)$ modelliert, wobei $\mathbf{P} \in \mathbb{R}^{n_{C,\mathrm{peec}}\times n_{C,\mathrm{peec}}}$ und $\mathbf{L}', \mathbf{R}' \in \mathbb{R}^{b_{L,\mathrm{peec}}\times b_{L,\mathrm{peec}}}$ gelte und $\mathbf{P}, \mathbf{L}'$ vollbesetzt sind. Die Berechnung der Koeffizienten der Matrizen $\mathbf{P}, \mathbf{L}', \mathbf{R}'$ erfolgt mittels 1D Diskretisierung (siehe z.B. [11, Abschnitt 3.3.3]). Die Berechnung der Koeffizienten von $\mathbf{P}, \mathbf{L}', \mathbf{R}'$ wurde zur Verfügung gestellt (siehe [25].

Für $\ell \in \{1, \ldots, n_{C,\mathrm{peec}}\}$ gilt $p_{\ell,\ell} > 0$. Es wird nun die Diagonalmatrix

$$\mathbf{F} := \mathrm{diag}\left(\frac{1}{p_{1,1}}, \ldots, \frac{1}{p_{n_{C,\mathrm{peec}},n_{C,\mathrm{peec}}}}\right) = \mathrm{diag}\big(C'_1, \ldots, C'_{n_{C,\mathrm{peec}}}\big) \tag{4.1}$$

definiert, dessen Hauptdiagonale aus den Pseudo-Kapazitäten C'_{i} der PEEC-Zellen besteht. Wird die Matrix $\mathbf{S} := \mathbf{P}\cdot\mathbf{F}$ definiert, dann gilt $\mathbf{S}^{-1} = \big(\mathbf{P}\cdot\mathbf{F}\big)^{-1} = \mathbf{F}^{-1}\cdot\mathbf{P}^{-1}$. Nachdem $\mathbf{P}$ invertierbar ist gilt $\mathbf{P}\cdot\mathbf{P}^{-1} = \mathbf{P}^{-1}\cdot\mathbf{P} = \mathbf{I}$ und somit folgt daraus $\mathbf{P}^\top\cdot\big(\mathbf{P}^{-1}\big)^\top = \big(\mathbf{P}^{-1}\cdot\mathbf{P}\big)^\top = \mathbf{I}^\top = \mathbf{I}$. Daher gilt $\big(\mathbf{P}^\top\big)^{-1} = \big(\mathbf{P}^{-1}\big)^\top$. Nachdem $\mathbf{P}^\top = \mathbf{P}$ gilt, folgt daraus $\mathbf{P}^{-1} = \big(\mathbf{P}^{-1}\big)^\top$. Die Matrix $\mathbf{S}$ ist nicht symmetrisch, aufgrund der Invertierbarkeit von $\mathbf{S}$ gilt aber ebenfalls $\big(\mathbf{S}^\top\big)^{-1} = \big(\mathbf{S}^{-1}\big)^\top$. Gemäß der Definition von $\mathbf{S}$, folgt $\mathbf{P}^{-1} = \mathbf{F}\cdot\mathbf{S}^{-1}$. Mit der Eigenschaft $\mathbf{F}^\top = \mathbf{F}$ folgt nun der Zusammenhang

$$\mathbf{F}\cdot\mathbf{S}^{-1} = \mathbf{P}^{-1} = \big(\mathbf{P}^{-1}\big)^\top = \big(\mathbf{F}\cdot\mathbf{S}^{-1}\big)^\top = \big(\mathbf{S}^{-1}\big)^\top\cdot\mathbf{F}^\top = \big(\mathbf{S}^\top\big)^{-1}\cdot\mathbf{F}\,. \tag{4.2}$$

Aufgrund der Definition von $\mathbf{S}$, gilt die Darstellung $\mathbf{S} = \left(\frac{p_{i,j}}{p_{j,j}}\right)_{1\leq i,j\leq n_{C,\mathrm{peec}}}$. Für $\mathrm{i} \in \{1, \ldots, n_{C,\mathrm{peec}}\}$ gilt gemäß Abbildung 4.1, für die Knotenspannung des i-ten PEEC-Knotens

$$e_{\mathrm{i}}(t) = u_{C'_{\mathrm{i}}}(t) + \sum_{\substack{\ell=1\\\ell\neq \mathrm{i}}}^{n_{C,\mathrm{peec}}} \frac{p_{\mathrm{i},\ell}}{p_{\ell,\ell}}\cdot u_{C'_\ell}(t) = \sum_{\ell=1}^{n_{C,\mathrm{peec}}} \frac{p_{\mathrm{i},\ell}}{p_{\ell,\ell}}\cdot u_{C'_\ell}(t) = \big[\mathbf{S}\cdot\mathbf{u}_{C'}(t)\big]_{\mathrm{i}}\,,$$

wobei $\mathbf{u}_{C'}(t) = \big(u_{C'_1}(t), \ldots, u_{C'_{n_{C,\mathrm{peec}}}}(t)\big)^\top$ gelte.
Wird $\mathbf{e}_{C,\mathrm{peec}}(t) = \big(e_{C_1,\mathrm{peec}}(t), \ldots, e_{C_{n_{C,\mathrm{peec}}},\mathrm{peec}}\big) := \big(e_1(t), \ldots, e_{n_{C,\mathrm{peec}}}(t)\big)$ definiert, dann gilt

$$\mathbf{e}_{C,\mathrm{peec}}(t) = \mathbf{S}\cdot\mathbf{u}_{C'}(t)\,, \quad \text{bzw.} \quad \mathbf{u}_{C'}(t) = \mathbf{S}^{-1}\cdot\mathbf{e}_{C,\mathrm{peec}}(t)\,. \tag{4.3}$$

Wird $\mathbf{i}_{C'}(t) = \left(i_{C'_1}(t), \ldots, i_{C'_{n_{C,\mathrm{peec}}}}(t)\right)^\top$ definiert, dann gilt aufgrund der Bauteilgleichung für lineare kapazitive Elemente, sowie der Beziehungen (4.1), (4.3) und (4.2), dass

$$\mathbf{i}_{C'}(t) = \mathrm{diag}\left(C'_1, \ldots, C'_{n_{C,\mathrm{peec}}}\right) \cdot \frac{\mathrm{d}\mathbf{u}_{C'}(t)}{\mathrm{d}t} = \mathbf{F} \cdot \mathbf{S}^{-1} \cdot \frac{\mathrm{d}\mathbf{e}_{C,\mathrm{peec}}(t)}{\mathrm{d}t} = \left(\mathbf{S}^\top\right)^{-1} \cdot \mathbf{F} \cdot \frac{\mathrm{d}\mathbf{e}_{C,\mathrm{peec}}(t)}{\mathrm{d}t} \,. \tag{4.4}$$

Werden die Vektoren $\mathbf{i}_{L,\mathrm{peec}}(t) = \left(i_{L_1,\mathrm{peec}}(t), \ldots, i_{L_{b_{L,\mathrm{peec}}},\mathrm{peec}}(t)\right)^\top := \left(i_{L'_1}(t), \ldots, i_{L'_{b_{L,\mathrm{peec}}}}(t)\right)^\top$ und $\mathbf{u}_{L,\mathrm{peec}}(t) := \left(u_{L_1,\mathrm{peec}}(t), \ldots, u_{L_{b_{L,\mathrm{peec}}},\mathrm{peec}}(t)\right)^\top$ definiert, dann gilt aufgrund der Bauteilgleichung für lineare Induktivitäten, für $m \in \{1, \ldots, b_{L,\mathrm{peec}}\}$

$$u_{L_m,\mathrm{peec}}(t) := u_{L'_{m,m}}(t) + \sum_{\substack{k=1 \\ k\neq m}}^{b_{L,\mathrm{peec}}} L'_{m,k} \cdot \frac{\mathrm{d}i_{L'_k}(t)}{\mathrm{d}t} = \sum_{k=1}^{b_{L,\mathrm{peec}}} L'_{m,k} \cdot \frac{\mathrm{d}i_{L'_k}(t)}{\mathrm{d}t} = \left[\mathbf{L}' \cdot \frac{\mathrm{d}\mathbf{i}_{L,\mathrm{peec}}(t)}{\mathrm{d}t}\right]_m .$$

Insbesondere gilt somit die Beziehung

$$\mathbf{u}_{L,\mathrm{peec}}(t) = \mathbf{L}' \cdot \frac{\mathrm{d}\mathbf{i}_{L,\mathrm{peec}}(t)}{\mathrm{d}t} \,. \tag{4.5}$$

Die Definition der nachfolgenden Matrizen $\mathbf{C}_{\mathrm{peec}}$ und $\mathbf{L}_{\mathrm{peec}}$, erfolgt für $\mathrm{i} \in \{1, \ldots, n_{C,\mathrm{peec}}\}$ und $m \in \{1, \ldots, b_{L,\mathrm{peec}}\}$, durch die Verwendung der Zeilenvektoren $\mathbf{C}^\top_{\mathrm{peec,i}}$ und $\mathbf{L}^\top_{\mathrm{peec},m}$, gemäß

$$\mathbf{C}_{\mathrm{peec}} = \begin{pmatrix} \mathbf{C}^\top_{\mathrm{peec},1} \\ \vdots \\ \mathbf{C}^\top_{\mathrm{peec},n_{C,\mathrm{peec}}} \end{pmatrix} := \left(\mathbf{S}^\top\right)^{-1} \cdot \mathbf{F} \,, \quad \mathbf{L}_{\mathrm{peec}} = \begin{pmatrix} \mathbf{L}^\top_{\mathrm{peec},1} \\ \vdots \\ \mathbf{L}^\top_{\mathrm{peec},b_{L,\mathrm{peec}}} \end{pmatrix} := \mathbf{L}' \,. \tag{4.6}$$

Mit der Definition der Vektoren $\mathbf{u}_{C,\mathrm{peec}}(t) = \left(u_{C_1,\mathrm{peec}}(t), \ldots, u_{C_{n_{C,\mathrm{peec}}},\mathrm{peec}}(t)\right)^\top := \mathbf{e}_{C,\mathrm{peec}}(t)$ und $\mathbf{i}_{C,\mathrm{peec}}(t) = \left(i_{C_1,\mathrm{peec}}(t), \ldots, i_{C_{n_{C,\mathrm{peec}}},\mathrm{peec}}(t)\right)^\top := \mathbf{i}_{C'}(t)$, gilt gemäß den Gleichungen (4.4) und (4.5)

$$\underbrace{\begin{pmatrix} \mathbf{i}_{C_1,\mathrm{peec}}(t) \\ \vdots \\ \mathbf{i}_{C_{n_{C,\mathrm{peec}}},\mathrm{peec}}(t) \end{pmatrix}}_{=\mathbf{i}_{C,\mathrm{peec}}(t)} = \begin{pmatrix} \mathbf{C}^\top_{\mathrm{peec},1} \\ \vdots \\ \mathbf{C}^\top_{\mathrm{peec},n_{C,\mathrm{peec}}} \end{pmatrix} \cdot \frac{\mathrm{d}\mathbf{u}_{C,\mathrm{peec}}(t)}{\mathrm{d}t} \,, \tag{4.7a}$$

$$\underbrace{\begin{pmatrix} \mathbf{u}_{L_1,\mathrm{peec}}(t) \\ \vdots \\ \mathbf{u}_{L_{b_{L,\mathrm{peec}}},\mathrm{peec}}(t) \end{pmatrix}}_{=\mathbf{u}_{L,\mathrm{peec}}(t)} = \begin{pmatrix} \mathbf{L}^\top_{\mathrm{peec},1} \\ \vdots \\ \mathbf{L}^\top_{\mathrm{peec},b_{L,\mathrm{peec}}} \end{pmatrix} \cdot \frac{\mathrm{d}\mathbf{i}_{L,\mathrm{peec}}(t)}{\mathrm{d}t} \,. \tag{4.7b}$$

Die Einführung der Vektoren $\mathbf{i}_{C,\mathrm{peec}}(t), \mathbf{u}_{C,\mathrm{peec}}(t), \mathbf{i}_{L,\mathrm{peec}}(t), \mathbf{u}_{L,\mathrm{peec}}(t)$ und $\mathbf{e}_{C,\mathrm{peec}}(t)$ sollen vor allem eine übersichtlichere Schreibweise bei der nachfolgenden Analyse durch das MKV ermöglichen. Aufgrund der bisherigen Definitionen und Eigenschaften, kann nun ein zu Abbildung 4.1 äquivalentes Ersatzschaltbild einer PEEC-Zelle angegeben werden, welches die kapazitive und induktive Kopplung gemäß den Beziehungen (4.7) verdeutlicht.

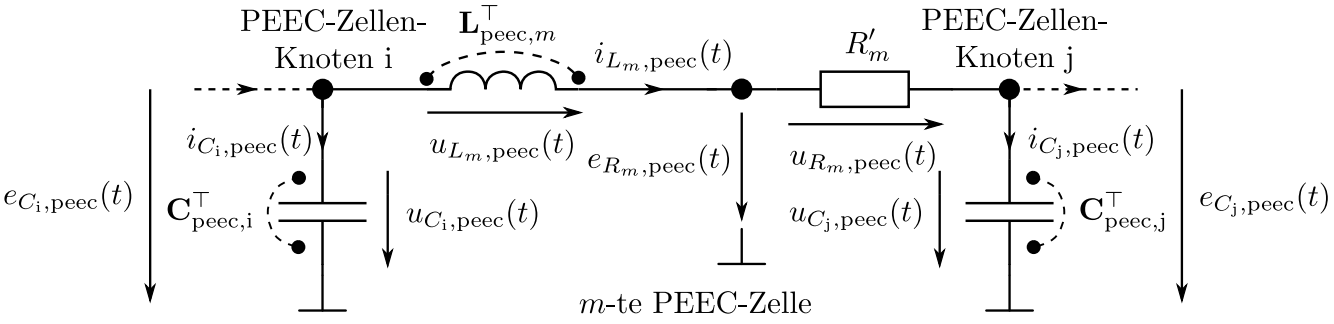


Abbildung 4.2: Äquivalentes Ersatzschaltbild der m-ten PEEC-Zelle einer leitfähigen Struktur, zwischen den PEEC-Knoten i, j, welches die kapazitive und induktive Kopplung durch entsprechende Zeilenvektoren der Matrizen $\mathbf{C}_{\text{peec}}$ und $\mathbf{L}_{\text{peec}}$ berücksichtigt.

