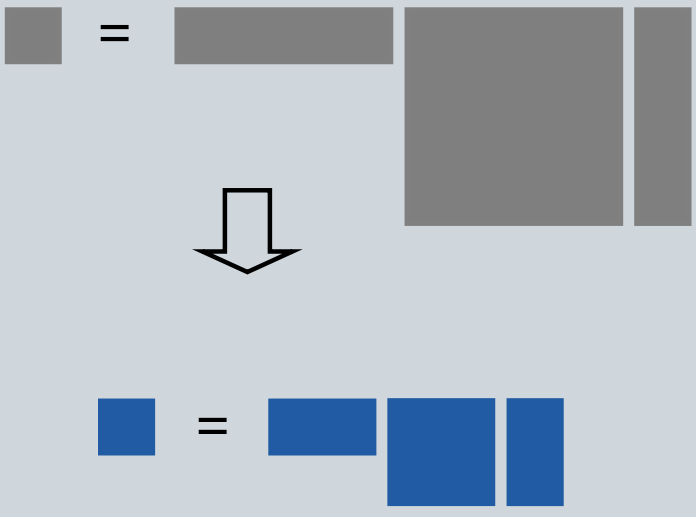
Tilmann Wittig

Zur Reduzierung der Modellordnung in elektromagnetischen Feldsimulationen



Cuvillier Verlag Göttingen

Zur Reduzierung der Modellordnung in elektromagnetischen Feldsimulationen

Vom Fachbereich Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt

zur Erlangung
der Würde eines Doktor-Ingenieurs (Dr.-Ing.)
genehmigte

Dissertation

von
Dipl.-Ing. Tilmann Wittig
geboren am 12. Mai 1972 in Leverkusen

Darmstadt 2003

Referent:	Prof. Dr.-Ing. Thomas Weiland
Korreferent:	Prof. Dr. Wilhelmus H. A. Schilders
Tag der Einreichung:	02. September 2003
Tag der mündlichen Prüfung:	31. Oktober 2003

D 17
Darmstädter Dissertation

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.
1. Aufl. - Göttingen : Cuvillier, 2004
Zugl.: Darmstadt, Techn. Univ., Diss., 2003
ISBN 3-86537-208-2

Nonnenstieg 8, 37075 Göttingen
Telefon: 0551-54724-0
Telefax: 0551-54724-21
www.cuvillier.de

1. Auflage, 2004
Gedruckt auf säurefreiem Papier

ISBN 3-86537-208-2

Inhaltsverzeichnis

1 Einleitung **1**

1.1 Einführung 1

1.2 Übersicht 3

2 Die Methode der Finiten Integration **5**

2.1 Die Maxwellschen Gleichungen 5

2.1.1 Materialeigenschaften 7

2.2 Diskretisierung der Maxwellschen Gleichungen 8

2.2.1 Die Gitter-Maxwellgleichungen 9

2.2.2 Materialdiskretisierung 13

2.2.3 Randbedingungen und Anregung des Rechengebiets 14

3 FIT in Systemdarstellungen **27**

3.1 Zustandsraumdarstellung der Impedanz 27

3.1.1 Klassischer Zustandsraum 27

3.1.2 Curl-Curl-Formulierung 30

3.1.3 Systeme höheren Grades 31

3.2 Systemeigenschaften 34

3.2.1 Kausalität 35

3.2.2 Stabilität 36

3.2.3 Passivität von Impedanzfunktionen 39

3.2.4 Steuerbarkeit und Beobachtbarkeit 41

3.3 Streuparameter 42

3.3.1 Grundlagen 42

3.3.2 Streuparameter aus Impedanzmatrizen 44

3.3.3 Zustandsraumdarstellung der Streuparameter 45

4 Reduzierung der Modellordnung **47**

4.1 Einführung 47

4.2 Mathematische Grundlagen 50

4.2.1 Ordnungsreduktion durch Projektion 50

4.2.2 Krylov-Unterraum-Verfahren 51

4.3 Verfahren zur Reduktion der Modellordnung 57

4.3.1 Partielle Realisierungen 57
4.3.2 Korrigierte Modalanalyse 70
4.3.3 Padé-Approximationen 72
4.3.4 Two-Step-Lanczos 85
4.3.5 Weitere Verfahren 90

5 Spektralschätzung aus Zeitbereichsdaten 93
5.1 FIT-Simulationen im Zeitbereich 93
5.2 Filterbasierte Spektralschätzung 95
5.2.1 ARMA-Modelle 96
5.2.2 Iterativer Prony und Verfahren nach Steiglitz-McBride 97
5.2.3 Ein Beispiel 98
5.3 4SID 100

6 Generierung von Ersatzschaltbildern 101
6.1 Einführung 101
6.2 Interpretation als Knotenanalysemodell 103
6.2.1 Lineare Systeme 104
6.2.2 Curl-Curl Systeme 105
6.3 Pol-Residuen-Darstellung 109

7 Anwendungsbeispiele 111
7.1 Langer-Filter 111
7.2 6-Kreis Hohlleiter-Filter 122
7.3 Patchantenne 125
7.4 Chip-Interconnect-Modell 127

8 Zusammenfassung und Ausblick 131

A Der Bi-Lanczos-Algorithmus 135

Symbolverzeichnis 137

Literaturverzeichnis 141

Danksagung 151

Wissenschaftlicher Werdegang 153

Kapitel 1

Einleitung

1.1 Einführung

Numerische Simulationsprogramme sind in den letzten Jahren zu unverzichtbaren Werkzeugen in vielen Bereichen der Ingenieurwissenschaften geworden. Simulationen ermöglichen beschleunigte Designoptimierung, ersparen häufig den Bau von Prototypen und entsprechen damit dem grundsätzlichen Trend zu immer kürzeren Entwicklungszyklen. Zudem ermöglicht die Simulation häufig Einblicke in Zusammenhänge, die aufgrund der Miniaturisierung in dieser Form messtechnisch gar nicht mehr erfasst werden können.

Einen besonders dynamischen Bereich stellt in diesem Zusammenhang die elektromagnetische Feldsimulation mit gitterbasierten Methoden dar. Dies beruht insbesondere auf dem rasanten Fortschreiten der Informationstechnologie in allen Bereichen der Technik und des alltäglichen Lebens sowie der damit einhergehenden Verwendung immer höherer Frequenzen in elektronischen Schaltungen.

Immer höher werdende Frequenzen haben in zweierlei Hinsicht Einfluss auf den Simulationsprozess: Zum Einen erfordern die kürzeren Wellenlängen die feinere räumliche Abtastung von Strukturen und führen auf immer größere Gleichungssysteme, die im Rahmen der Simulation gelöst werden müssen. Zum Anderen erhalten Feldeffekte auch immer größere Bedeutung in Bereichen, die bis vor kurzem komplett der Schaltkreissimulation zugerechnet wurden. So gewinnen Effekte wie Laufzeitverzögerungen, Nebensprechen oder Abstrahlung auch in klassischen Netzwerken oder sogar in einzelnen Teilen wie Chipzuleitungen und Steckverbindungen zunehmend an Bedeutung.

Da es auch in nächster Zukunft nicht möglich sein wird, ganze logische Schaltungen komplett durch Feldsimulationsprogramme zu erfassen, bleibt als Alternative die Verkopplung zweier separat arbeitender Simulationsprogramme zur Netzwerk- und zur Feldsimulation. Eine weitere und komfortablere Möglichkeit stellt die Generierung so genannter Makromodelle dar, welche das Verhalten des feldbehafteten Bauteils im interessierenden Frequenzbereich durch ein System möglichst niedriger Ordnung beschreiben. Diese Modelle können beispielsweise in Form eines Ersatzschaltbildes in das Netzwerksimulationsprogramm einbezogen werden, womit

zur Simulation des Gesamtsystems schließlich nur eine einzige Simulationsplattform benötigt wird.

Dass ein solches Modell mit stark reduzierter Modellordnung überhaupt existiert, wird anschaulich klar, wenn man bedenkt, dass das diskretisierte Modell Tausende bis hin zu Millionen von Unbekannten hat, während das Übertragungsverhalten im interessierenden Frequenzbereich oft nur eine kleine Anzahl von Polstellen besitzt.

Insbesondere wenn allein das Übertragungsverhalten einer feldbehafteten Struktur von Interesse ist, können Modelle reduzierter Ordnung jedoch auch direkt zur Lösung des diskretisierten Problems innerhalb der Feldsimulation herangezogen werden. Anstatt das Modell zu lösen, das sich aus der Diskretisierung ergibt und das häufig bis zu Millionen von Unbekannten hat, kann zunächst ein Reduzierungsschritt vorangestellt werden. Gelöst wird dann schließlich nur das System mit geringer Ordnung. Ein solches Vorgehen wird auch als *Fast Frequency Sweep* bezeichnet. Es ist offensichtlich, dass in diesem Fall die Rechenzeit zur Erstellung des Modells kleiner sein sollte als die Rechenzeit zur Lösung des unreduzierten Systems.

Makromodelle finden auch im Optimierungsprozess Anwendung, da das Verhalten bereits optimierter oder unveränderlicher Bereiche der Struktur durch ein nur einmalig zu erstellendes Modell erfasst werden kann, während der verbleibende Rest durch die Simulation beschleunigt optimiert werden kann. Auch wenn eine Teilstruktur innerhalb eines größeren Rechengebiets mehrfach vorkommt, kann der Einsatz eines Makromodells sinnvoll sein.

Es wird somit deutlich, dass je nach geplantem Einsatz des Modells die Schwerpunkte der Reduzierung unterschiedlich gewichtet sind: Soll das Modell für einen *Fast Frequency Sweep* verwendet werden, ist neben der Genauigkeit des Modells vor allem die Rechenzeit interessant, die zur Reduzierung benötigt wird. Wird das Modell jedoch nur ein einziges Mal erzeugt, um dann viele Male beispielsweise in einem Optimierungsprozess eingesetzt oder mit anderen Simulatoren verkoppelt zu werden, ist eine möglichst geringe Modellgröße und die Erhaltung physikalischer Eigenschaften wie Stabilität und Passivität von vorrangigem Interesse. Rechenzeit ist in diesem Fall nur von zweitrangiger Bedeutung.

Zum Auffinden eines reduzierten Modells wird im Rahmen dieser Arbeit ein sehr allgemeiner Projektionsansatz beschrieben und untersucht. Dieser ermöglicht unterschiedliche Varianten zur Wahl der Projektionsmatrizen. Als besonders geeignet erweisen sich dazu einzelne Feldlösungen, Eigenvektoren der betrachteten Polstellen, Taylormomente oder so genannte Krylov-Unterraum-Vektoren. Diese Varianten unterscheiden sich zum Teil deutlich in Rechenaufwand und resultierender Modellgröße.

Besonders die Kombination einer so genannten partiellen Realisierung basierend auf Krylov-Unterraum-Vektoren und einem momenten-basierten Verfahren zeigt sich als sehr effizient, was sowohl Rechenzeit als auch Modellgröße betrifft. Der unter dem Namen *Two-Step-Lanczos (TSL)* in dieser Arbeit vorgeschlagene Algorithmus erhält für resonante Systeme zusätzlich Stabilität und Passivität, läuft vollständig automatisiert ab und das resultierende Modell lässt sich auf einfache Weise als Ersatzschaltbild interpretieren.

Als alternatives Vorgehen wird ein Verfahren untersucht, bei dem das System zunächst mit einem sehr effizienten Zeitbereichslöser teilweise berechnet wird. Da bei resonanten Systemen die Signalamplitude jedoch nur sehr langsam abklingt, wird die Rechnung schließlich abgebrochen und das Übertragungsverhalten aus dem bereits berechneten Zeitsignal mittels moderner Signalverarbeitungsmethoden geschätzt.

1.2 Übersicht

Im Anschluss an diese Einleitung wird im Kapitel 2 die Methode der *Finiten Integration* vorgestellt. Da dieses Verfahren grundlegende physikalische Eigenschaften der kontinuierlichen Maxwellschen Gleichungen auch im Diskreten beibehält, ist es für die spätere Modellgenerierung besonders geeignet.

In Kapitel 3 werden die diskretisierten Modelle als Systeme betrachtet. Hierbei kommen klassische sowie erweiterte Zustandsraumdarstellungen zur Anwendung. Wichtige Systemeigenschaften wie Kausalität, Stabilität und Passivität werden ebenso betrachtet wie die Zusammenhänge zwischen den Impedanz- und den Streuparameter-Übertragungsfunktionen.

Kapitel 4 stellt den Kern der Arbeit dar und beschreibt den Prozess der Reduzierung der Ordnung mit Hilfe verschiedener Projektionsverfahren. Aus den jeweiligen Vorteilen einer partiellen Realisierung und einer momente-erhaltenden Padé-Approximation wird schließlich der *Two-Step-Lanczos*-Algorithmus abgeleitet und untersucht.

Verfahren zur Spektralschätzung aus Zeitsignalen werden als alternative Möglichkeit zur schnellen Berechnung des Übertragungsverhaltens einer Struktur in Kapitel 5 vorgestellt.

Kapitel 6 befasst sich mit unterschiedlichen Methoden, aus der Zustandsraumdarstellung des reduzierten Modells tatsächlich ein Netzwerk abzuleiten, das als Ersatzschaltbild verwendet werden kann. Das reduzierte System wird hierzu als Knotenanalysemodell interpretiert oder in eine Pol-Residuen-Darstellung überführt.

Die Vor- und Nachteile der verschiedenen Verfahren werden schließlich in Kapitel 7 anhand von vier praxisrelevanten Beispielen aus dem Hochfrequenzbereich verglichen. Dies sind zwei resonante Filterstrukturen, ein Antennenbeispiel sowie ein Chipgehäuse mit einer großen Zahl von Anschlüssen.

Die Arbeit schließt mit einer Zusammenfassung und einem Ausblick.

Kapitel 2

Die Methode der Finiten Integration

Die Maxwellschen Gleichungen und die zugehörigen Materialbeziehungen bilden die Grundlage der klassischen Elektrodynamik. Die Methode der Finiten Integration, die in diesem Kapitel vorgestellt werden soll, stellt eine Transformation dieser Gleichungen in einen diskreten Gitterraum dar, die wichtige physikalische Eigenschaften der Maxwellschen Gleichungen erhält.

Zunächst wird die Grundidee der Methode mit der Diskretisierunng der Gleichungen sowie der Materialbeziehungen beschrieben. Besondere Aufmerksamkeit erhalten frequenzabhängige Zusammenhänge wie dispersive Materialien und komplexe Berandungen des Rechengebiets wie Impedanzwände oder absorbierende Ränder, da diese bei der Interpretation des Modells als System von entscheidender Bedeutung sind.

2.1 Die Maxwellschen Gleichungen

Früheste Beobachtungen sowohl elektrischer als auch magnetischer Phänomene reichen bereits in die Antike zurück. Erste quantitative Untersuchungen erfolgten jedoch erst im 18. Jahrhundert auf dem Gebiet der Elektrostatik durch H. Cavendish (1773) und C. A. de Coulomb (1785). Die Verkopplung elektrischer und magnetischer Felder wurde erstmals 1826 von A.-M. Ampère in einer ersten Fassung des Durchflutungsgesetzes erfasst, 1831 folgte M. Faraday durch Aufstellung des Induktionsgesetzes. James Clark Maxwell erweiterte schließlich 1873 die bestehenden Ansätze zu einer einheitlichen elektromagnetischen Theorie [1]. Sein wesentlicher Beitrag hierbei war in Deutung der Ableitung der elektrischen Flussdichte als Stromdichte im Durchflutungsgesetz. Die vier - heute nach ihm benannten - Gleichungen bilden seitdem die Grundlage der klassischen makroskopischen Elektrodynamik.

Die Maxwellschen Gleichungen verknüpfen fünf vektorielle Grundgrößen[1], die elektrische Feldstärke $\vec{E}$, die elektrische Flussdichte $\vec{D}$, die elektrische Stromdichte $\vec{J}$, die magnetische Feldstärke $\vec{H}$ sowie die magnetische Flussdichte $\vec{B}$ und die skalare

[1] Ein Verzeichnis der verwendeten Symbole und Schreibweisen findet sich im Anhang.

Raumladungsdichte ρ. Die ersten beiden Gleichungen setzen jeweils die zeitliche Ableitung einer Flussgröße über einer beliebigen Fläche A mit der über ihren Rand ∂A abfallenden Spannung in Verbindung, während die anderen beiden Gleichungen die in einem Volumen V eingeschlossene Ladung mit den Flüssen durch deren geschlossene Oberfläche ∂V in Beziehung setzen, bzw. die Nichtexistenz der Ladung zeigen. Für eine ausführliche Darstellung siehe z.B. [2, 3]. Unter der Annahme ruhender Medien lauten die Gleichungen:

$$\oint_{\partial A} \vec{E}(\vec{r},t) \cdot \mathrm{d}\vec{s} = -\int_{A} \frac{\partial \vec{B}(\vec{r},t)}{\partial t} \cdot \mathrm{d}\vec{A} \qquad \forall A \subset \mathbb{R}^3, \tag{2.1.1a}$$

$$\oint_{\partial A} \vec{H}(\vec{r},t) \cdot \mathrm{d}\vec{s} = \int_{A} \left(\frac{\partial \vec{D}(\vec{r},t)}{\partial t} + \vec{J}(\vec{r},t) \right) \cdot \mathrm{d}\vec{A} \qquad \forall A \subset \mathbb{R}^3, \tag{2.1.1b}$$

$$\oint_{\partial V} \vec{D}(\vec{r},t) \cdot \mathrm{d}\vec{A} = \int_{V} \rho(\vec{r},t)\, dV \qquad \forall V \subset \mathbb{R}^3, \tag{2.1.1c}$$

$$\oint_{\partial V} \vec{B}(\vec{r},t) \cdot \mathrm{d}\vec{A} = 0 \qquad \forall V \subset \mathbb{R}^3. \tag{2.1.1d}$$

Nach ihren Urhebern werden Gleichung 2.1.1a auch als Faradaysches, Gl. 2.1.1b als Ampèresches und Gl. 2.1.1c als Gaußsches Gesetz bezeichnet.

Die elektrische Stromdichte

$$\vec{J}(\vec{r},t) \;=\; \vec{J}_e(\vec{r},t) + \vec{J}_l(\vec{r},t) + \vec{J}_k(\vec{r},t) \tag{2.1.2}$$

in Gl. 2.1.1b stellt die Summe aus einer von außen eingeprägten Stromdichte $\vec{J}_e$, der sich aufgrund einer Leitfähigkeit im elektrischen Feld einstellenden Leitungsstromdichte $\vec{J}_l$, sowie der Konvektionsstromdichte $\vec{J}_k$, hervorgerufen durch von elektromagnetischen Kräften bewegten Ladungen, dar.

Durch Anwendung der Integralsätze von Gauß und Stokes lassen sich die Gleichungen in die inhaltlich identische differenzielle Darstellung überführen:

$$\mathrm{rot}\, \vec{E}(\vec{r},t) = -\frac{\partial \vec{B}(\vec{r},t)}{\partial t}, \tag{2.1.3a}$$

$$\mathrm{rot}\, \vec{H}(\vec{r},t) = \frac{\partial \vec{D}(\vec{r},t)}{\partial t} + \vec{J}(\vec{r},t), \tag{2.1.3b}$$

$$\mathrm{div}\, \vec{D}(\vec{r},t) = \rho(\vec{r},t), \tag{2.1.3c}$$

$$\mathrm{div}\, \vec{B}(\vec{r},t) = 0. \tag{2.1.3d}$$

Die differenzielle Form der Maxwellschen Gleichungen ist nur vollständig, wenn zusätzlich die Stetigkeit der Vektorgrößen an Grenzflächen zwischen elektrisch oder magnetisch unterschiedlichen Medien eingehalten wird. Die entsprechenden Bedingungen lauten mit dem zur Grenzfläche normalen Vektor $\vec{n}_{12}$, der Oberflächenladung

σ_F und dem Oberflächenstrom $\vec{J}_F$

$$\vec{n}_{12} \times (\vec{E}_2 - \vec{E}_1) = 0, \qquad \vec{n}_{12} \times (\vec{H}_2 - \vec{H}_1) = \vec{J}_F, \tag{2.1.4a}$$

$$\vec{n}_{12} \cdot (\vec{D}_2 - \vec{D}_1) = \sigma_F, \qquad \vec{n}_{12} \cdot (\vec{B}_2 - \vec{B}_1) = 0. \tag{2.1.4b}$$

2.1.1 Materialeigenschaften

Eine Verkopplung der Maxwellschen Gleichungen und damit die Möglichkeit sie zu lösen ist erst gegeben, wenn zusätzlich zu den Gln. 2.1.1 oder 2.1.3 die Beziehungen zwischen den Feldstärken und Flussdichten bekannt sind. Diese sind materialabhängig und für den allgemeinen inhomogenen und anisotropen Fall durch die folgenden Materialgleichungen beschrieben:

$$\vec{D}(\vec{r},t) = \vec{D}(\vec{E}(\vec{r},t)) = \varepsilon_0 \vec{E}(\vec{r},t) + \vec{P}(\vec{E},\vec{r},t), \tag{2.1.5a}$$

$$\vec{B}(\vec{r},t) = \vec{B}(\vec{H}(\vec{r},t)) = \mu_0 \vec{H}(\vec{r},t) + \mu_0 \vec{M}(\vec{H},\vec{r},t), \tag{2.1.5b}$$

$$\vec{J}_l(\vec{r},t) = \vec{J}_l(\vec{E}(\vec{r},t)) = \kappa \vec{E}(\vec{r},t). \tag{2.1.5c}$$

Eine Flussgröße setzt sich demnach aus einem linearen Anteil des Vakuums zur Feldstärke sowie einem im Allgemeinen komplexen Einfluss des Materials zusammen, der durch die Polarisation $\vec{P}$ und die Magnetisierung $\vec{M}$ beschrieben wird. Hierbei wird zwischen dispersiven, anisotropen, nichtlinearen und frequenzabhängigen Effekten unterschieden. Für den linearen hysteresefreien Fall lassen sich die Materialeinflüsse mittels elektrischer und magnetischer Suszeptibilitätstensoren $\overleftrightarrow{\chi}_e$ und $\overleftrightarrow{\chi}_m$ angeben:

$$\vec{P}(\vec{r},t) \;=\; \varepsilon_0 \,\overleftrightarrow{\chi}_e(\vec{r},t) * \vec{E}(\vec{r},t) + \vec{P}_r(\vec{r})\,, \tag{2.1.6a}$$

$$\vec{M}(\vec{r},t) \;=\; \overleftrightarrow{\chi}_m(\vec{r},t) * \vec{H}(\vec{r},t) + \vec{M}_r(\vec{r})\,. \tag{2.1.6b}$$

Die Ausdrücke $\vec{P}_r$ und $\vec{M}_r$ beschreiben eine permanente Polarisation bzw. Magnetisierung, wie sie beispielsweise in Elektreten oder Permanentmagneten auftreten, sie sollen im Folgenden als Null angenommen werden. Die Faltungsausdrücke[2] tragen der Zeitabhängigkeit der Materialparameter Rechnung und lassen sich durch einen Übergang in den Frequenzbereich[3] in multiplikative, komplexwertige Ausdrücke[4] umwandeln

$$\vec{D}(\vec{r},t) = \varepsilon_0 \,\overleftrightarrow{\varepsilon}_r(\vec{r},\omega) \cdot \vec{E}(\vec{r},\omega)\,, \tag{2.1.7a}$$

$$\vec{B}(\vec{r},t) = \mu_0 \,\overleftrightarrow{\mu}_r(\vec{r},\omega) \cdot \vec{H}(\vec{r},\omega). \tag{2.1.7b}$$

Für die im Allgemeinen tensoriellen und komplexen relativen Permittivitäten und Permeabilitäten gilt $\overleftrightarrow{\varepsilon}_r = \overleftrightarrow{\chi}_e + 1$ und $\overleftrightarrow{\mu}_r = \overleftrightarrow{\chi}_m + 1$. Das frequenzabhängige Verhalten

[2] Der Faltungsoperator $*$ repräsentiert die Vorschrift $f(t) * g(t) = \int_{-\infty}^{\infty} f(\tau) g(t-\tau) d\tau$.

[3] Als *Frequenzbereich* wird das Ergebnis der Fouriertransformation $H(\omega) = \int_{-\infty}^{\infty} h(t)\, e^{-j\omega t} dt$ betrachtet.

[4] Da sich weite Teile dieser Arbeit mit Frequenzbereichsbetrachtungen befassen, wird auf eine spezielle Markierung komplexer Größen verzichtet. Sind Verwechslungen nicht ausgeschlossen, wird der komplexe Term mit einem hochgestellten $^{(c)}$ markiert.

wird durch Wechselwirkungen auf atomarer oder molekularer Ebene unter Einfluss eines eingeprägten Feldes hervorgerufen. Auf Basis der verschiedenen Mechanismen lassen sich wiederum makroskopische Beschreibungen finden.

Für den Fall von Orientierungspolarisation, der auftritt, wenn die zugrundeliegenden Moleküle eine Dipolstruktur aufweisen, ergibt sich mit der materialabhängigen Relaxationszeit τ die relative Permittivität zu

$$\varepsilon_r(\omega) \;=\; \varepsilon_\infty + \frac{(\varepsilon_s - \varepsilon_\infty)}{1 + j\omega\tau}\,. \tag{2.1.8}$$

wobei ε_s und ε_∞ die Permittivitäten im statischen Fall und für den Grenzwert unendlich hoher Frequenzen beschreiben. Dieser Polarisationseffekt wird häufig auch als Debye-Dispersion bezeichnet.

Für unpolarisierte Medien wirkt der Mechanismus der elektronischen Polarisation, der mit der ebenfalls materialabhängigen Dämpfung δ und der Resonanzfrequenz ω_0 durch die so genannte Lorentz-Dispersion beschrieben wird

$$\varepsilon_r(\omega) \;=\; \varepsilon_\infty + \frac{(\varepsilon_s - \varepsilon_\infty)\,\omega_0}{\omega_0^2 + 2\,\delta\,j\omega - \omega^2}\,. \tag{2.1.9}$$

Für Plasmen ergibt sich mit der Kollisionsfrequenz ν_c und der Plasmafrequenz ω_p die Drude-Dispersion zu

$$\varepsilon_r(\omega) \;=\; \varepsilon_\infty + \frac{\omega_p^2}{j\omega\nu_c - \omega^2}\,. \tag{2.1.10}$$

Eine ausführliche Darstellung dispersiver Materialien findet sich in [2, 11, 12]. Neben der eigentlichen Materialdispersion werden die genannten Modelle auch häufig eingesetzt, um andere physikalische Phänomene zu beschreiben. Beispiele hierfür sind die so genannten Metamaterialien [15], mikroskopische Resonatorstrukturen, die makroskopisch betrachtet in gewissen Frequenzbereichen negative Werte für die Permittivität und die Permeabilität annehmen und sich ebenfalls durch Lorentz-Dispersion beschreiben lassen. Eine Näherung für Substratmaterialien, die einen frequenzunabhängigen Verlustwinkel aufweisen, lässt sich durch ein angepasstes Debye-Modell finden.

In vielen Fällen kann jedoch einfach von frequenzunabhängigen und isotropen Materialien ausgegangen werden, die sich durch die einfachen Materialbeziehungen darstellen lassen, wobei ε_r und μ_r in Gl. 2.1.7 skalare zeitunabhängige Werte annehmen:

$$\vec{D} = \varepsilon_0\,\varepsilon_r\vec{E}\,, \tag{2.1.11a}$$

$$\vec{B} = \mu_0\,\mu_r\vec{H}. \tag{2.1.11b}$$

2.2 Diskretisierung der Maxwellschen Gleichungen

Mit den Maxwellgleichungen und den Materialbeziehungen ist eine Gesetzmäßigkeit gegeben, mit der unter Kenntnis aller aktuellen elektrischen und magnetischen

Größen sowie aller Quellen die entsprechenden zukünftigen Werte eindeutig bestimmt werden können. Dies gilt auch für einzelne Raumteile, sofern die Feldwerte am Rand im betrachteten Zeitraum bekannt sind. Die geschlossene Lösbarkeit der gegebenen Gleichungen ist allerdings auf einfache geometrische Anordnungen beschränkt.

Abhilfe hierbei schaffen numerische Verfahren, die es ermöglichen, die Feldverteilung bei komplexen Geometrien näherungsweise zu lösen, indem sie die Komplexität der kontinuierlichen Feldgrößen durch eine endliche Anzahl von Zustandsgrößen modellieren. Hierbei werden grundsätzlich unterschiedliche Ansätze verfolgt. Randelementmethoden (engl.: *Boundary Element Method, BEM*) [6] diskretisieren beispielsweise die Oberflächen leitender Strukturen innerhalb homogener Materialien und lösen numerisch eine äquivalente Formulierung der Maxwellgleichungen als Randintegralgleichung. Volumenbasierte Verfahren wie die *Methode der Finiten Elemente (FEM)* [7] hingegen schränken die Anzahl der Zustandsgrößen gleich in doppelter Hinsicht ein: Zunächst wird nur ein Teilgebiet des Gesamtraums als *Rechengebiet* definiert, dann werden in dessen Inneren durch ein Gitter Stützstellen für die einzelnen Feldwerte festgelegt.

Die im Rahmen dieser Arbeit verwendete und im Folgenden näher beschriebene *Methode der Finiten Integration* (engl. *Finite Integration Theory, FIT*) zählt ebenfalls zu den volumendiskretisierten Verfahren und basiert auf der direkten Anwendung der Maxwellgleichungen in ihrer integralen Form (Gl. 2.1.1) auf die durch Diskretisierung entstandenen Grundelemente im betrachteten Gebiet. Dies hat zur Folge, dass wichtige physikalische Eigenschaften der kontinuierlichen Feldlösungen direkt auf die numerischen Lösungen übertragen werden können. Das Verfahren wurde bereits 1977 erstmals von T. Weiland in [8], die heutige Notation mit Spannungen und Flüssen in [13] vorgestellt. Die folgende Darstellung orientiert sich an [9, 10, 12, 14].