Vorbereitungen für die Analyse des elektrischen Netzwerkes mit dem MKV

Für die nachfolgenden Betrachtungen, ist ein elektrisches Netzwerk mit dem MKV zu analysieren, wobei angenommen wird, dass die Voraussetzungen (i), (ii), (iii) des Modellproblems 3.1 auf Seite 36 erfüllt sind. Das elektrische Netzwerk besteht dabei aus einem äquivalenten elektrischen Netzwerk, welches durch die Anwendung der quasi-stationären PEEC Methode für leitfähige Strukturen entsteht, sowie einem externen elektrischen Netzwerk. Im Folgenden bezieht sich das Subskript 'ext' auf Elemente des externen Netzwerkes und das Subskript 'peec' auf Elemente der PEEC-Zellen. Weiters werden die Notationen des Abschnitts '*Quasi-stationäre 1D PEEC Methode für leitfähige Strukturen*' verwendet.
Das zu analysierende elektrische Netzwerk besitzt b_U unabhängige Spannungsquellen, b_I unabhängige Stromquellen, $b_C := n_{C,\text{peec}} + b_{C,ext}$ lineare kapazitive Elemente, $b_L := b_{L,\text{peec}} + b_{L,ext}$ lineare induktive Elemente, sowie $b_R := b_{R,\text{peec}} + b_{R,ext}$ resistive Elemente. Für die externen resistiven Elemente gelte $b_{R,ext} = b_{R,ext,lin} + b_{R,ext,NL}$, bestehend aus $b_{R,ext,lin}$ lineare und $b_{R,ext,NL}$ nichtlineare resistive Elemente. Die Anzahl der Knoten im elektrischen Netzwerk ergibt sich zu $n = 1 + n_{C,\text{peec}} + b_{L,\text{peec}} + n_{ext}$. Dabei wird der Bezugsknoten durch die PEEC Methode vorgegeben und in jeder PEEC-Zelle ist es für die Anwendung des MKV erforderlich, gemäß Abbildung 4.2, zwischen dem resistiven und dem gekoppelten induktiven Element einen zusätzlichen Knoten einzuführen.
Es wird nun eine Konvention bzgl. der Nummerierung der Knoten getroffen, welche sich insbesondere auf die Bedeutung der Zeilen der einzelnen Inzidenzmatrizen auswirkt. Die Knoten $1, \ldots, n_{C,\text{peec}}$ entsprechen gemäß Abbildung 4.2 den PEEC-Knoten an den gekoppelten kapazitiven Elementen. Die Knoten $n_{C,\text{peec}} + 1, \ldots, n_{C,\text{peec}} + b_{L,\text{peec}}$ entsprechen gemäß Abbildung 4.2, je PEEC-Zelle dem Knoten zwischen dem induktiven Element und dem Widerstand. Die Knotennummern $n_{C,\text{peec}} + b_{L,\text{peec}} + 1, \ldots, n_{C,\text{peec}} + b_{L,\text{peec}} + n_{ext}$ entsprechen den Knoten des externen elektrischen Netzwerkes. Mit dieser Konvention der Knotennummerierung können nun die (reduzierten) Teil-Inzidenzmatrizen der PEEC-Zellen Elemente angegeben werden, wobei für die (reduzierte) Inzidenzmatrix $\mathbf{A}$ folgende Aufteilung gewählt wird

$$\mathbf{A} := \big(\underbrace{\mathbf{A}_{R,peec}, \mathbf{A}_{R,ext,lin}, \mathbf{A}_{R,ext,NL}}_{=:\mathbf{A}_R}, \underbrace{\mathbf{A}_{C,peec}, \mathbf{A}_{C,ext}}_{=:\mathbf{A}_C}, \underbrace{\mathbf{A}_{L,peec}, \mathbf{A}_{L,ext}}_{=:\mathbf{A}_L}, \mathbf{A}_U, \mathbf{A}_I \big) ,$$

mit $\mathbf{A}_\ell \in \mathbb{R}^{(n-1)\times b_\ell}$ für $\ell \in \{R, C, L, U, I\}$, $\mathbf{A}_{R,ext,lin} \in \mathbb{R}^{(n-1)\times b_{\ell,ext,lin}}$,
$\mathbf{A}_{R,ext,NL} \in \mathbb{R}^{(n-1)\times b_{\ell,ext,NL}}$, $\mathbf{A}_{\ell,\text{peec}} \in \mathbb{R}^{(n-1)\times b_{\ell,\text{peec}}}$, $\mathbf{A}_{\ell,ext} \in \mathbb{R}^{(n-1)\times b_{\ell,ext}}$ für $\ell \in \{R, C, L\}$.

Es werden nun für die Elemente der PEEC-Zellen die (reduzierten) Inzidenzmatrizen $\mathbf{A}_{\ell,peec}$ für $\ell \in \{R, C, L\}$ angegeben. Im Folgenden werden die (Zählpfeil-) Richtungen der PEEC-Zellen

Zweigelemente, gemäß Abbildung 4.2 vorausgesetzt. Weiters beschreibt die j-te Spalte von $\mathbf{A}_{C,peec}$ das gekoppelte kapazitive Element am PEEC-Knoten j und die j-te Spalte von $\mathbf{A}_{L,peec}$ bzw. $\mathbf{A}_{R,peec}$ dem gekoppelten induktiven Element bzw. dem resistiven Element der j-ten PEEC-Zelle.
Die (reduzierte) Inzidenzmatrix $\mathbf{A}_{C,\mathrm{peec}} = \left(A_{C,\mathrm{peec}}[i,j]\right)_{\substack{i=1,\dots,n-1\\ j=1,\dots,n_{C,\mathrm{peec}}}} \in \mathbb{R}^{(n-1)\times n_{C,\mathrm{peec}}}$ wird durch

$$A_{C,\mathrm{peec}}[\mathrm{i,j}] := \begin{cases} 1 & \text{, wenn Knoten i der Startknoten des gerichteten kap. PEEC-Zweiges j ist} \\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten kap. PEEC-Zweiges j ist} \\ 0 & \text{, wenn der kap. PEEC-Zweig j nicht mit dem Knoten i inzident ist} \end{cases} .$$

definiert. Mit den (Zählpfeil-) Richtungen gemäß Abbildung 4.2 folgt damit

$$\mathbf{A}_{C,\mathrm{peec}} = \begin{pmatrix} \mathbf{I}_{n_{C,\mathrm{peec}}\times n_{C,\mathrm{peec}}} \\ \mathbf{0}_{(b_{L,\mathrm{peec}}+n_{ext})\times n_{C,\mathrm{peec}}} \end{pmatrix} \tag{4.8}$$

Die (reduzierte) Inzidenzmatrix $\mathbf{A}_{L,\mathrm{peec}} = \left(A_{L,\mathrm{peec}}[i,j]\right)_{\substack{i=1,\dots,n-1\\ j=1,\dots,b_{L,\mathrm{peec}}}} \in \mathbb{R}^{(n-1)\times b_{L,\mathrm{peec}}}$ wird durch

$$A_{L,\mathrm{peec}}[\mathrm{i,j}] := \begin{cases} 1 & \text{, wenn Knoten i der Startknoten des gerichteten ind. PEEC-Zweiges j ist} \\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten ind. PEEC-Zweiges j ist} \\ 0 & \text{, wenn der ind. PEEC-Zweig j nicht mit dem Knoten i inzident ist} \end{cases} .$$

definiert.
Die (reduzierte) Inzidenzmatrix $\mathbf{A}_{R,\mathrm{peec}} = \left(A_{R,\mathrm{peec}}[i,j]\right)_{\substack{i=1,\dots,n-1\\ j=1,\dots,b_{R,\mathrm{peec}}}} \in \mathbb{R}^{(n-1)\times b_{R,\mathrm{peec}}}$ wird durch

$$A_{R,\mathrm{peec}}[\mathrm{i,j}] := \begin{cases} 1 & \text{, wenn Knoten i der Startknoten des gerichteten res. PEEC-Zweiges j ist} \\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten res. PEEC-Zweiges j ist} \\ 0 & \text{, wenn der res. PEEC-Zweig j nicht mit dem Knoten i inzident ist} \end{cases} .$$

definiert.

Zur Beschreibung der zeitinvarianten resistiven, kapazitiven und induktiven Elemente im elektrischen Netzwerk werden nun gemäß Unterkapitel 2.1.2, für $\ell \in \{R, C, L\}$ die Elemente durch Funktionen $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \to \mathbb{R}^{b_\ell}$ beschrieben.
Das Verhalten der linearen kapazitiven Elemente im elektrischen Netzwerk, wird durch die Matrizen $\mathbf{C}_{\mathrm{peec}}$ (gemäß (4.6)) und $\mathbf{C}_{ext} \in \mathbb{R}^{b_{C,ext}\times b_{C,ext}}$, sowie der kapazitiven Zweigspannungen $\mathbf{u}_{C,\mathrm{peec}}$ und $\mathbf{u}_{C,ext}$ bestimmt und durch die folgende Funktion beschrieben

$$\boldsymbol{\gamma}_C : \mathbb{R}^{n_{C,\mathrm{peec}}+b_{C,ext}} \to \mathbb{R}^{n_{C,\mathrm{peec}}+b_{C,ext}} : \underbrace{\begin{pmatrix} \mathbf{u}_{C,\mathrm{peec}} \\ \mathbf{u}_{C,ext} \end{pmatrix}}_{=:\mathbf{u}_C} \mapsto \underbrace{\begin{pmatrix} \mathbf{C}_{\mathrm{peec}} & \mathbf{0}_{n_{C,\mathrm{peec}}\times b_{C,ext}} \\ \mathbf{0}_{b_{C,ext}\times n_{C,\mathrm{peec}}} & \mathbf{C}_{ext} \end{pmatrix}}_{=:\mathbf{C}} \cdot \begin{pmatrix} \mathbf{u}_{C,\mathrm{peec}} \\ \mathbf{u}_{C,ext} \end{pmatrix} .$$

Der Zusammenhang zwischen den Zweigströmen $\mathbf{i}_C(t)$ und den Zweigspannungen $\mathbf{u}_C(t)$ der linearen kapazitiven Elemente, ist durch folgende Beziehung gegeben

$$\mathbf{i}_C(t) = \begin{pmatrix} \mathbf{i}_{C,\mathrm{peec}}(t) \\ \mathbf{i}_{C,ext}(t) \end{pmatrix} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}_C(\mathbf{u}_C(t)) = \mathbf{C}\cdot\frac{\mathrm{d}\mathbf{u}_C(t)}{\mathrm{d}t} = \begin{pmatrix} \mathbf{C}_{\mathrm{peec}} & \mathbf{0}_{n_{C,\mathrm{peec}}\times b_{C,ext}} \\ \mathbf{0}_{b_{C,ext}\times n_{C,\mathrm{peec}}} & \mathbf{C}_{ext} \end{pmatrix} \cdot \frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \mathbf{u}_{C,\mathrm{peec}}(t) \\ \mathbf{u}_{C,ext}(t) \end{pmatrix} .$$

Das Verhalten der linearen induktiven Elemente im elektrischen Netzwerk, wird durch die Matrizen $\mathbf{L}_{\mathrm{peec}}$ (gemäß (4.6)) und $\mathbf{L}_{ext} \in \mathbb{R}^{b_{L,ext}\times b_{L,ext}}$, sowie der induktiven Zweigströme $\mathbf{i}_{L,\mathrm{peec}}$ und $\mathbf{i}_{L,ext}$ bestimmt und durch die folgende Funktion beschrieben

$$\boldsymbol{\gamma}_L : \mathbb{R}^{b_{L,\mathrm{peec}}+b_{L,ext}} \to \mathbb{R}^{b_{L,\mathrm{peec}}+b_{L,ext}} : \underbrace{\begin{pmatrix} \mathbf{i}_{L,\mathrm{peec}} \\ \mathbf{i}_{L,ext} \end{pmatrix}}_{=:\mathbf{i}_L} \mapsto \underbrace{\begin{pmatrix} \mathbf{L}_{\mathrm{peec}} & \mathbf{0}_{b_{L,\mathrm{peec}}\times b_{L,ext}} \\ \mathbf{0}_{b_{L,ext}\times b_{L,\mathrm{peec}}} & \mathbf{L}_{ext} \end{pmatrix}}_{=:\mathbf{L}} \cdot \begin{pmatrix} \mathbf{i}_{L,\mathrm{peec}} \\ \mathbf{i}_{L,ext} \end{pmatrix} .$$

Der Zusammenhang zwischen den Zweigspannungen $\mathbf{u}_L(t)$ und den Zweigströmen $\mathbf{i}_L(t)$ der linearen induktiven Elemente, ist durch folgende Beziehung gegeben

$$\mathbf{u}_L(t) = \begin{pmatrix} \mathbf{u}_{L,\text{peec}}(t) \\ \mathbf{u}_{L,ext}(t) \end{pmatrix} = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\gamma}_L(\mathbf{i}_L(t)) = \mathbf{L}\cdot\frac{\mathrm{d}\mathbf{i}_L(t)}{\mathrm{d}t} = \begin{pmatrix} \mathbf{L}_{\text{peec}} & \mathbf{0}_{b_{L,\text{peec}}\times b_{L,ext}} \\ \mathbf{0}_{b_{L,ext}\times b_{L,\text{peec}}} & \mathbf{L}_{ext} \end{pmatrix}\cdot\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} \mathbf{i}_{L,\text{peec}}(t) \\ \mathbf{i}_{L,ext}(t) \end{pmatrix}.$$

Das Verhalten der resistiven Elemente im elektrischen Netzwerk, wird durch die Matrizen $\mathbf{R}_{\text{peec}}^{-1} = (\mathbf{R}')^{-1} = \operatorname{diag}\big((R'_1)^{-1},\ldots,(R'_{b_{L,\text{peec}}})^{-1}\big) \in \mathbb{R}^{b_{L,\text{peec}}\times b_{L,\text{peec}}}$ und $\mathbf{R}_{ext,lin}^{-1} \in \mathbb{R}^{b_{R,ext,lin}\times b_{R,ext,lin}}$ und der nichtlinearen Funktion $\boldsymbol{\gamma}_{R,ext,NL} : \mathbb{R}^{b_{R,ext,NL}} \to \mathbb{R}^{b_{R,ext,NL}}$, sowie der resistiven Zweigspannungen $\mathbf{u}_{R,\text{peec}}$, $\mathbf{u}_{R,ext,lin}$ und $\mathbf{u}_{R,ext,NL}$ bestimmt und durch die folgende Funktion beschrieben

$$\boldsymbol{\gamma}_R : \mathbb{R}^{b_{L,\text{peec}}+b_{R,ext,lin}+b_{R,ext,NL}} \to \mathbb{R}^{b_{L,\text{peec}}+b_{R,ext,lin}+b_{R,ext,NL}} : \underbrace{\begin{pmatrix} \mathbf{u}_{R,\text{peec}} \\ \mathbf{u}_{R,ext,lin} \\ \mathbf{u}_{R,ext,NL} \end{pmatrix}}_{=:\,\mathbf{u}_R} \mapsto \begin{pmatrix} \mathbf{R}_{\text{peec}}^{-1}\cdot\mathbf{u}_{R,\text{peec}} \\ \mathbf{R}_{ext,lin}^{-1}\cdot\mathbf{u}_{R,ext,lin} \\ \boldsymbol{\gamma}_{R,ext,NL}(\mathbf{u}_{R,ext,NL}) \end{pmatrix}.$$

Hierbei sind $\mathbf{R}_{\text{peec}}^{-1}$ und $\mathbf{R}_{ext,lin}^{-1}$ die Inversen der Widerstands-Matrizen $\mathbf{R}_{\text{peec}}$ und $\mathbf{R}_{ext,lin}$, d.h. $\mathbf{R}_{\text{peec}}^{-1}$ und $\mathbf{R}_{ext,lin}^{-1}$ entsprechen den Leitwertsmatrizen.

Modellproblem unter Anwendung der PEEC Methode

Mit den Erläuterungen von Abschnitt '*Vorbereitungen für die Analyse des elektrischen Netzwerkes mit dem MKV*' ist es nun möglich das Modellproblem 3.1 anzuwenden, um das zu analysierende elektrische Netzwerk mit dem MKV zu beschreiben. Dabei setzt sich der Vektor der Unbekannten $\mathbf{x}(t)$ aus den folgenden $m := n-1+b_L+b_U = n_{C,\text{peec}}+b_{L,\text{peec}}+n_{ext}+b_{L,\text{peec}}+b_{L,ext}+b_U$ Komponenten zusammen

$$\mathbf{x}(t) := \big(\underbrace{\mathbf{e}_{C,\text{peec}}(t)^\top, \mathbf{e}_{R,\text{peec}}(t)^\top, \mathbf{e}_{ext}(t)^\top}_{=:\,\mathbf{e}(t)^\top}, \underbrace{\mathbf{i}_{L,\text{peec}}(t)^\top, \mathbf{i}_{L,ext}(t)^\top}_{=:\,\mathbf{i}_L(t)^\top}, \mathbf{i}_U(t)^\top\big)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m,$$

wobei $\mathbf{e}_{R,\text{peec}}(t)^\top = \big(e_{R_1,\text{peec}}(t),\ldots,e_{R_{b_{L,\text{peec}}},\text{peec}}\big)$ gemäß Abbildung 4.2 den Knotenspannungen an den partiellen PEEC-Widerständen $R'_1,\ldots,R'_{b_{L,\text{peec}}}$ entspricht, sowie $\mathbf{e}_{ext}(t)$ den Knotenspannungen der externen Knoten und $\mathbf{i}_{L,ext}(t)$ den induktiven Strömen der externen Induktivitäten.
Gemäß dem Modellproblem 3.1 ist in der Matrix $\mathbf{M}$ der Block $\mathbf{A}_C\cdot\mathbf{C}\cdot\mathbf{A}_C^\top$ enthalten, welcher sich u.a. gemäß (4.8) folgendermaßen darstellen lässt

$$\begin{aligned}
\mathbf{A}_C\cdot\mathbf{C}\cdot\mathbf{A}_C^\top &= \big(\mathbf{A}_{C,peec}, \mathbf{A}_{C,ext}\big)\cdot\begin{pmatrix} \mathbf{C}_{\text{peec}} & \mathbf{0}_{n_{C,\text{peec}}\times b_{C,ext}} \\ \mathbf{0}_{b_{C,ext}\times n_{C,\text{peec}}} & \mathbf{C}_{ext} \end{pmatrix}\cdot\begin{pmatrix} \mathbf{A}_{C,peec}^\top \\ \mathbf{A}_{C,ext}^\top \end{pmatrix} \\
&= \left(\begin{pmatrix} \mathbf{I}_{n_{C,\text{peec}}\times n_{C,\text{peec}}} \\ \mathbf{0}_{(b_{L,\text{peec}}+n_{ext})\times n_{C,\text{peec}}} \end{pmatrix}\ \mathbf{A}_{C,ext}\right)\cdot\begin{pmatrix} \big(\mathbf{S}^\top\big)^{-1}\cdot\mathbf{F}\cdot\mathbf{I}_{n_{C,\text{peec}}\times n_{C,\text{peec}}} & \mathbf{0}_{n_{C,\text{peec}}\times(b_{L,\text{peec}}+n_{ext})} \\ & \mathbf{C}_{ext}\cdot\mathbf{A}_{C,ext}^\top \end{pmatrix} \\
&= \begin{pmatrix} \big(\mathbf{S}^\top\big)^{-1}\cdot\mathbf{F} & \mathbf{0}_{n_{C,\text{peec}}\times(b_{L,\text{peec}}+n_{ext})} \\ \mathbf{0}_{(b_{L,\text{peec}}+n_{ext})\times n_{C,\text{peec}}} & \mathbf{0}_{(b_{L,\text{peec}}+n_{ext})\times(b_{L,\text{peec}}+n_{ext})} \end{pmatrix} + \mathbf{A}_{C,ext}\cdot\mathbf{C}_{ext}\cdot\mathbf{A}_{C,ext}^\top \in \mathbb{R}^{(n-1)\times(n-1)}.
\end{aligned}$$

Gemäß dem Modellproblem 3.1, können nun auch die ersten $n-1$ Komponenten der Funktion $\mathbf{q}(\mathbf{e},\mathbf{i}_L,\mathbf{i}_U)$, folgendermaßen angegeben werden

$$\mathbf{A}_C\cdot\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e}) = \left(\begin{pmatrix} \big(\mathbf{S}^\top\big)^{-1}\cdot\mathbf{F} & \mathbf{0}_{n_{C,\text{peec}}\times(b_{L,\text{peec}}+n_{ext})} \\ \mathbf{0}_{(b_{L,\text{peec}}+n_{ext})\times n_{C,\text{peec}}} & \mathbf{0}_{(b_{L,\text{peec}}+n_{ext})\times(b_{L,\text{peec}}+n_{ext})} \end{pmatrix} + \mathbf{A}_{C,ext}\cdot\mathbf{C}_{ext}\cdot\mathbf{A}_{C,ext}^\top\right)\cdot\mathbf{e}.$$

Sowohl die Matrix $\mathbf{A}_C\cdot\mathbf{C}\cdot\mathbf{A}_C^\top$ als auch die Funktion $\mathbf{A}_C\cdot\gamma_C(\mathbf{A}_C^\top\cdot\mathbf{e})$ haben den Nachteil, dass die Matrix $\big(\mathbf{S}^\top\big)^{-1}$ benötigt wird. Bei numerischen Implementierungen sollte allerdings stets die Berechnung einer inversen Matrix vermieden werden. Es ist nun zu bemerken, dass bei $\mathbf{A}_C\cdot\mathbf{C}\cdot\mathbf{A}_C^\top$

und $\mathbf{A}_C \cdot \gamma_C(\mathbf{A}_C^\top \cdot \mathbf{e})$ die Matrix $\left(\mathbf{S}^\top\right)^{-1}$ ganz links steht. Dies ist kein Zufall, sondern liegt an der Verwendung der äquivalenten Darstellung (4.2). Werden nun bei der Anwendung des Modellproblems 3.1 die ersten $n-1$ Zeilen der Gleichungen (3.1a) und (3.1b) mit der Matrix

$$\widetilde{\mathbf{S}} := \begin{pmatrix} \mathbf{S}^\top & \mathbf{0}_{n_{C,\mathrm{peec}} \times (b_{L,\mathrm{peec}} + n_{ext})} \\ \mathbf{0}_{(b_{L,\mathrm{peec}} + n_{ext}) \times n_{C,\mathrm{peec}}} & \mathbf{I}_{(b_{L,\mathrm{peec}} + n_{ext}) \times (b_{L,\mathrm{peec}} + n_{ext})} \end{pmatrix}$$

von links multipliziert, dann kann die Matrix $\left(\mathbf{S}^\top\right)^{-1}$ in den Gleichungen eliminiert werden. Um eine kompakte Schreibweise zu ermöglichen wird zudem folgende Matrix definiert

$$\begin{aligned} \widetilde{\mathbf{F}} &:= \widetilde{\mathbf{S}} \cdot \begin{pmatrix} \left(\mathbf{S}^\top\right)^{-1} \cdot \mathbf{F} & \mathbf{0}_{n_{C,\mathrm{peec}} \times (b_{L,\mathrm{peec}} + n_{ext})} \\ \mathbf{0}_{(b_{L,\mathrm{peec}} + n_{ext}) \times n_{C,\mathrm{peec}}} & \mathbf{0}_{(b_{L,\mathrm{peec}} + n_{ext}) \times (b_{L,\mathrm{peec}} + n_{ext})} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{F} & \mathbf{0}_{n_{C,\mathrm{peec}} \times (b_{L,\mathrm{peec}} + n_{ext})} \\ \mathbf{0}_{(b_{L,\mathrm{peec}} + n_{ext}) \times n_{C,\mathrm{peec}}} & \mathbf{0}_{(b_{L,\mathrm{peec}} + n_{ext}) \times (b_{L,\mathrm{peec}} + n_{ext})} \end{pmatrix} . \end{aligned}$$

Die bisherigen Überlegungen werden nun im Modellproblem 4.1 zusammengefasst, wobei berücksichtigt werden sollte, dass die Modellprobleme 3.1 und 4.1, unter den Voraussetzungen von Kapitel 4, mathematisch äquivalent sind und zu denselben Lösungen führen. Das Modellproblem 4.1 ist allerdings für den Anwendungsbereich dieses Kapitels, bei numerischen Berechnungen zu bevorzugen.

Modellproblem 4.1 (Spezialfall des MKV unter Verwendung der PEEC Methode)**:**
Gemäß dem Abschnitt 'Quasi-stationäre 1D PEEC Methode für leitfähige Strukturen' des Unterkapitels 4.1, seien Matrizen $\mathbf{P}, \mathbf{S}, \mathbf{F} \in \mathbb{R}^{n_{C,\mathrm{peec}} \times n_{C,\mathrm{peec}}}$ *und* $\mathbf{L}', \mathbf{R}' \in \mathbb{R}^{b_{L,\mathrm{peec}} \times b_{L,\mathrm{peec}}}$ *gegeben, welche ein elektromagnetisches Feldproblem unter Anwendung der PEEC Methode, durch ein äquivalentes elektrisches Netzwerk beschreiben. Dieses äquivalente elektrische Netzwerk sei mit einem externen elektrischen Netzwerk verbunden, wobei das resultierende elektrische Netzwerk die folgenden Eigenschaften erfülle.*

(i) Es gelten die grundlegenden Voraussetzungen und Notationen gemäß Abschnitt 'Zusammenfassung der Voraussetzungen und Notationen zur Erstellung des MKV' in Unterkapitel 2.2.1.

(ii) Die resistiven, kapazitiven und induktiven Elemente sind zeitinvariant, d.h. für $\ell \in \{R, C, L\}$ *werden die Elemente durch Funktionen* $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \to \mathbb{R}^{b_\ell}$ *beschrieben.*

(iii) Die induktiven und kapazitiven Elemente sind linear, d.h. es existieren Matrizen $\mathbf{C} \in \mathbb{R}^{b_C \times b_C}$ *und* $\mathbf{L} \in \mathbb{R}^{b_L \times b_L}$*, sodass* $\boldsymbol{\gamma}_C(\mathbf{u}_C) = \mathbf{C} \cdot \mathbf{u}_C$ *und* $\boldsymbol{\gamma}_L(\mathbf{i}_L) = \mathbf{L} \cdot \mathbf{i}_L$ *gelte.*

(iv) Es gelten die Notationen, Bezeichnungen und gewählten Nummerierungen gemäß dem Abschnitt 'Vorbereitungen für die Analyse des elektrischen Netzwerkes mit dem MKV' des Unterkapitels 4.1. Insbesondere gelte für $\ell \in \{R, C, L\}$ *die Aufteilung der Inzidenzmatrizen* $\mathbf{A}_\ell \in \mathbb{R}^{(n-1) \times b_\ell}$ *und die Definitionen der Funktionen* $\boldsymbol{\gamma}_\ell : \mathbb{R}^{b_\ell} \to \mathbb{R}^{b_\ell}$*.*

(v) Es sind nur unabhängige periodische Quellen mit Periode $T > 0$ *im elektrischen Netzwerk vorhanden, d.h. die Funktionen für die Quellen* $\mathbf{u}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_U}$ *und* $\mathbf{i}_Q : \mathbb{R} \mapsto \mathbb{R}^{b_I}$ *sind lediglich zeitabhängig und es gilt* $\mathbf{u}_Q(t) = \mathbf{u}_Q(t+T)$*,* $\mathbf{i}_Q(t) = \mathbf{i}_Q(t+T)$ *für alle* $t \in \mathbb{R}$*.*

Gemäß Abschnitt 'Struktur des MKV bei zeitinvarianten Elementen und unabhängigen Quellen' in Unterkapitel 2.2.1 und 'Modellproblem unter Anwendung der PEEC Methode' in Unterkapitel 4.1, wird mit $m := n - 1 + b_L + b_U = n_{C,\mathrm{peec}} + b_{L,\mathrm{peec}} + n_{ext} + b_{L,\mathrm{peec}} + b_{L,ext} + b_U$ *und dem Vektor der* m *Unbekannten*

$$\mathbf{x}(t) := \big(\underbrace{\mathbf{e}_{C,\mathrm{peec}}(t)^\top, \mathbf{e}_{R,\mathrm{peec}}(t)^\top, \mathbf{e}_{ext}(t)^\top}_{=: \mathbf{e}(t)^\top}, \underbrace{\mathbf{i}_{L,\mathrm{peec}}(t)^\top, \mathbf{i}_{L,ext}(t)^\top}_{=: \mathbf{i}_L(t)^\top}, \mathbf{i}_U(t)^\top \big)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m ,$$

das elektrische Netzwerk durch die Gleichung

$$\frac{\mathrm{d}}{\mathrm{d}t}\,\mathbf{q}(\mathbf{x}(t)) + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m\ , \tag{4.9a}$$

beschrieben, wobei das MKV System als nicht autonom vorausgesetzt wird und somit $\mathbf{b}$ *nicht der Nullvektorfunktion entspricht, bzw.* $\mathbf{b} \neq \mathbf{0}$ *gelte. Die Funktionen* $\mathbf{q}, \mathbf{f}$ *und* $\mathbf{b}$ *sind dabei durch*

$$\mathbf{q} : \mathbb{R}^{n-1} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \to \mathbb{R}^m : (\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) \mapsto \mathbf{q}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \begin{pmatrix} \left(\widetilde{\mathbf{F}} + \widetilde{\mathbf{S}} \cdot \mathbf{A}_{C,ext} \cdot \mathbf{C}_{ext} \cdot \mathbf{A}_{C,ext}^\top\right) \cdot \mathbf{e} \\ \gamma_L(\mathbf{i}_L) \\ \mathbf{0}_{b_U} \end{pmatrix}\ ,$$

$$\mathbf{f} : \mathbb{R}^{n-1} \times \mathbb{R}^{b_L} \times \mathbb{R}^{b_U} \to \mathbb{R}^m : (\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) \mapsto \mathbf{f}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \begin{pmatrix} \widetilde{\mathbf{S}} \cdot \left(\mathbf{A}_R \cdot \gamma_R(\mathbf{A}_R^\top \cdot \mathbf{e}) + \mathbf{A}_L \cdot \mathbf{i}_L + \mathbf{A}_U \cdot \mathbf{i}_U\right) \\ -\mathbf{A}_L^\top \cdot \mathbf{e} \\ \mathbf{A}_U^\top \cdot \mathbf{e} \end{pmatrix}\ ,$$

$$\mathbf{b} : \mathbb{R} \to \mathbb{R}^m : t \mapsto \mathbf{b}(t) = \begin{pmatrix} \widetilde{\mathbf{S}} \cdot \mathbf{A}_I \cdot \mathbf{i}_Q(t) \\ \mathbf{0}_{b_L} \\ -\mathbf{u}_Q(t) \end{pmatrix}\ .$$

gegeben. Wird in (3.1a) *die Kettenregel angewendet, so führt dies zu der äquivalenten Gleichung*

$$\mathbf{M} \cdot \frac{\mathrm{d}\mathbf{x}(t)}{\mathrm{d}t} + \mathbf{f}(\mathbf{x}(t)) + \mathbf{b}(t) = \mathbf{0}_m\ , \tag{4.9b}$$

wobei die Matrix $\mathbf{M}$ *die folgende Darstellung besitzt*

$$\mathbf{M} = \begin{pmatrix} \widetilde{\mathbf{F}} + \widetilde{\mathbf{S}} \cdot \mathbf{A}_{C,ext} \cdot \mathbf{C}_{ext} \cdot \mathbf{A}_{C,ext}^\top & \mathbf{0}_{(n-1)\times b_{L,\mathrm{peec}}} & \mathbf{0}_{(n-1)\times b_{L,ext}} & \mathbf{0}_{(n-1)\times b_U} \\ \mathbf{0}_{b_{L,\mathrm{peec}}\times(n-1)} & \mathbf{L}_{\mathrm{peec}} & \mathbf{0}_{b_{L,\mathrm{peec}}\times b_{L,ext}} & \mathbf{0}_{b_{L,\mathrm{peec}}\times b_U} \\ \mathbf{0}_{b_{L,ext}\times(n-1)} & \mathbf{0}_{b_{L,ext}\times b_{L,\mathrm{peec}}} & \mathbf{L}_{ext} & \mathbf{0}_{b_{L,ext}\times b_U} \\ \mathbf{0}_{b_U\times(n-1)} & \mathbf{0}_{b_U\times b_{L,\mathrm{peec}}} & \mathbf{0}_{b_U\times b_{L,ext}} & \mathbf{0}_{b_U\times b_U} \end{pmatrix} \in \mathbb{R}^{m\times m}\ .$$

Mit $\mathbf{x} := (\mathbf{e}^\top, \mathbf{i}_L^\top, \mathbf{i}_U^\top)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m$ *gilt zudem für die Jacobi-Matrix* $\mathbf{J}_{\mathbf{f}}(\mathbf{x}) \in \mathbb{R}^{m\times m}$,

$$\mathbf{J}_{\mathbf{f}}(\mathbf{x}) = \mathbf{J}_{\mathbf{f}}(\mathbf{e}, \mathbf{i}_L, \mathbf{i}_U) = \left(\mathbf{J}_{\mathbf{e};\mathbf{f}}, \mathbf{J}_{\mathbf{i}_L;\mathbf{f}}, \mathbf{J}_{\mathbf{i}_U;\mathbf{f}}\right) \tag{4.10}$$

$$= \begin{pmatrix} \widetilde{\mathbf{S}} \cdot \mathbf{A}_R \cdot \mathbf{J}_{\gamma_R}(\mathbf{A}_R^\top \cdot \mathbf{e}) \cdot \mathbf{A}_R^\top & \widetilde{\mathbf{S}} \cdot \mathbf{A}_L & \widetilde{\mathbf{S}} \cdot \mathbf{A}_U \\ -\mathbf{A}_L^\top & \mathbf{0}_{b_L\times b_L} & \mathbf{0}_{b_L\times b_U} \\ \mathbf{A}_U^\top & \mathbf{0}_{b_U\times b_L} & \mathbf{0}_{b_U\times b_U} \end{pmatrix}\ . \tag{4.11}$$

Mit der Leitwertsmatrix $\mathbf{G}$ *gemäß* (2.20) *gilt zudem* $\mathbf{G}(\mathbf{A}_R^\top \cdot \mathbf{e}) = \mathbf{J}_{\gamma_R}(\mathbf{A}_R^\top \cdot \mathbf{e})$.

4.2 Anwendungsbeispiel

Das Anwendungsbeispiel ist ein NFC System, bestehend aus zwei luftgekoppelten Spulen und einem gegebenen externen elektrischen Netzwerk. Die äußere Beschaltung bewirkt dabei, dass sich das NFC System in Resonanz befindet. Für die Berechnung des periodisch eingeschwungenen Zustands des NFC Systems, werden die numerischen Methoden aus Kapitel 3 zur Lösung der DAG (4.9) angewendet und untersucht.