2.2.1 Die Gitter-Maxwellgleichungen

Ausgangspunkt bei der numerischen Berechnung elektromagnetischer Felder nach der Methode der Finiten Integration ist eine geeignete lückenlose und überschneidungsfreie Zerlegung des Rechengebiets, so dass die Struktur im Inneren genügend gut approximiert wird[5]. Dies gelingt durch die Definition eines dreidimensionalen Gitters G, wie in Abb. 2.1 für den kartesischen Fall dargestellt. Die Wahl des Gittertyps lässt zahlreiche Varianten zu, beispielsweise allgemeine nichtorthogonale Gitter [14] oder Tetraedergitter (für eine Klassifizierung siehe [17]), die vorliegende Arbeit beschränkt sich jedoch ausschließlich auf einfach strukturierte Gitter parallel zu kartesischen oder kreiszylindrischen Koordinatensystemen.

Zur einfacheren Strukturierung werden die einzelnen Zellen entlang der Koordinatenrichtungen durchgezählt. Zu jedem Punkt P_n können damit eindeutig ein Zellvolumen V, drei Zellflächen A_{nu}, A_{nv} und A_{nw} sowie drei Kantenlängen L_{nu}, L_{nv}

[5] Die Dichte des Gitters bei dynamischen Feldsimulationen hängt neben der Größe von Strukturdetails auch von der Frequenz des Feldes ab, das innerhalb des Rechengebiets betrachtet werden soll.

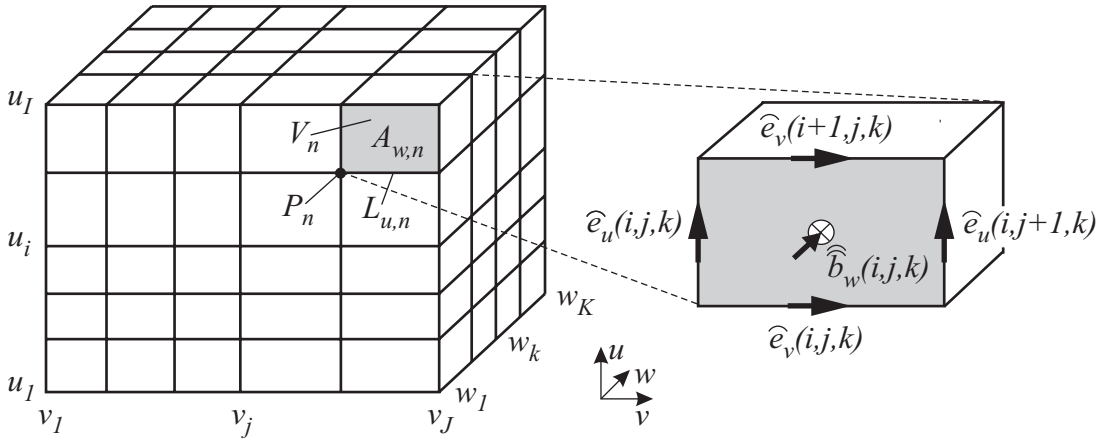


Abbildung 2.1: *Darstellung eines kartesischen Gitters zur Diskretisierung einer Struktur sowie die Indizierung der Primärfiguren. Für eine Fläche wird die Allokation der elektrischen Spannungen und des magnetischen Flusses hervorgehoben.*

und L_{nw} zugeordnet werden[6], wobei die jeweils in positive Koordinatenrichtung liegende Elementarfigur dem Punkt zugerechnet wird (siehe ebenfalls Abb. 2.1). Bei den vektoriellen Größen A und L, wird auch die Richtung des Vektors in positive Koordinatenrichtung angenommen. Mit der Anzahl von Punkten I, J und K in den drei Raumrichtungen und den zugehörigen Indizes i, j und k ergibt sich für die $N_P = I\,J\,K$ Gitterpunkte eine die Nummerierung in der Form

$$n(i,j,k) = i + (j-1)I + (k-1)IJ. \tag{2.2.1}$$

Die Diskretisierung der Maxwellgleichungen soll nun für eine beliebige Zellfläche (siehe auch Abb. 2.1, rechts) veranschaulicht werden. Mit der Definition der elektrischen Spannung $\widehat{e}_p$ mit $p \in \{u, v, w\}$ als Integral des elektrischen Feldes $\vec{E}$ über die Kantenlänge L_p und des magnetischen Flusses $\widehat{\widehat{b}}_p$ als Flächenintegral der magnetischen Flussdichte $\vec{B}$ über die Fläche A_p

$$\widehat{e}_p(i,j,k) = \int_{L_p(i,j,k)} \vec{E} \cdot d\vec{s}, \qquad \widehat{\widehat{b}}_p(i,j,k) = \int_{A_p(i,j,k)} \vec{B} \cdot d\vec{A}, \tag{2.2.2}$$

ergibt Gl. 2.1.1a für die Fläche $A_w(i,j,k)$

$$\widehat{e}_u(i,j,k) + \widehat{e}_v(i+1,j,k) - \widehat{e}_u(i,j+1,k) - \widehat{e}_v(i,j,k) = -\frac{d}{dt}\,\widehat{\widehat{b}}_w(i,j,k). \tag{2.2.3}$$

Durch die Verwendung der integralen Größen ist diese Darstellung exakt, also frei von Näherungen. Werden alle elektrischen Spannungen in einem Vektor $\widehat{\mathbf{e}}$ sowie alle magnetischen Flüsse im Vektor $\widehat{\widehat{\mathbf{b}}}$ zusammengefasst, lässt sich Gl. 2.2.3 für alle Flächen des Rechengebiets in einem Gleichungssystem darstellen:

$$\mathbf{C}\,\widehat{\mathbf{e}} = -\frac{d}{dt}\,\widehat{\widehat{\mathbf{b}}}. \tag{2.2.4}$$

[6]Punkten am Rand des Rechengebiets werden dabei nicht existierende Volumen, Flächen und Längen zugeordnet, die meist zur Beibehaltung der Indizierung jedoch nicht eliminiert werden.

Die Matrix $\mathbf{C}$ hat dieselbe Bedeutung wie der Rotationsoperator in Gl. 2.1.3a (engl. *curl-operator*). Sie enthält hierbei, wie in Gl. 2.2.3 zu erkennen, pro Zeile zwei Einträge $+1$ sowie zwei -1. Sie hat folglich rein topologischen Charakter und ist sehr dünn besetzt, darüberhinaus ist die Matrix singulär. Werden die Vektoren $\widehat{\mathbf{e}}$ und $\widehat{\widehat{\mathbf{b}}}$ nach Gl. 2.2.1 zunächst in u-, dann in v- und zuletzt in w-Richtung sortiert, hat $\mathbf{C}$ eine Bandstruktur. Definiert man die Hilfsmatrizen[7]

$$[\mathbf{P}_{u,v,w}]_{p,q} = \begin{cases} -1 & : \quad p = q \\ 1 & : \quad p = q + r \\ 0 & : \quad \text{sonst} \end{cases} \tag{2.2.5}$$

mit den Fällen $\mathbf{P}_u : \ r = 1$, $\mathbf{P}_v : \ r = I$ und $\mathbf{P}_w : \ r = IJ$, lässt sich $\mathbf{C}$ wie folgt angeben[8]:

$$\mathbf{C} = \begin{pmatrix} \mathbf{0} & -\mathbf{P}_w & \mathbf{P}_v \\ \mathbf{P}_w & \mathbf{0} & -\mathbf{P}_u \\ -\mathbf{P}_v & \mathbf{P}_u & \mathbf{0} \end{pmatrix}. \tag{2.2.6}$$

Auf entsprechende Weise wie in Gl. 2.2.3 kann auch die Nichtexistenz magnetischer Ladungen nach Gl. 2.1.1d diskretisiert werden. Betrachtet man das Oberflächenintegral über eine einzelne Zelle, ergibt sich

$$\begin{aligned} -\widehat{\widehat{b}}_u(i,j,k) + \widehat{\widehat{b}}_u(i+1,j,k) - \widehat{\widehat{b}}_v(i,j,k) + \widehat{\widehat{b}}_v(i,j+1,k) \\ - \widehat{\widehat{b}}_w(i,j,k) + \widehat{\widehat{b}}_w(i,j,k+1) = 0. \end{aligned} \tag{2.2.7}$$

Bezogen auf das ganze Gitter resultiert daraus

$$\mathbf{S}\widehat{\widehat{\mathbf{b}}} = \mathbf{0} \quad \text{mit} \quad \mathbf{S} = (\mathbf{P}_u \, \mathbf{P}_v \, \mathbf{P}_w). \tag{2.2.8}$$

Die Matrix $\mathbf{S}$ ist ebenfalls rein topologischer Natur und nur sehr dünn besetzt. Sie entspricht dem Divergenzoperator (engl. *source-operator*) in Gl. 2.1.3d.

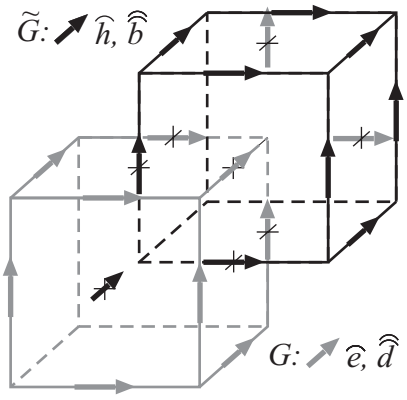


Abbildung 2.2: *Darstellung des dualen Gitters $\widetilde{G}$ (schwarz) relativ zum primären Gitter G (grau) sowie die Allokation der integralen Größen auf beiden Gittern.*

[7] Diese Matrizen können als diskretes Analogon zu den partiellen Differentiationsoperatoren interpretiert werden.

[8] Formell müssen die Hilfsmatrizen an den Rändern angepasst werden. Es zeigt sich aber, dass obige Definition in Verbindung mit den Randbedingungen, die später beschrieben werden, ebenfalls auf korrekte Lösungen führt.

Die Diskretisierung der verbleibenden Maxwellgleichungen erfolgt nach derselben Grundidee wie oben beschrieben, die Allokation aller Größen erfolgt jedoch auf einem zweiten, dem so genannten *dualen* Gitter[9] $\widetilde{G}$ [16]. Hierbei durchstößt jede Kante des dualen Gitters eine Fläche des primären Gitters (und umgekehrt) senkrecht und jede Zelle des einen Gitters nimmt genau einen Knoten des anderen auf. Alle auf das duale Gitter bezogenen Größen und Operatoren werden im Folgenden mit einer Tilde gekennzeichnet. In Analogie zu elektrischen Spannungen werden magnetische Spannungen $\widehat{h}_x$ als exakte Integrale über die duale Kante $\widetilde{L}_x$ definiert, ebenso die elektrischen Flüsse $\widehat{\widetilde{d}}_x$ und Ströme $\widehat{\widetilde{j}}_x$ über die duale Fläche $\widetilde{A}_x$. Elektrische Ladungen q werden als Volumenintegral der Raumladungsdichte über eine duale Gitterzelle bestimmt.

Werden nun die Umlaufintegrale aus den Gln. 2.1.1b und 2.1.1c auf dem dualen Gitter ausgewertet, die lokalen Komponenten erneut in Vektoren $\widehat{\mathbf{h}}$, $\widehat{\widetilde{\mathbf{d}}}$, $\widehat{\widetilde{\mathbf{j}}}$ und $\mathbf{q}$ zusammengefasst und die entsprechenden Operatoren $\widetilde{\mathbf{C}}$ und $\widetilde{\mathbf{S}}$ definiert, lassen sich *Gitter-Maxwellgleichungen* als diskrete Formulierung der kontinuierlichen Maxwellgleichungen für das Rechengebiet angeben:

$$\mathbf{C}\,\widehat{\mathbf{e}} = -\frac{d}{dt}\,\widehat{\widetilde{\mathbf{b}}}, \tag{2.2.9a}$$

$$\widetilde{\mathbf{C}}\,\widehat{\mathbf{h}} = \frac{d}{dt}\,\widehat{\widetilde{\mathbf{d}}} + \widehat{\widetilde{\mathbf{j}}}, \tag{2.2.9b}$$

$$\widetilde{\mathbf{S}}\,\widehat{\widetilde{\mathbf{d}}} = \mathbf{q}, \tag{2.2.9c}$$

$$\mathbf{S}\,\widehat{\widetilde{\mathbf{b}}} = \mathbf{0}. \tag{2.2.9d}$$

Eine Eigenschaft dieser Diskretisierung, der später bei der Betrachtung von Stabilität und Passivität eine wesentliche Bedeutung zukommt, ist die Dualitätsbeziehung

$$\mathbf{C}^T = \widetilde{\mathbf{C}} \tag{2.2.10}$$

zwischen den numerischen Rotationsoperatoren auf dem primären und dualen Gitter. Auch für die Maxwellgleichungen grundlegende vektoranalytische Beziehungen gelten in vollem Umfang ebenfalls für das diskrete Analogon [9, 10]:

$$\mathbf{S}\mathbf{C} = \widetilde{\mathbf{S}}\widetilde{\mathbf{C}} = \mathbf{0} \quad \Leftrightarrow \quad \operatorname{div}\operatorname{rot} \equiv 0 \tag{2.2.11a}$$

$$\mathbf{C}\widetilde{\mathbf{S}}^T = \widetilde{\mathbf{C}}\mathbf{S}^T = \mathbf{0} \quad \Leftrightarrow \quad \operatorname{rot}\operatorname{grad} \equiv 0. \tag{2.2.11b}$$

Wirbelfelder sind folglich stets divergenzfrei und Gradientenfelder[10] stets wirbelfrei. Diese Beziehungen gewährleisten somit zugleich, dass wichtige physikalische Eigenschaften wie Energie- und Ladungserhaltung, Gültigkeit der Kontinuitätsgleichung und des Poyntingschen Satzes sowie die Eindeutigkeit des Lösungsraums auch im diskreten Modell erhalten bleiben.

Die meisten Ansätze der klassischen Elektrodynamik wie beispielsweise Potentialansätze lassen sich daher direkt auf den diskretisierten Fall übertragen. Es sei auch

[9] In Verbindung mit dem Leapfrog-Zeitintegrationsverfahren ist FIT auf kartesischen Gittern für transiente Simulationen damit äquivalent zum *Finite Difference Time Domain (FDTD)*-Verfahren.

[10] Es kann gezeigt werden, dass die Gradientenoperatoren auf dem primären bzw. dualen Gitter durch $-\widetilde{\mathbf{S}}^T$ und $-\mathbf{S}^T$ repräsentiert werden.

erneut betont, dass durch die Diskretisierung lediglich der Gültigkeitsbereich der Maxwellgleichungen auf den Gitterraum eingeschränkt wurde, jedoch keinerlei Näherungen gemacht wurden.

2.2.2 Materialdiskretisierung

Die in jedem numerischen Verfahren unvermeidlichen Näherungen werden durch die Definition der *Gitter-Materialgleichungen* in das diskrete System eingebracht. Diese vervollständigen die FI-Methode, indem die Beziehungen zwischen Fluss- und Spannungsgrößen definiert und zugleich eine Kopplung zwischen primärem und dualem Gitter eingeführt wird. Entsprechend den Gln. 2.1.11 können die Materialbeziehungen für lineare Materialien ohne permanente Polarisation wie folgt angegeben werden, wobei $\widehat{\widetilde{\mathbf{j}}}_e$ eingeprägte Ströme beschreibt:

$$\widehat{\widetilde{\mathbf{d}}} = \mathbf{M}_\varepsilon \, \widehat{\mathbf{e}}, \tag{2.2.12a}$$

$$\widehat{\widetilde{\mathbf{j}}} = \mathbf{M}_\kappa \, \widehat{\mathbf{e}} + \widehat{\widetilde{\mathbf{j}}}_e, \tag{2.2.12b}$$

$$\widehat{\mathbf{h}} = \mathbf{M}_\mu^{-1} \, \widehat{\widetilde{\mathbf{b}}}. \tag{2.2.12c}$$

Für dual orthogonale Gittersysteme ist jede Spannungsgröße direkt einer Flussgröße zugeordnet, die Matrizen $\mathbf{M}_\varepsilon$, $\mathbf{M}_\kappa$ und $\mathbf{M}_\mu$ haben daher in der Standardformulierung Diagonalgestalt. Die Berechnung der Einträge erfolgt in zwei Schritten. Zum einen müssen Spannungs- und Flussgrößen in Feldstärken und Flussdichten umgewandelt werden, zum anderen müssen die im Allgemeinen ortsabhängigen und unstetigen Materialeigenschaften im Berechnungspunkt, dem Schnittpunkt von Kante und Fläche, geeignet gemittelt werden. Diese Schritte können nicht näherungsfrei durchgeführt werden, es muss aber zumindest sichergestellt sein, dass das Verfahren für beliebig feine Diskretisierung gegen die kontinuierliche Lösung konvergiert.

Zudem ist zu fordern, dass die tangentiale Stetigkeit des elektrischen Feldes und die normale des magnetischen Flusses an Materialgrenzen gewährleistet wird. Dies lässt sich am einfachsten erfüllen, indem die Zellvolumina des primären Gitters homogen gefüllt werden. Werden die Felder und Flussdichten als konstante Mittelwerte über die entsprechenden Längen und Flächen angenommen, folgt für die Materialbeziehungen mit einer Materialverteilung

$$(\mathbf{M}_\varepsilon)_{p,p} = \frac{\int_{\widetilde{A}_p} \varepsilon \, dA}{L_p}, \qquad (\mathbf{M}_\kappa)_{p,p} = \frac{\int_{\widetilde{A}_p} \kappa \, dA}{L_p}, \tag{2.2.13a}$$

$$(\mathbf{M}_\mu^{-1})_{p,p} = \frac{\int_{\widetilde{L}_p} \mu^{-1} \, ds}{A_p}. \tag{2.2.13b}$$

Haben die betrachteten Permittivitäten dispersiven Charakter, muss analog zum Vorgehen in Gl. 2.2.13 auch $(\varepsilon_s - \varepsilon_\infty)$ diskretisiert werden, was auf ein frequenzabhängiges $\mathbf{M}_\varepsilon(\omega)$ führt. Dieses lässt sich mit den Diagonalmatrizen $\mathbf{M}_{\varepsilon\infty}$ und $\mathbf{M}_{\varepsilon d}$ als

$$\mathbf{M}_\varepsilon(\omega) = \mathbf{M}_{\varepsilon\infty} + \frac{1}{\alpha_0 + j\omega\alpha_1 + (j\omega)^2} \mathbf{M}_{\varepsilon d} \tag{2.2.14}$$

angeben. Die Parameter α_0, α_1 und die Einträge der Diagonalmatrizen $\mathbf{M}_{\varepsilon\infty}$ und $\mathbf{M}_{\varepsilon d}$ können je nach Dispersionsmodell nach den Gln. 2.1.8-2.1.10 bestimmt werden.

Erweiterungen, die eine verbesserte Strukturapproximation ermöglichen, bilden Dreieckszellen, Tetraederfüllungen oder so genannte Adapterzellen [17]. Eine deutliche Verbesserung des Verfahrens bringen teilgefüllte Zellen, bei denen idealleitende Gebiete als spannungs- und flussfrei angenommen werden und folglich nicht in den Mittelungsprozess Gl. 2.2.13 einfließen [18].

Zusammenfassend sollen erneut die beiden entscheidenden Eigenschaften, die für spätere Betrachtungen und die Effizienz von Lösungsstrategien von Bedeutung sind, betont werden:

- Die Materialmatrizen $\mathbf{M}_\varepsilon$ und $\mathbf{M}_\mu$ sind für den betrachteten Gittertyp diagonal, sie sind folglich trivial zu invertieren.
- Alle Matrixeinträge sind positiv, $\mathbf{M}_\varepsilon$ bzw. $\mathbf{M}_{\varepsilon\infty}$ und $\mathbf{M}_\mu$ sind somit positiv definit, $\mathbf{M}_\kappa$ sowie $\mathbf{M}_{\varepsilon d}$ positiv semi-definit.

Materialbeziehungen höherer Ordnung werden in [19] beschrieben. Hierbei muss jedoch wie auch bei nichtorthogonalen Gittern [14] die Diagonalgestalt der Materialbeziehungen aufgegeben werden.

Um die Zuverlässigkeit einer numerischen Methode einschätzen zu können, ist es von elementarer Bedeutung, die durch Näherungen verursachten Fehler abschätzen zu können. Eine ausführliche Fehlerbetrachtung und Konvergenzuntersuchung findet sich beispielsweise in [12, 14].

2.2.3 Randbedingungen und Anregung des Rechengebiets

An der Berandung des Rechengebiets sind nur die Operatoren $\mathbf{C}$ und $\mathbf{S}$ des normalen Gitters definiert, während die entsprechenden Operatoren des dualen Gitters $\widetilde{\mathbf{C}}$ und $\widetilde{\mathbf{S}}$ auf Komponenten zugreifen, die außerhalb des Rechengebiets liegen. Die entsprechenden Umläufe und Oberflächenintegrale können folglich am Rand nicht geschlossen werden.

Analog zur kontinuierlichen Feldtheorie müssen daher Randbedingungen definiert werden, die das Verhalten der Felder am Rand beschreiben bzw. die Wirkung des Außenraums modellieren. Dabei wird zwischen *geschlossenen* Berandungen wie idealen elektrischen oder magnetischen Rändern, bei denen kein Energieaustausch mit dem Außenraum stattfindet, und *offenen* Randbedingungen unterschieden, wobei verlustbehafteten elektrischen Rändern eine Zwischenstellung zukommt. Zunächst werden die geschlossenen Ränder betrachtet.

2.2.3.1 Magnetische Randbedingungen

Bei magnetischen Berandungen wird formal eine unendliche magnetische Leitfähigkeit angenommen, wodurch die tangentialen magnetischen Spannungen sowie die

normalen elektrischen Flüsse verschwinden. Dieser Zustand stellt sich automatisch ein, wenn die fehlenden dualen Kanten in der Umlaufmatrix $\widetilde{\mathbf{C}}$ sowie die fehlenden dualen Flächen in der Quellenmatrix $\widetilde{\mathbf{S}}$ einfach unberücksichtigt bleiben.

Auch wenn zu dieser Randbedingung kein physikalisch sinnvolles Pendant existiert, kommt ihr in der Praxis dennoch große Bedeutung als Symmetriebedingung in Strukturen mit symmetrischem Feldverlauf zu.

2.2.3.2 Elektrische Randbedingungen

Ebenfalls idealisiert, stellt die elektrische Randbedingung eine unendliche elektrische Leitfähigkeit dar, wodurch analog alle tangentialen elektrischen Spannungen und alle normalen magnetischen Flüsse zu Null werden. Die entsprechenden primären Kanten und Flächen werden entbehrlich und die entsprechenden Zeilen und Spalten können daher in $\mathbf{C}$ und $\mathbf{S}$ eliminiert werden. Häufig soll jedoch die Bandstruktur dieser Matrizen aufrecht erhalten werden, weswegen alternativ die entsprechenden Einträge der Materialmatrizen $\mathbf{M}_\varepsilon$ und $\mathbf{M}_\mu$ zu Null gesetzt werden[11]. Die elektrische Randbedingung findet sowohl zur Simulation von Strukturen in leitfähigen Gehäusen als auch als Symmetrieebene Verwendung.

2.2.3.3 Impedanz-Randbedingungen

Nicht immer ist die Annahme unendlicher Leitfähigkeiten für Metalle zulässig. Beispielsweise bei der Güteberechnung hochresonanter Filter können die Verluste, die durch die tatsächliche Leitfähigkeit des Gehäuses entstehen, von entscheidender Bedeutung sein. Zwar ist es möglich, mit Hilfe eines Störansatzes [4] die Güte oder Dämpfung eines Resonators aus den mit idealer Berandung gewonnenen Ergebnissen nachträglich zu berechnen, in bestimmten Fällen ist es jedoch wünschenswert, die entstehenden Verluste direkt bei der Rechnung zu erfassen.

Soll Metall hingegen als klassisches Material modelliert werden, zeigt sich für die Wellenlänge λ einer ebenen Welle der Frequenz ω im Leiter mit der Leitfähigkeit $\kappa \gg \omega\,\varepsilon$

$$\varepsilon = \varepsilon + \frac{\kappa}{j\omega} \approx -j\frac{\kappa}{\omega}, \tag{2.2.15a}$$

$$k = \omega\sqrt{\mu\varepsilon} \approx \sqrt{\frac{\omega\mu\kappa}{2}}\,(1-j) = \frac{1-j}{\delta}, \tag{2.2.15b}$$

$$\lambda = \frac{2\pi}{\mathrm{Re}(k)} \approx 2\pi\delta, \tag{2.2.15c}$$

wobei $\delta = \sqrt{\frac{2}{\omega\mu\kappa}}$ als äquivalente Leitschichtdicke bezeichnet wird. Damit verkürzt sich beispielsweise die Wellenlänge einer Welle mit der Frequenz 1 GHz von etwa

[11] Hierbei ist zu beachten, dass die Matrizen in diesem Fall nicht mehr invertierbar sind. Bei Bedarf muss eine Pseudoinverse gebildet werden, bei der nur Einträge ungleich Null invertiert werden.

30 cm im Vakuum auf etwa 0.6 μm in Kupfer. Es wird offensichtlich, dass diese Wellenlänge nicht ohne großen Aufwand vom Gitter abgetastet werden kann.

Zugleich wird deutlich, dass die Welle im Metall stark gedämpft wird. Im Allgemeinen wird $\delta \ll \Delta z$ im Leiter gelten, womit die Amplitude $\sim e^{-z/\delta}$ innerhalb einer halben Gitterschrittweite $\Delta z/2$ auf nahezu Null abgeklungen sein wird. Wird zusätzlich der schräge Einfall einer ebenen Welle auf den gut leitfähigen Halbraum betrachtet, zeigt sich nach dem Brechungsgesetz für den Ausfallswinkel φ_a in Abhängigkeit des Einfallswinkels φ_e (jeweils zur Flächennormalen)

$$\sin(\varphi_a) = \sqrt{\frac{1}{1+\frac{\kappa}{j\omega\varepsilon}}}\sin(\varphi_e) \approx 0 \quad \text{für} \quad \kappa \gg \omega\varepsilon. \tag{2.2.16}$$

Die Welle breitet sich demnach unabhängig vom Einfallswinkel im Metall nahezu senkrecht zur Leiteroberfläche aus. Damit kann das Verhalten des Feldes im Metall durch ein Oberflächenimpedanzmodell [20] unter den folgenden Annahmen beschrieben werden:

- Die normalen Komponenten der magnetischen Flussdichte sind im Metall und aufgrund der Stetigkeit auch im angrenzenden Material Null.
- Im Gegensatz zur ideal elektrischen Randbedingung werden elektrische Komponenten in der Grenzfläche zugelassen.
- Die aufgrund der gedämpften Welle im Metall auftretende Stromdichte kann unter Verwendung von Gl. 2.2.15b durch eine äquivalente Oberflächenstromdichte j_F in der Grenzschicht modelliert werden. Es wird eine senkrecht (in w-Richtung) in das Metall einfallende ebene Welle (siehe auch Abb. 2.3) angenommen:

$$j_F = \frac{\widehat{\widetilde{j}}}{\widetilde{\Delta u}} = \kappa\frac{\widehat{e}_v}{\Delta v}\int_0^\infty e^{-w\,j\,(1-j)/\delta}\,dw = \sqrt{\frac{\kappa}{2\omega\mu}}\,(1-j)\frac{\widehat{e}_v}{\Delta v} = \frac{1}{\sqrt{j\omega}}\sqrt{\frac{\kappa}{\mu}}\,\frac{\widehat{e}_v}{\Delta v}. \tag{2.2.17}$$

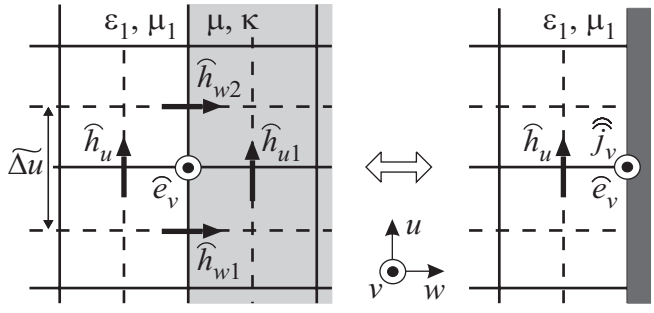


Abbildung 2.3: *Feldkomponenten am Übergang zwischen nicht- und gut leitendem Gebiet (links) und äquivalente Anordnung mit Oberflächenstrom (rechts).*

Damit ergibt sich für den in Abb. 2.3 dargestellten Umlauf unter der vereinfachten Annahme homogener ε- und κ-Verteilung

$$-\widehat{h}_u = j\omega\,\underbrace{\frac{\varepsilon\,\widetilde{A}_v}{2\Delta v}}_{M_\varepsilon}\,\widehat{e}_v + \frac{1}{\sqrt{j\omega}}\,\underbrace{\sqrt{\frac{\kappa}{\mu}}\,\frac{\widetilde{\Delta u}}{\Delta v}}_{K}\,\widehat{e}_v\,, \tag{2.2.18}$$

oder unter Berücksichtigung der Materialverteilung in allgemeiner Matrixschreibweise:

$$\widetilde{\mathbf{C}}\,\widehat{\mathbf{h}} = j\omega\,\mathbf{M}_\varepsilon\;\widehat{\mathbf{e}} + \frac{1}{\sqrt{j\omega}}\,\mathbf{K}_e\;\widehat{\mathbf{e}}\,. \tag{2.2.19}$$

Die Einträge in $\mathbf{M}_\varepsilon$ berücksichtigen hierbei in den Randzellen zur Integration nur den nichtleitenden Flächenanteil, die Einträge der Matrix entsprechen demnach denen, die sich bei ideal magnetischen Randbedingungen ergeben würden. Bei der Matrix $\mathbf{K}_e$ handelt es sich um eine Diagonalmatrix, die nur an den Stellen Einträge hat, deren zugehörige Komponente $\widehat{e}$ transversal in einer Impedanzwand liegt. Die Diskretisierung des Induktionsgesetzes (Gl. 2.2.9a) bleibt unverändert im Vergleich zu ideal elektrischen Randbedingungen.