NFC Spulenanordnung und Modellierung

Abbildung 4.3 zeigt die Anordnung der NFC Spulen, dessen z-Abstand 5 mm beträgt. Spule 1 entspricht der Primärspule und Spule 2 der Sekundärspule.

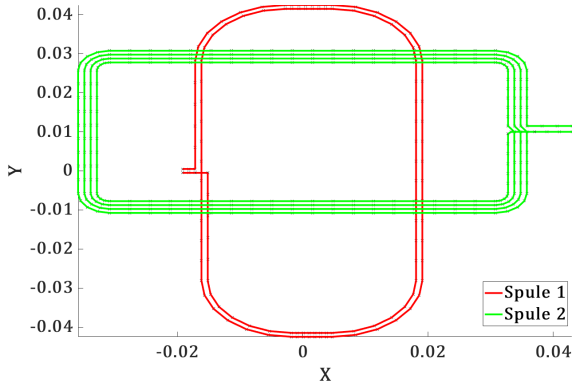


Abbildung 4.3: NFC Spulen mit einem z-Abstand von 5mm.

Die NFC Primärpule wird über ein Anpassnetzwerk durch eine Sinusquelle mit $3\,V$ Amplitude und einer Frequenz von $f_0 = 13.56\,\text{MHz}$ versorgt. Über die Gleichung $c_0 = \lambda_0 \cdot f_0$ ergibt sich somit eine Wellenlänge von $\lambda_0 \approx 22,1\,\text{m}$, wobei c_0 der Lichtgeschwindigkeit im freien Raum entspricht. Aufgrund der geringen geometrischen Ausdehnung der Spulenanordnung, ist es somit zulässig den Verschiebestrom zu vernachlässigen und die quasi-stationäre Näherung der Maxwell Gleichungen, für die Beschreibung des elektromagnetischen Verhaltens zu verwenden. Zur Modellierung dieses elektromagnetischen Feldproblems wurde die quasi-stationäre PEEC Methode gemäß Unterkapitel 4.1 verwendet. Spule 1 besitzt dabei 126 PEEC-Zellen mit 127 PEEC-Knoten und Spule 2 besitzt dabei 313 PEEC-Zellen mit 314 PEEC-Knoten. Die Modellierung der Spulenanordnung des NFC Systems erfolgt somit gemäß Abbildung 4.4, durch ein äquivalentes elektrisches Netzwerk, bestehend aus 441 PEEC-Knoten und 439 PEEC-Zellen, d.h. es gilt $n_{C,\text{peec}} = 441$, $b_{L,\text{peec}} = 439$.

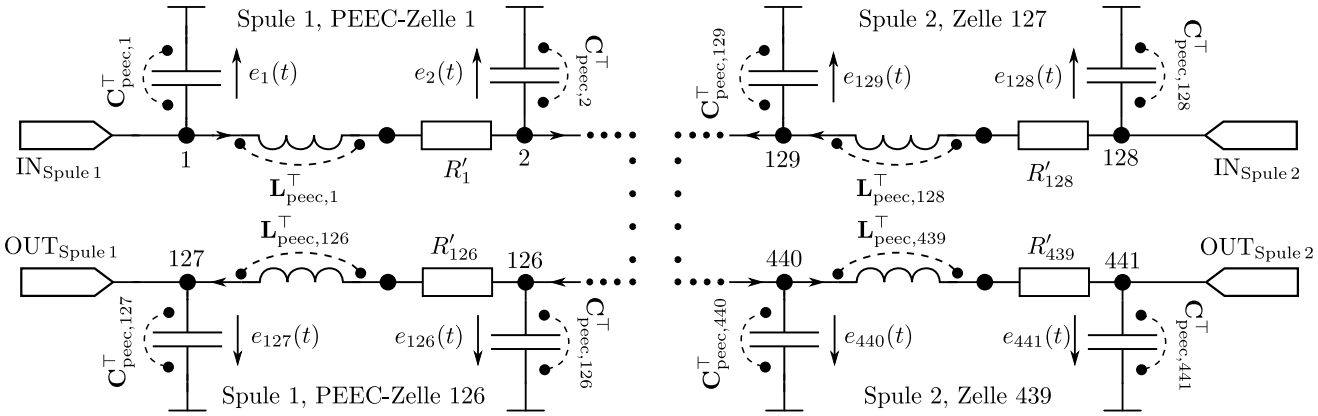


Abbildung 4.4: Äquivalentes elektrisches Netzwerk der NFC Antennen

Berechnung des periodisch eingeschwungenen Zustands

Als externes elektrisches Netzwerk der NFC Spulenanordnung, wurde die Beschaltung gemäß Abbildung 4.5 gewählt. Die Elemente C_{s1}, C_{s1}, C_p, R_p auf der Primärseite und der 60.23 pF Kondensator auf der Sekundärseite des Spulenpaars, sind Teil eines Anpassnetzwerkes, welches gewährleistet, dass sich das NFC System in Resonanz befindet. In Abbildung 4.5 sind zudem die $n_{ext} = 6$ externen Knoten mit der fortlaufenden Nummer 881 bis 886 eingetragen, sowie die Terminalknoten 1 und 127 der Primärspule und die Terminalknoten 128 und 441 der Sekundärspule.

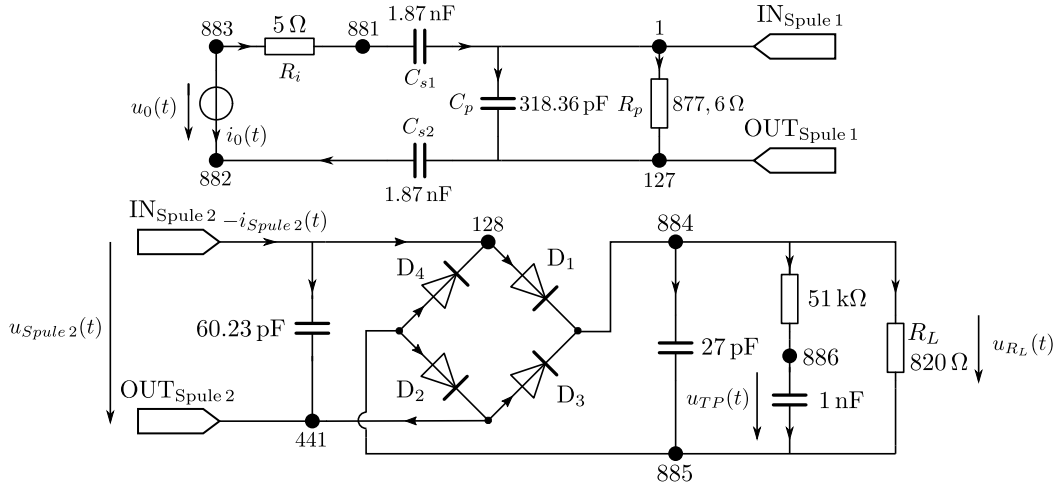


Abbildung 4.5: Externes Netzwerk des NFC Systems.

Das beschreibende elektrische Netzwerk des NFC Systems erfüllt dabei die Voraussetzungen des Modellproblems 4.1. Nachdem die Modellprobleme 3.1 und 4.1 äquivalent sind, werden im Folgenden die numerischen Zeitbereichsmethoden transiente Analyse und einfaches Shooting, basierend auf dem BDF3 Verfahren und die Frequenzbereichsmethode Harmonic Balance (gemäß Kapitel 3) angewendet, um die DAG (4.9) zu lösen bzw. um den periodisch eingeschwungenen Zustand numerisch zu ermitteln.
Das vollständige elektrische Netzwerk besitze gemäß Modellproblems 4.1

$$\begin{aligned} m := n - 1 + b_L + b_U &= n_{C,\text{peec}} + b_{L,\text{peec}} + n_{ext} + b_{L,\text{peec}} + b_{L,ext} + b_U \\ &= 441 + 439 + 6 + 439 + 0 + 1 = 1326 \end{aligned}$$

Unbekannte und der Vektor der Unbekannten ist

$$\mathbf{x}(t) := \big(\underbrace{\mathbf{e}_{C,\text{peec}}(t)^\top, \mathbf{e}_{R,\text{peec}}(t)^\top, \mathbf{e}_{ext}(t)^\top}_{=:\,\mathbf{e}(t)^\top}, \underbrace{\mathbf{i}_{L,\text{peec}}(t)^\top}_{=:\,\mathbf{i}_L(t)^\top}, i_0(t) \big)^\top \in \mathbb{R}^{n-1+b_L+b_U} = \mathbb{R}^m = \mathbb{R}^{1326} \ .$$

Das externe elektrische Netzwerk wird dabei im Modellproblem 4.1 folgendermaßen berücksichtigt: Die Diagonalmatrix

$$\mathbf{C}_{ext} = \operatorname{diag}\big(318.36\,\text{pF}, 1.87\,\text{nF}, 1.87\,\text{nF}, 60.23\,\text{pF}, 27\,\text{pF}, 1\,\text{nF}\big)$$

enthält die externen Kapazitäten, wobei die Reihenfolge der Elemente in $\mathbf{C}_{ext}$, auch die Element-Zuordnung der Spalten der (reduzierten) Inzidenzmatrix $\mathbf{A}_{C,ext} = \big(A_{C,ext}[i,j]\big)_{\substack{i=1,\dots,n-1 \\ j=1,\dots,b_{C,ext}}}$ bestimmt. Tabelle 4.1 zeigt jene Zeilen von $\mathbf{A}_{C,ext}$, welche Einträge ungleich Null enthalten können, wobei die Komponenten von $A_{C,ext}[i,j]$ folgendermaßen gegeben sind

$$A_{C,ext}[\text{i,j}] := \begin{cases} 1 & \text{, wenn Knoten i der Startknoten des gerichteten ext. kap. Zweiges j ist} \\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten ext. kap. Zweiges j ist} \\ 0 & \text{, wenn der ext. kap. Zweig j nicht mit dem Knoten i inzident ist} \end{cases} .$$

Knoten	318.36 pF	1.87 nF	1.87 nF	60.23 pF	27 pF	1 nF
1	1	−1	0	0	0	0
127	−1	0	1	0	0	0
128	0	0	0	1	0	0
441	0	0	0	−1	0	0
881	0	1	0	0	0	0
882	0	0	−1	0	0	0
883	0	0	0	0	0	0
884	0	0	0	0	1	0
885	0	0	0	0	−1	−1
886	0	0	0	0	0	1

Tabelle 4.1: Ausgewählte Zeilen der (reduzierten) Inzidenzmatrix $\mathbf{A}_{C,ext}$.

Die Diagonalmatrix

$$\mathbf{R}_{ext,lin} = \mathrm{diag}\big(5\,\Omega, 877.6\,\Omega, 51\,\mathrm{k}\Omega, 820\,\Omega\big)$$

enthält die externen Widerstände, wobei die Reihenfolge der Elemente in $\mathbf{R}_{ext,lin}$, auch die Element-Zuordnung der Spalten der (reduzierten) Inzidenzmatrix $\mathbf{A}_{R,ext,lin} = \big(A_{R,ext,lin}[i,j]\big)_{\substack{i=1,\dots,n-1\\ j=1,\dots,b_{R,ext,lin}}}$ bestimmt. Tabelle 4.2 zeigt jene Zeilen von $\mathbf{A}_{R,ext,lin}$, welche Einträge ungleich Null enthalten können, wobei die Komponenten von $A_{R,ext,lin}[i,j]$ folgendermaßen gegeben sind

$$A_{R,ext,lin}[\mathrm{i,j}] := \begin{cases} 1 & \text{, wenn Knoten i der Startknoten des gerichteten ext. res. Zweiges j ist} \\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten ext. res. Zweiges j ist} \\ 0 & \text{, wenn der ext. res. Zweig j nicht mit dem Knoten i inzident ist} \end{cases} .$$

Knoten	5 Ω	877.6 Ω	51 kΩ	820 Ω
1	0	1	0	0
127	0	−1	0	0
128	0	0	0	0
441	0	0	0	0
881	−1	0	0	0
882	0	0	0	0
883	1	0	0	0
884	0	0	1	1
885	0	0	0	−1
886	0	0	−1	0

Tabelle 4.2: Ausgewählte Zeilen der (reduzierten) Inzidenzmatrix $\mathbf{A}_{R,ext,lin}$.

Die Dioden $D_1, \dots, D_4$ wurden gemäß der Shockley Gleichung

$$i_D(u_D) = I_S \cdot \left(\exp\left(\frac{u_D}{n \cdot U_T}\right) - 1\right)$$

modelliert und als Schottky Dioden BAR43S implementiert. Die gewählten Parameter sind

$$k_B = 1.380649{\cdot}10^{-23}\,J/K\;,\quad T = 300\,K\;,\quad e = 1.602176634{\cdot}10^{-19}\,C\;,\quad U_T = \frac{k_B \cdot T}{e}\;,$$
$$n = 1.4622\;,\quad I_S = 0.4345{\cdot}10^{-6}\,A\;,$$

welche gemäß Typ BAR43SFILM, der PSPICE Bibliothek [24] entnommen wurden. Die resistiven nichtlinearen Elemente, d.h. die Dioden, werden gemäß der Schockley Gleichung durch die Funktion

$$\boldsymbol{\gamma}_{R,ext,NL}(\mathbf{u}_D) = \boldsymbol{\gamma}_{R,ext,NL}(u_{D_1},\ldots,u_{D_4}) = \big(i_{D_1}(u_{D_1}),\ldots,i_{D_4}(u_{D_4})\big)^\top$$

beschrieben. Die Beschreibung der resistiven Elemente erfolgt durch die Funktion

$$\boldsymbol{\gamma}_R : \mathbb{R}^{b_{L,\text{peec}}+b_{R,ext,lin}+b_{R,ext,NL}} \to \mathbb{R}^{b_{L,\text{peec}}+b_{R,ext,lin}+b_{R,ext,NL}} : \underbrace{\begin{pmatrix}\mathbf{u}_{R,\text{peec}}\\ \mathbf{u}_{R,ext,lin}\\ \mathbf{u}_D\end{pmatrix}}_{=:\,\mathbf{u}_R} \mapsto \begin{pmatrix}\mathbf{R}_{\text{peec}}^{-1}\cdot\mathbf{u}_{R,\text{peec}}\\ \mathbf{R}_{ext,lin}^{-1}\cdot\mathbf{u}_{R,ext,lin}\\ \boldsymbol{\gamma}_{R,ext,NL}(\mathbf{u}_D)\end{pmatrix}\;.$$

Die (reduzierte) Inzidenzmatrix $\mathbf{A}_{R,ext,NL} = \big(A_{R,ext,NL}[i,j]\big)_{\substack{i=1,\ldots,n-1\\ j=1,\ldots,b_{R,ext,NL}}}$ der nichtlinearen resistiven Elemente, ist gemäß Tabelle 4.3 bzw. der nachfolgenden Definition gegeben.

$$A_{R,ext,lin}[\mathrm{i,j}] := \begin{cases}1 & \text{, wenn Knoten i der Startknoten des gerichteten ext. nl. res. Zweiges j ist}\\ -1 & \text{, wenn Knoten i der Endknoten des gerichteten ext. nl. res. Zweiges j ist}\\ 0 & \text{, wenn der ext. nl. res. Zweig j nicht mit dem Knoten i inzident ist}\end{cases}.$$

Knoten	D_1	D_2	D_3	D_4
1	0	0	0	0
127	0	0	0	0
128	1	0	0	-1
441	0	−1	1	0
881	0	0	0	0
882	0	0	0	0
883	0	0	0	0
884	−1	0	−1	0
885	0	1	0	1
886	0	0	0	0

Tabelle 4.3: Ausgewählte Zeilen der (reduzierten) Inzidenzmatrix $\mathbf{A}_{R,ext,NL}$.