Vergleicht man in Gl. 2.2.18 die Beiträge der beiden Summanden zu $\widehat{h}_u$, zeigt sich, dass der Anteil des Oberflächenstromes wesentlich größer als der Beitrag des Verschiebungsstroms ist, für typische Beispiele im Hochfrequenzbereich ergibt sich ein Verhältnis von über 10^8. Dieser Unterschied führt zu einer schlechten Konditionierung von Gl. 2.2.19, die die Erzeugung von Modellen reduzierter Ordnung erheblich erschwert.

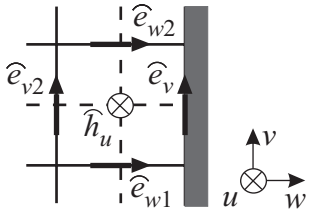


Abbildung 2.4: *Dieselbe Konstellation wie Abb. 2.3 rechts, jedoch mit um 90° gedrehtem Koordinatensystem. Betrachtung eines normalen Umlaufs vor der Impedanzwand.*

Es erweist sich daher als vorteilhaft, anstelle einer hohen elektrischen Leitfähigkeit im Metallrand eine äquivalente kleine magnetische Leitfähigkeit in der ersten Zellschicht innerhalb des Rechengebiets in Verbindung mit einer ideal elektrischen Randbedingung anzunehmen. Vernachlässigt man den Anteil des Verschiebungsstroms in Gl. 2.2.18 und betrachtet nach Abb. 2.4 den Umlauf

$$-j\omega\,\widehat{\overline{b}}_u = -\,\widehat{e}_v + \,\widehat{e}_{v2} - \,\widehat{e}_{w1} + \,\widehat{e}_{w2}, \tag{2.2.20a}$$

$$-j\omega\frac{\mu\,A_u}{\widetilde{\Delta u}}\;\widehat{h}_u = \sqrt{j\omega}\,\sqrt{\frac{\mu}{\kappa}}\,\frac{\Delta v}{\widetilde{\Delta u}}\;\widehat{h}_u + \,\widehat{e}_{v2} - \,\widehat{e}_{w1} + \,\widehat{e}_{w2}, \tag{2.2.20b}$$

$$-j\omega\,\underbrace{\left(\frac{\mu\,A_u}{\widetilde{\Delta u}} + \sqrt{\frac{\mu}{j\omega\,\kappa}}\,\frac{\Delta v}{\widetilde{\Delta u}}\right)}_{M_\mu}\;\widehat{h}_u = \,\widehat{e}_{v2} - \,\widehat{e}_{w1} + \,\widehat{e}_{w2}, \tag{2.2.20c}$$

lässt sich ein äquivalentes komplexes M_μ angeben:

$$M_\mu = \frac{\mu\,A_u}{\widetilde{\Delta u}}\left(1 + \frac{1}{\sqrt{j\omega}}\,\underbrace{\frac{1}{\Delta w}\,\frac{1}{\sqrt{\mu\kappa}}}_{K_m}\right). \tag{2.2.21}$$

Die rechte Seite in Gl. 2.2.20c entspricht hierbei tatsächlich einem Umlauf unter Annahme eines idealen elektrischen Randes, der Imaginärteil von M_μ entspricht

einer magnetischen Leitfähigkeit. Neben M_μ wird in der Praxis häufig auch der inverse Term M_μ^{-1} benötigt. Da der rechte Summand in Gl. 2.2.21 wesentlich kleiner als Eins ist, lässt dieser sich näherungsweise bestimmen:

$$M_\mu^{-1} \approx \frac{\widetilde{\Delta u}}{\mu\, A_u} \left(1 - \frac{1}{\sqrt{j\omega}} \frac{1}{\Delta w} \frac{1}{\sqrt{\mu\kappa}} \right). \tag{2.2.22}$$

Bei Betrachtung des gesamten Rechengebiets ergibt sich eine Diagonalmatrix $\mathbf{K}_m$. Mit dieser lassen sich mit den Gln. 2.2.21 und 2.2.22 komplexe frequenzabhängige Diagonalmatrizen

$$\mathbf{M}_\mu^{(c)}(\omega) = \mathbf{M}_\mu(\mathbf{I} + \frac{1}{\sqrt{j\omega}}\, \mathbf{K}_m) \quad \text{und} \tag{2.2.23a}$$

$$\mathbf{M}_\mu^{-1\,(c)}(\omega) = \mathbf{M}_\mu^{-1}(\mathbf{I} - \frac{1}{\sqrt{j\omega}}\, \mathbf{K}_m) \tag{2.2.23b}$$

angeben.

2.2.3.4 Offene Randbedingungen

Die im vorigen Abschnitt beschriebenen elektrischen bzw. Impedanz-Randbedingungen setzen voraus, dass sich die betrachtete Struktur stets in einem ideal oder endlich leitfähigen Gehäuse befindet. Während diese Annahme für Filter oder Schaltungen, die aus Gründen der elektromagnetischen Verträglichkeit gekapselt sind, erfüllt ist, ist es häufig wünschenswert, auch freie Abstrahlung in den als unendlich angenommenen Freiraum zu simulieren. Typische Beispiele hierfür sind u.a. Antennen, aber auch Verbindungsstrukturen innerhalb integrierter Schaltungen (engl. *Interconnects*), die eine von der Simulation zu berücksichtigende parasitäre Abstrahlung aufweisen. Aufgrund der finiten Größe des Rechengebiets ist bei solchen Fällen in Kombination mit idealen Randbedingungen mit unphysikalischen stehenden Wellen zu rechnen.

Die Modellierung des Freiraums im Zusammenhang mit FIT ist auf drei grundsätzlich unterschiedliche Arten möglich.

- Das FIT-Rechengebiet kann in Form einer Hybridmethode mit einer Randelementmethode gekoppelt werden.
- Durch die Lösung eines 2D-Eigenwertproblems auf einer Oberfläche des Rechengebiets werden die Feldbilder eines oder mehrerer Moden berechnet. Werden diese Modenbilder am Rand eingeprägt, kann die unendliche homogene Fortsetzung der 2D-Struktur als Wellenleiter simuliert werden.
- Das Rechengebiet kann mit einem absorbierenden Material umgeben werden, in das jede Welle reflexionsfrei eindringt und in dem sie anschließend vollständig gedämpft wird. Das absorbierende Material kann daher wieder mit einer idealen Randbedingung abgeschlossen werden. Dieser Ansatz soll im Weiteren genauer betrachtet werden.

Erste offene Randbedingungen (engl. *Absorbing Boundary Condition, ABC*), basierend auf absorbierenden Halbräumen, wurden als *Mur* Randbedingung [21] bekannt. Ein Durchbruch auf diesem Gebiet wurde durch die so genannte *Perfectly Matched Layer* Randbedingung (PML) erzielt. Hierbei wird ein unphysikalisches Medium eingeführt, in das Wellen theoretisch unabhängig von ihrer Frequenz, Polarisation und ihrem Einfallswinkel refexionsfrei eindringen können. In der ursprünglichen Formulierung von Berenger [22] müssen dazu die Feldkomponenten innerhalb des PML-Materials aufgespalten werden, um die Zahl der Freiheitsgrade zu erhöhen. Neuere Formulierungen lassen die Interpretation des PML-Mediums als anisotropes verlustbehaftetes Material zu. Seine Materialeigenschaften $\overleftrightarrow{\varepsilon}$ und $\overleftrightarrow{\mu}$ sind wie folgt definiert:

$$\overleftrightarrow{\varepsilon} = \varepsilon_0 \, \Lambda \quad \text{und} \quad \overleftrightarrow{\mu} = \mu_0 \, \Lambda \quad \text{mit} \quad \Lambda = \begin{pmatrix} c_u & 0 & 0 \\ 0 & c_v & 0 \\ 0 & 0 & c_w \end{pmatrix}. \tag{2.2.24}$$

Es lässt sich zeigen [23], dass in eine Schicht, die in w-Richtung an das Rechengebiet angefügt wurde, bei einer Wahl von

$$c_u = c_v = \frac{1}{c_w}. \tag{2.2.25}$$

eine ebene Welle unabhängig von der Polarisation und vom Einfallswinkel reflexionsfrei in das Medium eindringen kann.

Bei der Wahl von $c_{u,v}$ muss nun beachtet werden, dass $\mathrm{Re}(c_{u,v}) \geq 1$ gewählt werden sollte, um die Dämpfung evaneszenter Wellenanteile beizubehalten, und $\mathrm{Im}(c_{u,v})$ einen negativen Wert annehmen muss, um die ausbreitungsfähigen Wellen im PML-Material ebenfalls zu dämpfen. Für den einfachsten Fall wird in [23] daher

$$c_{u,v} = \alpha - j \, \beta \tag{2.2.26}$$

mit Konstanten α und β vorgeschlagen. Hierbei ist jedoch zu beachten, dass es sich um ein rein mathematisches Modell handelt, das einem komplexen frequenzunabhängigen $\overleftrightarrow{\varepsilon}$ und $\overleftrightarrow{\mu}$ entspricht. Für die Simulation im Frequenzbereich (d.h. der Betrachtung eines stationär eingeprägten Anregungssignals bestehend aus einer harmonischen Schwingung mit fester Frequenz) kann dieses Material effizient eingesetzt werden, bei einer Simulation im Zeitbereich (mit beliebigen transienten Verläufen des Anregungssignals) treten jedoch unphysikalische Ergebnisse auf[12]. Daher wird in der Literatur häufig ein frequenzabhängiger Zusammenhang bevorzugt [24]

$$c_{u,v}(\omega) = 1 + \frac{p}{j\omega}. \tag{2.2.27}$$

Das resultierende, perfekt angepasste und absorbierende Material entspricht damit in transversaler Richtung einem elektrisch wie magnetisch leitfähigem Medium

[12] Voraussetzung für einen reellen Zusammenhang im Zeitbereich ist, dass $\mathrm{Re}(c)$ eine gerade und $\mathrm{Im}(c)$ eine ungerade Funktion von ω darstellt. Wird mit der Signumfunktion $c(\omega) = \alpha - j \, \mathrm{sgn}(\omega) \, \beta$ gewählt, ergibt sich im Frequenzbereich für positive ω-Werte dasselbe Ergebnis wie mit Gl. 2.2.26 bei gleichzeitig reellen Werten für $\overleftrightarrow{\varepsilon}(t)$ und $\overleftrightarrow{\mu}(t)$. Für eine Zeitbereichsrechnung eignet sich die Formulierung dennoch nicht, da $\overleftrightarrow{\varepsilon}(t)$ und $\overleftrightarrow{\mu}(t)$ in diesem Fall nichtkausale Medien beschreiben.

(abhängig von p) mit der Permittivität und Permeabilität des Vakuums. Es lässt sich zeigen, dass sowohl c als auch $\frac{1}{c}$ die Kramers-Kronig-Bedingung zwischen Real- und Imaginärteil erfüllen [25], das Medium damit kausal ist und somit sowohl im Zeit- als auch im Frequenzbereich physikalische Resultate liefert[13].

PML-Realisierung in FIT

Obwohl theoretisch jede Welle in das im vorigen Abschnitt hergeleitete PML-Medium reflexionsfrei eindringen kann, ist dies im diskreten Modell dennoch nur bedingt erfüllt, da an Grenzflächen zwischen großen Materialvariationen trotz Wellenanpassung aufgrund der Gitterdispersion[14] [20, 13] Reflexionen auftreten. Um diesen Effekt gering zu halten, werden anstelle eines homogenen Absorbermaterials mehrere, typischerweise vier bis acht, angepasste Absorberschichten an das Rechengebiet angefügt, wobei die Leitfähigkeiten von Schicht zu Schicht nach einem vorgegebenen Dämpfungsprofil vergrößert werden. Als Profil wird hierbei neben geometrischen Funktionen meist eine Potenzfunktion [22]

$$\sigma(w) = \sigma_{max} \left(\frac{w}{\Delta w}\right)^q \qquad \text{mit} \qquad \sigma_{max} = -\frac{\varepsilon_0 c}{2\,\Delta w}\,\frac{q+1}{N_{lay}}\,\ln(R) \tag{2.2.28}$$

mit dem Exponenten q, der Anzahl der PML-Schichten N_{lay} und dem minimal erreichbaren Reflexionsfaktor $R = \exp\left(\frac{2}{c}\int_0^\delta \sigma(w)\,dw\right)$ gewählt, der sich für eine senkrecht auf das Medium einfallende Welle aus der endlichen Dicke δ des Absorbermaterials ergibt. Eine ausführliche Studie über die Wahl dieser Parameter findet sich in [26]. Die Werte p in Gl. 2.2.27 ergeben sich damit für jede Schicht durch Integration von Gl. 2.2.28 über die jeweilige Gitterschrittweite. Hierbei wird im Fall der elektrischen Verluste über die normale Zelle (Δw), für die magnetischen Leitfähigkeiten über die duale Zelle ($\widetilde{\Delta w}$) integriert. Der schematische Aufbau der resultierenden Anordnung, die das Verhalten des offenen Raumes in FIT für eine Raumrichtung realisiert, ist in Abb. 2.5 zusammengefasst.

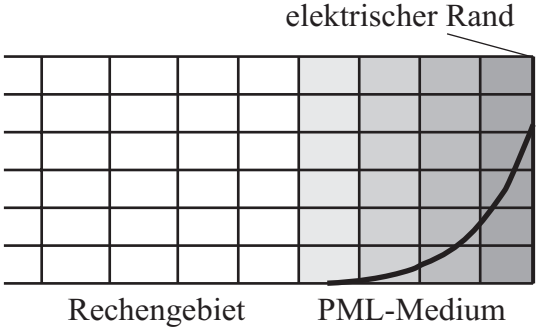


Abbildung 2.5: *4-Schichten-Flächen-PML mit ansteigendem Leitfähigkeitsprofil. Das PML-Medium selbst wird mit einer elektrischen Randbedingung abgeschlossen.*

[13] In [25] wird $a(\omega) = 1+(p\,\beta)/(1+j\omega\,\beta)$ vorgeschlagen, das für niedrige Frequenzen ($\omega \to 0$) ein günstigeres Verhalten erzielt und ebenfalls ein kausales Medium darstellt. Um die Komplexität für den späteren Ordnungsreduktions-Prozess geringer zu halten, wird in dieser Arbeit jedoch die Beziehung 2.2.27 verwendet. Beide Formulierungen weisen für hohe Frequenzen denselben Grenzwert auf.

[14] Selbst innerhalb homogenen Materials treten bei inhomogenen Gittern an großen Gittersprüngen Reflexionen aufgrund der Gitterdispersion auf.

PML-Absorber höherer Ordnung

Wird das Rechengebiet in mehrere Raumrichtungen durch PML-Material abgeschlossen, ergeben sich in den Überschneidungszonen Bereiche, in denen neue Materialien auftreten, die wiederum reflexionsfrei an die angrenzenden Absorbermaterialien angepasst sein müssen. Man unterscheidet daher insgesamt (siehe Abb. 2.6) die Absorbertypen für Flächen-PML, Kanten-PML und Ecken-PML. Es lässt sich zeigen [23], dass der sich ergebende Operator Λ_{uw} in $\vec{\vec{\varepsilon}}$ und $\vec{\vec{\mu}}$ im Überschneidungsbereich eines PML-Materials in u- und in w-Richtung als

$$\Lambda_{uw} = \operatorname{diag}\left\{\frac{c_w}{c_u},\, c_w\, c_u,\, \frac{c_u}{c_w}\right\} = \Lambda_u \Lambda_w \tag{2.2.29}$$

dargestellt werden kann. Analog ergibt sich Λ_{uvw} für das Material der Ecken-PML zu

$$\Lambda_{uvw} = \operatorname{diag}\left\{\frac{c_v\, c_w}{c_u},\, \frac{c_u\, c_w}{c_v},\, \frac{c_u\, c_v}{c_w}\right\} = \Lambda_u \Lambda_v \Lambda_w. \tag{2.2.30}$$

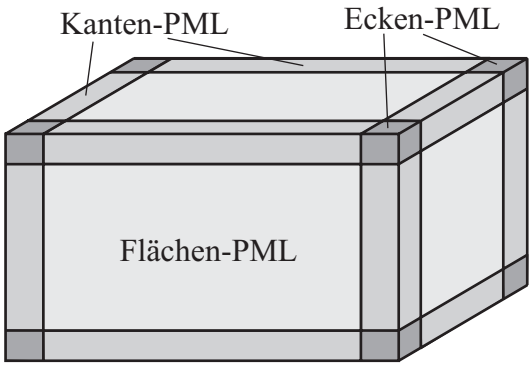


Abbildung 2.6: *Auftreten von PML-Absorbern höherer Ordnung in Überschneidungszonen der Flächen-PML.*

2.2.3.5 Wellenleiterrandbedingung

Eine besondere Bedeutung zur Berandung aber insbesondere auch zur Anregung des Rechengebiets kommt den so genannten *Wellenleiterports* zu. Diese basieren auf der Tatsache, dass innerhalb eines Hohlleiters nur spezielle Wellenformen die Randbedingungen erfüllen und sich folglich ausbreiten können. Unter Annahme einer unendlichen Ausdehnung des Wellenleiters lassen sich die entsprechenden *Moden* alleine durch die Feldbilder in einer beliebigen Querschnittsfläche sowie die frequenzabhängige Ausbreitungskonstante k beschreiben. Die zugehörigen Phasoren $\vec{E}_m$ und $\vec{H}_m$ sind im Fall eines transversal homogenen und verlustfreien Materials frequenzunabhängig. Die Feldbilder unterschiedlicher Moden sind zudem nach einem geeigneten Skalarprodukt zueinander orthogonal [3].

Übertragen auf das diskrete Modell lässt sich ebenfalls jedes Feldbild innerhalb eines Wellenleiters durch die Superposition aller Moden darstellen, wobei jeder Mode in

einen hin- und einen rücklaufenden Anteil mit den Amplituden a und b getrennt werden kann, hier für die Ausbreitungsrichung w angegeben:

$$\widehat{\mathbf{e}}(w) = \sum_{m=1}^{M} \widehat{\mathbf{e}}_{t,m} \left(a_m e^{-jk_w w} + b_m e^{+jk_w w} \right), \tag{2.2.31a}$$

$$\widehat{\mathbf{h}}(w) = \sum_{m=1}^{M} \widehat{\mathbf{h}}_{t,m} \left(a_m e^{-jk_w w} - b_m e^{+jk_w w} \right). \tag{2.2.31b}$$

Die Phasoren $\widehat{\mathbf{e}}_{t,m}$ und $\widehat{\mathbf{h}}_{t,m}$ sind hierbei 2D-Vektoren, die nur die transversalen Elemente $\widehat{e}$ und $\widehat{h}$ des Modes m beinhalten, analog sind auch $\widehat{\mathbf{e}}(w)$ und $\widehat{\mathbf{h}}(w)$ auf die transversalen Komponenten beschränkt. Die Berechnung der Phasoren bzw. 2D-Moden $\widehat{\mathbf{e}}_{t,m}$ in längshomogenen Strukturen mit Hilfe der Methode der Finiten Integration soll an dieser Stelle nur kurz skizziert werden. Eine erste Darstellung findet sich bereits in [28], eine ausführliche Herleitung in [14] und [27]. Durch den Grenzübergang $\Delta w \to 0$ lässt sich ein Gleichungssystem der Dimension $2 \cdot I \cdot J$ in der folgenden Form angeben [14]:

$$\left(\mathbf{B}_e^{-1} \mathbf{A}_e - \omega^2 \mathbf{B}_e^{-1} \right) \widehat{\mathbf{e}}_t = k_w^2 \, \widehat{\mathbf{e}}_t. \tag{2.2.32}$$

Wird die Frequenz ω beispielsweise auf den Mittelwert des interessierenden Frequenzintervalls festgelegt, entspricht dies einem einfachen Eigenwertproblem mit den Lösungen für $\widehat{\mathbf{e}}_{t,m}$ und $k_{w,m}$. Die zugehörigen magnetischen Spannungen $\widehat{\mathbf{h}}_{t,m}$ können über das Induktionsgesetz einfach bestimmt werden.

Mit dem diskreten Operator $\mathbf{N}_w$

$$\mathbf{N}_w = \begin{pmatrix} \mathbf{0} & -\mathbf{I} \\ \mathbf{I} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{N}_w \widehat{\mathbf{h}}_t = \begin{pmatrix} -\widehat{\mathbf{h}}_v \\ \widehat{\mathbf{h}}_u \end{pmatrix}. \tag{2.2.33}$$

gelten die folgenden Entsprechungen mit dem Kontinuierlichen:

$$\vec{n}_w \times \vec{H} \qquad \to \qquad \mathbf{N}_w \widehat{\mathbf{h}}_t, \tag{2.2.34a}$$

$$\vec{n}_w \cdot (\vec{E} \times \vec{H}) = \vec{E} \cdot (-(\vec{n}_w \times \vec{H})) \qquad \to \qquad \widehat{\mathbf{e}}_t^{\,T} (-\mathbf{N}_w) \widehat{\mathbf{h}}_t. \tag{2.2.34b}$$

Die frei wählbare Amplitude der Phasoren wird damit üblicherweise so festgelegt, dass die durch den Wellenleiter fließende Leistung den Wert 1 W annimmt. Für das Integral über den Poyntingschen Vektor $\int \vec{n}_w \cdot (\vec{E} \times \vec{H}) \, dA$ folgt im Diskreten also

$$\widehat{\mathbf{e}}_{t,m}^{\,T} (-\mathbf{N}_w) \widehat{\mathbf{h}}_{t,m} = 1. \tag{2.2.35}$$

Mit den Vektoren $\widehat{\mathbf{e}}_{t,m}$ und $\widehat{\mathbf{h}}_{t,m}$ ist es nun möglich, einen offenen Wellenleiterrand zu konstruieren, über den Energie das Rechengebiet verlassen, aber auch zur Anregung eingebracht werden kann. Für genauere Beschreibungen siehe [13, 14, 20].

2.2.3.6 Das Äquivalenzprinzip

Anstelle eines wirklich offenen Wellenleiterrandes kann die unendliche Fortsetzung des Hohlleiters aber auch durch ein äquivalentes geschlossenes Problem mit eingeprägtem Randstrom modelliert werden. Dies entspricht einer Modellierung des

kontinuierlichen Äquivalenzprinzips $\vec{J}_F = \vec{n} \times \vec{H}$ im Diskreten. Durch die Allokation der magnetischen Komponenten auf dem dualen Gitter ist jedoch das Magnetfeld in der Randebene selbst nicht bekannt, was durch einen Korrekturterm berücksichtigt werden kann. Betrachtet man jeweils den Umlauf um die grau markierte duale Fläche in der Originalanordnung (Abb. 2.7a) und der äquivalenten Anordnung (Abb. 2.7b), folgt für $\widehat{\widehat{j}}$

$$\widehat{\widehat{j}} = \frac{1}{2}(\widehat{h}_2 + \widehat{h}_4). \tag{2.2.36}$$

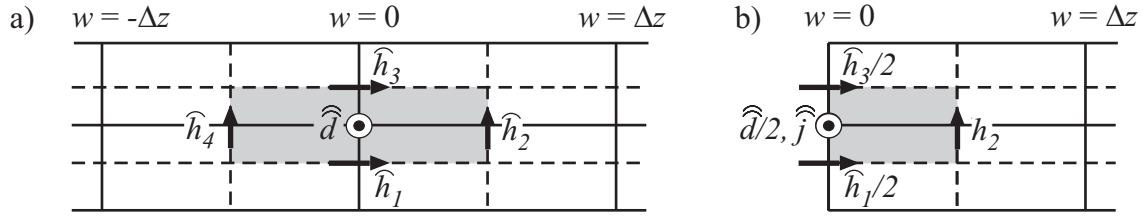


Abbildung 2.7: *Originalanordnung a) und äquivalente Anordnung b) zur Herleitung der Wellenleiterrandbedingung durch Stromeinprägung.*

Wird diese lokale Betrachtung auf ganze Vektoren erweitert, bedeutet die Drehung der Orientierung die Anwendung der Kreuzprodukts $\vec{J}_F = \vec{n} \times \vec{H}$.

$$\widehat{\widehat{\mathbf{j}}}_e = \frac{1}{2}\mathbf{N}_w \left(\widehat{\mathbf{h}} \left(\frac{\Delta w}{2} \right) + \widehat{\mathbf{h}} \left(-\frac{\Delta w}{2} \right) \right). \tag{2.2.37}$$

Sinnvoll ist an dieser Stelle zusätzlich die Definition verallgemeinerter Ströme $\mathbf{i}$ und Spannungen $\mathbf{u}$ in der Randfläche nach:

$$\mathbf{u} = \mathbf{Z}_L^{1/2}(\mathbf{a} + \mathbf{b}), \qquad \mathbf{a} = \frac{1}{2}\left(\mathbf{Z}_L^{-1/2}\mathbf{u} + \mathbf{Z}_L^{1/2}\mathbf{i} \right), \tag{2.2.38a}$$

$$\mathbf{i} = \mathbf{Z}_L^{-1/2}(\mathbf{a} - \mathbf{b}), \qquad \mathbf{b} = \frac{1}{2}\left(\mathbf{Z}_L^{-1/2}\mathbf{u} - \mathbf{Z}_L^{1/2}\mathbf{i} \right). \tag{2.2.38b}$$

Die Vektoren $\mathbf{a}$ und $\mathbf{b}$ enthalten im Fall mehrerer Moden oder Ports alle Einträge a_m und b_m aus Gl. 2.2.31. Durch die symmetrische Normierung mit $\mathbf{Z}_L^{1/2} = \mathrm{diag}(\sqrt{Z_{L,m}})$ wird erreicht, dass für die Leistung eines Ports für den reflexionsfreien Fall ($a_m = 1$, $b_m = 0$) ebenfalls gilt: $u_m \cdot i_m = 1$. Die Wahl der Einträge der Diagonalmatrix $\mathbf{Z}_L$ ist hierbei zunächst beliebig, im Fall von Mehrleitermoden ist es jedoch naheliegend, die tatsächlichen Leitungswiderstände zu wählen.

Die w-Abhängigkeit eines einzelnen Modes $\widehat{\mathbf{h}}_t$ ergibt sich nach Gl. 2.2.31b schließlich zu

$$\widehat{\mathbf{h}}(w) = \widehat{\mathbf{h}}_t \left(a\, e^{-jk_w w} - b\, e^{jk_w w} \right) \tag{2.2.39a}$$

$$= \frac{1}{2}\,\widehat{\mathbf{h}}_t \left(\left(\frac{u}{\sqrt{Z_L}} + i\sqrt{Z_L} \right) e^{-jk_w w} - \left(\frac{u}{\sqrt{Z_L}} - i\sqrt{Z_L} \right) e^{jk_w w} \right) \tag{2.2.39b}$$

$$= \widehat{\mathbf{h}}_t \left(i\sqrt{Z_L}\,\cos(k_w w) - j\,\frac{u}{\sqrt{Z_L}}\sin(k_w w) \right). \tag{2.2.39c}$$

Die bisherige Betrachtung von $\widehat{\mathbf{h}}(w)$ gilt für die in einer 2D-Ebene unter der Annahme $\Delta w \to 0$ berechneten Moden $\widehat{\mathbf{h}}_t(w)$. Durch den Übergang auf das 3D-Gitter muss nun die 2D-Ausbreitungskonstante k_w an den Dispersionsfehler des 3D-Gitters angepasst werden, um einen Energiefehler zu vermeiden. Diese ergibt sich nach [14] zu

$$k_{w,3D} = \frac{2}{\Delta w} \arcsin\left(\frac{k_w \, \Delta w}{2}\right). \tag{2.2.40}$$

Für den Anregungsvektor ergibt sich nach den Gln. 2.2.37 und 2.2.39c:

$$\widehat{\widehat{\mathbf{j}}}_{e,m} = i_m \sqrt{Z_{L,m}} \cos(k_{w,3D}\Delta w/2) \, \mathbf{N}_w \widehat{\mathbf{h}}_{t,m}, \tag{2.2.41a}$$

$$= i_m \sqrt{Z_{L,m}} \cos(k_{w,3D}\Delta w/2) \, \mathbf{b}_m^{(2D)}. \tag{2.2.41b}$$

Der Kosinusterm korrigiert folglich den Energiefehler, der entsteht, da die Normierung 2.2.35 der Portmoden in derselben Ebene $w = 0$ erfolgt, während die Größen tatsächlich auf räumlich getrennten Gittern allokiert sind. Für feine Diskretisierungen ist der Korrekturfaktor $\cos(k_{w,3D}\Delta w/2)$ jedoch meist sehr nahe an Eins und kann häufig vernachlässigt werden.

Der Anregungsvektor $\widehat{\widehat{\mathbf{j}}}_e$ ist zunächst noch allein in der 2D-Ebene definiert und muss noch auf das 3D-Gitter erweitert werden, was durch einen portabhängigen Operator $\mathbf{L}_k$ erfolgt. Die Anregungsvektoren $\widehat{\widehat{\mathbf{j}}}'_{e,m}$ bzw. $\mathbf{b}_m$ werden damit wie folgt eingeführt:

$$\widehat{\widehat{\mathbf{j}}}'_{e,m} = \mathbf{L}_k \widehat{\widehat{\mathbf{j}}}_{e,m}, \qquad \mathbf{b}_m = \mathbf{L}_k \mathbf{b}_m^{(2D)}. \tag{2.2.42}$$

Erfolgt die Anregung bei $w_{\max}$, muss das Vorzeichen entsprechend geändert werden.

Aufgrund der Orthogonalität der Moden lässt sich mit den Definitionen Gl. 2.2.38 und der Normierung Gl. 2.2.35 auch die verallgemeinerte Portspannung u_m angeben:

$$u_m = \sqrt{Z_{L,m}} \, (a_m + b_m) = \sqrt{Z_{L,m}} \, \mathbf{b}_m^T \, \widehat{\mathbf{e}}. \tag{2.2.43}$$

Die Ein- und Auskopplung am Wellenleiterrand ist also bis auf den Korrekturfaktor symmetrisch.

Frequenzabhängige Referenzimpedanzen

Die Moden in Gl. 2.2.31 lassen sich in drei Grundtypen TE, TM und TEM einteilen, wobei die Buchstaben E (elektrisch) und M (magnetisch) diejenigen Komponenten beschreiben, die ausschließlich in transversaler Richtung existieren. Es ist zu beachten, dass die beiden erstgenannten Typen nur oberhalb einer gewissen Grenzfrequenz ω_c ausbreitungsfähig sind, während TEM-Moden für alle Frequenzen existieren, dafür aber getrennte Hin- und Rückleiter voraussetzen.

Die Normierung nach Gl. 2.2.35 ist nur breitbandig gültig, wenn die Portimpedanz $Z_w = E^T/H^T$ frequenzunabhängig ist. Dies ist im Fall von TEM-Moden exakt, für Quasi-TEM-Moden (z. B. Mikrostreifenleiter) näherungsweise erfüllt. Für TE- und TM-Moden kann jedoch eine einfache Korrektur erfolgen, da die Portimpedanz analytisch bekannt ist, sie lautet

$$Z_{wTE} = \frac{\omega \, \mu}{k_{w,m}}, \qquad Z_{wTM} = \frac{k_{w,m}}{\omega \, \varepsilon}. \tag{2.2.44}$$

Mit der Cutoff-Kreisfrequenz $\omega_{c,m}$, die mit Hilfe der Dispersionsgleichung bestimmt werden kann, ergibt sich die frequenzabhängige Ausbreitungskonstante $k_{w,m}(\omega)$ aus dem für den Wert ω_0 durch Lösung der Eigenwertgleichung bestimmten $k_{0,w,m}$

$$k_{w,m}(\omega) = \sqrt{\frac{\left(\omega^2 - \omega_{c,m}^2\right)\, k_{0,w,m}^2}{\left(\omega_0^2 - \omega_{c,m}^2\right)}}. \tag{2.2.45}$$

Damit ergibt sich der Korrekturfaktor F schließlich zu:

$$F_{TE} = \frac{\omega_0\, k_{w,m}}{\omega\, k_{0,w,m}}, \qquad F_{TM} = \frac{\omega\, k_{0,w,m}}{\omega_0\, k_{w,m}} = \frac{1}{F_{TE}}. \tag{2.2.46}$$

Die Korrektur muss für jeden betrachteten Frequenzpunkt gesondert berechnet werden. Zur Bestimmung wird die 2D-Ausbreitungskonstante k_w herangezogen.

2.2.3.7 Electromagnetic-Circuit-Element-Randbedingungen

Die oben beschriebenen Wellenleiterrandbedingungen bilden die physikalische Realität für diverse Wellenleitertypen im diskreten Modell sehr genau ab. Dazu gehört insbesondere auch, dass Welleneffekte auftreten. Wegunabhängige Spannungen und Stromflüsse ausschließlich durch Metallkontakte, wie in der klassischen Kirchhoffschen Theorie gefordert, existieren daher nur für den TEM-Fall sowie näherungsweise für *Quasi-TEM-Wellen* wie beispielsweise Mikrostreifenleiter. Soll das diskrete Modell mit einem klassischen Kirchhoffschen Netzwerk verbunden werden, besteht die Möglichkeit, eindeutig definierte Ströme und Spannungen mit Hilfe der so genannten *Electric-Circuit-Element, ECE* Randbedingung [29] grundsätzlich zu erzwingen.

Hierzu wird die einfach geschlossene Oberfläche Σ der betrachteten Struktur in einen isolierenden Anteil S_0 sowie in K nichtüberlappende Kontaktstellen S_k mit k=1...K zerlegt (siehe Abb. 2.8).

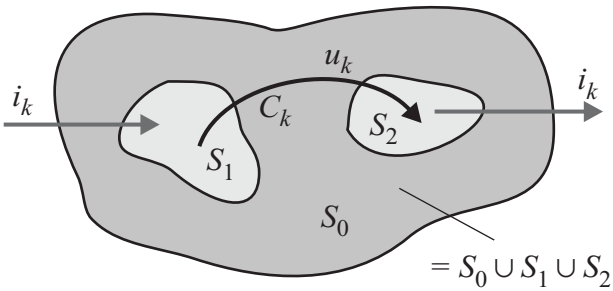


Abbildung 2.8: *Ein Electromagnetic-Circuit-Element mit zwei Kontaktflächen.*

Auf der Oberfläche müssen die folgenden ECE-Bedingungen gelten:

$$\int_C \vec{E}\, d\vec{r} = 0, \quad \text{für jede Kurve } C \subset S_k, k = 1 \ldots K, \tag{2.2.47}$$

$$\int_{\partial S} \vec{E}\, d\vec{r} = 0, \quad \text{für jedes } S \subset \Sigma, \tag{2.2.48}$$

$$\int_{\partial S} \vec{H}\, d\vec{r} = 0, \quad \text{für jedes } S \subset S_0. \tag{2.2.49}$$

Diese Beziehungen bedingen

- ein konstantes Potenzial innerhalb der Kontaktflächen,
- magnetische Entkopplung sowie eine eindeutige Definition der Spannung auf der Oberfläche,
- den alleinigen Stromfluss durch die Kontaktflächen.

Ist außerdem ein Kontakt geerdet, womit ein Referenzpotenzial für die gesamte Oberfläche definiert ist, und werden alle Feldkomponenten im Anfangszustand $\vec{E}$ und $\vec{H}$ als Null angenommen, hat das Feldproblem mit den Anregungen

$$u_k(t) = \int_{C_k \subset S_0} \vec{E}\, d\vec{r}, \quad k = 1 \ldots K_1 \quad \text{bzw.} \tag{2.2.50}$$

$$i_k(t) = \int_{\partial S_k} \vec{H}\, d\vec{r}, \quad k = K_1 + 1 \ldots K - 1 \tag{2.2.51}$$

eine eindeutige Lösung. Die Kurve C_k wird wie in Abb. 2.8 dargestellt gewählt.

Eine ausführliche Darstellung, wie die ECE-Randbedingungen mit Hilfe zweier dualer Bäume auf der Oberfläche des Rechengebiets mit der FIT-Formulierung verkoppelt werden können, findet sich in [99]. Praktische Vergleiche haben gezeigt, dass für TEM- und Quasi-TEM-Wellenleiter mit ideal leitenden Kontaktflächen ECE-Randbedingungen und die klassischen Wellenleiterränder nahezu identische Ergebnisse liefern.

2.2.3.8 Diskrete Ports

Mit den Wellenleiter- und den ECE-Randbedingungen wurden bereits zwei Möglichkeiten beschrieben, das Rechengebiet vom Rand aus anzuregen. Häufig ist es jedoch auch wünschenswert, im Inneren des Rechengebiets eine Anregung zu definieren, die einer Strom- oder Spannungsquelle bzw. einer Dipolantenne entspricht. Grundsätzlich kann hierzu jede Feld- oder Flussgröße auf beliebigen normalen, bzw. dualen Kanten oder Flächen (auch Kombinationen dieser) auf vorgegebene Werte gesetzt werden. Man unterscheidet zwischen *harter* Anregung, bei der dieser Wert auf der entsprechenden Elementarfigur fest eingeprägt wird, und *weicher* Anregung, bei der das Anregungssignal zum bestehenden Feld oder Fluss lediglich addiert wird. In Anlehnung an die physikalische Realität beschränkt man sich üblicherweise auf das Einprägen von Strömen, Spannungen oder Wellenamplituden.

Wird eine einzelne primäre Kante oder ein zusammenhängender Pfad von Kanten zur Anregung genutzt, spricht man von einem *diskreten Port*. Dem Pfad wird hierbei eine Impedanz zugeordnet und es kann wahlweise eine Spannung oder ein Strom[15] vorgegeben werden. Aus praktischen Erwägungen wird hierbei meist eine Stromeinprägung bevorzugt.

[15]Im Falle der Stromeinprägung erfolgt diese nicht auf der primären Kante, sondern auf der zugehörigen dualen Fläche. Da für dual orthogonale Gitter die Flächennormale auf der primären Kante liegt, kann dieser Strom auch als *durch die primäre Kante fließend* interpretiert werden.

Kapitel 3

FIT in Systemdarstellungen

Mit der Methode der Finiten Integration wurde im vorhergehenden Kapitel ein leistungsfähiges Verfahren zur Berechnung elektromagnetischer Felder vorgestellt. Häufig sind jedoch gar nicht die kompletten Feldlösungen, sondern nur das Verhältnis definierter Eingangs- und Ausgangsgrößen von Interesse. Der FI-Methode kommt in diesem Zusammenhang die Rolle eines Systems zu, das diese Größen miteinander verbindet.

Es existieren vielfältige Möglichkeiten, dieses System zu definieren. Abhängig vom Charakter der Anregungsgrößen wird zunächst zwischen Impedanz- und Streuparameterdarstellungen unterschieden, wobei beide Formulierungen ineinander überführt werden können. Zudem variieren die Darstellungen in der Anzahl der betrachteten inneren Zustände und damit zugleich im Grad der zugehörigen Differentialgleichung. Hierzu sollen verschiedene Varianten beschrieben werden.

Große praktische Bedeutung kommt auch speziellen Systemeigenschaften wie Kausalität, Stabilität, Passivität, Steuer- und Beobachtbarkeit zu. Diese sollen für unterschiedliche Systeme anhand der Systemmatrizen untersucht werden. Die aufgestellten Systeme bilden die Grundlage für ordnungsreduzierte Modelle, die im nächsten Kapitel näher beschrieben werden.

3.1 Zustandsraumdarstellung der Impedanz

3.1.1 Klassischer Zustandsraum

Kombiniert man die ersten beiden Gitter-Maxwellgleichungen in Gl. 2.2.9 mit den Materialbeziehungen Gl. 2.2.12 zu einem gemeinsamen System mit sämtlichen Einträgen aus $\widehat{\mathbf{e}}$ und $\widehat{\mathbf{h}}$ als Unbekannten, ergibt sich ein in der Frequenz lineares System:

$$\underbrace{\begin{pmatrix} \mathbf{M}_\varepsilon & \mathbf{0} \\ \mathbf{0} & \mathbf{M}_\mu \end{pmatrix}}_{\mathbf{M}} \frac{d}{dt} \begin{pmatrix} \widehat{\mathbf{e}} \\ \widehat{\mathbf{h}} \end{pmatrix} = - \begin{pmatrix} \mathbf{M}_\kappa & -\widetilde{\mathbf{C}} \\ \mathbf{C} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \widehat{\mathbf{e}} \\ \widehat{\mathbf{h}} \end{pmatrix} + \begin{pmatrix} \widehat{\widehat{\mathbf{j}}}_s \\ \mathbf{0} \end{pmatrix}. \tag{3.1.1}$$

Die Systemordnung ist hierbei $n = n_e + n_h$[1]. Den verbleibenden Gitter-Maxwellgleichungen kommt die Rolle von Nebenbedingungen zu. Durch die Beziehungen Gl. 2.2.11 werden sie bei geeigneten Verfahren, exakter Rechnung und konsistenter Anregung ebenfalls erfüllt.

Multipliziert man Gl. 3.1.1 mit der invertierten Matrix $\mathbf{M}^{-1}$, was aufgrund der Diagonalgestalt trivial möglich ist, setzt $\widetilde{\mathbf{C}} = \mathbf{C}^T$ nach Gl. 2.2.10 und führt die normierten Feldstärken[2] $\widehat{\mathbf{e}}' = \mathbf{M}_\varepsilon^{1/2}\,\widehat{\mathbf{e}}$ mit $\mathbf{M}_\varepsilon^{1/2}\mathbf{M}_\varepsilon^{1/2} = \mathbf{M}_\varepsilon$ und entsprechend $\widehat{\mathbf{h}}' = \mathbf{M}_\mu^{1/2}\widehat{\mathbf{h}}$ ein, erhält man die Systemmatrix $\mathbf{A}$, die im verlustfreien Fall $\mathbf{M}_\kappa = \mathbf{0}$ schiefsymmetrisch ist. Die normierten Feldstärken haben damit zudem den Vorteil, dass $\widehat{\mathbf{e}}'$ und $\widehat{\mathbf{h}}'$ und somit der gesamte Vektor der Unbekannten die gleiche physikalische Einheit ($\sqrt{VAs/m}$) besitzen. Wird die zu simulierende Struktur an m Eingangsports angeregt, kann zusätzlich eine Einkoppelmatrix $\mathbf{R}'$ mit $\mathbf{M}_\varepsilon^{1/2}\,\widehat{\widehat{\mathbf{j}}}_s = \mathbf{M}_\varepsilon^{-1/2}\,\mathbf{R}\,\mathbf{i} = \mathbf{R}'\,\mathbf{i}$ definiert werden, die den Zusammenhang zwischen den verallgemeinerten Strömen der Ports, repräsentiert durch den m-zeiligen Vektor $\mathbf{i}$, und dem Stromvektor $\widehat{\widehat{\mathbf{j}}}_s$ darstellt. Schließlich ergibt sich:

$$\underbrace{\frac{d}{dt}\begin{pmatrix}\widehat{\mathbf{e}}'\\ \widehat{\mathbf{h}}'\end{pmatrix}}_{\dot{\mathbf{x}}} = -\underbrace{\begin{pmatrix}\mathbf{M}_\varepsilon^{-1/2}\mathbf{M}_\kappa\mathbf{M}_\varepsilon^{-1/2} & -\mathbf{M}_\varepsilon^{-1/2}\widetilde{\mathbf{C}}\,\mathbf{M}_\mu^{-1/2}\\ \mathbf{M}_\mu^{-1/2}\mathbf{C}\,\mathbf{M}_\varepsilon^{-1/2} & \mathbf{0}\end{pmatrix}}_{\mathbf{A}}\underbrace{\begin{pmatrix}\widehat{\mathbf{e}}'\\ \widehat{\mathbf{h}}'\end{pmatrix}}_{\mathbf{x}} + \underbrace{\begin{pmatrix}\mathbf{R}'\\ \mathbf{0}\end{pmatrix}}_{\mathbf{B}}\mathbf{i}. \tag{3.1.2}$$

Analog zu $\mathbf{R}'$ kann auch eine Auskoppelmatrix $\mathbf{L}' = \mathbf{L}\,\mathbf{M}_\varepsilon^{-1/2}$ definiert werden, die an l Auskoppelports die l äquivalenten Spannungen $\mathbf{u}$ aus dem normierten Spannungsvektor $\widehat{\mathbf{e}}'$ nach

$$\mathbf{u} = \mathbf{L}'\,\widehat{\mathbf{e}}' \qquad \text{oder äquivalent} \qquad \mathbf{u} = \underbrace{(\mathbf{L}'\quad \mathbf{0})}_{\mathbf{C}}\begin{pmatrix}\widehat{\mathbf{e}}'\\ \widehat{\mathbf{h}}'\end{pmatrix} \tag{3.1.3}$$

extrahiert. In kombinierter Schreibweise der Gln. 3.1.2 und 3.1.3 entspricht dies von der Struktur der klassischen Zustandsraumdarstellung[3] [34, 35]

$$\dot{\mathbf{x}} = -\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{i} \tag{3.1.4a}$$

$$\mathbf{u} = \mathbf{C}\mathbf{x} + \mathbf{D}\mathbf{i}. \tag{3.1.4b}$$

Hierbei stellt $\mathbf{x}$ den n-dimensionalen Zustandsvektor und $\mathbf{A}$ die $n \times n$ Systemmatrix dar. Die Einkoppelmatrix $\mathbf{B}$ und die Auskoppelmatrix $\mathbf{C}$ haben die Dimensionen $n \times m$ und $l \times n$ und die Matrix $\mathbf{D}$ schließlich $l \times m$. Diese Dimensionen werden bildlich auch in Abb. 3.1 verdeutlicht. Für diesen speziellen Fall nach Gl. 3.1.3 gilt $\mathbf{D} = \mathbf{0}$, es besteht also keine direkte Kopplung zwischen Eingang und Ausgang.

[1] In der Standardformulierung gilt $n_e = n_h$. Unter Umständen kann es aber sinnvoll sein, die durch ideale Randbedingungen entstandenen Nullzeilen und Spalten des System zu eliminieren, was auf $n_e \neq n_h$ führt.

[2] Aufgrund der Diagonalgestalt der Materialmatrizen lassen sich die Wurzelmatrizen sehr einfach durch Radizieren jedes einzelnen Diagonaleintrags bilden.

[3] In Abweichung zu einigen Lehrbüchern wird die Matrix $\mathbf{A}$ hier negativ definiert.

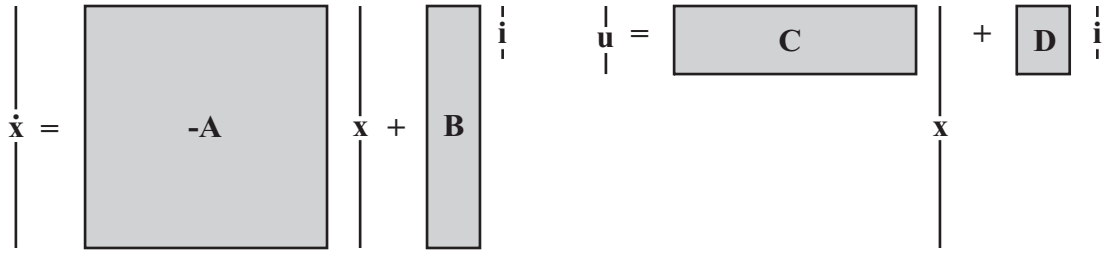


Abbildung 3.1: *Bildliche Darstellung der Matrix- und Vektordimensionen der Zustandsraumgleichungen Gl. 3.1.4.*

Eine alternative Möglichkeit einen Zustandsraum für eine diskretisierte Struktur aufzustellen, basiert auf der Idee der ECE-Randbedingungen. Dies erfordert die Aufstellung zweier zueinander dualer Bäume auf der Randoberfläche. Eine detaillierte Darstellung hierzu findet sich in [99].

Als Lösung des Systems folgt für den allgemeinsten Fall mit den Anfangswerten des Zustandsvektors $\mathbf{x}(0)$ für ein kausales Eingangssignal $\mathbf{i}(t)$:

$$\mathbf{u}(t) = \mathbf{C}e^{-\mathbf{A}t}\mathbf{x}(0) + \mathbf{C}\int_0^t e^{-\mathbf{A}(t-\tau)}\mathbf{B}\mathbf{i}(\tau)\mathrm{d}\tau + \mathbf{D}\mathbf{i}(t). \tag{3.1.5}$$

Wird nun generell $\mathbf{D} = \mathbf{0}$ und $\mathbf{x}(0) = \mathbf{0}$ angenommen und das System durch einen Dirac-Impuls angeregt, folgt als Impulsantwort des FIT-Systems

$$\mathbf{h}(t) = \mathbf{C}e^{-\mathbf{A}t}\mathbf{B} = \sum_{k=0}^{\infty} \frac{1}{k!} t^k \, \mathbf{B}(-\mathbf{A})^k\mathbf{C}. \tag{3.1.6}$$

Die Koeffizienten der Reihendarstellung $\mathbf{B}(-\mathbf{A})^k\mathbf{C}$ werden auch als *Markovparameter* bezeichnet und erhalten später im Rahmen der Ordnungsreduzierung einige Bedeutung.

Neben der Lösung im Zeitbereich ist häufig das Systemverhalten für harmonische Anregungen interessant. Sind alle vorkommenden Signale harmonischer Natur, kann $\mathbf{x}$ durch einen komplexen Phasor und damit $\dot{\mathbf{x}}$ durch $j\omega\,\mathbf{x}$ dargestellt werden. Dies ermöglicht die Eliminierung des Zustandsvektors, was auf die Übertragungsfunktion $\mathbf{H}$ mit

$$\mathbf{u}(j\omega) = \mathbf{H}(j\omega)\,\mathbf{i}(j\omega), \tag{3.1.7a}$$

$$\mathbf{H}(j\omega) = \mathbf{C}(j\omega\mathbf{I} + \mathbf{A})^{-1}\mathbf{B} \tag{3.1.7b}$$

führt. Die Gleichungen 3.1.6 und 3.1.7b bilden somit auch ein Transformationspaar der Fouriertransformation.

Üblicherweise erfolgt Anregung und Signalauskopplung an denselben Ports, womit $m = l$ gilt. Zudem ist es stets möglich, die Ports auf eine Weise zu definieren, dass $\mathbf{C} = \mathbf{B}^T$ gilt, was das System Gl. 3.1.4, bzw. 3.1.7b zusätzlich symmetrisiert. Mit der Normierung nach Abschnitt 2.2.3.5 ergibt sich damit $\mathbf{R} = \mathbf{L}^T = [\mathbf{b}_1, \ldots, \mathbf{b}_m]$, bei diskreten oder ECE-Ports hat jede Spalte von $\mathbf{R}$ nur genau einen Eintrag.

$\mathbf{H}(j\omega)$ bildet in diesem Fall eine Impedanzfunktion $\mathbf{Z}(j\omega)$:

$$\mathbf{Z}(j\omega) = \mathbf{B}^T(j\omega\mathbf{I} + \mathbf{A})^{-1}\mathbf{B}\,, \tag{3.1.8}$$

die die verallgemeinerten Spannungen nach

$$Z_{ij} = \left.\frac{u_i}{i_j}\right|_{i_k=0\;\forall\;k\neq j} \tag{3.1.9}$$

verkoppelt. Diese normierte Impedanz bildet die Basis für die Modelle reduzierter Ordnung im weiteren Verlauf der Arbeit. $\mathbf{Z}$ entspricht jedoch im Allgemeinen nicht der physikalischen Impedanzmatrix des Bauteils, da die Bezugswiderstände der Ports und diverse Korrekturfaktoren unberücksichtigt sind. Die Bezugswiderstände werden durch die Diagonalmatrix $\mathbf{Z}_L = \mathrm{diag}(Z_{L,m})$ beschrieben, für Wellenleiterports werden hierbei typischerweise die Feld- (TE, TM) oder Leitungswiderstände (TEM) gewählt, für diskrete Ports die Innenwiderstände der Quelle.

Ebenso müssen im Fall von Wellenleiterrändern bei TE- oder TM-Anregung die Energie-Korrekturfaktoren nach Gl. 2.2.46 in Form von $\mathbf{F} = \mathrm{diag}(F_m)$ berücksichtigt werden, wobei die Einträge für TEM-Wellen oder diskrete Ports Eins sind. Letzter zu betrachtender Faktor ist die diagonale Gitterkorrekturmatrix $\mathbf{D}_\beta$, deren Einträge die Kosinusterme nach Gl. 2.2.41b enthalten, die ebenfalls nur für Wellenleiterränder ungleich Eins sind. Für die physikalische Impedanzmatrix $\mathbf{Z}_r$ gilt damit schließlich:

$$\begin{aligned}\mathbf{Z}_r(j\omega) &= \mathbf{Z}_L^{1/2}\,\mathbf{F}^{1/2}\,\mathbf{B}^T(j\omega\mathbf{I} + \mathbf{A})^{-1}\mathbf{B}\,\mathbf{F}^{1/2}\,\mathbf{Z}_L^{1/2}\,\mathbf{D}_\beta \\ &= \mathbf{Z}_L^{1/2}\,\mathbf{F}^{1/2}\,\mathbf{Z}(j\omega)\,\mathbf{F}^{1/2}\,\mathbf{Z}_L^{1/2}\,\mathbf{D}_\beta.\end{aligned} \tag{3.1.10}$$

3.1.2 Curl-Curl-Formulierung

Eine Alternative zum linearen System[4] findet sich, wenn einer der Vektoren $\widehat{\mathbf{e}}$ oder $\widehat{\mathbf{h}}$ in Gl. 3.1.1 eliminiert wird. In Anlehnung an die kontinuierliche Wellengleichung mit den beiden Rotations- oder englisch *Curl*-Operatoren wird das resultierende System auch als *Curl-Curl-Formulierung* bezeichnet. Von der Struktur her entspricht das System nicht dem klassischen Zustandsraum, viele Zustandsraumverfahren können aber in leicht abgewandelter Form dennoch auch auf dieses verwandte System angewendet werden. Wählt man $\widehat{\mathbf{e}}$ als verbleibenden Unbekanntenvektor, ergibt sich das Curl-Curl-System zu

$$\mathbf{M}_\varepsilon\frac{d^2}{dt^2}\,\widehat{\mathbf{e}} + \mathbf{M}_\kappa\frac{d}{dt}\,\widehat{\mathbf{e}} + \underbrace{\mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}}_{\mathbf{A}'_{CC}}\,\widehat{\mathbf{e}} = \frac{d}{dt}\mathbf{R}\mathbf{i} \tag{3.1.11a}$$

$$\mathbf{u} = \mathbf{L}\,\widehat{\mathbf{e}} + \mathbf{D}\mathbf{i}. \tag{3.1.11b}$$

Die Anzahl der Unbekannten wird hierbei auf $n = n_e$ in etwa halbiert, was einen bedeutenden Vorteil dieser Formulierung darstellt. Erneut lässt sich das System

[4] Im Folgenden wird unter einem *linearen System* stets eine in der Frequenz lineare Formulierung verstanden.

mit $\mathbf{C}^T = \mathbf{B} = \mathbf{M}_\varepsilon^{-1/2}\mathbf{R} = \mathbf{R}'$ und durch Einführung eines neuen Zustandsvektors $\mathbf{x} = \mathbf{M}_\varepsilon^{1/2}\,\widehat{\mathbf{e}}$ symmetrisieren. Mit $\mathbf{D} = \mathbf{0}$, $\mathbf{A}_{CC} = \mathbf{M}_\varepsilon^{-1/2}\mathbf{A}'_{CC}\mathbf{M}_\varepsilon^{-1/2}$ und $\mathbf{K} = \mathbf{M}_\varepsilon^{-1}\mathbf{M}_\kappa$ folgt im Frequenzbereich

$$\left((j\omega)^2\mathbf{I} + j\omega\mathbf{K} + \mathbf{A}_{CC}\right)\mathbf{x} = (j\omega)\mathbf{B}\mathbf{i} \tag{3.1.12a}$$

$$\mathbf{u} = \mathbf{C}\mathbf{x}. \tag{3.1.12b}$$

Die Impedanzfunktion ergibt sich für Curl-Curl-Systeme damit zu:

$$\mathbf{Z}(j\omega) = (j\omega)\mathbf{B}^T\left((j\omega)^2\mathbf{I} + j\omega\mathbf{K} + \mathbf{A}_{CC}\right)^{-1}\mathbf{B}. \tag{3.1.13}$$

In diese Darstellung kann auf einfachem Wege auch der Term für Impedanzwände aus Abschnitt 2.2.3.3 eingebracht werden. Dies führt mit dem komplexen $\mathbf{M}_\mu^{-1}$ aus Gl. 2.2.22 auf ein $\mathbf{A}_{CC}^{(c)} = \mathbf{A}_{CC} - \frac{1}{\sqrt{j\omega}}\mathbf{P}$ mit $\mathbf{P} = \mathbf{M}_\varepsilon^{-1/2}\mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{K}_m\mathbf{C}\mathbf{M}_\varepsilon^{-1/2}$ auf

$$\mathbf{Z}(j\omega) = (j\omega)\mathbf{B}^T\left((j\omega)^2\mathbf{I} + j\omega\mathbf{K} - \frac{1}{\sqrt{j\omega}}\mathbf{P} + \mathbf{A}_{CC}\right)^{-1}\mathbf{B}. \tag{3.1.14}$$

Alle beschriebenen Impedanzen entsprechen wieder einer normierten Impedanz, für die tatsächliche Impedanz müssen Korrekturfaktoren entprechend Gl. 3.1.10 hinzugefügt werden.

Es ist zu beachten, dass neben der Eliminierung von Unbekannten umgekehrt auch durch Hinzufügen von inneren Zuständen ein Curl-Curl-System in eine lineare Formulierung transformiert werden kann. Transformiert man Gl. 3.1.12 (erneut ohne den Beitrag der Impedanzwände) in den Frequenzbereich und wählt $\mathbf{y} = \frac{q}{j\omega}\mathbf{x}$ mit einer zunächst beliebigen Konstante q, ergibt sich ein System der Gestalt

$$j\omega\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} = -\underbrace{\begin{pmatrix}\mathbf{K} & \frac{1}{q}\mathbf{A}_{CC}\\ -q\mathbf{I} & \mathbf{0}\end{pmatrix}}_{\mathbf{A}'_l}\begin{pmatrix}\mathbf{x}\\ \mathbf{y}\end{pmatrix} + \begin{pmatrix}\mathbf{B}\\ \mathbf{0}\end{pmatrix}\mathbf{i}. \tag{3.1.15}$$

Dieses System beschreibt dasselbe dynamische Verhalten wie auch Gl. 3.1.2, $\mathbf{A}'_l$ hat folglich dieselben dynamischen Eigenwerte wie $\mathbf{A}$ aus Gl. 3.1.2, obwohl die Symmetrieeigenschaften von Gl. 3.1.2 in Gl. 3.1.15 nicht erhalten sind. Die Konstante q sollte so gewählt werden, dass alle Blöcke in $\mathbf{A}_l$ etwa dieselbe Norm aufweisen. Es zeigt sich, dass mit $q = \sqrt{\|\mathbf{A}_{CC}\|}$ eine gute Konditionierung erreicht werden kann.