Die Jacobi-Matrix $\mathbf{J}_{\gamma_R}(\mathbf{u}_{R,\text{peec}}, \mathbf{u}_{R,ext,lin}, \mathbf{u}_D)$, welche für die Implementierung der numerischen Methoden benötigt wird, berechnet sich zu der Diagonalmatrix

$$\mathbf{J}_{\gamma_R}(\mathbf{u}_{R,\text{peec}}, \mathbf{u}_{R,ext,lin}, \mathbf{u}_D) = \begin{pmatrix}\mathbf{R}_{\text{peec}}^{-1} & \mathbf{0}_{b_{R,\text{peec}}\times b_{R,ext,lin}} & \mathbf{0}_{b_{R,\text{peec}}\times b_{R,ext,NL}}\\ \mathbf{0}_{b_{R,ext,lin}\times b_{R,\text{peec}}} & \mathbf{R}_{ext,lin}^{-1} & \mathbf{0}_{b_{R,ext,lin}\times b_{R,ext,NL}}\\ \mathbf{0}_{b_{R,ext,NL}\times b_{R,\text{peec}}} & \mathbf{0}_{b_{R,ext,NL}\times b_{R,ext,lin}} & \operatorname{diag}\left(\frac{\mathrm{d}i_{D_1}(u_{D_1})}{\mathrm{d}t},\ldots,\frac{\mathrm{d}i_{D_4}(u_{D_4})}{\mathrm{d}t}\right)\end{pmatrix},$$

wobei sich für $\ell \in \{1, \ldots, 4\}$, die Ableitung des Diodenstroms folgendermaßen berechnet

$$\frac{\mathrm{d}i_{D_\ell}(u_{D_\ell})}{\mathrm{d}t} = \frac{I_S}{n \cdot U_T} \cdot \exp\left(\frac{u_{D_\ell}}{n \cdot U_T}\right) .$$

Bei der Implementierung der Methoden aus Kapitel 3, war es erforderlich nichtlineare Gleichungssysteme zu lösen. Diese wurden stets mit dem gedämpften Newton Verfahren nach Bank-Rose, gemäß Algorithmus 2 aus Unterkapitel 1.3 gelöst. Für die transiente Analyse wurde $\mathrm{Tol}_{\mathrm{TA}} = 10^{-7}$, für das Shooting Verfahrens $\mathrm{Tol}_{\mathrm{Sh}} = 10^{-5}$ und für das Harmonic Balance Verfahren $\mathrm{Tol}_{\mathrm{HB}} = 10^{-5}$ gewählt, sodass die Methoden mit folgenden Paramtetern implementiert wurden

$$\mathrm{Tol}_{abs} = \mathrm{Tol}_\ell \ , \quad \mathrm{Tol}_{rel} = \mathrm{Tol}_\ell \ , \quad \mathrm{Tol}_{Fkt} = \mathrm{Tol}_\ell \ , \quad \text{für } \ell \in \{\mathrm{TA}, \mathrm{Sh}, \mathrm{HB}\} \ , \quad \text{sowie} \quad \omega = 0.9 \ .$$

Die nachfolgenden Berechnungen werden die Eingangsspannung $u_0(t) = 3\,V \cdot \sin(2 \cdot \pi \cdot f_0 \cdot t)$ durchgeführt, wobei sich bei $f_0 = 13.56\,\mathrm{MHz}$ eine Periodendauer von $T_0 = \frac{1}{f_0} \approx 0.07375\,\mu\mathrm{s}$ ergibt.

Transiente Analyse (BDF3):

Die Abbildungen 4.6 und 4.7 zeigen das Verhalten einiger ausgewählter Größen, innerhalb der ersten 10 Perioden.

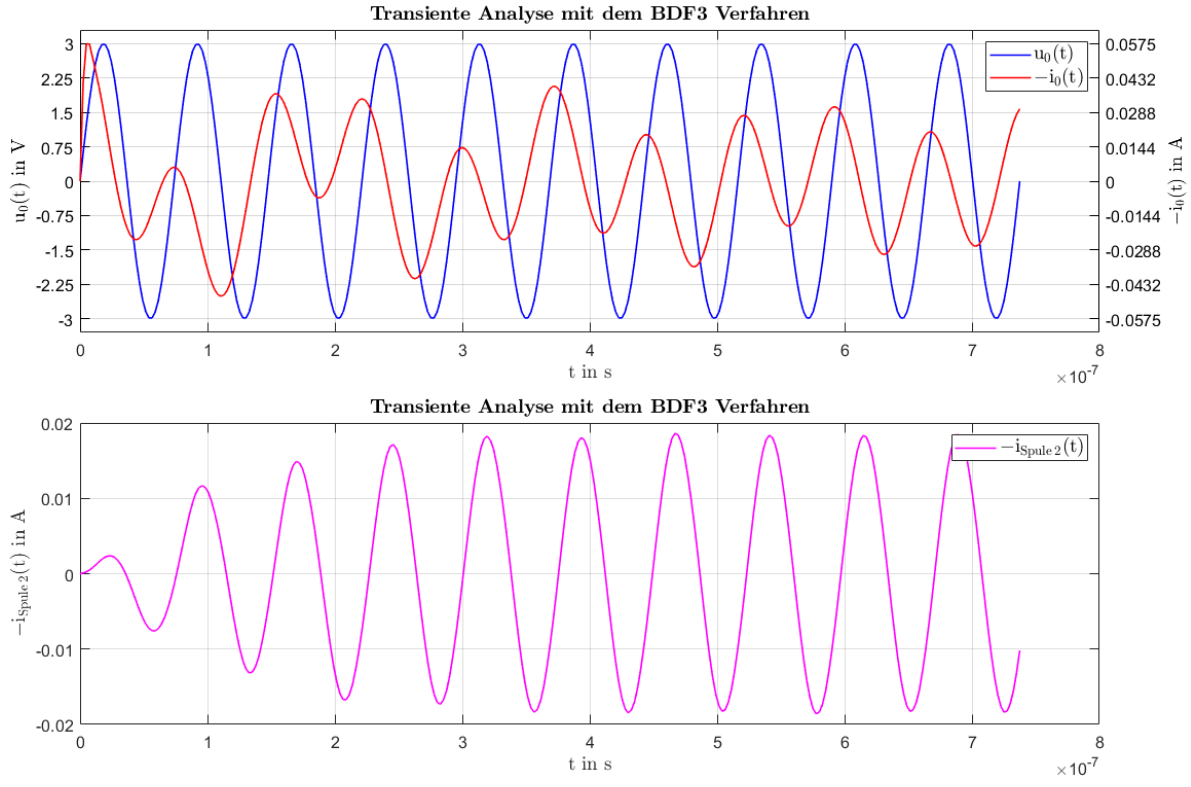


Abbildung 4.6: Berechnung von $u_0(t)$, $-i_0(t)$ und $-i_{Spule\,2}(t)$ für $t \in [0, 10 \cdot T_0]$, unter Verwendung der transienten Analyse (BDF3) mit 30 Stützstellen/Periode.

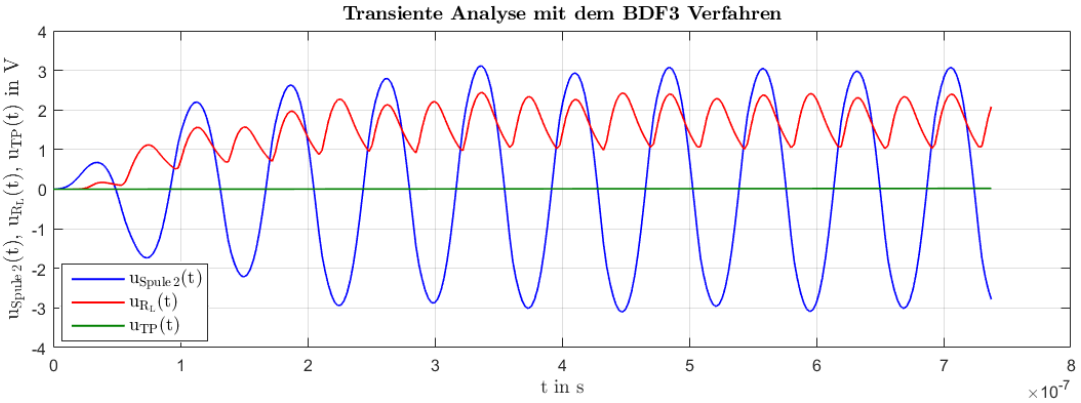


Abbildung 4.7: Berechnung von $u_{Spule\,2}(t)$, $u_{R_L}(t)$ und $u_{TP}(t)$ für $t \in [0, 10{\cdot}T_0]$, unter Verwendung der transienten Analyse (BDF3) mit 30 Stützstellen/Periode.

Von besonders großem Interesse sind hierbei die Signale $u_{R_L}(t)$ und $u_{TP}(t)$, da die Frage der verfügbaren Energie bzw. Leistung an der Last R_L, in vielen Anwendungen von großer Bedeutung ist. Beispielsweise könnte die induzierte Spannung, an den Anschlüssen von Spule 2, die einzige Energiequelle eines Schaltkreises sein.
Gemäß Abbildung 4.7 ist ersichtlich, dass sich bei $u_{R_L}(t)$ nach wenigen Perioden der eingeschwungene Zustand einstellt. Die Spannung an $u_{TP}(t)$ verändert sich im Vergleich zur Periodendauer T_0 nur sehr langsam. Dies lässt sich dadurch begründen, dass $u_{TP}(t)$ die Ausgangsspannung eines RC-Tiefpasses ist, bestehend aus einem $51\,\mathrm{k\Omega}$ Widerstand und einem $1\,\mathrm{nF}$ Kondensator. Wird die Zeitkonstante dieser beiden Komponenten berechnet, so führt dies auf einen Wert von $\tau_{TP} = 51\,\mathrm{k\Omega} \cdot 1\,\mathrm{nF} = 51\,\mu s$. Mit der Faustformel, dass sich ein ungeladener Kondensator eines RC-Tiefpasses nach $5{\cdot}\tau_{TP}$ auf den DC-Wert aufgeladen hat, so entspricht dies $255\,\mu s$ oder $3457.8 \cdot T_0$. D.h. um den eingeschwungenen Zustand von $u_{TP}(t)$ zu berechnen, müssten mindestens 3458 Perioden mittels transienter Analyse berechnet werden. Dass diese Überlegung zulässig ist, zeigt Abbildung 4.8, welche den zeitlichen Verlauf der Spannung $u_{TP}(t)$ innerhalb der ersten 3600 Perioden darstellt.

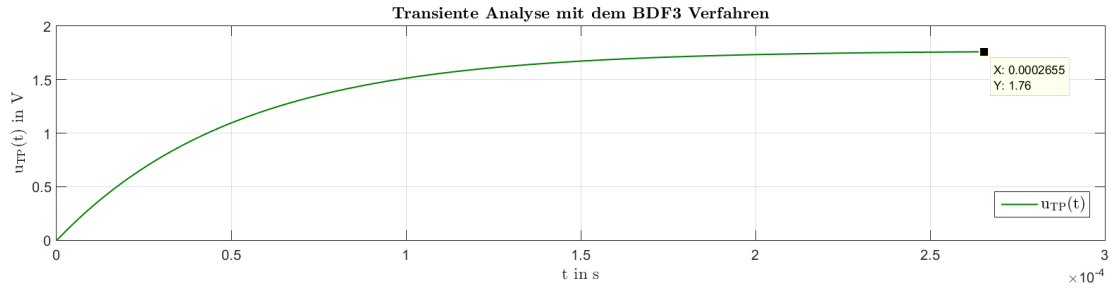


Abbildung 4.8: Verlauf von $u_{R_L}(t)$ innerhalb der ersten 3600 Perioden, unter Verwendung der transienten Analyse (BDF3) mit 30 Stützstellen/Periode.

Dieses Verhalten kann durch einen RC-Tiefpass mit $R = 51\,\mathrm{k\Omega}$ und $C = 1\,\mathrm{nF}$ erklärt werden, welcher durch eine Gleichspannung U_{DC} versorgt wird, wobei U_{DC} dem Mittelwert von $u_{R_L}(t)$ im eingeschwungenen Zustand entspricht. $u_C(t)$ verändert sich dann gemäß

$$u_C(t) = U_{DC} \cdot \left(1 - \exp\left(-\frac{t}{R{\cdot}C}\right)\right).$$

Diese lange Simulationzeit, im Vergleich zur Periodendauer T_0, hat auch eine direkte Auswirkung auf die Laufzeit der Berechnung. Für 10 Perioden und 30 Stützstellen je Periode, benötigte

MATLAB 44.25 s und für 3600 Perioden und 30 Stützstellen je Periode, benötigte MATLAB 6 Stunden und 59 Minuten.
Für $t \in [3599{\cdot}T_0, 3600{\cdot}T_0]$ zeigen die Abbildungen 4.9 und 4.10 das Verhalten einiger ausgewählter Größen, wobei die Signalverläufe dem eingeschwungenen Zustand entsprechen.

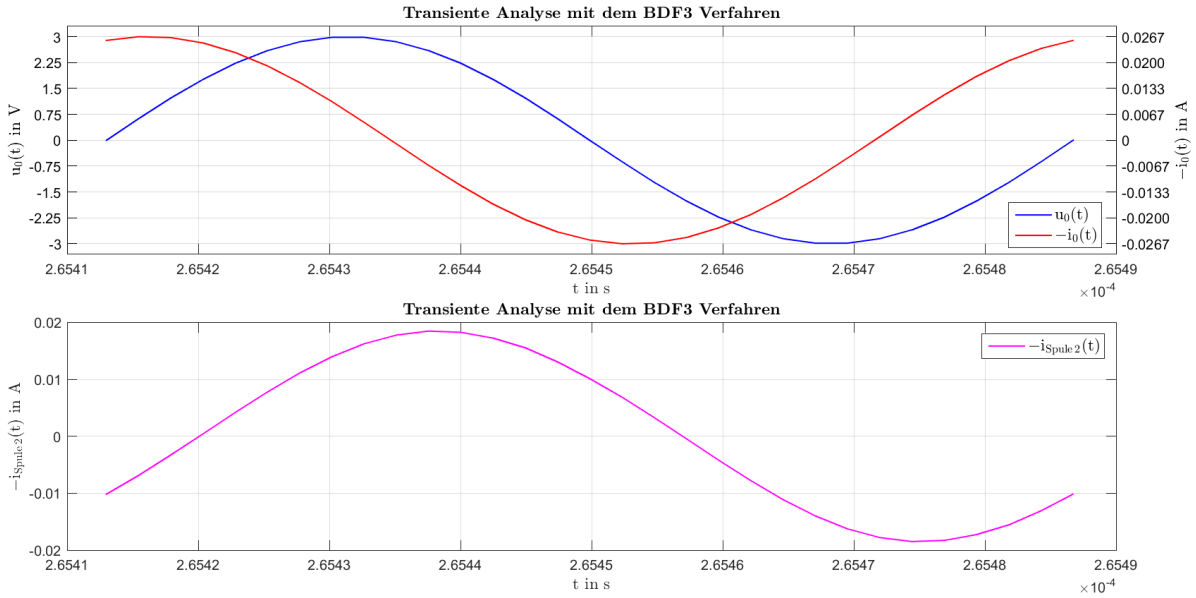


Abbildung 4.9: Berechnung von $u_0(t)$, $-i_0(t)$ und $-i_{Spule\,2}(t)$ für $t \in [3599{\cdot}T_0, 3600{\cdot}T_0]$, unter Verwendung der transienten Analyse (BDF3) mit 30 Stützstellen/Periode.

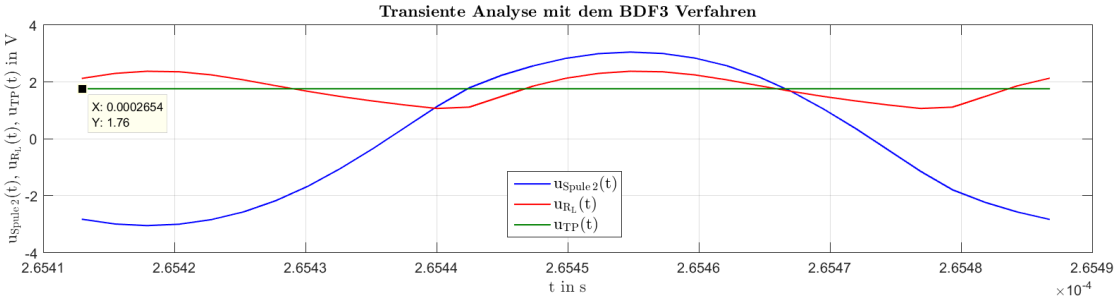


Abbildung 4.10: Berechnung von $u_{Spule\,2}(t)$, $u_{R_L}(t)$ und $u_{TP}(t)$ für $t \in [3599{\cdot}T_0, 3600{\cdot}T_0]$, unter Verwendung der transienten Analyse (BDF3) mit 30 Stützstellen/Periode.

Einfaches Shooting (BDF3):

Die lange Simulationzeit bzw. Laufzeit zur Berechnung von $u_{TP}(t)$, zeigt die Bedeutung von numerischen Methoden, welche den periodisch eingeschwungenen Zustand effizienter berechnen können. Die Abbildungen 4.11 und 4.12 zeigen für einige ausgewählte Größen, die Ergebnisse des einfachen Shooting Verfahrens.