3.1.3 Systeme höheren Grades

Die im vorigen Abschnitt vorgestellte Curl-Curl-Formulierung Gl. 3.1.13 entspricht einem System zweiten Grades. Mit zunehmend komplexeren Materialmodellierungen treten allgemein auch Systeme höheren Grades auf, die sich im Frequenzbereich durch

$$\sum_{k=0}^{N_A}(j\omega)^k\mathbf{A}_k\mathbf{x} = \sum_{k=0}^{N_R}(j\omega)^k\mathbf{R}_k\,\mathbf{i} \tag{3.1.16a}$$

$$\mathbf{u} = \mathbf{L}\mathbf{x} \tag{3.1.16b}$$

oder in Impedanzformulierung durch

$$\mathbf{Z}(j\omega) = \mathbf{L}\left(\sum_{k=0}^{N_A}(j\omega)^k\mathbf{A}_k\right)^{-1}\sum_{k=0}^{N_R}(j\omega)^k\mathbf{R}_k \tag{3.1.17}$$

beschreiben lassen. Hierbei stellt $\max(N_A, N_R)$ den Grad des Systems dar. Die Gesamtordnung des Systems ergibt sich aus der Multiplikation des Grades mit der Dimension des Zustandsvektors $\mathbf{x}$. Je nach Anwendungsfall kann $\mathbf{x}$ hierbei entweder allein elektrische Feldkomponenten $\mathbf{x} = \widehat{\mathbf{e}}$ oder elektrische *und* magnetische $\mathbf{x} = (\widehat{\mathbf{e}}, \widehat{\mathbf{h}})^T$ enthalten. Grundsätzlich lassen sich Systeme beliebigen Grades entsprechend Gl. 3.1.15 stets in lineare Systeme umwandeln, in denen die Dimension des Zustandsvektors der Gesamtordnung entspricht. Dies soll im Folgenden für zwei wichtige Fälle einer Anordnung mit dispersiven Material und einer PML-berandeten Struktur verdeutlicht werden.

3.1.3.1 System mit dispersivem Dielektrikum

Alle in Abschnitt 2.1.1 auftretenden dispersiven dielektrischen Materialien lassen sich nach Gl. 2.2.14 als $\mathbf{M}_\varepsilon(\omega)$ darstellen. Wird dieser Zusammenhang in Gl. 3.1.11 eingesetzt, ergibt sich im Frequenzbereich:

$$(j\omega)^2\left(\mathbf{M}_{\varepsilon\infty} + \frac{1}{\alpha_0 + j\omega\alpha_1 + (j\omega)^2}\mathbf{M}_{\varepsilon d}\right)\widehat{\mathbf{e}} + \mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}\,\widehat{\mathbf{e}} = j\omega\mathbf{R}\,\mathbf{i} \tag{3.1.18a}$$

$$\mathbf{u} = \mathbf{R}^T\,\widehat{\mathbf{e}}. \tag{3.1.18b}$$

Die Diagonalmatrix $\mathbf{M}_{\varepsilon\infty}$ hat stets vollen Rang und enthält die Beiträge aller nichtdispersiven Dielektrika sowie die Grenzwerte der Permittivität des dispersiven Materials für Frequenzen gegen Unendlich. Die ebenfalls diagonale Matrix $\mathbf{M}_{\varepsilon d}$ enthält nur für die Komponenten Einträge, die im dispersiven Material liegen. Treten in der Struktur mehrere dispersive Dielektrika auf, erfordert dies die Summierung mehrerer Matrizen $\mathbf{M}_{\varepsilon di}$, worauf an dieser Stelle nicht weiter eingegangen werden soll. Wird Gl. 3.1.18a nun mit $(\alpha_0 + j\omega\alpha_1 + (j\omega)^2)$ multipliziert, führt dies auf ein System nach Gl. 3.1.16 mit $N_A = 4$ und $N_R = 3$.

Das System kann jedoch auch in ein lineares überführt werden. Dies erfordert die Definition neuer Zustände nach

$$\mathbf{x}_1 = \mathbf{M}_{\varepsilon\infty}\,\widehat{\mathbf{e}}, \tag{3.1.19a}$$

$$j\omega\mathbf{x}_2 = \mathbf{x}_3, \tag{3.1.19b}$$

$$j\omega\mathbf{x}_3 = -\alpha_0\mathbf{x}_2 - \alpha_1\mathbf{x}_3 + \mathbf{M}_{\varepsilon d}\,\widehat{\mathbf{e}}, \tag{3.1.19c}$$

$$j\omega\mathbf{x}_4 = \mathbf{x}_1. \tag{3.1.19d}$$

Mit $\mathbf{x} = (\mathbf{x}_1^T\ \mathbf{x}_2^T\ \mathbf{x}_3^T\ \mathbf{x}_4^T)^T$ ergibt sich schließlich das lineare System

$$j\omega\dot{\mathbf{x}} = \underbrace{\begin{pmatrix} \mathbf{0} & -\mathbf{I} & \mathbf{0} & -\mathbf{A}'_{CC}\mathbf{M}_{\varepsilon\infty}^{-1} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{M}_{\varepsilon d}\mathbf{M}_{\varepsilon\infty}^{-1} & -\alpha_0\mathbf{I} & -\alpha_1\mathbf{I} & \mathbf{0} \\ \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}}_{\mathbf{A}_{l\,\mathrm{disp}}} \mathbf{x} + \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix} \mathbf{i} \tag{3.1.20a}$$

$$\mathbf{u} = (\mathbf{R}^T\ \mathbf{0}\ \mathbf{0}\ \mathbf{0})\ \mathbf{x}. \tag{3.1.20b}$$

Im allgemeinen Fall hat das lineare System also die vierfache Ordnung des Systems nach Gl. 3.1.18. Es ist aber zu beachten, dass die Zustände in $\mathbf{x}_2$ und $\mathbf{x}_3$ nur dort Werte ungleich Null annehmen können, wo $\mathbf{M}_{\varepsilon d}$ ungleich Null ist. Die zu Nullzuständen gehörigen Zeilen des Systems nach Gl. 3.1.20 und die entsprechenden Spalten der Matrix $\mathbf{A}_{l\,\mathrm{disp}}$ können folglich gestrichen werden, was die Ordnung des Systems unter Umständen massiv reduziert, falls nur einzelne Teilgebiete der Struktur dispersive Eigenschaften aufweisen. Ist die Beschreibung des dispersiven Materials nur erster Ordnung, z. B. ein Debye-Modell, wird der Zustandsvektor $\mathbf{x}_2$ vollständig entbehrlich. Dieser Fall wird ausführlich in [98] betrachtet.

Das entsprechende systematische Vorgehen für dispersive magnetische Materialien sowie Modelle höherer Ordnung oder die Kombination mehrerer unterschiedlicher dispersiver Materialien wird ausführlich in [11] beschrieben.

3.1.3.2 System mit PML-Randbedingungen

In der Praxis ebenfalls sehr große Bedeutung haben Strukturen mit *offenen* bzw. perfekt angepassten absorbierenden Rändern. Diese werden beispielsweise zur Berechnung von Antennen benötigt. Wird das absorbierende Material nach der Relation Gl. 2.2.26 gewählt, kann ohne weitere Veränderungen ein in der Frequenz lineares oder ein Curl-Curl-System nach den Gln. 3.1.2 oder 3.1.12 aufgestellt werden. Allerdings sind die Materialparameter innerhalb des PML-Materials nun komplex, was auf ebenfalls komplexe Systemmatrizen führt. Solche Systeme führen zwar auf korrekte Frequenzbereichsergebnisse, haben aber keinen physikalischen Sinn im Zeitbereich.

Um ein gleichermaßen im Frequenz- und im Zeitbereich nutzbares Modell zu erhalten, wird üblicherweise die Materialrelation nach Gl. 2.2.27 verwendet. Dies führt jedoch auf eine kompliziertere Frequenzabhängigkeit des Systems verglichen zu einer ideal berandeten Struktur. In gängigen kommerziellen Frequenzbereichslösern [30] wird dieses Problem meist umgangen, indem für jeden betrachteten Frequenzpunkt die Systemmatrix neu aufgestellt wird. Für den späteren Zweck der Berechnung ordnungsreduzierter Modelle ist es jedoch erforderlich, die Frequenzabhängigkeit des Systems korrekt zu erfassen [56]. Zerlegt man die Materialmatrizen des anisotropen PML-Materials nach Frequenzabhängigkeit in die konstanten Matrizen $\mathbf{M}_\mu$, $\mathbf{M}_\sigma$ und $\mathbf{M}_{\sigma e}$ bzw. $\mathbf{M}_\varepsilon$, $\mathbf{M}_\kappa$ und $\mathbf{M}_{\kappa e}$, lassen sich die Maxwellgleichungen im Fall

reiner Flächen-PML wie folgt formulieren [56]:

$$-\left(\mathbf{I}+\frac{1}{j\omega}\mathbf{M}_{\sigma e}\right)\mathbf{C}\,\widehat{\mathbf{e}} = (j\omega\mathbf{M}_{\mu}+\mathbf{M}_{\sigma})\,\widehat{\mathbf{h}}, \tag{3.1.21a}$$

$$\left(\mathbf{I}+\frac{1}{j\omega}\mathbf{M}_{\kappa e}\right)\mathbf{C}^{T}\,\widehat{\mathbf{h}} = (j\omega\mathbf{M}_{\varepsilon}+\mathbf{M}_{\kappa})\,\widehat{\mathbf{e}}. \tag{3.1.21b}$$

Diese beiden Gleichungen lassen sich erneut in Matrixform bringen, was auf ein System zweiten Grades

$$((j\omega)^2\mathbf{I}+j\omega\mathbf{A}_1+\mathbf{A}_0)\mathbf{x} = j\omega\mathbf{B} \tag{3.1.22a}$$

$$\mathbf{u} = \mathbf{B}^{T}\mathbf{x}. \tag{3.1.22b}$$

mit den Systemmatrizen

$$\mathbf{A}_0 = \begin{pmatrix} \mathbf{0} & -\mathbf{M}_{\varepsilon}^{-1/2}\mathbf{M}_{\kappa e}\mathbf{C}\mathbf{M}_{\mu}^{-1/2} \\ \mathbf{M}_{\mu}^{-1/2}\mathbf{M}_{\sigma e}\mathbf{C}^{T}\mathbf{M}_{\varepsilon}^{-1/2} & \mathbf{0} \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} \mathbf{R} \\ \mathbf{0} \end{pmatrix}, \tag{3.1.23a}$$

$$\mathbf{A}_1 = \begin{pmatrix} \mathbf{M}_{\varepsilon}^{-1/2}\mathbf{M}_{\kappa}\mathbf{M}_{\varepsilon}^{-1/2} & -\mathbf{M}_{\varepsilon}^{-1/2}\mathbf{C}^{T}\mathbf{M}_{\mu}^{-1/2} \\ \mathbf{M}_{\mu}^{-1/2}\mathbf{C}\mathbf{M}_{\varepsilon}^{-1/2} & \mathbf{M}_{\mu}^{-1/2}\mathbf{M}_{\sigma}\mathbf{M}_{\mu}^{-1/2} \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} \mathbf{M}_{\varepsilon}^{-1/2}\,\widehat{\mathbf{e}} \\ \mathbf{M}_{\mu}^{-1/2}\,\widehat{\mathbf{h}} \end{pmatrix}. \tag{3.1.23b}$$

führt. Da der zugehörige Zustandsvektor $\mathbf{x}$ sowohl $\widehat{\mathbf{e}}$ als auch $\widehat{\mathbf{h}}$ als Unbekannte enthält, handelt es sich von der Struktur um kein System in Curl-Curl-Formulierung. Es ist aber ebenfalls zweiten Grades und kann somit auch nach Gl. 3.1.15 in ein lineares System umwandeln lassen. Hierbei weist $\mathbf{A}_1$ zahlreiche Nullspalten auf, welche erneut eliminiert werden können. Die Ordnung wird im Verhältnis zum linearen System daher nicht verdoppelt, sondern nur um die tatsächliche Anzahl neuer Zustände im PML-Medium vergrößert.

Bei zusätzlicher Berücksichtigung von Kanten- und Flächen-PML treten in Gl. 3.1.21 noch höhere Potenzen von $j\omega$ auf, die auf entsprechende Weise linearisiert werden können.

Die Bildung eines Curl-Curl-Systems mit ausschließlich $\widehat{\mathbf{e}}$ als Unbekanntenvektor ist ineffizient, da das Ausmultiplizieren der benötigten Matrix $\mathbf{M}_{\mu}^{-1}(w)$ auf ein Polynom deutlich höheren Grades führt, abhängig von der Anzahl der PML-Schichten.

3.2 Systemeigenschaften

Im vorigen Abschnitt wurden unterschiedliche mathematische Systeme zur Lösung von Feldproblemen für einige wichtige Materialeigenschaften und Anwendungsfälle formuliert. Diese Systeme sind stets konzentriert, haben also eine endliche Zahl von Zustandsvariablen, sind zeitinvariant, d. h. die Systemmatrizen ändern sich nicht mit der Zeit und sind linear (engl. *linear time-invariant, LTI*). *Linear* ist in diesem Zusammenhang nicht auf den Grad des Zustandsraums bezogen, sondern beschreibt die Tatsache, dass die Amplitude des Ausgangssignals $\mathbf{u}$ stets linear von der des Eingangssignals $\mathbf{i}$ abhängt. Im Weiteren sollen weitere wichtige Systemeigenschaften untersucht werden. Diese sind *Kausalität*, *Stabilität*, *Passivität* sowie *Steuer-* und *Beobachtbarkeit*.

3.2.1 Kausalität

Unter Kausalität wird die Eigenschaft eines Systems verstanden, dass die Antwort auf ein Signal nicht vor dessen Beginn startet[5]. Setzt man den Startpunkt auf den Zeitpunkt $t = 0$ muss für die Reaktion $h(t)$

$$h(t) = 0 \quad \text{für} \quad t < 0 \tag{3.2.1}$$

gelten. Für alle existierenden physikalischen Systeme ist die Kausalität stets gegeben. Auch die kontinuierlichen Maxwellgleichungen beschreiben Feldvorgänge als kausales System.

Zu Verletzungen des Kausalitätsprinzips kann es jedoch kommen, wenn das Übertragungsverhalten im Frequenzbereich definiert wird. Ein bekanntes Beispiel stellt im diesem Zusammenhang der ideale Tiefpass dar, dessen Impulsantwort für Zeitpunkte $t < 0$ ungleich Null ist. Da die Methode der Finiten Integration (Gl. 2.2.9) Kausalität grundsätzlich erhält, muss nur gezeigt werden, dass die im Frequenzbereich definierten Übertragungsfunktionen, z.B. konstitutive Materialbeziehungen, ebenfalls kausal sind. Dies kann auf unterschiedliche Arten erfolgen: Es zeigt sich, dass Real- und Imaginärteil der Frequenzrepräsentation eines kausalen Systems nicht unabhängig voneinander sind, sondern in fester Beziehung zueinander stehen. Das Verhältnis kann für dispersive Materialien mit Hilfe der Kramers-Kronig-Relationen [2], im allgemeinen Fall mit der verwandten Hilbert-Transformation [34] überprüft werden. Sofern die Fourierrücktransformierte $f(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(j\omega)\, e^{j\omega t}\, d\omega$ einfach bestimmbar ist, kann diese mit Gl. 3.2.1 direkt auf Kausalität überprüft werden.

Bisher wurde in drei Fällen ein Verhalten im Frequenzbereich definiert, deren Kausalität an dieser Stelle gezeigt werden soll:

- **dispersive Materialien:** Es lässt sich zeigen [2], dass die Kramers-Kronig-Bedingungen erfüllt sind. Dies wird auch insbesondere anschaulich, da die in Gln. 2.1.8 - 2.1.10 beschriebenen Zusammenhänge durch einfache RLC-Schwingkreise modelliert werden können [11].
- **Oberflächenimpedanzen:** Hier findet sich die direkte Fourier-Transformationskorrespondenz mit der Sprungfunktion $g(t)$:

$$\frac{K}{\sqrt{\pi t}}\, g(t) \quad \circ\!\!-\!\!\bullet \quad \frac{K}{\sqrt{j\omega}}. \tag{3.2.2}$$

- **PML-Randbedingungen:** Wie bereits in Abschnitt 2.2.3.4 angesprochen, hängt die Kausalität von der Wahl des Faktors $c_{u,v}$ in Gl. 2.2.24 ab. Für eine Wahl nach Gl. 2.2.27 wurde die Kausalität in [25] gezeigt.

Es bleibt zu beachten, dass bei reinen Frequenzbereichsbetrachtungen auch nichtkausale Systeme wie Gl. 2.2.26 physikalisch sinnvolle Ergebnisse liefern können. Im Falle transienter Simulationen ist Kausalität jedoch eine entscheidende Voraussetzung.

[5] Kausalität wird an dieser Stelle im Sinne eines Netzwerks interpretiert und berücksichtigt keine Laufzeiten und Ausbreitungsgeschwindigkeiten.

3.2.2 Stabilität

Eine elementare Eigenschaft aller Systeme ist die Stabilität. Dabei werden grundsätzlich zwei Arten unterschieden:

- Die **Übertragungsstabilität** fordert, dass das Ausgangssignal eines entspannten Systems (d. h. $\mathbf{x}(t=0)=\mathbf{0}$) im Fall einer endlichen Anregung stets endlich bleibt. Häufig wird auch von BIBO-Stabilität (engl.: *bounded input-bounded output*) gesprochen. Für die Impulsantwort $h(t)$ gilt folglich:

$$\int_0^\infty |h(t)|\,dt \leq M < \infty. \tag{3.2.3}$$

- Die **interne Stabilität** untersucht das Systemverhalten im anregungsfreien Fall $\dot{\mathbf{x}}(t)=\mathbf{A}\mathbf{x}(t)$. Ausgehend von einem beliebigen endlichen Anfangszustand $\mathbf{x}(t_0)$ aller inneren Variablen wird das Verhalten für große Zeiten $t\to\infty$ untersucht. Klingen hierbei alle Zustandsgrößen auf Null ab, wird von *asymptotischer* Stabilität gesprochen. Konvergieren ein oder mehrere Zustände nicht gegen Null, bleiben aber für alle Zeitpunkte mit $|x_i|\leq M_i<\infty$ begrenzt, wird dies als *marginale* Stabilität bezeichnet. Die interne Stabilität ist die tiefergreifende Forderung, da auch Instabilitäten aufdeckt werden, die durch das Eingangssignal möglicherweise nicht angeregt werden.

Durch die Superposition beider Fälle lässt sich jedes System unter beliebiger Anregung auf Stabilität überprüfen.

3.2.2.1 Übergang zur komplexen Laplace-Ebene

Zur Untersuchung der Stabilität und später auch der Passivität eines Systems ist es vorteilhaft, die Kreisfrequenzvariable ω zu einem komplexen Frequenzparameter s, auch als *Laplace-Variable* bezeichnet[6], zu erweitern

$$s = j\omega + \sigma\,, \tag{3.2.4}$$

um zur Systemanalyse auf funktionentheoretische Ansätze zurückgreifen zu können. Formell bedeutet dies den Übergang von der Fourier- zur Laplace-Transformation mit den korrespondierenden Transformationsvorschriften [34]:

$$F(s) = \int_0^\infty f(t)\,e^{-st}\,\mathrm{d}t \quad \bullet\!\!-\!\!\circ \quad f(t) = \frac{1}{2\pi j}\int_{\sigma-j\infty}^{\sigma+j\infty} F(s)\,e^{st}\,\mathrm{d}s. \tag{3.2.5}$$

[6]Formell entspricht f der Frequenz, ω der Kreisfrequenz und s der Laplace-Variablen. Solange jedoch keine Verwechslungsgefahr besteht, werden im Folgenden der Einfachheit halber alle als Frequenz bezeichnet.

Für diese *einseitige* Laplace-Transformation mit der unteren Integrationsgrenze Null in Gl. 3.2.5 wird unmittelbar die Kausalität des Zeitsignals vorausgesetzt.

Sofern eine (Fourier-)Frequenzbeschreibung ebenfalls eine kausale Zeitfunktion repräsentiert, können in vielen Fällen s und $j\omega$ direkt ausgetauscht werden. Es ist aber zu beachten, dass dies nur gültig ist, wenn die imaginäre Achse $j\omega$ *innerhalb* der Konvergenzebene von $F(s)$ liegt. Diese Forderung entspricht im Zeitbereich der Bedingung, dass $f(t)$ quadratisch integrabel ist. Ein Gegenbeispiel liefert die Funktion $x(t) = \sin(t)\, g(t)$, deren Fouriertransformierte zwei Polstellen auf der imaginären Achse aufweist, wobei die $j\omega$-Achse zugleich den Rand der Konvergenzebene bildet. In diesem Fall lassen sich Laplace- und Fourier-Transformation nicht ohne weiteres ineinander überführen. Für eine genauere Betrachtung dieser Sonderfälle sei erneut auf [34] verwiesen. Für alle im Rahmen dieser Arbeit betrachteten Frequenzbereichsrepräsentationen ist der direkte Austausch von $j\omega$ und s möglich.

Nach Übergang in den komplexen Frequenzbereich lassen sich beide oben genannten Arten von Stabilität direkt durch die Lage der Polstellen untersuchen. Die Übertragungsfunktion Gl. 3.1.4 lässt sich mit der adjungierten Matrix [42] wie folgt schreiben:

$$\mathbf{Z}(s) = \frac{1}{\det(s\mathbf{I} + \mathbf{A})}\mathbf{C}[\mathrm{Adj}(s\mathbf{I} + \mathbf{A})]\mathbf{B}. \tag{3.2.6}$$

Es wird offensichtlich, das jeder Pol von $\mathbf{Z}(s)$ ein Eigenwert von $-\mathbf{A}$ ist. Stabilität lässt sich damit folgendermaßen definieren [34, 35]

- Ein System ist intern asymptotisch stabil, wenn alle Eigenwerte von $-\mathbf{A}$ einen negativen Realteil aufweisen[7].
- Das System ist intern marginal stabil, wenn alle Eigenwerte von $-\mathbf{A}$ einen Realteil kleiner oder gleich Null besitzen. Die Eigenwerte auf der imaginären Achse müssen zudem einfache Nullstellen des charakteristischen Polynoms $\det(s\mathbf{I} + \mathbf{A})$ sein.
- Ein System ist übertragungsstabil, wenn alle Eigenwerte von $-\mathbf{A}$ (Nullstellen des Nenners), die nicht durch Kürzung mit Nullstellen des Zählers wegfallen, einen negativen Realteil haben. Kürzen sich keine Nullstellen (das System wird dann steuerbar genannt), entspricht dies asymptotischer Stabilität.

3.2.2.2 Eigenwerte in FIT-Systemen

Grundlegende Aussagen über die Lage der Pole des Systems, bzw. der Eigenwerte der Systemmatrix, lassen sich direkt aus der Matrixstruktur ableiten. So handelt es sich bei der Systemmatrix $\mathbf{A}_l$ des linearen Systems im verlustfreien Fall um eine schiefsymmetrische reelle Matrix, wobei die Pole über $s_i = -\lambda_i$ mit den Eigenwerten zusammenhängen. Es lässt sich leicht zeigen [42], dass eine solche Matrix nur

[7] Da nach [36] auch kausale Systeme keine Pole in der rechten Halbebene haben dürfen, wird in [25] aus der Kausalität des PML-Mediums nach Gl. 2.2.27 auf dessen Stabilität geschlossen.

rein imaginäre, konjugiert komplexe Eigenwerte enthält. Zudem zählen schiefsymmetrische Systeme zur Klasse der diagonalisierbaren Matrizen, zu einem n-fachen Eigenwert gehören damit stets n zueinander linear unabhängige Eigenvektoren. Jeder n-fache Eigenwert geht damit nur als einfache Nullstelle in das charakteristische Polynom der Matrix ein, was gleichbedeutend mit der Forderung ist, dass Systempole auf der imaginären Achse nur einfach sein dürfen.

Verlustfreie Systeme sind somit garantiert intern marginal stabil, jedoch nicht generell übertragungsstabil, da die Anregung mit beispielsweise einem harmonischen Signal, dessen Frequenz einer Resonanzfrequenz auf der imaginären Achse entspricht, auf ein unbegrenztes Ausgangssignal führt.

Addiert man zur schiefsymmetrischen Matrix eine positiv semidefinite Diagonalmatrix wie im verlustbehafteten System, verschiebt das die Eigenwerte der Matrix ausschließlich in die rechte s-Halbebene [42], was wiederum Polen in ausschließlich der linken Halbebene entspricht.

Entsprechende Aussagen lassen sich auch für die Curl-Curl-Formulierung finden. In diesem Fall gilt für die Systempole $s_i = \sqrt{-\lambda_i}$. Wird die symmetrische Systemmatrix $\mathbf{A}_{CC}$ wie folgt umformuliert,

$$\mathbf{A}_{CC} = (\mathbf{M}_\varepsilon^{-1/2}\widetilde{\mathbf{C}}\mathbf{M}_\mu^{-1/2})(\mathbf{M}_\varepsilon^{-1/2}\widetilde{\mathbf{C}}\mathbf{M}_\mu^{-1/2})^T \tag{3.2.7}$$

zeigt sich unmittelbar, dass diese positiv semi-definit [10, 31] ist, also nur Eigenwerte $\lambda_i \geq 0$ besitzt. Da auch symmmetrische Matrizen diagonalisierbar sind, führt auch diese Formulierung stets auf imaginäre konjugiert komplexe Pole, die nur einfach auftreten.

Wie bereits im Rahmen der Kausalität festgestellt, gilt auch für instabile Systeme, dass Frequenzbereichsrechnungen trotz gewisser Instabilitäten sinnvoll sein können. Dies gilt beispielsweise, wenn allein das Übertragungsverhalten für einzelne Frequenzen gesucht ist, das System aber aufgrund der numerischen Umformungen (nicht physikalisch bedingte) Instabilitäten aufweist. Zeitbereichssimulationen führen jedoch auch bei geringen Instabilitäten meist zu unbrauchbaren Ergebnissen.

Wenn auch für die Stabilität nicht von Relevanz, ist für die effiziente Implementierung numerischer Löser eine Klassifizierung der auftretenden Eigenwerte interessant. Grundsätzlich wird zwischen *statischen* und *dynamischen* Moden unterschieden [31].

Die statischen Moden haben einen Eigenwert von Null und treten aufgrund der denkbaren Ladungen im Rechengebiet auf. Ihre Anzahl ist in Curl-Curl-Formulierung folglich auf die Anzahl der nicht innerhalb leitfähiger Gebiete liegenden Gitterpunkte begrenzt, es gilt $N_S \approx N_P$ mit der Gitterpunktzahl N_P. Die zugehörigen Felder, die durch die Eigenwerte repräsentiert werden, sind rotationsfrei und können daher leicht detektiert werden, da $\mathbf{C}\,\widehat{\mathbf{e}}_S = 0$ gilt.

Die dynamischen Moden hingegen sind divergenzfrei und lassen sich durch $\widetilde{\mathbf{S}}\,\widehat{\widetilde{\mathbf{d}}}_D = 0$ erkennen. Ihre Anzahl entspricht etwa $N_D \approx 2N_P$. Eine Kombination der beiden beschriebenen Fälle bilden die so genannten *Mehrleitermoden*, die die Felder zwischen leitenden, voneinander isolierten Körpern beschreiben. Diese Felder sind sowohl rotations- als auch divergenzfrei. Ihre Anzahl entspricht der isolierter leitfähiger Körper minus Eins.

Da für die Beschreibung des dynamischen Verhaltens eines Systems allein die dynamischen Moden ausschlaggebend sind, wird deutlich, dass die Dimension des mathematischen Systems um rund ein Drittel gesenkt werden könnte, ohne das Übertragungsverhalten einer Struktur zu verändern. Dies bildet damit einen ersten Ansatz, ein ordnungsreduziertes Modell zu erzeugen, auch wenn ein resultierendes Modell der Größe $\approx 2N_P$ noch nicht wirklich niedriger Ordnung wäre. Ein auf einer *Tree-Cotree-Eichung* beruhender Ansatz wird hierzu in [27] beschrieben. Auch sogennannte Block-Jakobi-Vorkonditionierer [47] zur iterativen Lösung des Systems basieren auf dieser Idee. Bei der iterativen Berechnung der technisch interessanten niederfrequentesten dynamischen Eigenwerte, die ebenfalls durch die Existenz der statischen Moden erschwert wird, behilft man sich wiederum mit einem so genannten *Grad-Div-Term* [10], der die statischen Eigenwerte in höhere Frequenzbereiche verschiebt.

3.2.3 Passivität von Impedanzfunktionen

Neben der Stabilität kommt der Passivität bei Systembetrachtungen eine bedeutende Rolle zu. Unter Passivität wird im allgemeinen verstanden, dass ein System nicht mehr Energie abgeben kann, als es zuvor aufgenommen hat, es also keine Energie generiert. Für die Energie $w(t)$ gilt folglich mit den Portvektoren $\mathbf{u}$ und $\mathbf{i}$ der Zusammenhang

$$w(t) = \int_{-\infty}^{t} \mathbf{u}^T(\tau)\mathbf{i}(\tau)\, d\tau \geq 0 \tag{3.2.8}$$

für jeden beliebigen Zeitpunkt t. Dieser Eigenschaft kommt eine besondere Bedeutung bei der Verknüpfung von Schaltgruppen zu. Während bei der Zusammenschaltung zweier stabiler Bauelemente die Stabilität des Gesamtsystems nicht generell garantiert werden kann, ist das Ergebnis einer Verschaltung zweier passiver Systeme grundsätzlich ebenfalls passiv. Da Passivität, wie unten gezeigt wird, auch Stabilität beinhaltet, kann somit die Stabilität des Gesamtsystems in jedem Fall erreicht werden. Ein einfaches Beispiel für eine instabiles System aus der Verschaltung zweier stabiler Schaltungen bildet ein invertierender Verstärker mit sehr hoher Verstärkung $K \to \infty$ in Verbindung mit einem stabilen aber nicht phasenminimalen Zweipol $H(s)$, der Nullstellen in der rechten Halbebene aufweist. Die resultierende Übertragungsfunktion ergibt sich zu $1/H(s)$ und ist folglich trotz der Stabilität beider Komponenten instabil.

Auch im Rahmen transienter Simulationen bietet die Passivität des diskreten Modells entscheidende Vorteile. Werden unterschiedliche numerische Verfahren, z.B. zur Berechnung linearer und nichtlinearer Teilbereiche, auf Simulatorebene verkoppelt, ist Passivität eine hinreichende Bedingung für die Vermeidung numerischer Instabilitäten [37].

Da Passivität eine Eigenschaft ist, die von den meisten elektromagnetischen Bauteilen in der Realität erfüllt wird, ist es folglich wünschenswert, diese Eigenschaft auch im diskretisierten Modell und darüber hinaus auch in einem Makromodell reduzierter Ordung zu erhalten. Für idealberandete FIT-Systeme wurde die grundsätzliche

Passivität bereits in [32] gezeigt, sie soll wegen Ihrer Bedeutung im Rahmen dieser Arbeit aber in einem allgemeineren Rahmen dargestellt werden.

Betrachtet man die Forderung Gl. 3.2.8 im Frequenzbereich, lautet die äquivalente Passivitätsbedingung $\mathrm{Re}\{\mathbf{u}^*\mathbf{i}\} = 0.5\,(\mathbf{i}^*(\mathbf{Z}+\mathbf{Z}^*)\mathbf{i}) \geq 0$. Wird einer der Vektoren $\mathbf{u}(t)$ oder $\mathbf{i}(t)$ vorgegeben und sind alle seine Einträge quadratisch integrabel, muss auch das Integral über die Energie endlich bleiben, folglich muss auch die abhängige Größe quadratisch integrabel sein. Aus den Eigenschaften der Laplace-Transformation folgt damit, dass $\mathbf{Z}(s)$ keine Pole in der rechten Halbebene $\mathrm{Re}\{s\} > 0$ haben darf, Passivität beinhaltet damit folglich Stabilität. Als weitere Bedingung ergibt sich, dass im Zeitbereich die Impulsantwort der Impedanz reell sein muss. Dies erfordert, dass auch $\mathbf{Z}(s)$ für reelle Werte von s ebenfalls reell bleiben muss.

Diese Bedingungen für die Passivität einer Impedanz $\mathbf{Z}(s)$ lassen sich damit durch die folgenden drei Punkte notwendig wie hinreichend definieren:

- Jedes Element von $\mathbf{Z}(s)$ ist analytisch für $\mathrm{Re}\{s\} > 0$. (3.2.9a)
- $\mathbf{Z}(s^*) = \mathbf{Z}^*(s)$ für $\mathrm{Re}\{s\} > 0$. (3.2.9b)
- $\mathbf{Z}^H(s) + \mathbf{Z}(s) \geq 0$ für $\mathrm{Re}\{s\} > 0$. (3.2.9c)

Eine Matrix, die diese Forderungen erfüllt, wird auch positiv reell genannt. Das Konzept wurde erstmals von O. Brune 1930 aufgestellt. Eine genaue funktionentheoretische Herleitung dieser Beziehungen sowie ausführliche Beschreibungen der Eigenschaften, die sich damit für positiv reelle Systeme ergeben, sind in [36, 38] angegeben. Insbesondere einige der Eigenschaften in [39] sind für die Untersuchung der Passivität für die oben hergeleiteten Systeme von entscheidender Bedeutung:

- Theorem 1: Ist $\mathbf{V}$ eine reelle konstante $m \times n$-Matrix und $\mathbf{G}(s)$ eine positiv reelle $m \times m$-Matrix, dann ist auch $\mathbf{V}^T\mathbf{G}(s)\mathbf{V}$ eine positiv relle $n \times n$-Matrix.
- Theorem 2: Sind $\mathbf{F}(s)$ und $\mathbf{G}(s)$ positiv relle Matrizen, so ist auch $\mathbf{F}(s)+\mathbf{G}(s)$ positiv reell.
- Theorem 3: Wenn $\mathbf{G}(s)$ positiv reell und $\mathbf{G}^H(s)+\mathbf{G}(s) > 0$ für $\mathrm{Re}\{s\} > 0$ ist, dann existiert $\mathbf{G}^{-1}(s)$ und ist ebenfalls positiv reell.

Solange sich ein FIT-diskretisiertes System in der Form

$$\mathbf{Z}(s) = \mathbf{B}^T(\mathbf{Y}_1(s) + \mathbf{Y}_2(s) + \ldots)^{-1}\mathbf{B} \tag{3.2.10}$$

darstellen lässt[8], kann die Passivität hinreichend getestet werden, indem alle $\mathbf{Y}_k$ auf positive Reellheit überprüft werden. Die ersten beiden Bedingungen sind hierbei üblicherweise direkt erfüllt. Es bleibt folglich nur die dritte Bedingung zu überprüfen. Da bei der Standard-FIT-Formulierung alle Materialmatrizen diagonal sowie positiv semi-definit sind, darüberhinaus aufgrund der Symmetrien auch $\mathbf{A}_{CC}^H + \mathbf{A}_{CC} \geq 0$,

[8]Streng genommen ist dies aufgrund der Korrekturmatrix $\mathbf{D}_\beta$ in Gl. 3.1.10 nicht möglich. Für eine genügend feine Diskretisierung gilt allerdings $\mathbf{D}_\beta \approx \mathbf{I}$. Im Fall einer symmetrischen Systemmatrix $\mathbf{A}$ kann mit $\mathbf{D}_\beta^{1/2}$ auch eine symmetrische Korrektur angewendet werden.

$\mathbf{A}'^{H}_{CC} + \mathbf{A}'_{CC} \geq 0$ und $\mathbf{A}^{H}_{l} + \mathbf{A}_{l} \geq 0$ gilt, reduziert sich die Untersuchung auf die Passivität der speziellen Abhängigkeit der Laplace-Variablen $s = j\omega + \sigma$ innerhalb der Materialdefinition. Im Curl-Curl-Fall Gl. 3.1.12 ergibt sich beispielsweise

$$\mathbf{Y}(s)^H + \mathbf{Y}(s) = \left(s\mathbf{I} + \mathbf{K} + \frac{1}{s}\mathbf{A}_{CC}\right)^H + \left(s\mathbf{I} + \mathbf{K} + \frac{1}{s}\mathbf{A}_{CC}\right) \tag{3.2.11a}$$

$$= 2\left(\sigma\mathbf{I} + \mathbf{K} + \frac{\sigma}{\sigma^2 + \omega^2}\mathbf{A}_{CC}\right) \geq 0 \quad \text{für} \quad \mathrm{Re}\{s\} > 0. \tag{3.2.11b}$$

Genauso kann zur Überprüfung der Passivität dispersiver Systeme nach Gl. 2.1.8 - 2.1.10 vorgegangen werden, indem die Positivität von $s\,\mathbf{M}_\varepsilon(s)$ nach Gl. 2.2.14 getestet wird. Es zeigt sich auch hier, dass es sich um passive Systeme handelt.

Auch für Systeme mit Flächen-PML-Randbedingungen kann die Passivität ohne weiteres gezeigt werden. Wird jedoch das Kanten-PML-Material in der Überschneidungszone zweier Flächen-PML betrachtet, ergibt sich nach Gl. 2.2.29 ein Materialtyp mit der Abhängigkeit $c_{uw} = c_u c_w = (1 + \sigma_u/s)(1 + \sigma_w/s)$. Für den Grenzwert $\sigma \to 0$ ergibt sich

$$c^*_{uw} + c_{uw} = 2 - \frac{2\sigma_u\sigma_w}{\omega^2}. \tag{3.2.12}$$

Das Material ist also nur passiv, wenn $\sigma_u\sigma_w < \omega^2$ gilt. Da in praktischen Anwendungen die Werte des Dämpfungsprofils σ durchaus in der Größenordnung von ω liegen, können tatsächlich aktive Elemente auftreten. Da im Allgemeinen jedoch das dämpfende Verhalten der übrigen Komponenten überwiegen wird, ist davon auszugehen, dass sich das System dennoch passiv verhält, zumal die Überprüfung nach Gl. 3.2.10 nur eine hinreichende, nicht aber notwendige Bedingungen darstellt. Die Passivität müsste folglich im Einzelfall durch die Überprüfung der Positivität der Eigenwerte der Matrix $\mathbf{A}^H_1 + \mathbf{A}_1$ nach Gl. 3.1.23b erfolgen. Aufgrund der Matrixgröße wird dies im Allgemeinen allerdings schwer realisierbar sein.

Auch die Passivität der Impedanzwände mit der Abhängigkeit $1/\sqrt{s}$ in Gl. 2.2.19 kann gezeigt werden, nicht jedoch die von $1/(s\sqrt{s})$ der Gln. 2.2.20c bzw. 3.1.14, da diese einen doppelten Pol bei $s = 0$ aufweist. Dies beruht auf der angewandten Näherung von Gl. 2.2.21 nach Gl. 2.2.22. Auch hier ist jedoch davon auszugehen, dass die Passivität dennoch erhalten ist. Zur Erstellung eines Ersatzschaltbilds muss der Zusammenhang $1/(s\sqrt{s})$ ohnehin durch eine passive rationale Funktion approximiert werden.

Abschließend lässt sich sagen, dass die FI-Technik aufgrund des Zusammenhangs $\widetilde{\mathbf{C}}^T = \mathbf{C}$ und positiv semi-definiten Materialmatrizen in den meisten Fällen auf positiv reelle und folglich passive Systeme führt. Dies ist keineswegs eine Selbstverständlichkeit und wird von zahlreichen anderen Simulationsverfahren, wie beispielsweise einigen FDTD-Erweiterungen (z. B. bestimmten Untergittern) nicht erfüllt.

3.2.4 Steuerbarkeit und Beobachtbarkeit

Weitere in der Regelungstechnik sehr bedeutende Systemeigenschaften sind die Steuerbarkeit und Beobachtbarkeit interner Zustände. Unter Steuerbarkeit wird anschau-

lich verstanden, dass ein System von einem beliebigen Anfangszustand $\mathbf{x}_0$ über ein endliches Eingangssignal $\mathbf{i}(t)$ in einen definierten Endzustand $\mathbf{x}_e$ gebracht werden kann. Unterschiedliche Möglichkeiten, die Steuerbarkeit nachzuweisen, bieten die folgenden Punkte [35]:

1. Die $n \times np$ Steuerbarkeitsmatrix $\mathcal{C} = [\mathbf{B}, \mathbf{AB}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{n-1}\mathbf{B}]$ hat den vollen Zeilenrang n.
2. Die $n \times (n+p)$ Matrix $[\mathbf{A} - \lambda_k\mathbf{I}, \mathbf{B}]$ hat für jeden Eigenwert λ_k der Matrix $\mathbf{A}$ vollen Rang.
3. Die Gramsche Steuerbarkeitsmatrix $\mathbf{W}_C$ mit $\mathbf{AW}_C + \mathbf{W}_C\mathbf{A}^T = -\mathbf{BB}^T$ ist positiv definit.

Die Erfüllung eines Punktes bedingt auch die Erfüllung aller übrigen. Ein System wird zudem beobachtbar genannt, wenn bei Kenntnis aller Eingangs- $\mathbf{i}(t)$ und aller Ausgangssignale $\mathbf{u}(t)$ eindeutig auf den Anfangszustand $\mathbf{x}_0$ zurückgeschlossen werden kann. Aufgrund einer Dualitätsbeziehung kann die Beobachtbarkeit gezeigt werden, indem die Steuerbarkeit des Matrizenpaars $(\mathbf{A}^T, \mathbf{C}^T)$ nachgewiesen wird [35].

Ist ein System sowohl voll steuerbar als auch voll beobachtbar, enthält es keine entbehrlichen Zustände und wird auch als *minimale Realisierung* bezeichnet. Wie bereits in Abschnitt 3.2.2.2 im Rahmen der Eigenwertbetrachtung dargestellt, enthält die Systemmatrix den $\sim N_P$-fachen Eigenwert 0. Dies widerspricht folglich der obigen Bedingung 2, weswegen FIT-Systeme grundsätzlich keine minimalen Realisierungen darstellen. Nach Anwendung einer *Tree-Cotree-Eichung* kann das System jedoch als minimale Realisierung aufgefasst werden. Auch Modelle reduzierter Ordnung, auf deren Berechnung im nächsten Kapitel eingegangen wird, sollten stets minimale Realisierungen darstellen. Das Reduzierungsverfahren *Balanced Realization* nutzt direkt die Steuer- und Beobachtbarkeits-Gramians aus Punkt 3 und erzeugt ein reduziertes Modell allein aus den am Besten erreichbaren Zuständen. Verfahren, die auf Krylov-Unterräumen basieren, besitzen wiederum eine enge Verbindung zur Steuerbarkeitsmatrix in Punkt 1.

3.3 Streuparameter

3.3.1 Grundlagen

Alle bisherigen Betrachtungen bezogen sich stets auf Impedanzmatrizen, die verallgemeinerte oder auch tatsächliche Ströme und Spannungen miteinander verkoppeln. In der Praxis der Hochfrequenztechnik kommt den Streuparametern S, die die Wellenamplituden a und b nach

$$S_{ij} = \left.\frac{b_i}{a_j}\right|_{a_k=0 \,\forall\, k\neq j} \tag{3.3.1}$$

in Relation setzen, jedoch häufig eine größere Bedeutung zu. Dies basiert zum Großteil auf der messtechnischen Notwendigkeit, wonach sich ideale Leerläufe oder Kurzschlüsse, wie zur Bestimmung der Impedanz bzw. Admittanz notwendig, für Frequenzen im Mikrowellenbereich aufgrund von Abstrahlung und parasitäten Effekten nur schwer realisieren lassen. Der zur Bestimmung der Streuparameter erforderliche reflexionsfreie Abschluss ist hingegen messtechnisch weit einfacher zu erzielen. Der Umstand, dass zur Berechnung der Streuparameter ein offenes System verwendet wird, ist auch für Zeitbereichssimulationen ein entscheidender Vorteil, da einmal ins Rechengebiet eingespeiste Energie auch im verlustfreien Fall über die Ports entweichen kann. Für Frequenzbereichsrechnungen bietet jedoch die Impedanzdarstellung günstigere Matrixeigenschaften und wird daher häufig bevorzugt. Ein weiterer Vorteil bei Verwendung von Streumatrizen besteht darin, dass diese in jedem Fall existieren, während dies für Impedanz- oder Admittanzmatrizen nicht grundsätzlich gewährleistet ist.

In Matrixschreibweise für alle Ports ergibt sich für die Streu- oder kurz S-Matrix

$$\mathbf{b} = \mathbf{S}\,\mathbf{a}. \tag{3.3.2}$$

Die S-Matrix weist zahlreiche charakteristische Eigenschaften auf, die im Folgenden kurz zusammengefasst werden sollen:

- Der Betrag jedes Eintrags der S-Matrix kann für passive Systeme Werte zwischen Null und Eins annehmen. Die Diagonaleinträge S_{kk} repräsentieren den Reflexionsfaktor des Ports k, die übrigen Elemente S_{kl} die Transmission von Port l zu Port k.
- Im üblichen reziproken Fall ist die S-Matrix symmetrisch, es gilt folglich $\mathbf{S}^T = \mathbf{S}$.
- Die S-Matrix verlustfreier Systeme ist zusätzlich unitär $\mathbf{S}^T = \mathbf{S}^{-1}$. Als Energiebilanz ergibt sich daraus für die l. Spalte $\sum_k |S_{kl}|^2 = 1$. Im Zweiportfall gilt darüberhinaus $|S_{11}| = |S_{22}|$.

Äquivalent zu positiv reellen Matrizen, die die Passivität von Impedanz- oder Admittanzmatrizen mathematisch erfassen, lässt sich die Passivität von Streumatrizen durch die Theorie *begrenzt reeller* Matrizen beschreiben. Während die Reellwertigkeits- und Stabilitätsforderung erhalten bleiben, wird die Positivitätsbedingung (Gl. 3.2.9c) durch die Begrenzungsvorschrift

$$\mathbf{I} - \mathbf{S}^H(s)\mathbf{S}(s) \geq \mathbf{0} \quad \text{für} \quad \mathrm{Re}\{s\} > 0 \tag{3.3.3}$$

ersetzt [39].

Der klassische Weg, S-Parameter in einer Feldsimulation zu bestimmen, erfolgt über die Analyse der elektrischen oder magnetischen Feldkomponenten in zwei unterschiedlichen Gitterebenen. Da die Signallaufzeit $e^{-jk_w\,\Delta w}$ aus der Ausbreitungskonstanten des Wellenleiterrands bekannt ist, können die Wellenanteile a_k und b_k nach Gl. 2.2.31a bestimmt werden, siehe auch Abb. 3.2. Während dieser Ansatz

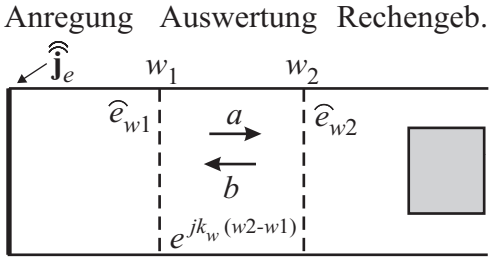


Abbildung 3.2: *Bestimmung der Wellenamplituden a und b durch Auswertung des elektrischen Feldes in zwei unterschiedlichen Gitterebenen.*

bei Zeitbereichssimulationen sehr erfolgreich verwendet wird, führt er bei Frequenzbereichsrechnungen (mit abgeschlossenen Ports) häufig zu numerischen Problemen aufgrund schlechter Konditionierung. Es erweist sich daher häufig als robuster, die S-Parameter aus der Impedanz zu bestimmen.

3.3.2 Streuparameter aus Impedanzmatrizen

Sofern die Impedanzmatrix existiert, kann die S-Matrix direkt daraus berechnet werden. Mit der Referenzimpedanzmatrix der Anschlüsse $\mathbf{Z}_L$ lässt sich allgemein die normierte Impedanz $\bar{\mathbf{Z}}$ bestimmen

$$\bar{\mathbf{Z}} = \mathbf{Z}_L^{-1/2}\, \mathbf{Z}_r\, \mathbf{Z}_L^{-1/2}. \tag{3.3.4}$$

Der Zusammenhang für die Streumatrix $\mathbf{S}$ ergibt sich damit zu

$$\mathbf{S} = \left(\bar{\mathbf{Z}} - \mathbf{I}\right)\left(\bar{\mathbf{Z}} + \mathbf{I}\right)^{-1}, \tag{3.3.5a}$$

$$= \mathbf{I} - 2\left(\bar{\mathbf{Z}} + \mathbf{I}\right)^{-1}. \tag{3.3.5b}$$

Im Gegenzug lässt sich auch die Impedanz aus den Streuparametern berechnen:

$$\mathbf{Z}_r = \mathbf{Z}_L^{1/2}\left(\mathbf{I} + \mathbf{S}\right)\left(\mathbf{I} - \mathbf{S}\right)^{-1}\mathbf{Z}_L^{1/2}. \tag{3.3.6}$$

Sollen die Streuparameter direkt aus der normierten Impedanz nach Gl. 3.1.8 bestimmt werden, sind die Referenzimpedanzen bereits Eins, es muss aber zusätzlich der Gitterkorrekturterm $\mathbf{D}_\beta$ berücksichtigt werden, was zu

$$\mathbf{S} = \left(\bar{\mathbf{Z}} - \mathbf{D}_\beta^{-1}\right)\left(\bar{\mathbf{Z}} + \mathbf{D}_\beta^{-1}\right)^{-1}. \tag{3.3.7}$$

führt. Es ist beachtenswert, dass bei der Bestimmung der S-Parameter über den Impedanzansatz demnach die Portimpedanzen an keiner Stelle explizit benötigt werden, sondern implizit durch die Normierung der Portmoden einfließen.

Da die numerische Berechnung der Impedanz nach der Methode der Finiten Integration niemals fehlerfrei erfolgen kann, ist es erforderlich zu überprüfen, wie sich der Fehler der Impedanzdarstellung auf die Streuparameter fortpflanzt. Dies erfolgt durch Bildung des absoluten Integrals in Gl. 3.3.5 bzw. Gl. 3.3.7. Besonders anschaulich ist die Fehlerbetrachtung im Einportfall. Für den absoluten Fehler ΔS des S-Parameters gilt mit dem relativen Fehler $\delta Z = \Delta\bar{Z}/\bar{Z}$ der normierten Impedanz

$$\Delta S = \left|\frac{2\,\Delta\bar{Z}}{(1+\bar{Z})^2}\right| = \left|\frac{2\,\delta\bar{Z}\cdot\bar{Z}}{(1+\bar{Z})^2}\right|. \tag{3.3.8}$$

Es ist leicht zu zeigen, dass ΔS damit sein Maximum für den reellen Widerstand $\bar{Z} = 1$ und damit für $Z_r = Z_L$ annimmt, wobei $\Delta S = \delta Z$ gilt, der absolute Fehler des Streuparameters also dem relativen Fehler der Impedanz entspricht. Für alle übrigen Werte von $\bar{Z}$ ist der Fehler ΔS stets kleiner als $\delta\bar{Z}$. Der absolute Fehler ist für drei Impedanzen auch in Abb. 3.3 dargestellt. Für große Werte von $\bar{Z}$, z. B. in der Nähe von Singularitäten, gilt gar $\Delta S \approx 2\,\delta Z\ /\bar{Z}$. Da insbesondere in der Nähe von Polstellen größere Fehler in der Impedanz auftreten, werden diese bei der Umrechnung in Streuparameter deutlich verringert.

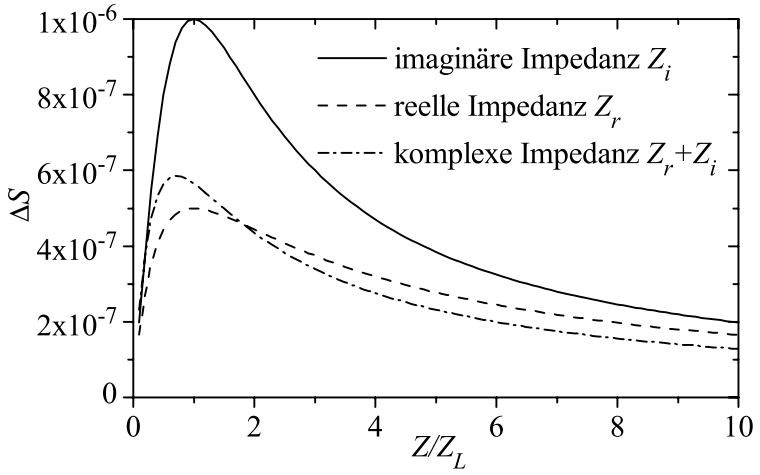


Abbildung 3.3: *Der absolute Fehler des Streuparameters S für drei Impedanzen Z: eine rein imaginäre, eine rein reelle und eine komplexe mit betragsgleichem Real- und Imaginärteil. Der relative Fehler der Impedanz δZ beträgt dabei jeweils* 10^{-6}.

Teilt man Gl. 3.3.8 durch den Wert von S, dessen Betrag zwischen 0 und 1 liegt, ergibt sich der relative Fehler δS zu

$$\delta S = \left| \frac{2\,\delta\bar{Z} \cdot \bar{Z}}{-1 + \bar{Z}^2} \right|. \tag{3.3.9}$$

Für kleine Beträge von S kann der Fehler δS folglich größere Werte als ΔS annehmen, für $S = 0$ bzw. $\bar{Z} = \vec{Z}_L$ sogar Unendlich. Im in der Praxis bedeutsamen Fall verlustarmer Systeme mit nahezu rein imaginären Werten von $\bar{Z}$ gilt jedoch $\|S\| \approx 1$ und damit $\delta S \approx \Delta S \leq \delta\bar{Z}$.

Im Mehrportfall ist die Fehlerbetrachtung weit aufwendiger und von einer größeren Anzahl von Impedanzwerten $\bar{Z}_{ij}$ und deren Fehlern abhängig. Es lässt sich jedoch auch im Zweiportfall zeigen, dass $\Delta\mathbf{S} \leq \max(\delta\bar{Z}_{ij})$ gilt. Praktische Tests für resonante Mehrportsysteme zeigen, dass der relative Fehler der Streumatrizen rund 0,5 bis 2 Größenordnungen unterhalb des Impedanzfehlers liegt.

3.3.3 Zustandsraumdarstellung der Streuparameter

Anstelle der Berechnung der Streuparameter aus der Impedanzmatrix ist es alternativ auch möglich, direkt eine Zustandsraumdarstellung der Streumatrix anzuge-

ben. Aus Gl. 3.3.5 lässt sich nach [40] der Zustandsraum der Streuparameter mit den Matrizen des linearen Impedanzzustandsraums (Gl. 3.1.4) wie folgt direkt im Laplace-Bereich angeben

$$s\mathbf{x} = -\mathbf{A}_S\mathbf{x} + \mathbf{B}_S\mathbf{a} \tag{3.3.10a}$$

$$\mathbf{b} = \mathbf{C}_S\mathbf{x} - \mathbf{I} \tag{3.3.10b}$$

$$\text{mit} \quad \mathbf{A}_S = \mathbf{A} - \mathbf{BC},\ \mathbf{B}_S = \mathbf{B} \ \text{ und } \ \mathbf{C}_S = 2\mathbf{C}. \tag{3.3.10c}$$

Es ist zu beachten, dass dieser Zustandsraum in jedem Fall über eine Matrix $\mathbf{D}_S = -\mathbf{I}$ verfügt. Zudem ist offensichtlich, dass das System selbst für $\mathbf{C} = \mathbf{B}^T$ seine Symmetrie verliert $\mathbf{C}_S \neq \mathbf{B}_S^T$. Eine entsprechende Formulierung findet sich für das verlustfreie Curl-Curl-System:

$$s^2\mathbf{x} + s\mathbf{K}_S\mathbf{x} + \mathbf{A}_{CC}\mathbf{x} = s\mathbf{B}_S\mathbf{a} \tag{3.3.11a}$$

$$\mathbf{b} = \mathbf{C}_S\mathbf{x} - \mathbf{I} \tag{3.3.11b}$$

$$\text{mit} \quad \mathbf{K}_S = \mathbf{BC},\ \mathbf{B}_S = \mathbf{B} \ \text{ und } \ \mathbf{C}_S = 2\mathbf{C}. \tag{3.3.11c}$$

Die Matrix $\mathbf{K}_S$ stellt hierbei keine Verluste im ohmschen Sinne dar (das betrachtete System ist verlustfrei), sondern beschreibt die Energie, die durch die Ports austritt.

Alle Methoden zur Reduzierung der Ordnung, die im folgenden Kapitel vorgestellt werden, können somit gleichermaßen auf eine Zustandsraumdarstellung der Impedanz oder der Streuparameter angewandt werden. Aufgrund der einfacheren Matrixstruktur, insbesondere im Fall verlustfreier Systeme in Verbindung mit einer Curl-Curl-Formulierung, wird jedoch im Standardfall die Impedanzformulierung verwendet. Zudem ist meist die Bedingung positiv reeller Funktionen einfacher zu überprüfen als die begrenzt reeller Matrizen und die bedeutsame Eigenschaft der Passivitätserhaltung bei Projektion nach Theorem 1 gilt nur für positiv reelle aber nicht für begrenzt reelle Systeme.

Einen sinnvollen Einsatz können die Systeme Gln. 3.3.10 und 3.3.11 jedoch nach der Reduzierung der Modellordnung finden, beispielsweise zur Berechnung der Streuparameter-Übertragungsfunktion anstelle von Gl. 3.3.5 in einem *Frequencysweep* oder zur Bestimmung der Güte einzelner Resonanzen in Filterstrukturen. Eine mögliche Vorgehensweise zur Güteberechnung lautet dabei folgendermaßen: Ein verlustfreies System wird in Curl-Curl-Formulierung in der Ordnung reduziert, das reduzierte Curl-Curl-Modell wird nach Gl. 3.1.15 in ein lineares Impedanzsystem transformiert und dieses wiederum nach Gl. 3.3.10 in Streuparameterdarstellung gebracht. Die Eigenwerte dieser Systemmatrix $\mathbf{A}_S$ enthalten nun den Einfluss der Kopplung zu den Ports. Aus Real- und Imaginärteil dieser Eigenwerte s_i lässt sich die Güte eines Modes nach [33] mit

$$Q_k = \frac{\mathrm{Im}\{s_k\}}{2\mathrm{Re}\{s_k\}} \tag{3.3.12}$$

angeben. Dieses Verfahren kann als Erweiterung des in [33] angegebenen Ersatzschaltbildes auf den Mehrportfall gesehen werden und hat insbesondere bei hohen Güten Vorteile gegenüber Verfahren, die diese aus dem Spektrum bestimmen.

Kapitel 4

Reduzierung der Modellordnung

Die im vorigen Kapitel aufgestellten Systeme beschreiben das Verhalten einer elektromagnetischen Struktur im Rahmen der Diskretisierungsfehler vollständig, sie sind für realistische Beispiele jedoch meist von sehr hoher Ordnung, die im Bereich von Millionen von Unbekannten liegen kann.

Es zeigt sich jedoch, dass das Übertragungsverhalten im interessierenden Frequenzbereich meist mit Funktionen weit geringerer Ordnung beschrieben werden kann. Das folgende Kapitel zeigt mit Hilfe eines sehr allgemeinen Projektionsansatzes verschiedene Möglichkeiten, solche Modelle reduzierter Ordnung vollständig automatisiert zu erzeugen. Im Mittelpunkt stehen hierbei so genannte partielle Realisierungen sowie Ansätze, die auf Taylormomenten bzw. Padé-Approximationen beruhen. Während momentenbasierte Verfahren in vielen Bereichen der numerischen Simulation verbreitet sind und daher ein starkes Echo in wissenschaftlichen Veröffentlichungen finden, sind partielle Realisierungen bisher weit weniger beachtet. Dies liegt zum einen an der größeren resultierenden Modelldimension, zum anderen aber daran, dass sich die besondere Effizienz des Verfahrens nur unter der Bedingung entfaltet, dass die Materialmatrizen leicht zu invertieren sind. Dies ist bei FIT-Systemen gegeben, bei vielen FE-Verfahren oder numerischen Schaltkreissimulationen jedoch nicht gewährleistet. Die Verknüpfung der beiden Ansätze führt schließlich auf den Two-Step-Lanczos (TSL) Algorithmus, der sowohl im Hinblick auf Rechenzeit als auch auf die Modellgröße besonders effizient ist.