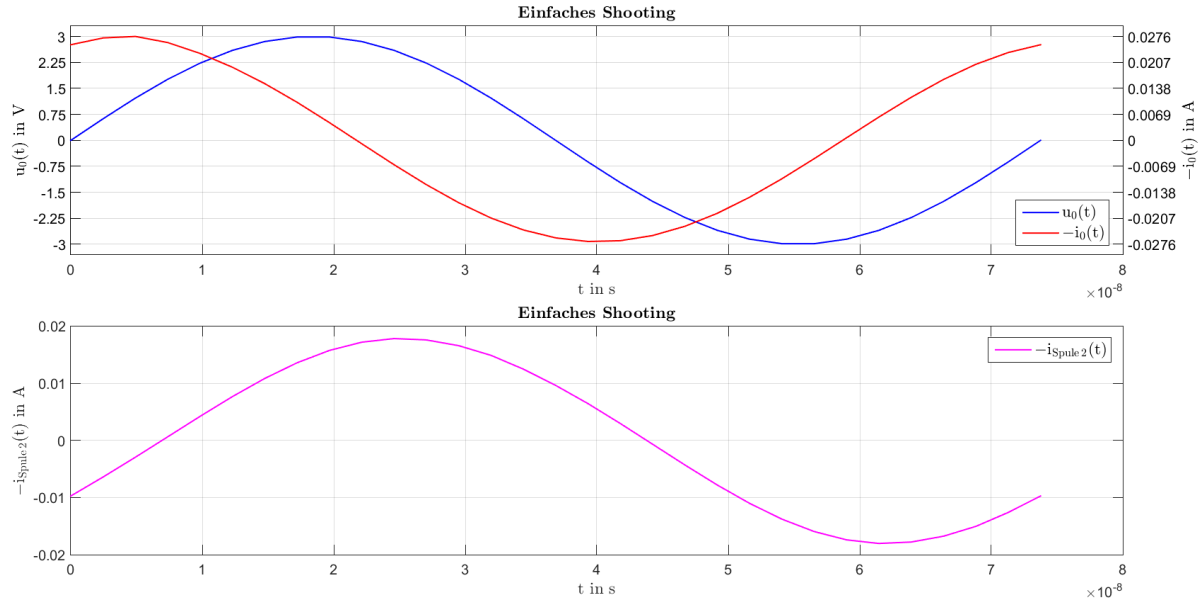


Abbildung 4.11: Berechnung von $u_0(t)$, $-i_0(t)$ und $-i_{Spule\,2}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 30 Stützstellen/Periode.

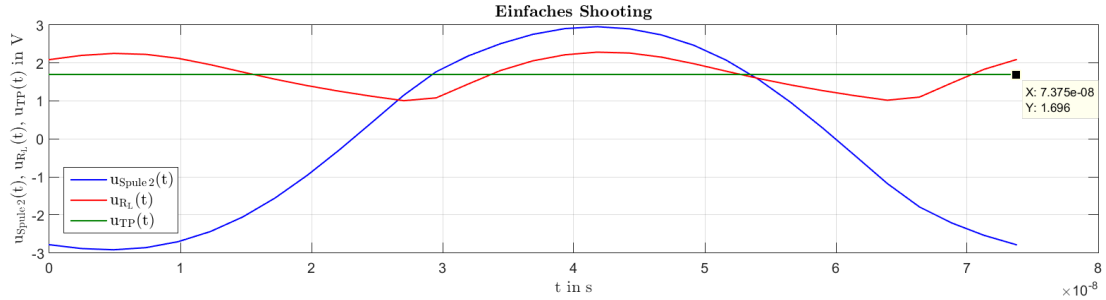


Abbildung 4.12: Berechnung von $u_{Spule\,2}(t)$, $u_{R_L}(t)$ und $u_{TP}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 30 Stützstellen/Periode.

MATLAB benötigte für die Berechnung des einfachen Shooting Verfahrens (BDF3) 219.6 s, wobei hierfür 5 Iterationen im Shooting Verfahren erforderlich waren. Die Laufzeiten der einzelnen Iterationen sind in Tabelle 4.4 zusammengefasst.

Iteration	1	2	3	4	5
Laufzeit in s	30.63	23.71	23.82	22.86	98.99

Tabelle 4.4: Laufzeit je Iteration im einfachen Shooting Verfahren (BDF3).

In Abbildung 4.13 ist exemplarisch der Verlauf der ersten drei Shooting-Iterierten von $u_{R_L}(t)$ dargestellt, sowie der periodisch eingeschwungene Zustand von $u_{TP}(t)$ und $u_{R_L}(t) = u_{R_L}^{(5)}(t)$.

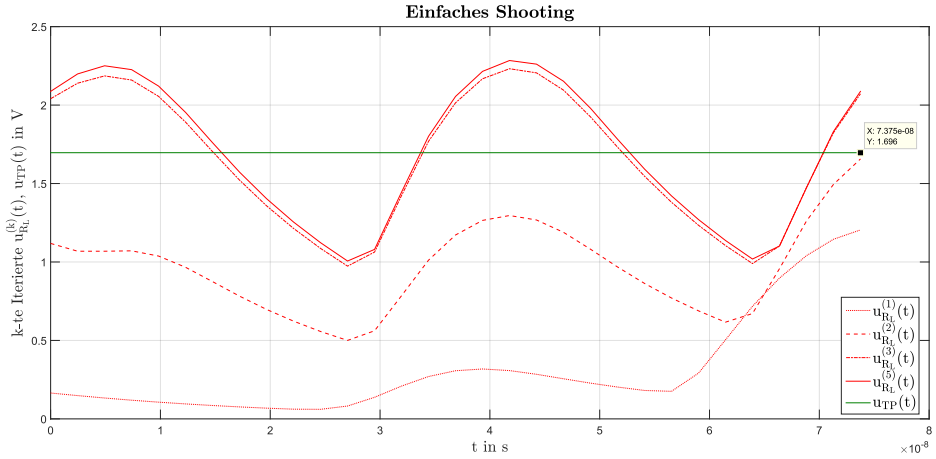


Abbildung 4.13: Verlauf von $u_{TP}(t)$ und $u_{R_L}(t) = u_{R_L}^{(5)}(t)$, sowie der ersten Iterierten $u_{R_L}^{(1)}(t)$, $u_{R_L}^{(2)}(t)$ und $u_{R_L}^{(3)}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 30 Stützstellen/Periode.

Abbildung 4.14 zeigt den Amplitudengang mit den ersten 15 Harmonischen der Signale $i_0(t)$ und $u_{R_L}(t)$. Für die Ermittlung des Amplitudengangs wurde zunächst das Shooting Verfahrens (BDF3) mit 31 Stützstellen je Periode berechnet und anschließend die reelle DFT angewendet, um die DFT Koeffizienten zu erhalten.

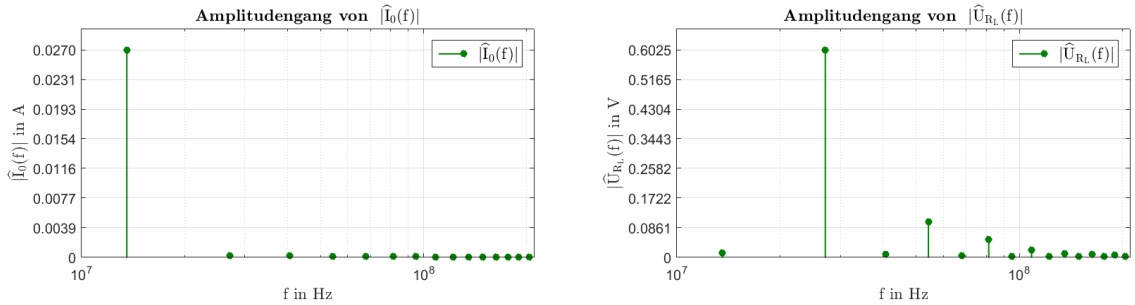


Abbildung 4.14: Amplitudengang der ersten 15 Harmonischen von $i_0(t)$ und $u_{R_L}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 31 Stützstellen/Periode.

Ein Vergleich der Abbildungen 4.10 und 4.12 zeigt, dass die wesentlichen Spannungsverläufe der nichtlinearen Last, bzw. der Brückengleichrichterschaltung, übereinstimmen. Tabelle 4.5 zeigt zudem einen Vergleich des Wertes $u_{TP}(0)$ im eingeschwungenen Zustands, in Abhängigkeit von den Methoden transiente Analyse (BDF3) und einfaches Shooting (BDF3) (gemäß der Abbildungen 4.10 und 4.12).

	$u_{TP}(0)$ in V
Transiente Analyse	1.76
Einfaches Shooting	1.696

Tabelle 4.5: Vergleich von $u_{TP}(0)$ im eingeschwungenen Zustand, bzgl. der transienten Analyse (BDF3) und des einfachen Shooting Verfahrens (BDF3), mit jeweils 30 Stützstellen/Periode.

Die Abweichung der Werte von $u_{TP}(0)$ in Tabelle 4.5 kann damit begründet werden, dass die transiente Analyse bei einer langen Simulationszeit, im Vergleich zur Periodendauer T_0, einen grö-

ßeren Fehler, als das einfache Shooting Verfahren, welches die transiente Analyse lediglich für eine Periodendauer verwendet. Die Abweichung sollte auch geringer werden, wenn in der transienten Analyse die Stützstellen je Periode vergrößert werden, wobei dies wieder zu einer längeren Laufzeit führt, da mehr Schritte berechnet werden müssen.

Bei der genauen Betrachtung von Abbildung 4.11, scheint es so, als ob der Strom $-i_0(t)$ einen kleinen DC-Offset besitzt. Dieser Effekt zeigt allerdings eine Abhängigkeit von der gewählten Anzahl an Stützstellen je Periode. So zeigt Abbildung 4.15, dass dieser Effekt bei einer größeren Anzahl an Stützstellen je Periode stark reduziert wird. Diese Auffälligkeit zeigt allerdings auch, dass die numerischen Berechnungen stets kritisch betrachtet werden sollten und ggf. auf physikalische Sinnhaftigkeit überprüft werden müssen. Bezüglich des Stroms $-i_0(t)$ bedeutet dies, dass zwar ein transienter Vorgang vorhanden sein kann, da eine Sinusquelle im Zeitpunkt $t = 0\,s$ auf ein ungeladenes System geschaltet wird, aber der transiente Vorgang muss abklingen. Es kann auch empfohlen werden, dass das Ergebnis des einfachen Shooting Verfahrens mit anderen Methoden verglichen wird, wie z.B. mit der transienten Analyse oder der Harmonic Balance Methode.

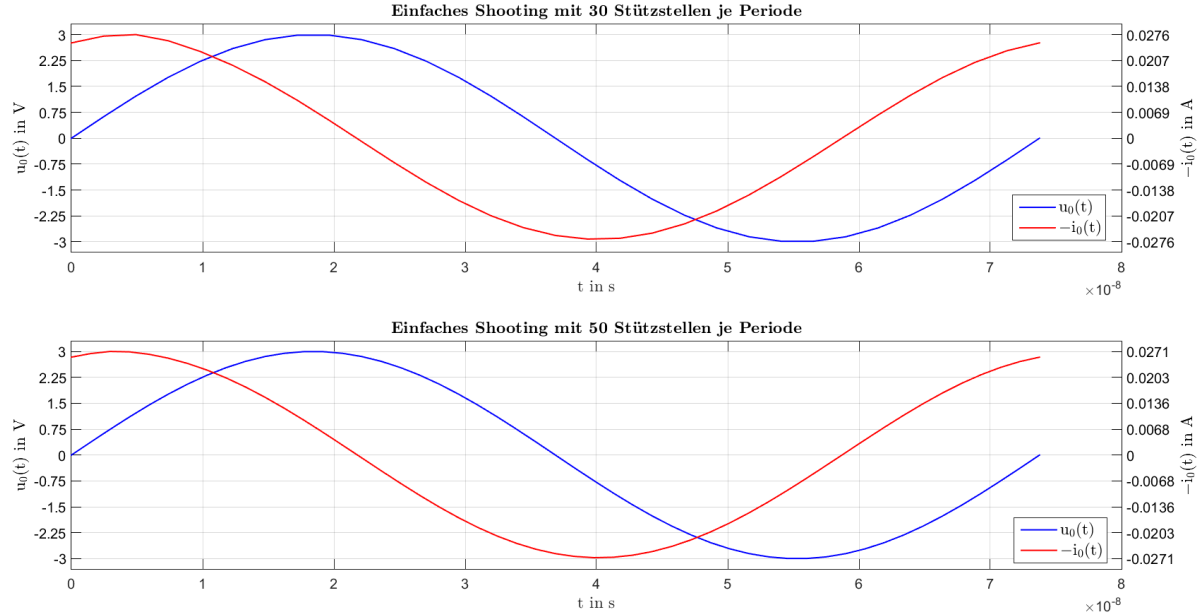


Abbildung 4.15: Vergleich des Stroms $-i_0(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), bei 30 und 50 Stützstellen/Periode.

Harmonic Balance:

Die Berechnung des NFC Systems, welches die externe Beschaltung gemäß Abbildung 4.5 vorsieht, konnte mit der Harmonic Balance Methode nicht berechnet werden. Grund dafür war, dass das lineare Gleichungssystem (3.31) in der ersten Newtoniteration des HB Verfahrens, nicht gelöst werden konnte. MATLAB lieferte dabei die Fehlermeldung, dass das Gleichungssystem nicht nur sehr schlecht konditioniert ist, sondern auch zu singulär ist, um es zu lösen, wobei als Startbedingung der Nullvektor verwendet wurde.
Wie der nächste Abschnitt zeigt, verursachen vor allem die Kondensatoren C_{s1} und C_{s2} (gemäß Abbildung 4.5) Probleme. Es wird vermutet, dass es einerseits am Aufbau der Matrizen des HB Verfahrens liegen könnte, sowie dass die PEEC-Zellen gekoppelte Kapazitäten in der Größenordnung von fF besitzen und C_{s1} und C_{s2} Werte im nF Bereich. Dieser Größenunterschied könnte z.B. aufgrund der endlichen Rechengenauigkeit, zur Singularität der Matrix führen.

Weiters sollte auch stets berücksichtigt werden, dass die resultierenden Gleichungen bei der Anwendung des MKV's zu DAG führen, welche numerisch schwer zu lösen sein können. Bei Zeitschrittverfahren gibt es z.B. mit Satz 2.3 eine Aussage, wie sich bestimmte Eigenschaften des elektrischen Netzwerkes auf den Index der DAG 4.9 und somit auf die Auswahl numerischer Zeitschrittverfahren auswirkt. Bei der Erstellung dieser Masterarbeit war ein entsprechendes Resultat bzgl. dem HB Verfahren nicht bekannt, wobei vermutet wird, dass sich versteckte Nebenbedingungen aufgrund von DAG (4.9) mit höherem Index auch negativ bemerkbar machen sollten. Es ist zudem festzustellen, dass die Kondensatoren zwischen den Knoten 1 und 127 bzw. 128 und 441, zusammen mit den gekoppelten Kapazitäten der PEEC-Zellen der entsprechenden Knoten (gemäß Abbildung 4.4), Schleifen bilden, welche nur aus kapazitiven Elementen bestehen. Dadurch kann gemäß Satz 2.3 davon ausgegangen werden, dass das modellierende elektrische Netzwerk der NFC Systeme, einen DAG Index von 2 besitzt. Eine Abhilfe könnte hierfür z.B. sein, wenn bei den Kondensatoren zwischen den Knoten 1 und 127 bzw. 128 und 441 ein äquivalenter Serienwiderstand hinzugefügt wird.

Überprüfung der Funktionalität der Harmonic Balance Methode

Es zeigte sich, dass vor allem die Kondensatoren C_{s1} und C_{s2} dafür verantwortlich waren, dass die Berechnung des NFC Systems, welches die externe Beschaltung gemäß Abbildung 4.5 enthält, durch die Harmonic Balance Methode nicht funktionierte. In diesem Abschnitt wird daher das externe elektrische Netzwerk des NFC Systems, gemäß Abbildung 4.16 angepasst. Mit dieser Modifikation ist es nun möglich, das Harmonic Balance Verfahren zur Ermittlung des eingeschwungenen Zustands für das vollständige elektrische Netzwerk anzuwenden, wobei die Ergebnisse im Anschluss mit dem einfachen Shooting Verfahren verifiziert werden.

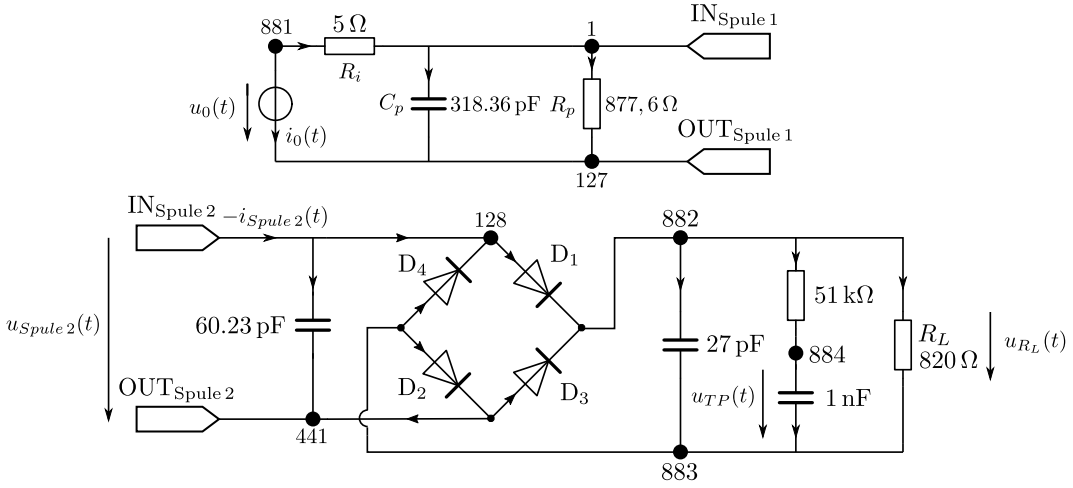


Abbildung 4.16: Modifiziertes externes Netzwerk des NFC Systems.

Harmonic Balance:

Das HB Verfahren liefert unmittelbar die DFT Koeffizienten der gesuchten Lösung $\mathbf{x}(t)$. Die zeitliche Darstellung einzelner Komponenten von $\mathbf{x}(t)$, erhält man dabei folgendermaßen durch die approximierte Fourierreihe. Wurde das HB Verfahren für $K \in \mathbb{N}$ Harmonische berechnet, dann erhält man für jede Komponente der gesuchten Lösung $\mathbf{x}(t)$, eine approximierte Fourierreihe gemäß (3.18). Diese approximierte Fourierreihe wird dann in 100 äquidistanten Stützstellen ausgewertet und dies entspricht dem zeitlichen Verlauf der Lösungskomponenten von $\mathbf{x}(t)$.

Die Abbildungen 4.17 und 4.18 zeigen für einige ausgewählte Größen, die Ergebnisse des HB Verfahrens.

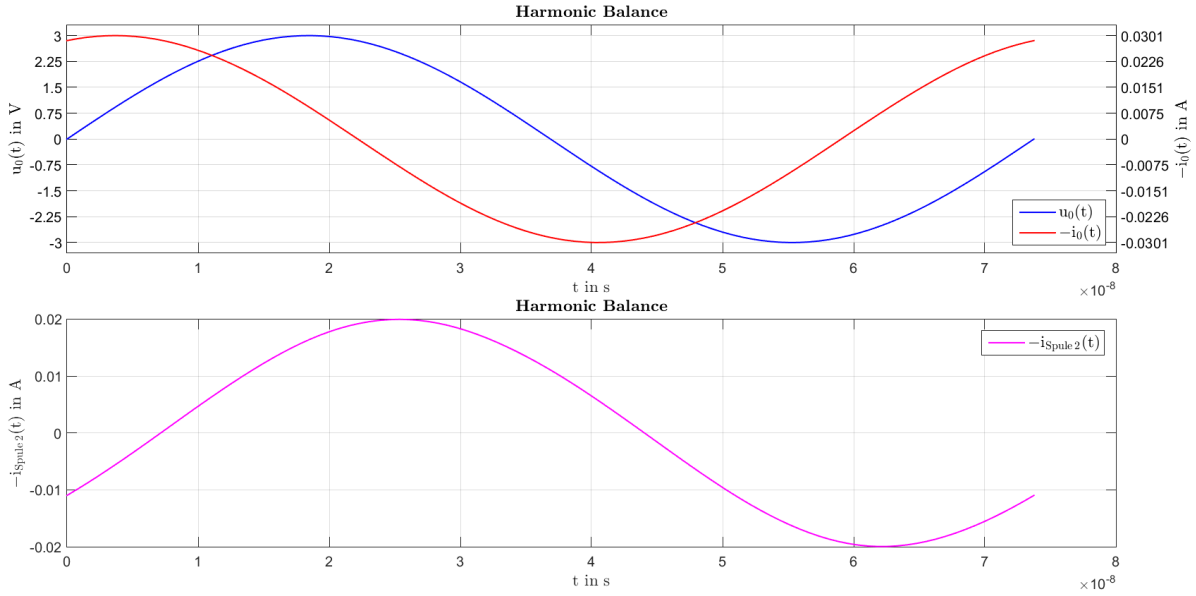


Abbildung 4.17: Berechnung von $u_0(t)$, $-i_0(t)$ und $-i_{Spule\,2}(t)$, unter Verwendung des HB Verfahrens, mit 8 Harmonischen.

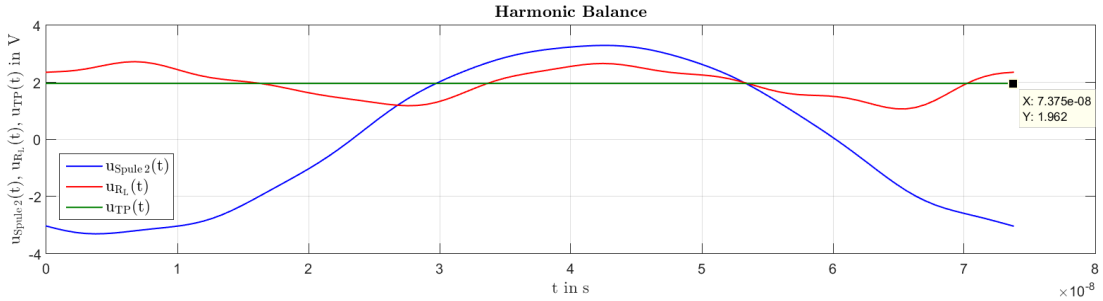


Abbildung 4.18: Berechnung von $u_{Spule\,2}(t)$, $u_{R_L}(t)$ und $u_{TP}(t)$, unter Verwendung des HB Verfahrens, mit 8 Harmonischen.

In Abbildung 4.18 erkennt man für 8 Harmonische bereits eine gute Approximation des Signals $u_{R_L}(t)$. Mit der Wahl $K > 8$ Harmonischen ist zu erwarten, dass die Welligkeit von $u_{R_L}(t)$ aufgrund der approximierenden Fourierreihe geringer wird. Allerdings reichten bei MATLAB bereits bei 10 Harmonischen ein Arbeitsspeicher von 8 GB nicht mehr aus, um den direkten Löser in Kombination mit einer LU- Zerlegung auszuführen.

MATLAB benötigte bei $K = 8$ Harmonischen, für die Initialisierungsphase des HB Verfahrens $75\,s$ und im Durchschnitt 11 Minuten und $58.5\,s$ je HB Iteration, wobei 19 HB Iterationen bis zur Konvergenz benötigt wurden. Für die Gesamtlaufzeit bis zur Konvergenz benötigte das HB Verfahren etwa 3 Stunden, 48 Minuten und $47\,s$.
Es sei angemerkt, dass die Laufzeit je HB Iteration sehr stark von der Vorgabe der Harmonischenanzahl K abhängt. Versuchsweise wurde bei einem Computer mit 16 GB RAM, 13 Harmonische berechnet und hierbei benötigte eine HB Iteration etwa 45 Minuten, wobei die Konvergenzgeschwindigkeit ebenfalls langsam war, bzw. nach 30 Iterationen noch keine Konvergenz gemäß der Algorithmusabbruchbedingung stattgefunden hat.

In Abbildung 4.19 ist exemplarisch der Verlauf ausgewählter Iterierter von $u_{R_L}(t)$ dargestellt, sowie der periodisch eingeschwungene Zustand von $u_{TP}(t)$ und $u_{R_L}(t) = u_{R_L}^{(19)}(t)$.

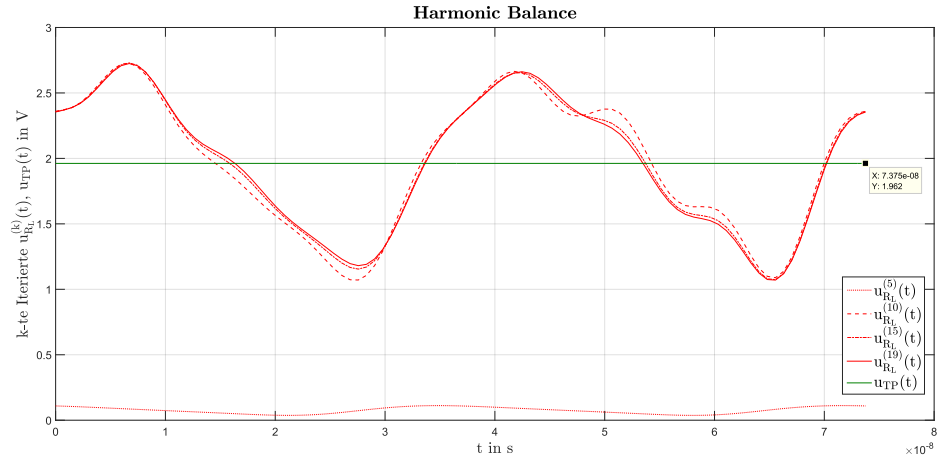


Abbildung 4.19: Verlauf von $u_{TP}(t)$ und $u_{R_L}(t) = u_{R_L}^{(19)}(t)$, sowie der Iterierten $u_{R_L}^{(5)}(t)$, $u_{R_L}^{(10)}(t)$ und $u_{R_L}^{(15)}(t)$, unter Verwendung des HB Verfahrens, mit 8 Harmonischen.

Abbildung 4.20 zeigt den Amplitudengang mit den ersten 8 Harmonischen der Signale $i_0(t)$ und $u_{R_L}(t)$, welche mit dem HB Verfahren berechnet wurden.

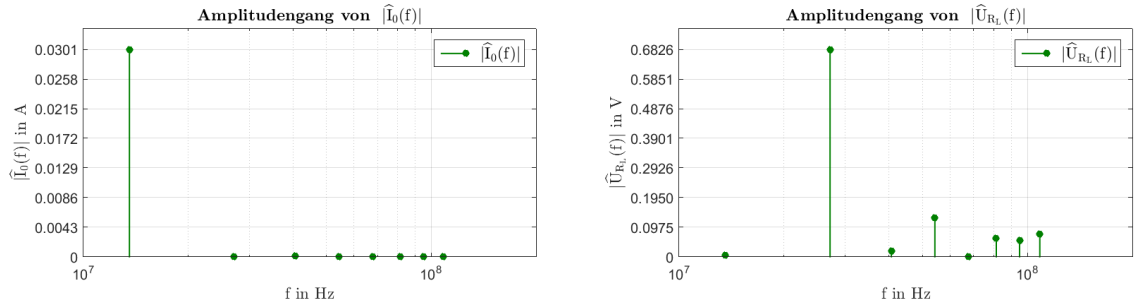


Abbildung 4.20: Amplitudengang von $i_0(t)$ und $u_{R_L}(t)$, unter Verwendung des HB Verfahrens, mit 8 Harmonischen.

Für die gegebene NFC Anwendung zeigt sich, dass eine Konvergenz bei der Vorgabe von geraden Harmonischen besser gelingt, als bei ungeraden. Eine Begründung könnte hierfür z.B. der Amplitudengang von Abbildung 4.20 liefern. So ist die dominante Harmonische der Ladekurve von $u_{R_L}(t)$, die doppelte Grundfrequenz und ungerade Harmonische haben kaum einen Beitrag im Signal $u_{R_L}(t)$. Weiters zeigte sich, dass sich die letzten 2 bis 3 Harmonischen kaum auf den idealen Wert einstellen, sofern nicht ausreichend Harmonische zur Berechnung berücksichtigt wurden.

Verifikation mit dem einfachen Shooting Verfahren (BDF3):

Die Abbildungen 4.21 und 4.22 zeigen für einige ausgewählte Größen, die Ergebnisse des einfachen Shooting Verfahrens.

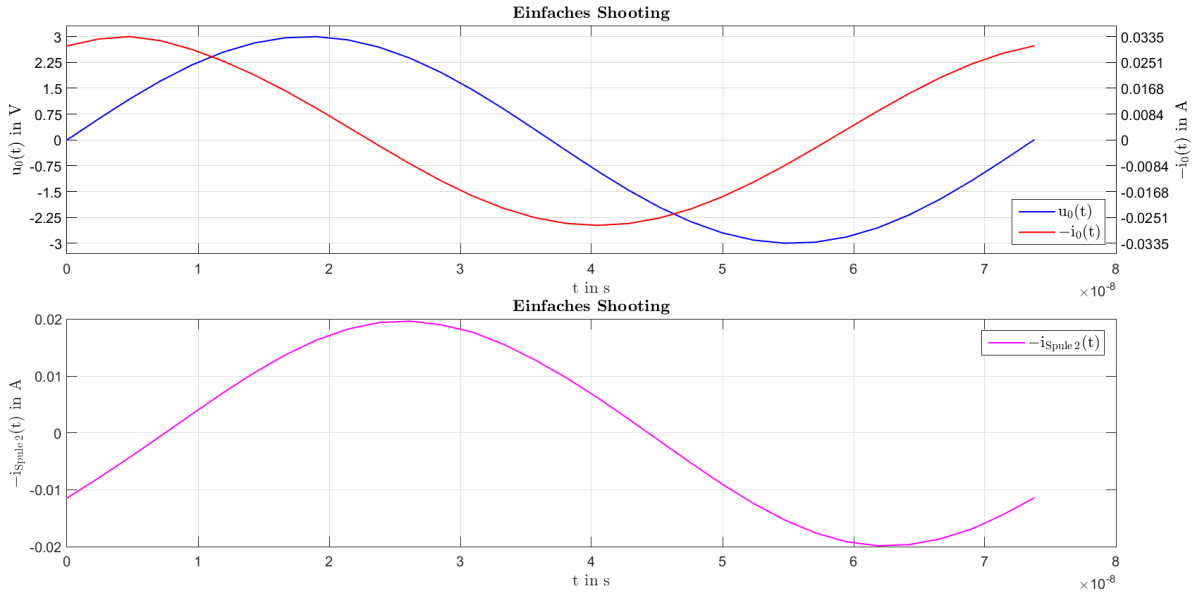


Abbildung 4.21: Berechnung von $u_0(t)$, $-i_0(t)$ und $-i_{Spule\,2}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 31 Stützstellen/Periode.

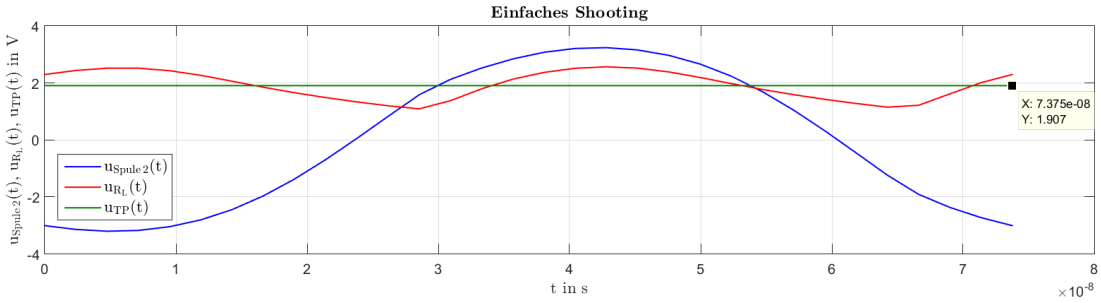


Abbildung 4.22: Berechnung von $u_{Spule\,2}(t)$, $u_{R_L}(t)$ und $u_{TP}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 31 Stützstellen/Periode.

Abbildung 4.23 zeigt den Amplitudengang mit den ersten 15 Harmonischen der Signale $i_0(t)$ und $u_{R_L}(t)$.