Die engen Verknüpfungen zwischen den unterschiedlichen Ansätzen werden in der folgenden Einführung genauer betrachtet.

4.1 Einführung

Zur Lösung der in den vorigen Kapiteln vorgestellten FIT-Systeme existiert eine Vielzahl von Lösungsstrategien. Insbesondere explizite Zeitintegrationsverfahren, auf die später kurz eingegangen werden soll, stellen extrem effiziente Verfahren dar, da für jeden folgenden Zeitschritt nur eine Matrix-Vektor-Multiplikation

durchgeführt werden muss. Auch frequenzabhängige Ergebnisse wie Übertragungsfunktionen können mit Hilfe einer diskreten Fouriertransformation einfach aus den Zeitsignalen berechnet werden. Eine bedeutende Einschränkung stellt allerdings die maximale Größe des zu wählenden Zeitschritts dar: Um die Stabilität der Methode zu garantieren, muss der Zeitschritt das Courant-Friedrichs-Levi-Kriterium erfüllen, wodurch das Verfahren auf hochfrequente Anwendungen beschränkt wird, in denen die Bauteilabmessungen in der Größenordnung der betrachteten Wellenlängen liegen. Für niedrigere Frequenzen muss auf implizite Verfahren zurückgegriffen werden, bei denen pro Zeitschritt ein lineares Gleichungssystem zu lösen ist. Doch selbst bei Erfüllung des Stabilitätskriteriums kann es bei hochresonanten Systemen wie beispielsweise Filtern zu Problemen kommen, da die Energie im Rechengebiet nur sehr langsam abklingt und die Zeitsignale folglich über lange Zeiträume berechnet werden müssen.

Alternativ kann das System auch für eine harmonische Anregung im Frequenzbereich gelöst werden. Dies erfordert pro Frequenzpunkt ebenfalls die Lösung eines linearen Gleichungssystems, das bis zu Millionen von Unbekannten beinhalten kann. Um scharfe Resonanzen im Übertragungsverhalten zu lokalisieren, kann zudem die Anzahl der Frequenzpunkte große Werte annehmen. Zur Lösung des Gleichungssystems werden üblicherweise Methoden verwendet, die zunächst einen Krylov-Unterraum $\mathcal{K}(\mathbf{A}, \mathbf{b})$ aus der Systemmatrix $\mathbf{A}$ und dem Anregungsvektor $\mathbf{b}$ aufstellen. Es zeigt sich, dass sich die Systemmatrizen für unterschiedliche Frequenzen im Fall von FIT lediglich durch eine diagonalen Term unterscheiden, und Krylov-Unterräume, wie unten näher gezeigt wird, invariant zu diagonalen Verschiebungen $\mathcal{K}(\mathbf{A} - s\mathbf{I}, \mathbf{b}) = \mathcal{K}(\mathbf{A}, \mathbf{b})$ sind. Unter Vernachlässigung moderner Vorkonditionierer muss also grundsätzlich für jeden Frequenzpunkt stets *derselbe* Unterraum erneut erzeugt werden, da es bedauerlicherweise selbst für Probleme mittlerer Größe nicht möglich ist, den betreffenden Unterraum im Speicher zu halten. Ist allerdings allein das Übertragungsverhalten von Interesse, ermöglicht es der *Lanczos-Algorithmus* [48], ein spezielles Krylov-Unterraumverfahren, auf das ebenfalls noch genauer eingegangen wird, ein Modell reduzierter Ordnung durch Projektion des Originalsystems auf den betreffenden Unterraum zu erzeugen. Dies erfolgt erneut ausschließlich durch Matrix-Vektor-Multiplikationen und Orthogonalisierungsschritte. Nur die Anzahl der notwendigen Iterationsschritte stellt zunächst eine Unbekannte dar. Modelle dieser Art werden als *partielle Realisierungen* bezeichnet und ihre speziellen Eigenschaften in Verbindung mit Taylormomenten der Originalübertragungsfunktion sind seit rund 20 Jahren bekannt [57]. In [58] und [60] wurden partielle Realisierungen auch bereits im Zusammenhang mit elektromagnetischen Simulationen betrachtet.

Verwendet man die gleiche Anzahl von Iterationsschritten in einem Krylov-basierten Eigenwertlöser, zeigt sich, dass üblicherweise alle Eigenwerte im betrachteten Frequenzintervall gut approximiert sind. Dies bestätigt erneut, dass die essentielle spektrale Information des Originalsystems von dem Krylov-Unterraum erfasst wird und ermöglicht es, ein Abbruchkriterium über die Konvergenz der Eigenwerte des reduzierten Modells zu definieren. Zudem rechtfertigt es ein als *Modalanalyse* [27] bezeichnetes Vorgehen, bei dem anstelle des großen Krylov-Unterraums direkt die Eigenvektoren verwendet werden, die zu den im interessierenden Frequenzband lie-

genden Eigenwerten gehören. Diese wesentlich kleinere Matrix kann ohne Weiteres im Speicher gehalten werden, was neben der reinen Übertragungsfunktion auch die Berechnung von Feldlösungen ermöglicht. Allerdings zeigt sich, dass die vollständige Vernachlässigung der höheren Moden auf einen Offset-Fehler führt, der geeignet kompensiert werden muss. Durch eine Kombination von Krylov-Unterräumen und Methoden der Eigenwertberechnung wie beispielsweise die Anwendung von Tschebyscheff-Beschleunigungspolynomen lässt sich ein nahezu fließender Übergang zwischen den angesprochenen Verfahren finden.

Einen auf den ersten Blick völlig unterschiedlichen Ansatz liefern Verfahren zur Ordnungsreduktion, die auf der Anpassung von Taylormomenten (engl.: *moment matching*) basieren und zu einer Padé-Approximation [51] der Übertragungsfunktion um einen oder mehrere Entwicklungspunkte führen. In der Regelungstechnik werden explizite Moment-Matching-Verfahren bereits seit rund 30 Jahren verwendet [61] und erfuhren im Bereich der Netzwerk- und Feldsimulation durch ein Verfahren namens *Asymptotic Waveform Evaluation (AWE)* [62] große Aufmerksamkeit. Eine wesentlich effizientere implizite Formulierung führt erneut zu Krylov-Unterräumen, diesmal auf die invertierte Systemmatrix angewandt. Diese Verbindung wurde in den späten achtziger Jahren erkannt und hatte vor rund zehn Jahren ihren Durchbruch mit einem *Padé Via Lanczos (PVL)* [63, 64, 66] genannten Algorithmus. Dieses Verfahren sowie dazu verwandte wie das passivitätserhaltende PRIMA [65] haben sich in der Praxis als äußerst robust erwiesen, erfordern jedoch die numerisch aufwendige Inversion der Systemmatrix oder alternativ erneut die Lösung zahlreicher linearer Systeme.

Der im Rahmen dieser Arbeit vorgeschlagene und intensiv untersuchte Two-Step-Lanczos (TSL) Algorithmus beruht auf der sukzessiven Anwendung von partieller Realisierung und Padé-Approximation in Kombination mit einer vollautomatischen Steuerung. Im Zusammenhang mit FIT und seinen im üblichen Fall dual orthogonaler Gitter einfach zu invertierenden Materialmatrizen führt dies auf ein besonders effizientes Verfahren zur schnellen und passivitätserhaltenden Erzeugung eines ordnungsreduzierten Modells, insbesondere für resonante Systeme. Zudem wird eine Erweiterung des Algorithmus für leicht verlustbehaftete Systeme betrachtet.

Ein letztes Verfahren zur Ordnungsreduktion soll zum Abschluss des Kapitels kurz vorgestellt werden: *Balanced Truncation* [67]. Es basiert auf der Theorie von Steuer- und Beobachtbarkeit. Das Originalsystem wird durch eine Äquivalenztransformation in ein System überführt, in dem jeder Zustand gleich gut bzw. schlecht gesteuert und beobachtet werden kann. Anschließend werden die schwer erreichbaren Zustände eliminiert. Dieses Verfahren hat den Vorteil, über den gesamten Frequenzbereich einen Maximalfehler garantieren zu können.

Bei der Bewertung aller Verfahren sollen stets die beiden Hauptanwendungsfälle der reduzierten Modelle betrachtet werden. Diese sind:

- Die schnelle Berechnung des Übertragungsverhaltens (engl.: *Fast Frequency Sweep*): Neben der Genauigkeit ist die Rechenzeit zur Erzeugung des Modells von primärem Interesse.

- Makromodellierung oder Ersatzschaltbildgenerierung: Das Modell soll unter Umständen häufig wiederverwendet werden, weswegen die Modellgröße entscheidend ist. Zudem sollen Stabilität und Passivität des Originalsystems erhalten bleiben, Rechenzeit ist hingegen zweitrangig.

4.2 Mathematische Grundlagen

4.2.1 Ordnungsreduktion durch Projektion

Wie im vorigen Abschnitt bereits angeklungen, ist das primäre Ziel bei der Erstellung ordnungsreduzierter Modelle, ein System von bedeutend geringerer Ordnung im Vergleich zum Originalsystem zu erzeugen, das die Übertragungsfunktion in einem vorgegebenen Frequenzband approximiert.

Nahezu alle modernen Reduktionstechniken können grundsätzlich als Projektion des Originalsystems auf zwei rechteckige Matrizen $\mathbf{V}$ und $\mathbf{W}$ der Dimension $n \times p$ mit $p \ll n$ interpretiert werden. Mit $\mathbf{V}$ wird zunächst ein neuer reduzierter Zustandsvektor $\mathbf{y}$ nach der Relation

$$\tilde{\mathbf{x}} = \mathbf{V}\mathbf{y} \tag{4.2.1}$$

definiert, im Anschluss wird die Systemgleichung von links mit $\mathbf{W}^T$ multipliziert. Es ergibt sich das reduzierte System, hier beispielhaft für das verlustbehaftete Curl-Curl-System (Gl. 3.1.12) angegeben:

$$\left(s^2\mathbf{W}^T\mathbf{V} + s\mathbf{W}^T\mathbf{K}\mathbf{V} + \mathbf{W}^T\mathbf{A}\mathbf{V}\right)\mathbf{y} = s\mathbf{W}^T\mathbf{B}\mathbf{i} \tag{4.2.2a}$$

$$\mathbf{u} = \mathbf{C}\mathbf{V}\mathbf{y} + \mathbf{D}\mathbf{i}. \tag{4.2.2b}$$

Entsprechend ergibt sich die Übertragungsfunktion des reduzierten Systems, ebenfalls für das verlustbehaftete Curl-Curl-System formuliert:

$$\mathbf{Z}_{\mathrm{red}}(s) = s\,\mathbf{C}\mathbf{V}(s^2\mathbf{W}^T\mathbf{V} + s\mathbf{W}^T\mathbf{K}\mathbf{V} + \mathbf{W}^T\mathbf{A}\mathbf{V})^{-1}\mathbf{W}^T\mathbf{B}. \tag{4.2.3}$$

Der Effekt der Reduktion wird anschaulich auch in Abb. 4.1 dargestellt.

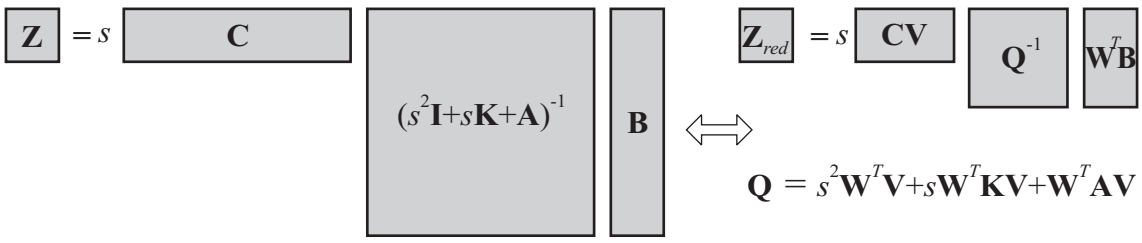


Abbildung 4.1: *Reduzierung der Systemordnung durch Projektion auf die Matrizen* $\mathbf{V}$ *und* $\mathbf{W}$.

Die Wahl von $\mathbf{V}$ und $\mathbf{W}$ ist in diesem Zusammenhang zunächst grundsätzlich frei. Um jedoch ein Modell niedriger Ordnung auf zuverlässige Weise zu erhalten, müssen

die von den beiden Projektionsmatrizen aufgespannten Unterräume mit bestimmten Eigenschaften des interessierenden Frequenzbereichs verknüpft sein. Im einfachsten Fall kann bereits eine gewisse Anzahl von Feldlösungen an unterschiedlichen Frequenzen zu einem befriedigenden Ergebnis führen. Dieses Verfahren entspricht einer *rationalen Interpolation*. Die Qualität der Approximation ist hierbei jedoch stark von der Auswahl der Frequenzpunkte abhängig und damit schwer vorhersehbar.

Für die Methode der *Balanced Truncation* werden die Hankel-Singulärvektoren zur Projektion herangezogen, im Fall der Modalanalyse sind es die Eigenvektoren zu den im interessierenden Frequenzintervall liegenden Eigenwerten. Partielle Realisierung und Padé-Approximation verwenden Krylov-Unterräume der Systemmatrix bzw. ihrer Inversen.

Auch wenn die Projektion nach Gl. 4.2.3 die größtmögliche Flexibilität in der Auswahl von Systemcharakteristiken bietet, werden im Allgemeinen Stabilität und Passivität nicht erhalten. Um diesen Nachteil zu umgehen, kann die Projektion nach Theorem 1 in Abschnitt 3.2.3 auf eine symmetrisch reelle beschränkt werden, was unter Umständen die Dimension des reduzierten Modells jedoch vergrößert, bzw. bei gleicher Größe die Genauigkeit verringert:

$$\mathbf{V} = \mathbf{W}, \qquad \mathbf{V} \in \mathbb{R}^{(n \times p)}. \tag{4.2.4}$$

Eine symmetrisch reelle Projektion erhält die Passivität immer dann, wenn alle zugrundeliegenden Systemmatrizen positiv reelle Matrizen nach den Gln. 3.2.9 darstellen. Dies ist für die meisten der im letzten Kapitel beschriebenen Formulierungen gewährleistet, Ausnahmen bilden jedoch die PML-Systeme nach Gl. 3.1.23a sowie linearisierte Systeme nach den Gl. 3.1.15 und 3.1.20.

Aus Gründen der numerischen Stabilität ist es zusätzlich empfehlenswert, die Unterräume orthonormal $\mathbf{V}^T\mathbf{V} = \mathbf{I}$, bzw. biorthogonal $\mathbf{W}^T\mathbf{V} = \mathbf{I}$ zu wählen.

4.2.2 Krylov-Unterraum-Verfahren

Wie in der Einführung bereits erwähnt, spielen so genannte *Krylov-Unterräume* eine zentrale Rolle bei der Erstellung ordnungsreduzierter Modelle. Unter einem Krylov-Unterraum wird der Raum verstanden, der durch wiederholte Anwendung einer Matrix $\mathbf{A}$ auf einen Startvektor $\mathbf{b}$ entsteht:

$$\mathcal{K}_p(\mathbf{A}, \mathbf{b}) = \mathrm{span}\{\mathbf{b}, \mathbf{A}\mathbf{b}, \mathbf{A}^2\mathbf{b}, \ldots, \mathbf{A}^{p-1}\mathbf{b}\}. \tag{4.2.5}$$

Für den Rang von $\mathcal{K}_p$ gilt: $\mathrm{rank}(\mathcal{K}_p) \leq \mathrm{rank}(\mathbf{A}) \leq n$. Aus 4.2.5 lässt sich leicht zeigen, dass Krylov-Unterräume invariant gegenüber Matrixskalierungen und diagonalen Verschiebungen sind

$$\mathcal{K}_p(t\mathbf{A} - s\mathbf{I}, \mathbf{b}) = \mathcal{K}_p(\mathbf{A}, \mathbf{b}). \tag{4.2.6}$$

Die Definition lässt sich auch auf mehrere Startvektoren erweitern. Dies führt mit $\mathbf{B} = [\mathbf{b}_1 \ldots \mathbf{b}_m]$ zu Block-Krylov-Unterräumen

$$\mathcal{K}_{pm}(\mathbf{A}, \mathbf{B}) = \mathrm{span}\{\mathbf{B}, \mathbf{A}\mathbf{B}, \mathbf{A}^2\mathbf{B}, \ldots, \mathbf{A}^{p-1}\mathbf{B}\}. \tag{4.2.7}$$

Im Gegensatz zur eindimensionalen Definition kann bei der Blockformulierung der Fall auftreten, dass ein linear abhängiger Vektor auftritt, ohne dass der Unterraum vollständig ausgeschöpft ist. Dies führt dazu, dass die betroffene Spalte in $\mathbf{B}$ eliminiert werden kann, während die übrigen $m-1$ Spalten den Krylov-Unterraum weiter aufbauen. Dieser Vorgang wird *Deflationierung* genannt.

Die Formulierung für $\mathcal{K}_{pm}(\mathbf{A}, \mathbf{B})$ entspricht genau der Definition der Steuerbarkeitsmatrix und entsprechend $\mathcal{K}_{pm}(\mathbf{A}^T, \mathbf{C}^T)$ der Beobachtbarkeitsmatrix des Systems in Abschnitt 3.2.4. Die Projektion auf Krylov-Unterräume bedeutet folglich in systemtheoretischer Interpretation die Projektion auf auf die ersten p Vektoren der Steuer- und/oder Beobachtbarkeitsmatrizen. Dies bedeutet zugleich, dass auf diese Weise erzeugte reduzierte Modelle stets minimale Realisierungen sind.

Grundsätzlich zeigt sich bei den oben angegebenen Krylov-Unterräumen, dass die Vektoren mit hoher Matrixpotenz zunehmend gegen den größten Eigenvektor von $\mathbf{A}$ konvergieren, was die Vektoren numerisch nahezu linear abhängig macht. Es ist folglich ratsam, die Vektoren bei ihrer Erstellung unmittelbar zu orthogonalisieren, um die volle Information des Unterraums zu erhalten. In den folgenden Abschnitten werden drei Algorithmen vorgestellt, die iterativ orthonormierte Krylov-Unterräume erstellen und implizit die Matrix $\mathbf{A}$ darauf projizieren.

4.2.2.1 Der symmetrische Lanczos Algorithmus

Das 1950 von C. Lanczos vorgeschlagene Verfahren [48] führt iterativ die Tridiagonalisierung einer symmetrischen bzw. hermiteschen Matrix $\mathbf{A}$ durch. Dies geschieht durch die implizite Projektion der Matrix $\mathbf{A}$ auf den iterativ aufgebauten und orthonormalisierten Krylov-Unterraum $\mathbf{V}$[1]. Nach p Iterationsschritten folgt damit:

$$\mathbf{V}_p^H \mathbf{A} \mathbf{V}_p = \mathbf{T}_p. \tag{4.2.8}$$

Nach $p = n$ Schritten sind die Matrizen $\mathbf{A}$ und $\mathbf{T}_n$ zueinander ähnlich, d. h. sie besitzen dieselben Eigenwerte. Der durchschlagende Erfolg des Verfahrens beruht aber auf der Erfahrung, dass gewisse Eigenwerte der Matrix $\mathbf{T}_p$, auch *Ritz Werte* genannt, bereits für $p \ll n$ Iterationsschritte gute Approximationen der entsprechenden Eigenwerte des Originalsystems $\mathbf{A}$ bilden.

Dieser Effekt lässt sich erklären, indem das charakteristische Polynom P_T der Matrix $\mathbf{T}_p$ betrachtet wird. In [44] wird gezeigt, dass beim Lanczos-Verfahren mit der Matrix $\mathbf{A}$ und dem Startvektor $\mathbf{b}$ für das Polynom P_T im Vergleich zu allen möglichen Polynomen derselben Ordnung stets gilt:

$$\|P_T(\mathbf{A})\mathbf{b}\| = \min. \tag{4.2.9}$$

Wird der Vektor $\mathbf{b}$ als Überlagerung der Anteile aller n Eigenvektoren $\mathbf{x}_\eta (\eta = 1 \ldots n)$ der Matrix $\mathbf{A}$ dargestellt

$$\mathbf{b} = \sum_{\eta=1}^{n} \alpha_\eta \mathbf{x}_\eta, \tag{4.2.10}$$

[1] Streng genommen bildet $\mathbf{V}$ eine Basis des Krylov-Unterraums $\mathcal{K}$. Aus Gründen der Einfachheit soll im folgenden $\mathbf{V}$ aber auch direkt als Krylov-Unterraum bezeichnet werden.

zeigt sich, dass die Aussage in 4.2.9 gleichbedeutend mit der Forderung ist, dass

$$\left\| \sum_{\eta=1}^{n} \alpha_\eta P_T(\lambda_\eta) \mathbf{x}_\eta \right\| = \min \tag{4.2.11}$$

gilt. Das Polynom P_T muss folglich an den Stellen jedes Eigenwerts der Matrix $\mathbf{A}$ einen kleinen Wert annehmen. Dies kann auf zwei Arten erfolgen: Entweder nimmt das Polynom generell im Frequenzintervall, in dem alle Eigenwerte liegen, geringe Werte an, was bedeutet, dass die Nullstellen des charakteristischen Polynoms etwa Tschebyscheff-verteilt [46] sind. Die andere Möglichkeit ist, dass P_T an der Stelle des Eigenwerts von $\mathbf{A}$ eine Nullstelle und $\mathbf{T}_p$ folglich selbst einen Eigenwert hat. Sind die Eigenwerte der Matrix $\mathbf{A}$ in etwa Tschebyscheff-verteilt, tritt primär der erste Fall in Kraft und der Lanczos-Algorithmus wird nur langsam gegen die Eigenwerte von $\mathbf{A}$ konvergieren. Sind die Eigenwerte jedoch anders verteilt, z. B. stärker oder schwächer separiert, was in praktischen technischen Systemen meist der Fall ist, tritt der zweite Fall ein, und die Eigenwerte von $\mathbf{T}_p$ konvergieren entsprechend schnell gegen die entsprechenden von $\mathbf{A}$. Ist ein Eigenwert besonders stark separiert, führt dies gar auf geometrische Konvergenz.

Zugleich wird offensichtlich, dass das Lanczos-Verfahren nur Eigenvektoren finden kann, deren Eigenwerte in $\mathbf{b}$ enthalten sind. In vielen Anwendungen erfolgt die Anregung daher mittels Zufallsvektoren. Dieser Effekt kann aber gezielt ausgenutzt werden, indem das Verfahren mit einem divergenzfreien Vektor angeregt wird. In diesem Fall werden die ca. N_P statischen Eigenwerte nicht berücksichtigt.

Zahlreiche Details zu Implementierungen des klassischen Lanczos Algorithmus für einzelne Startvektoren finden sich in [43, 45]. Da im Rahmen dieser Arbeit meist Mehrportsysteme betrachtet werden, soll an dieser Stelle unmittelbar eine Bandvariante des Verfahrens vorgestellt werden, die in Abb. 4.2 in einfacher Notation aufgeführt ist. Sie unterscheidet sich von einer Blockimplementierung, indem pro Iterationsschritt nur ein einzelner Vektor anstelle der Block-Matrix prozessiert wird. Mathematisch sind beide Varianten identisch, die Bandversion ermöglicht jedoch eine flexiblere Deflationierung linear abhängig gewordener Vektoren. Wird eine Blockgröße von m Vektoren angenommen, bekommt $\mathbf{T}_p$ die Gestalt einer Block-Tridiagonalmatrix mit m Bändern ober- und unterhalb der Hauptdiagonale, für $m = 1$ ergibt sich der eindimensionale Standardalgorithmus. Die jeweils m neusten Unterraumvektoren werden als Hilfsvektoren $\hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m$ gebraucht, bevor sie nach vollständiger Orthonormierung zu regulären Lanczos-Vektoren werden. Mit ihnen folgt für das Verfahren:

$$\mathbf{A}\mathbf{V}_p = \mathbf{V}_p\mathbf{T}_p + [\underbrace{\mathbf{0} \dots \mathbf{0}}_{p-m}\ \hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m], \tag{4.2.12a}$$

$$\mathbf{V}_p^H\mathbf{V}_p = \mathbf{I}, \quad \mathbf{V}_p^H[\hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m] = \mathbf{0}. \tag{4.2.12b}$$

Zur Initialisierung des Algorithmus werden die Hilfsvektoren gleich der Rechte-Seite-Matrix gesetzt $[\hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m] = \mathbf{B}$. In jedem Iterationsdurchlauf wird im Folgenden ein neuer Krylov-Vektor durch eine Matrix-Vektor-Multiplikation mit $\mathbf{A}$ erzeugt und

gegeben: $\mathbf{A}$, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$, p

for $i = 1 : p + m$ *Erzeuge Krylov- und Hilfsvektoren*
  if $i \leq m$
    $\mathbf{v}_i = \mathbf{b}_i$
  else
    $\mathbf{v}_i = \mathbf{A}_i \mathbf{v}_{i-m}$
  end

  $j_0 = \max(1, i - 2m)$ *Orthonormalisierung*
  for $j = j_0 : i - 1$
    $t_{j,i} = \mathbf{v}_j^T \mathbf{v}_i$
    $\mathbf{v}_i = \mathbf{v}_i - t_{j,i} \mathbf{v}_j$
  end
  $t_{i,i} = \|\mathbf{v}_i\|_2$
  $\mathbf{v}_i = \mathbf{v}_i / t_{i,i}$
end

$\mathbf{B}_p = [t_{1..p,1..m}]$ *endgültige Lanczos-Matrizen*
$\mathbf{T}_p = [t_{1..p,m+1..m+p}]$
$\mathbf{V}_p = [\mathbf{v}_{1..p}]$, $(\hat{\mathbf{V}} = [\mathbf{0..0}\ \mathbf{v}_{p+1..p+m}])$

Abbildung 4.2: *Bandlanczos-Algorithmus.*

durch das modifizierte Gram-Schmidt-Verfahren zu den bestehenden Vektoren orthonormalisiert. Unter Ausnutzung der Symmetrie ist es ausreichend, jeden neuen Vektor allein zu den vorhergehenden m Vektoren zu orthogonalisieren. Die Erfahrung im Zuge dieser Arbeit hat jedoch gezeigt, dass sich aufgrund endlicher Rechengenauigkeit Rundungsfehler in diesem Fall stark akkumulieren und das Ergebnis innerhalb von oft weniger als 500 Iterationen zur Ordnungsreduktion vollständig unbrauchbar wird. Das Verfahren erweist sich hingegen als erheblich robuster, wenn stattdessen, trotz mathematischer Redundanz, über $2m$ Vektoren orthogonalisiert wird. Ein vergleichbarer Effekt wurde im Zusammenhang mit dem verwandten Bi-Lanczos Verfahren bereits in [49] erkannt und beschrieben.

Zeigt sich, dass ein Vektor zum bestehenden Unterraum linear abhängig ist, wird dieser zur Deflationierung aus der Liste der Hilfsvektoren gestrichen. Die Blockdimension m verringert sich damit formal um Eins.

Durch die Orthogonalisierung über $2m$ Vektoren ist die Symmetrie der Matrix $\mathbf{T}_p$ bei endlicher Rechengenauigkeit nicht mehr exakt gegeben. Diese kann jedoch durch

$$\mathbf{T}'_p = \frac{1}{2}(\mathbf{T}_p^T + \mathbf{T}_p) \tag{4.2.13}$$

wieder erreicht werden.

Neben $\mathbf{T}_p$ liefert der Algorithmus auch eine Matrix $\mathbf{B}_p = \mathbf{V}_p^H \mathbf{B}$. Die Matrix $\mathbf{V}_p$ selbst wird damit in einigen Fällen, auf die später genauer eingegangen wird, im Projektionsprozess nach Gl. 4.2.3 entbehrlich, was einen entscheidenden Vorteil des Lanczos

Algorithmus darstellt. Von der Matrix $\mathbf{V}_p$ der Größe $n \times p$ müssen zum Fortschritt des Algorithmus nur die zur Gram-Schmidt-Rekursion erforderlichen letzten $2m+1$ Vektoren tatsächlich im Speicher gehalten werden. Dies ermöglicht es, das Verfahren auch bei großen Iterationszahlen zu verwenden. Wird $\mathbf{V}_p$ nicht komplett gespeichert, ist es allerdings unmöglich, nach der Projektion durch Gl. 4.2.1 Rückschlüsse auf den Lösungsvektor $\mathbf{x}$, beispielsweise den Feldvektor, zu ziehen.

Auch trotz des beschriebenen erhöhten Orthogonalisierungsaufwands zeigt sich jedoch, dass der Unterraum seine Orthogonalität mit steigender Iterationszahl verliert. Wurde der Eigenvektor eines dominanten Eigenwerts bis auf Rechengenauigkeit approximiert, sollte dessen Anteil in den folgenden Vektoren nicht mehr enthalten sein. Aufgrund von endlicher Rechengenauigkeit und Rundungsfehlern tritt diese Raumrichtung allerdings erneut auf, wenn zunächst auch nur in kleinen Beträgen. Im Laufe der folgenden Iterationen werden diese aufgrund der Dominanz des zugehörigen Eigenwerts verstärkt, bis der Eigenvektor von Neuem im Unterraum auftritt. Bei hohen Iterationszahlen ist es daher ein typisches Verhalten des Lanczos Algorithmus, dass dominante Eigenwerte mehrfach in $\mathbf{T}_p$ enthalten sind. Sie werden als *Geistereigenwerte* bezeichnet.