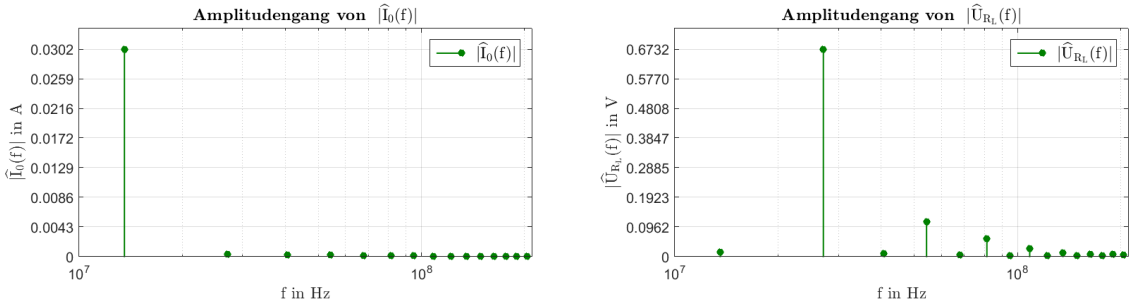


Abbildung 4.23: Amplitudengang der ersten 15 Harmonischen von $i_0(t)$ und $u_{R_L}(t)$, unter Verwendung des einfachen Shooting Verfahrens (BDF3), mit 31 Stützstellen/Periode.

Werden die Abbildungen 4.17 und 4.18 des HB Verfahrens mit den Abbildungen 4.21 und 4.22 des einfachen Shooting Verfahrens verglichen, bzw. Abbildung 4.20 mit Abbildung 4.23, so zeigt sich eine gute Übereinstimmung der Ergebnisse. Es ist aber zu berücksichtigen, dass das HB Verfahren

mit 8 Harmonischen, die Lösung noch nicht ausreichend genau approximiert. Eine größere Vorgabe von Harmonischen wirkt sich aber sehr nachteilig auf die Laufzeit des HB Verfahrens aus, sofern direkte Löser verwendet werden. Es wird daher empfohlen das HB Verfahren zu verbessern, in dem iterative Löser in Kombination mit einer passenden Vorkonditionierung implementiert werden. Für diese NFC Anwendung, mit der externen Beschaltung gemäß Abbildung 4.16, besitzt die Lösung $\mathbf{x}(t)$ 1324 Unbekannte. Bereits bei 8 Harmonischen sind im HB Verfahren lineare Gleichungssyssteme der Form $\mathbf{A}\cdot\mathbf{x}=\mathbf{b}$ zu lösen, wobei $\mathbf{A}\in\mathbb{R}^{N\times N}$ mit $N=1324\cdot(2\cdot 8+1)=22508$ gelte.

Diskussion

Mit den Zeitbereichsmethoden transiente Analyse und einfaches Shooting, sowie der Frequenzbereichsmethode Harmonic Balance, stehen für nichtlineare elektrische Netzwerke, numerische Methoden zur Verfügung, mit denen der periodisch eingeschwungene Zustand berechnet werden kann. Diese numerischen Methoden haben jeweils Vor- und Nachteile. Welche Methode sich für eine Anwendung am besten eignet, ist insbesondere von den Quellfunktionen und dem elektrischen Netzwerk abhängig. Besitzt beispielsweise das elektrische Netzwerk große Zeitkonstanten im Vergleich zur Periodendauer, dann kann der transiente Vorgang im Vergleich zur Periodendauer lange dauern und die transiente Analyse erweist sich als ineffizient. Mit dem einfachen Shooting Verfahren wird gezielt der eingeschwungene Zustand ermittelt, wodurch man von dem transienten Vorgang weniger abhängig ist. Die gemeinsame Basis der betrachteten Zeitbereichsverfahren, bilden die Zeitschrittverfahren BDF und Radau IIA. Diese Zeitschrittverfahren sind dafür optimiert, um polynomiale Funktionen mit bestimmten Polynomgrad exakt berechnen zu können. Sinusähnliche Funktionen lassen sich durch Polynome aber sehr schlecht approximieren, wodurch hierbei die Anzahl der Stützstellen groß gewählt werden muss, um eine bestimmte Genauigkeit gewährleisten zu können.
Es gibt viele Anwendungen bei denen ein elektrisches Netzwerk durch eine sinusförmige Quelle angeregt wird und die Lösung durch eine Fourierreihe mit wenigen Harmonischen gut approximiert werden kann. Für diesen Anwendungsfall besitzt das Harmonic Balance Verfahren den Vorteil, dass Fourierreihen mit wenigen Harmonischen, am besten durch eine trigonometrische Interpolation dargestellt werden können.

In Abschnitt 3.4 wurde die Übereinstimmung der verwendeten Methoden anhand zweier Gleichrichterschaltungen verifiziert. Die Implementierung wurde in MATLAB realisiert, wobei nichtlineare Gleichungssysteme mit dem gedämpften Newton Verfahren nach Bank-Rose gelöst wurde, da hiermit die Abhängigkeit durch den Startwert im Newton Verfahren reduziert wird und dieses auf den Berechnungen des lokalen Newton Verfahrens aufbaut.
Die spezielle Wahl des Modellproblems ermöglichte eine übersichtliche Anwendung der ausgewählten numerischen Methoden. Wie die Herleitung des MKV aber zeigt, ist es u.a. auch möglich nichtlineare kapazitive oder induktive Elemente, sowie gesteuerte Quellen zu berücksichtigen. Sind solche nichtlinearen Elemente im elektrischen Netzwerk enthalten, so wird allerdings zur Modellierung das ladungsorientierte MKV empfohlen, da hiermit durch die numerischen Verfahren, beispielsweise das Prinzip der Ladungserhaltung besser erfüllt werden kann.

Die NFC Systeme von Unterkapitel 4.2, bestehend aus einer vorgegebenen NFC Spule und den äußeren Beschaltungen gemäß Abbildung 4.5 bzw. Abbildung 4.16, bilden weitere Anwendungsbeispiele zur Überprüfung der betrachteten numerischen Methoden aus Kapitel 3. Das Anwendungsbeispiel mit der externen Beschaltung gemäß Abbildung 4.5, konnte mit den Zeitbereichsmethoden transiente Analyse und einfaches Shooting, basierend auf dem BDF3 Verfahren, gelöst werden. Aufgrund großer Unterschiede der Zeitkonstanten in diesem Anwendungsbeispiel, benötigte die transiente Analyse mit mehr als 6 Stunden, deutlich länger als das einfache Shooting Verfahren, welches zur Berechnung des periodisch eingeschwungenen Zustands etwa 3 Minuten und 40 Sekunden benötigte. Nachdem das Shooting Verfahren die transiente Analyse ledigleich

zur Ermittlung einer Periodenzeit einsetzt, ist im Vergleich zur transienten Analyse mit einer langen Simulationszeit, eine größere Genauigkeit zu erwarten, da der Fehler der Zeitschrittverfahren zwischen wahrer und berechneter Lösung, bei einem kleinen Zeitintervall kleiner ausfällt. Die Ergebnisse für das Anwendungsbeispiel zeigten allerdings auch, dass die Anzahl der Stützstellen pro Periode, im einfachen Shooting Verfahren nicht zu klein gewählt werden sollte und es auch vorteilhaft sein kann, wenn das Ergebnis mit anderen numerischen Verfahren überprüft wird. Im Rahmen der Masterarbeit konnte eine detaillierte Fehleranalyse bzw. Überprüfung der (numerischen) Konvergenzordnung nicht durchgeführt werden, wobei dies bei einer weiterführenden Verifikation der numerischen Methoden empfohlen wird.
Bei der Anwendung des Harmonic Balance Verfahrens musste die externe Beschaltung gemäß Abbildung 4.16 modifiziert werden. Das HB Verfahren konnte bei Vorhandensein der Kondensatoren C_{s1} und C_{s2} nicht berechnet werden. Hierfür konnte keine eindeutige Begründung gefunden werden, wobei vermutet wird, dass der Index der DAG und der Größenunterschied der Kapazitäten im Anwendungsbeispiel dafür verantwortlich sind. Bei Zeitschrittverfahren gibt es z.B. mit Satz 2.3 eine Aussage, wie sich bestimmte Eigenschaften des elektrischen Netzwerkes auf den Index der DAG (4.9) und somit auf die Auswahl numerischer Zeitschrittverfahren auswirkt. Bei der Erstellung dieser Masterarbeit war ein entsprechendes Resultat bzgl. dem HB Verfahren nicht bekannt, wobei vermutet wird, dass sich versteckte Nebenbedingungen, aufgrund einer DAG (4.9) mit höherem Index, auch negativ bemerkbar machen sollten. Es ist zudem festzustellen, dass die Kondensatoren zwischen den Knoten 1 und 127 bzw. 128 und 441, zusammen mit den gekoppelten Kapazitäten der PEEC-Zellen der entsprechenden Knoten (gemäß Abbildung 4.4), Schleifen bilden, welche nur aus kapazitiven Elementen bestehen. Dadurch kann gemäß Satz 2.3 davon ausgegangen werden, dass das modellierende elektrische Netzwerk der NFC Systeme, einen DAG Index von 2 besitzt. Eine Abhilfe könnte hierfür z.B. sein, wenn bei den Kondensatoren zwischen den Knoten 1 und 127 bzw. 128 und 441 ein äquivalenter Serienwiderstand hinzugefügt wird.
Das Harmonic Balance Verfahrens konnte allerdings bei der Verwendung der externen Beschaltung gemäß Abbildung 4.16 berechnet werden, wobei die Laufzeit mit der Anzahl der Harmonischen stark zunimmt. So benötigte die Berechnung bei 8 Harmonischen etwa 12 Minuten je HB Iteration und bei 13 Harmonischen bereits etwa 45 Minuten je HB Iteration, wobei für eine Konvergenz mehr als 15 HB Iterationen benötigt werden. Bei einer Vorgabe ab 10 Harmonischen war zudem ein Arbeitsspeicher von 8 GB nicht mehr ausreichend und für die Berechnung von 13 Harmonischen war ein Arbeitsspeicher von 16 GB erforderlich, um das HB Verfahren mit direkten Lösern zu berechnen. Dabei sei zu berücksichtigen, dass bei einer Vorgabe von K Harmonischen im HB Verfahren, für das Anwendungsbeispiel lineare Gleichungssyssteme der Form $\mathbf{A}\cdot\mathbf{x} = \mathbf{b}$ zu lösen sind, wobei $\mathbf{A} \in \mathbb{R}^{N\times N}$ mit $N = 1324\cdot(2\cdot K + 1)$ gelte. Bei 8 Harmonischen entspricht dies schon $N = 22508$. Diese Größenordnungen machen direkte Löser zunehmend ineffizienter und daher kann für weiterführende Entwicklungen der HB Implementierung empfohlen werden, iterative Löser mit geeigneter Vorkonditionierung zu implementieren. Zudem könnten in weiteren Entwicklungen auch Verfahren zur Modellordnungsreduktion eingesetzt werden, um die Anzahl der Unbekannten zu reduzieren.

Literaturverzeichnis

[1] Uri M. Ascher and Linda R. Petzold. *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, USA, 1st edition, 1998. ISBN 0898714125.

[2] Bardia Bandali. Steady-State Analysis of Nonlinear Circuits using the Harmonic Balance on GPU. Master's thesis, University of Ottawa, 2013. https://ruor.uottawa.ca/bitstream/10393/26251/1/Bandali_Bardia_2013_Thesis.pdf [Online verfügbar am 09.05.2020].

[3] R. E. Bank and D. J. Rose. Global Approximate Newton Methods. *Numerische Mathematik*, 37:279–295, 1981.

[4] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. Society for Industrial and Applied Mathematics, 1995. ISBN 978-0-89871-353-4.

[5] Francois E. Cellier and Ernesto Kofman. *Continuous System Simulation*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 978-0-387-26102-7.

[6] Hajdin Ceric. *Numerical Techniques in Modern TCAD*. PhD thesis, Technische Universität Wien, 2005. https://www.iue.tuwien.ac.at/phd/ceric/thesis.html [Online verfügbar am 02.05.2020].

[7] L. Chua. Dynamic nonlinear networks: State-of-the-art. *IEEE Transactions on Circuits and Systems*, 27(11):1059–1087, 1980.

[8] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer-Verlag, 2008. 2. korrigierte Auflage, ISBN 978-3-540-76492-2.

[9] P. Deuflhard. *Newton Methods for Nonlinear Problems*. Springer, 2011. 978-3-642-23898-7.

[10] Thomas Dorn. Abbruch einer Iteration. Website. http://www.dorn.org/uni/sls/kap10/j02_05.htm [Online verfügbar am 01.05.2020].

[11] Jonas Ekman. *Electromagnetic Modeling Using the Partial Element Equivalent Circuit Method*. PhD thesis, Lulea University of Technology, 2003. http://www.diva-portal.org/smash/get/diva2:990875/FULLTEXT01.pdf [Online verfügbar am 28.08.2020].

[12] M. Günther, U. Feldmann, and E.J.W. Maten, ter. *Modelling and discretization of circuit problems*. Handbook of Numerical Analysis. Elsevier, Netherlands, 2005. DOI 10.1016/S1570-8659(04)13006-8, ISBN 0-444-51375-2.

[13] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Systems by Runge-Kutta Methods*. Springer, 1989. 1. Edition, ISBN 978-3-540-51860-0.

[14] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II, Stiff and Differential-Algebraic Problems*. Springer, 1996. Second Revised Edition, ISBN 978-3-642-05220-0.

[15] Ken Kundert. Simulation Methods for RF Integrated Circuits. In *Proceedings of the 1997 IEEE/ACM International Conference on Computer-Aided Design*, ICCAD '97, page 752–765, USA, 1997. IEEE Computer Society.

[16] Kenneth S. Kundert, Jacob K. White, and Alberto Sangiovanni-Vincentelli. *Steady-State Methods for Simulating Analog and Microwave Circuits*. Springer, 1990. https://doi.org/10.1007/978-1-4757-2081-5 [Online verfügbar am 09.05.2020], ISBN 978-1-4419-5121-2.

[17] Ricardo Riaza. *Differential-Algebraic Systems, Analytical Aspects and Circuit Applications*. World Scientific, 2008. ISBN-13 978-981-279-180-1.

[18] Albert Ruehli, Giulio Antonini, and Li Jiang. *Circuit Oriented Electromagnetic Modeling Using the PEEC Techniques*. 07 2017. ISBN 978-1-118-43664-6.

[19] Diana Estévez Schwarz. *Consistent initialization for index-2 differential algebraic equations and its application to circuit simulation*. PhD thesis, Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, 2000. http://dx.doi.org/10.18452/14512 [Online verfügbar am 17.04.2020].

[20] Diana Estévez Schwarz. *Consistent initial values for DAE systems in circuit simulation*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2005.

[21] Diana Estévez Schwarz and Caren Tischendorf. *Structural analysis for electric circuits and consequences for MNA*. Humboldt-Universität zu Berlin, Mathematisch-Naturwissenschaftliche Fakultät II, Institut für Mathematik, 2005. http://dx.doi.org/10.18452/2687 [Online verfügbar am 08.04.2020].

[22] H. Schwetlick. *Numerische Lösung nichtlinearer Gleichungen*. VEB Deutscher Verlag der Wissenschaften, 1979. ISBN 3-486-21731-3.

[23] Hans Wilhelm Schüßler. *Netzwerke, Signale und Systeme: Band 1, Systemtheorie linearer elektrischer Netzwerke*. Springer Verlag, zweite, neubearbeitete und erweiterte auflage edition, 1988. ISBN 978-3-540-52987-3.

[24] STMicroelectronics. ST_SIGNAL_SCHOTTKY_V6.LIB, Rev. 6.0, Date Dec. 2019. Website. enthalten in en.ST_signal_schottky_diodes_pspice_models_v4.zip und abrufbar unter https://www.st.com/en/diodes-and-rectifiers/bar43.html#resource Abschnitt 'HW Model, CAD Libraries & SVD' [Online verfügbar am 28.05.2020].

[25] Riccardo Torchio. P.E.E.C. modelling of large-scale fusion reactor magnets. Master's thesis, Universitá degli Studi di Padova, Dipartimento di Ingegneria Industriale, 2016. http://tesi.cab.unipd.it/52819/1/Torchio_Riccardo_tesi.pdf [Online verfügbar am 10.09.2020].

[26] M. Ulbrich and S. Ulbrich. *Nichtlineare Optimierung*. Birkhäuser, 2012. 978-3-0346-0142-9.

[27] Wolfgang Walter. *Analysis 2*. Springer, 1995. 4. durchgesehene und ergänzte Auflage, ISBN 978-3-540-58666-1.