Dieser Effekt lässt sich verlangsamen, indem der Gram-Schmidt Algorithmus doppelt ausgeführt wird, wahlweise in jedem Iterationsschritt oder nur, falls sich die Norm eines Vektors durch die Orthogonalisierung stark ändert, da in solchen Fällen das Gram-Schmidt Verfahren zu Ungenauigkeiten neigt. Alternativ wurden Methoden entwickelt [52], die mehrfache Eigenwerte zunächst hinnehmen und diese nach der Berechnung detektieren, um sie von tatsächlichen mehrfach vorkommenden entarteten Eigenwerten zu unterscheiden.

Im Zusammenhang mit der Ordnungsreduktion im Rahmen dieser Arbeit zeigt sich, dass mehrfach vorkommende Eigenvektoren in den Projektionsmatrizen $\mathbf{V}_p$ das Ergebnis nicht beeinträchtigen. Der Aufwand der doppelten Orthogonalisierung ist im Vergleich zur nur leicht verbesserten Konvergenz meist nicht gerechtfertigt.

4.2.2.2 Der Arnoldi Algorithmus

Handelt es sich bei der Matrix $\mathbf{A}$ um eine unsymmetrische oder nicht-hermitesche Matrix, ist die resultierende Matrix nicht mehr (Block-)tridiagonal wie in Gl. 4.2.12a, sondern erhält eine obere (Block-)Hessenbergform. Die Orthonormalitätsbeziehungen gelten auch weiterhin:

$$\mathbf{A}\mathbf{V}_p = \mathbf{V}_p\mathbf{H}_p + [\underbrace{\mathbf{0}\ldots\mathbf{0}}_{p-m}\ \hat{\mathbf{v}}_1\ldots\hat{\mathbf{v}}_m], \tag{4.2.14a}$$

$$\mathbf{V}_p^H\mathbf{V}_p = \mathbf{I}, \quad \mathbf{V}_p^H[\hat{\mathbf{v}}_1\ldots\hat{\mathbf{v}}_m] = \mathbf{0}. \tag{4.2.14b}$$

Die Berechnung erfolgt analog zum Bandlanczos-Algorithmus (Abb. 4.2), wobei die Orthogonalisierung eines neuen Krylov-Vektors nicht auf die vorangegangenen m oder $2m$ Vektoren beschränkt werden kann, sondern grundsätzlich über den vollständigen, bereits aufgestellten Unterraum erfolgen muss. Damit überwiegt bei einer großen Zahl von Iterationsschritten schnell der Orthogonalisierungsaufwand

über den der Matrix-Vektor-Multiplikation, der numerische Aufwand wächst näherungsweise quadratisch mit $\mathcal{O}(p^2)$ an. Schwerer als der numerische Aufwand wiegt in der Praxis jedoch meist der Nachteil, dass die Matrix $\mathbf{V}_p$ tatsächlich komplett im Speicher gehalten werden muss, was für große Systemmatrizen und hohe Iterationszahlen oft nicht möglich ist.

Dieses Verfahren wurde erstmals 1951 von W.E. Arnoldi [50] formuliert. Die Eigenschaften zum Auffinden von Eigenwerten sind ähnlich denen des Lanczos Algorithmus. Die Methode findet auch bei der Lösung unsymmetrischer linearer Gleichungssysteme Verwendung, ein bekanntes Verfahren nennt sich *Quasi-Minimale-Residuen, QMR* [43].

Für den Fall symmetrischer Matrizen fällt der Arnoldi Algorithmus formell, wie bereits angedeutet, mit dem Lanczos Algorithmus zusammen. Wird bei nicht exakter Arithmetik jedoch die Orthogonalisierung über alle vorhergehenden Arnoldi-Vektoren beibehalten, stoppt dies den Orthogonalitätsverlust, der beim Lanczos Verfahren häufig beobachtet wird. Ein Auftreten von Geistereigenwerten ist daher nahezu ausgeschlossen, allerdings um den Preis eines deutlich erhöhten numerischen Aufwands.

4.2.2.3 Der Bi-Lanczos Algorithmus

Um die Vorteile der kurzen Rekursion des Lanczos Verfahrens auch für unsymmetrische Matrizen erhalten zu können, kann alternativ zum Arnoldi Algorithmus auch eine unsymmetrische (Block-)Tridiagonalisierung durchgeführt werden. Diese führt mit den bi-orthogonalen Unterräumen $\mathbf{V}_p$ und $\mathbf{W}_p$ auf

$$\mathbf{W}_p^H \mathbf{A} \mathbf{V}_p = \mathbf{T}_p, \qquad \text{mit} \qquad \mathbf{W}_p^H \mathbf{V}_p = \mathbf{I}. \tag{4.2.15}$$

Das Verfahren wird als Bi-Lanczos Algorithmus bezeichnet [43]. Die Matrizen $\mathbf{V}_p$ und $\mathbf{W}_p$ spannen mit den Startmatrizen $\mathbf{B}$ und $\mathbf{C}$ jeweils die Krylov-Unterräume $\mathcal{K}_p(\mathbf{A}, \mathbf{B})$ und $\mathcal{K}_p(\mathbf{A}^H, \mathbf{C}^H)$ auf. In einer Band-Formulierung können die Blockdimensionen m und l der Startmatrizen unterschiedlich gewählt werden, was zusätzlich die Deflationierung ermöglicht, da linear abhängige Vektoren in beiden Krylov-Unterräumen nicht grundsätzlich im selben Iterationsschritt auftreten werden. Selbst bei gleichen Startmatrizen können folglich im Verlauf des Algorithmus unterschiedliche Blockgrößen auftreten. Mit zwei Sätzen Hilfsvektoren $\hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m$ und $\hat{\mathbf{w}}_1 \dots \hat{\mathbf{w}}_l$ ergeben sich die Relationen:

$$\mathbf{A}\mathbf{V}_p = \mathbf{V}_p \mathbf{T}_p + [\underbrace{\mathbf{0} \dots \mathbf{0}}_{p-m}\ \hat{\mathbf{v}}_1 \dots \hat{\mathbf{v}}_m], \tag{4.2.16a}$$

$$\mathbf{A}^H \mathbf{W}_p = \mathbf{W}_p \mathbf{T}_p^H + [\underbrace{\mathbf{0} \dots \mathbf{0}}_{p-l}\ \hat{\mathbf{w}}_1 \dots \hat{\mathbf{w}}_l]. \tag{4.2.16b}$$

Neben der Matrix $\mathbf{T}_p$ erzeugt der Algorithmus auch zwei Matrizen $\mathbf{B}_p = \mathbf{W}_p^T \mathbf{B}$ und $\mathbf{C}_p = \mathbf{C}\mathbf{V}_p$. Ein Problem des Bi-Lanczos Verfahrens ergibt sich, wenn die Vektoren $\mathbf{v}_p$ und $\mathbf{w}_p$, deren Produkt $\mathbf{w}_p^H \mathbf{v}_p$ auf Eins normiert werden muss, zueinander orthogonal sind. Dies wird als ernster Zusammenbruch (engl.: *serious breakdown*)

bezeichnet [43, 53] und kann auftreten, ohne dass die Krylov-Unterräume selbst ausgeschöpft sind. Abhilfe schafft eine vorausschauende Technik (engl.: *look-ahead*), die bei drohendem Zusammenbruch auf die vektorweise Orthogonalisierung verzichtet und alternativ $\mathbf{v}_p$ und $\mathbf{w}_p$ in Blöcken mit dynamisch angepasster Größe orthogonalisiert. Dies ändert lokal die Bandstruktur in $\mathbf{T}_p$. Für nähere Details siehe [53]. Die praktische Erfahrung zeigt jedoch, dass Zusammenbrüche dieser Art extrem selten auftreten.

Erneut erweist sich, dass eine Realisierung mit zusätzlichen Orthogonalisierungsschritten numerisch robuster ist. Die Rekursionslänge beträgt $m+l+1$ Vektoren. Der numerische Aufwand des Verfahrens bleibt damit linear mit der Systemgröße und ist im Vergleich zum symmetrischen Lanczos (unter Annahme von $l = m$) nahezu exakt verdoppelt. Eine einfache Variante des Algorithmus findet sich in Anhang A, ein detaillierter Band-Bi-Lanczos Algorithmus mit Deflationierung und *look-ahead* findet sich, mit im Vergleich zu dieser Darstellung leicht veränderter Notation, in [53, 64].

4.3 Verfahren zur Reduktion der Modellordnung

4.3.1 Partielle Realisierungen

Unter einer partiellen Realisierung des Systems $\mathbf{Z}(s)$ wird ein System $\mathbf{Z}_p(s)$ mit $p \ll n$ verstanden, dessen erste $2p$ Markov-Parameter im Einportfall mit denen des Originalsystems übereinstimmen. Für allgemeine Mehrportsysteme mit m Eingangs- und l Ausgangsports müssen sich mit $m_p = \mathrm{floor}(p/m)$ und $l_p = \mathrm{floor}(p/l)$ mindestens $m_p + l_p$ Markov-Parameter decken.

Für in der Frequenz lineare Systeme können die Markov-Parameter $\mathbf{m}_k = \mathbf{C}(-\mathbf{A})^k\mathbf{B}$ nach Gl. 3.1.6 als Taylormomente der Impulsantwort des Systems zum Zeitpunkt $t = 0$ angesehen werden. Im Frequenzbereich ergibt sich mit der geometrischen Reihe

$$\mathbf{Z}(s) = \mathbf{C}(s\mathbf{I} + \mathbf{A})^{-1}\mathbf{B} \tag{4.3.1a}$$

$$= \sum_{k=0}^{\infty} \mathbf{C}(-\mathbf{A})^k\mathbf{B}\,\frac{1}{s^k} = \sum_{k=0}^{\infty} \frac{\mathbf{m}_k}{s^k}\,. \tag{4.3.1b}$$

Verlustfreie Curl-Curl-Systeme mit $\mathbf{M}_\kappa = 0$ lassen sich durch die Transformation $s' = s^2$ ebenfalls mit den Methoden eines linearen Systems beschreiben.

4.3.1.1 Partielle Realisierungen durch Krylov-Unterräume

Im Folgenden soll gezeigt werden, dass die Projektion eines linearen Systems auf die Matrizen $\mathbf{V}_p$ und $\mathbf{W}_p$ des beschriebenen Bi-Lanczos Algorithmus, die die Krylov-Unterräume $\mathcal{K}_p(\mathbf{A}, \mathbf{B})$ und $\mathcal{K}_p(\mathbf{A}^H, \mathbf{C}^H)$ aufspannen, einer partiellen Realisierung des Originalsystems entspricht. Das reduzierte System lautet analog zu 4.2.3 für

den allgemeinen Fall eines linearen Systems:

$$\mathbf{Z}_p(s) = \mathbf{C}\mathbf{V}_p \left(s\mathbf{W}_p^H\mathbf{V}_p + \mathbf{W}_p^H\mathbf{A}\mathbf{V}_p\right)^{-1}\mathbf{W}_p^H\mathbf{B} \tag{4.3.2a}$$

$$= \mathbf{C}_p \left(s\mathbf{I} + \mathbf{T}_p\right)^{-1}\mathbf{B}_p \tag{4.3.2b}$$

$$= \sum_{k=0}^{\infty} \mathbf{C}_p \left(-\mathbf{T}_p\right)^k \mathbf{B}_p \frac{1}{s^k}\,. \tag{4.3.2c}$$

Zur Beweisführung wird zunächst erneut Beziehung 4.2.16a betrachtet. Mit der Hilfsvektorenmatrix einschließlich der Nullvektoren $\hat{\mathbf{V}} = [\,\mathbf{0}\ldots\mathbf{0}\ \hat{\mathbf{v}}_1\ldots\hat{\mathbf{v}}_m]$ gilt:

$$\mathbf{A}^i\mathbf{V}_p = \mathbf{V}_p\mathbf{T}_p^i + \underbrace{\hat{\mathbf{V}}\mathbf{T}^{i-1} + \mathbf{A}\hat{\mathbf{V}}\mathbf{T}^{i-2} + \ldots + \mathbf{A}^{i-1}\hat{\mathbf{V}}}_{\bar{\mathbf{V}}}, \tag{4.3.3}$$

hier exemplarisch für $i > 2$ aufgeschrieben. Mit der Bandstruktur von $\mathbf{T}_p$ und der Struktur von $\hat{\mathbf{V}}$ lässt sich zeigen, dass die Summanden in $\bar{\mathbf{V}}$ keine Einträge in den ersten m Spalten haben. Diese Eigenschaft ist grundsätzlich für $0 \leq i < m_p$ gültig. Zudem gilt durch die Initialisierung des Bi-Lanczos Algorithmus $\mathbf{B} = \mathbf{V}_p\mathbf{B}_p$, wobei $\mathbf{B}_p$ nur im oberen $m \times m$ Block Einträge ungleich Null hat. Wird nun 4.3.3 von rechts mit $\mathbf{B}_p$ multipliziert, führt dies auf:

$$\mathbf{A}^i\mathbf{B} = \mathbf{V}_p\mathbf{T}_p^i\mathbf{B}_p \qquad \text{für} \qquad 0 \leq i < m_p. \tag{4.3.4}$$

Analog lässt sich aus 4.2.16b

$$\mathbf{C}\mathbf{A}^i = \mathbf{C}_p\mathbf{T}_p^i\mathbf{W}_p^H \qquad \text{für} \qquad 0 \leq i < l_p \tag{4.3.5}$$

herleiten. Multipliziert man schließlich 4.3.4 und 4.3.5 miteinander, ergibt sich mit der Biorthogonalitätsbeziehung $\mathbf{W}_p^H\mathbf{V}_p = \mathbf{I}$

$$\left.\begin{aligned} &\mathbf{C}\mathbf{A}^{i'}\mathbf{A}^{i''}\mathbf{B} = \mathbf{C}\mathbf{A}^i\mathbf{B} = \mathbf{C}_p\mathbf{T}_p^i\mathbf{B}_p \\ \Longrightarrow\quad &\mathbf{C}(-\mathbf{A})^i\mathbf{B} = \mathbf{C}_p(-\mathbf{T}_p)^i\mathbf{B}_p \end{aligned}\right\} \quad \text{für} \quad 0 \leq i < m_p + l_p - 1. \tag{4.3.6}$$

Diese Beziehung lässt sich nach [64] für $i = m_p + l_p - 1$ ebenfalls beweisen, womit $\mathbf{Z}(s)$ und $\mathbf{Z}_p(s)$ tatsächlich die ersten $m_p + l_p$ Markov-Parameter gemeinsam haben. $\mathbf{Z}_p(s)$ nach Gl. 4.3.2c ist demnach eine partielle Realisierung von $\mathbf{Z}(s)$.

Da s^k anstelle des Bruchs in 4.3.1b eine Taylorreihe um $s = 0$ bilden würde, wird bei partiellen Realisierungen in Anlehnung auch von *moment-matching* um Unendlich gesprochen. Überraschend mag zunächst erscheinen, dass eine partielle Realisierung auch für niedrige Frequenzen gute Approximationseigenschaften zum Originalsystem aufweist, obwohl die geometrische Reihe nur für $\|\mathbf{A}\|/|s| < 1$, also für Frequenzen oberhalb des betragsgrößten Eigenwerts von $\mathbf{A}$ und damit weit außerhalb des technisch interessanten Bereichs, definiert ist. Eine partielle Realisierung stellt aber keine Taylorreihe dar, deren Konvergenzradius stets nur bis zur nächstgelegenen Polstelle reicht, sondern bildet eine gebrochen rationale Funktion mit derselben Struktur wie die Originalfunktion, lediglich mit deutlich verringerter Ordnung. Eine solche gebrochen rationale Funktion, die in einem oder mehreren Entwicklungspunkten eine Anzahl von Taylormomenten mit der Originalfunktion gemeinsam hat, wird allgemein

auch als *Padé-Approximation* [51] bezeichnet. Zu den grundlegenden Eigenschaften dieser Klasse von Approximationen gehört es, Funktionen auch über Polstellen hinweg anzunähern. Partielle Realisierungen stellen mit einer Entwicklungsfrequenz von Unendlich lediglich einen Sonderfall einer solchen Padé-Approximation dar.

Da alle benötigten Matrizen in Gl. 4.3.2b durch das Bi-Lanczos Verfahren direkt berechnet werden, entspricht der Aufwand zur Erstellung der partiellen Realisierung exakt dem Aufwand des Algorithmus, also primär 2 Matrix-Vektor-Multiplikationen und $2(m+l+1)$ Orthogonalisierungsschritte pro Iterationsschritt. Über die benötigte Modellgröße lassen sich im Voraus keine Aussagen machen, da diese von der Größe des Originalsystems aber auch von dessen innerer Dynamik (d. h. der Lage der Polstellen) abhängt. Typische Werte für p liegen zwischen 500 und 5000 Iterationen.

Die obige Herleitung ist sehr allgemein gehalten. Üblicherweise werden Eingangs- und Ausgangsports identisch sein, womit auch bei unsymmetrischen Systemen $\mathbf{C} = \mathbf{B}^T$ gelten wird.

Eine wichtige Untergruppe bilden zudem die symmetrischen Systeme mit $\mathbf{A}^T = \mathbf{A}$ wie das verlustfreie Curl-Curl-System, da in diesem Fall der symmetrische Lanczos Algorithmus verwendet werden kann. Dies ist auch bei schiefsymmetrischen Systemen $\mathbf{A}^T = -\mathbf{A}$ wie dem verlustfreien linearen System möglich, da mit der Skalierungsbedingung 4.2.6 $\mathcal{K}(\mathbf{A}^T, \mathbf{B}) = \mathcal{K}(-\mathbf{A}, \mathbf{B}) = \mathcal{K}(\mathbf{A}, \mathbf{B})$ gilt. Die Verwendung der symmetrischen Methode halbiert zunächst den numerischen Aufwand zur Erstellung des Modells. Als entscheidender Vorteil zeigt sich aber vor allem, dass die Projektion symmetrisch reell wird und damit die Stabilität und Passivität des Originalsystems auch im reduzierten Modell erhält. Dies kann für unsymmetrische Systeme bei Projektion nach 4.3.2a nicht garantiert werden und muss im Einzelfall überprüft werden, was die Verwendung als Makromodell einschränkt.

Eine symmetrische und somit passivitätserhaltende Projektion auch im Fall unsymmetrischer Systeme ist nur möglich, wenn die Matrix $\mathbf{V}_p$ während des Algorithmus komplett im Speicher gehalten und die Projektion im Anschluss explizit symmetrisch durchgeführt wird oder indem anstelle des Bi-Lanczos der Arnoldi-Algorithmus verwendet wird, der die symmetrische Projektion implizit durchführt. In beiden Fällen ist die Anzahl der übereinstimmenden Markov-Parameter im Vergleich zur unsymmetrischen Version halbiert, da nur Gl. 4.3.4 und nicht Gl. 4.3.5 erfüllt ist. Beide genannten Varianten sind jedoch für große Systemmatrizen in der Praxis meist aufgrund von Speicherproblemen nicht durchführbar, was die passive Ordnungsreduktion mittels partiellen Realisierungen de facto zunächst auf symmetrische Systeme beschränkt.

Diese stellen in Form verlustloser Systeme wie beispielsweise Filterstrukturen tatsächlich einen häufigen Anwendungsfall partieller Realisierungen für einen *Fast Frequency Sweep* dar, da deren spektrale Auswertung sowohl mit Zeitbereichs- als auch mit Frequenzbereichsmethoden sehr aufwändig ist. Mit der schiefsymmetrischen linearen und der symmetrischen Curl-Curl-Formulierung stehen zwei konkurrierende Ansätze zur Verfügung, deren Aufwand näher verglichen werden soll.

Grundsätzlich ist die Dimension des linearen Systems etwa doppelt so groß wie die des zugehörigen Curl-Curl-Systems. Zugleich besitzt die lineare Systemmatrix aber

nur vier Einträge pro Zeile, während dies bei der Curl-Curl-Matrix 13 sind (Randzellen jeweils ausgenommen), jede Matrix-Vektor-Multiplikation hat demnach den rund $1,5$-fachen Aufwand im Vergleich zum linearen Fall. Eine Einsparung an Multiplikationen kann erzielt werden, wenn in jedem Aufruf anstelle des Produkts mit der Systemmatrix die Operatormatrizen $\mathbf{C}$ und $\mathbf{C}^T$ sukzessive ausgeführt werden. In diesem Fall können die Curl-Operatoren direkt durch je vier Additionen ausgeführt werden und es bleiben nur zwei, bzw. drei Multiplikationen mit den Materialmatrizen pro Zeile für die lineare bzw. die Curl-Curl-Formulierung.

Jeder Orthogonalisierungsschritt, die Anzahl ist bei beiden Formulierungen identisch, erfordert je ein volles Vektor-Vektor-Produkt und eine Vektor-Vektor-Subtraktion. Bei halber Systemgröße im Curl-Curl-Fall ist der Aufwand hierfür ebenfalls halbiert. Als Abschätzung des Gesamtaufwands ergibt sich mit der Dimension des Curl-Curl-Systems N und der Portzahl m pro Iterationsschritt:

$$\text{lineares System:} \quad 2N(4\mathcal{M} + 3\mathcal{A} + (2m+1)\mathcal{M} + (2m+1)\mathcal{A}), \tag{4.3.7a}$$

$$\text{Curl-Curl-System:} \quad N(13\mathcal{M} + 12\mathcal{A} + (2m+1)\mathcal{M} + (2m+1)\mathcal{A}), \tag{4.3.7b}$$

$$\text{lineare Operatoren:} \quad 2N(2\mathcal{M} + 4\mathcal{A} + (2m+1)\mathcal{M} + (2m+1)\mathcal{A}), \tag{4.3.7c}$$

$$\text{Curl-Curl-Operatoren:} \quad N(3\mathcal{M} + 8\mathcal{A} + (2m+1)\mathcal{M} + (2m+1)\mathcal{A}). \tag{4.3.7d}$$

$\mathcal{M}$ steht für Multiplikationen, $\mathcal{A}$ für Additionen, die linken Angaben in der Klammer beschreiben jeweils den Aufwand für die Matrix-Vektor-Multiplikation, die rechten den für die $2m+1$ Orthogonalisierungsschritte. Es zeigt sich beispielsweise, dass der Aufwand pro Iterationsschritt bei Verwendung der Systemmatrizen für $m = 2$ Ports nach Anzahl der Multiplikationen in beiden Fällen nahezu identisch ist. Bei Verwendung der Operatoren ist die Curl-Curl-Formulierung nach Multiplikationen stets weniger aufwendig. Zudem ist bei Operatorverwendung der Aufwand für die Orthogonalisierung meist größer als der für das Matrix-Vektor-Produkt.

Ausschlaggebend für die Entscheidung ist aber vielmehr die Anzahl der benötigten Iterationsschritte. Diese kann, wie schon gesagt, nicht deterministisch bestimmt werden, die Erfahrung zeigt jedoch, dass im linearen Fall rund doppelt so viele Schritte erforderlich sind wie im Curl-Curl-Fall, um dieselbe Approximationsgüte zu erzielen. Dies ist anschaulich nachvollziehbar, weil im linearen Fall sowohl der positive als auch der negative Anteil des Spektrums betrachtet wird und damit auch die doppelte Anzahl von Polstellen approximiert werden muss. Die Modellgenerierung aus der Curl-Curl-Formulierung erweist sich damit in der Summe als nur etwa halb so aufwendig wie im linearen Fall. Ein weiterer erheblicher Vorteil des Curl-Curl-Falls ergibt sich, wenn die partielle Realisierung nur als Zwischenergebnis genutzt und durch eine Padé-Approximation weiter reduziert wird, worauf später eingegangen werden soll.

4.3.1.2 Abbruchkriterium

Einen bislang unbeachteten Punkt stellt die Frage des Abbruchs der Iteration dar, nach wie vielen Iterationsschritten also davon auszugehen ist, dass mit der partiellen Realisierung eine Approximation ausreichender Genauigkeit im interessierenden Frequenzintervall vorliegt.

Sofern die Übertragungsfunktion $\mathbf{Z}(s_i)$ des Originalsystems an einem oder mehreren Frequenzpunkten bekannt ist, kann dies mit der Definition einer Fehlerschranke ε_Z erfolgen. In periodischen Abständen wird die numerisch wenig aufwendige Lösung des reduzierten Systems $\mathbf{Z}_p(s_i)$ an denselben Frequenzpunkten berechnet und

$$\delta_{Zi} = \frac{\|\mathbf{Z}_p(s_i) - \mathbf{Z}(s_i)\|}{\|\mathbf{Z}(s_i)\|} < \varepsilon_Z \tag{4.3.8}$$

überprüft. Diese Vorgehensweise ist jedoch nur bedingt zuverlässig, da einige Teile des Spektrums möglicherweise bereits gut angenähert sind, während andere Teile des betrachteten Frequenzbereichs, insbesondere die Polstellen, noch nicht approximiert sind. Da zur Berechnung von $\mathbf{Z}(s_i)$ meist ebenfalls Krylov-Unterraumverfahren herangezogen werden, liegt zudem der Aufwand zur Berechnung dieser Referenzlösung im gleichen Bereich wie der zur Berechnung der gesamten partiellen Realisierung.

Als Richtwert lässt sich sagen, dass die Iterationszahl der partiellen Realisierung mit der maximalen Unterraumgröße aller iterativen Rechnungsdurchläufe einer Frequenzbereichsrechnung über den interessierenden Frequenzbereich übereinstimmen sollte. Wird das reduzierte Modell für einen *Fast Frequency Sweep* verwendet, möchte man sich allerdings genau diese Frequenzbereichsrechnung ersparen, weswegen die maximale Zahl im Allgemeinen unbekannt ist.

Eine alternative Herangehensweise bietet die Untersuchung der Polstellen. Insbesondere bei resonanten Systemen zeigt sich, dass die Übertragungsfunktion gut approximiert ist, wenn deren Polstellen gut angenähert sind. Es bietet sich daher an, ein Abbruchkriterium über die Eigenwerte des Systems zu definieren. Der enge Zusammenhang zwischen Eigenwert- und Übertragungsapproximation unter Verwendung von Krylov-Unterräumen wird schon deutlich, da sich diese Klasse von Verfahren sowohl zur iterativen Lösung von Gleichungssystemen als auch zur Eigenwertabschätzung mit großem Erfolg verwenden lässt. Die Eigenwert-Approximationseigenschaften der Krylov-Verfahren liefern damit neben den Markov-Parametern eine zweite Erklärung für die Qualität der partiellen Realisierungen.

Der Eigenwertfehler kann für den Eigenwert λ_i mit

$$\delta_{Ei} = \frac{\|\lambda_{i,p} - \lambda_i\|}{\|\lambda_i\|} < \varepsilon_{\text{eig}} \tag{4.3.9}$$

definiert werden. Da die Bestimmung der Referenzeigenwerte wiederum ein numerisch aufwendigeres Problem darstellt, das ohnehin meist ebenfalls über Krylov-Unterraumverfahren gelöst wird, ist es vorzuziehen, die Eigenwerte während der Berechnung der partiellen Realisierung direkt mitzubestimmen. Hierzu wird der Lanczos-Algorithmus beginnend nach einer Startiterationszahl p_0 in periodischen Abständen gestoppt, die Eigenwerte bzw. Ritz-Werte der Matrix $\mathbf{T}_p$ werden mit klassischen Methoden[2] bestimmt, was aufgrund der Bandstruktur der Matrix nur

[2]Direkte Methoden im engeren Sinne zur Bestimmung von Eigenwerten existieren nicht, da das Auffinden von Eigenwerten ähnlich der Nullstellensuche von Polynomen stets iterativ erfolgen muss. Allerdings werden sehr schnell konvergierende Verfahren wie der QR Algorithmus [43] häufig als *direkte* Methoden bezeichnet.

geringen numerischen Aufwand erfordert. Zur Analyse der Genauigkeit der Eigenwerte existieren zwei Möglichkeiten:

- Eine Abschätzung der Konvergenz eines Eigenwerts $\lambda_{i,p}$ lässt sich mit dem zugehörigen Ritz-Vektor $\theta_{i,p}$ und der Hilfsmatrix $\hat{\mathbf{V}}$ angeben [54]:

$$|\hat{\mathbf{V}}\theta_{i,p}| < \varepsilon_{\text{eig}}. \tag{4.3.10}$$

- Es kann ein relativer Fehler der Eigenwerte im Vergleich zur letzten Berechnung durchgeführt werden

$$\delta_{Di} = \frac{\|\lambda_{i,p+\Delta p} - \lambda_{i,p}\|}{\|\lambda_{i,p}\|} < \varepsilon_{\delta\text{eig}}. \tag{4.3.11}$$

  Bleiben die Eigenwerte konstant bzw. die Differenz unter einer Schwelle $\varepsilon_{\delta\text{eig}}$, werden sie als auskonvergiert betrachtet.

Die Iteration des Lanczos Algorithmus zur Modellgenerierung wird endgültig gestoppt, wenn alle im interessierenden Frequenzintervall liegenden Eigenwertfehler unterhalb der Grenze $\varepsilon_{\delta\text{eig}}$ liegen. Die Anzahl der Eigenwerte, die im betrachteten Frequenzbereich liegen, muss dabei während der Eigenwertberechnung dynamisch bestimmt werden. Sollte kein Eigenwert im Frequenzbereich selbst liegen, werden die nächstgelegen außerhalb betrachtet.

Die Wirkungsweise des Abbruchkriteriums wird anhand eines Beispiels verdeutlicht. Es handelt sich hierbei um eine verlustlose Filterstruktur, die über zwei Ports angeregt wird und im betrachteten Frequenzbereich von 350 MHz bis 650 MHz zwei Pole aufweist. Die Diskretisierung ist mit 2400 Unbekannten in Curl-Curl-Darstellung verhältnismäßig grob gewählt, so dass die exakte Lösung noch mit einem direkten Verfahren bestimmt werden kann. Dasselbe Beispiel soll auch in den folgenden Abschnitten zur Veranschaulichung herangezogen werden.

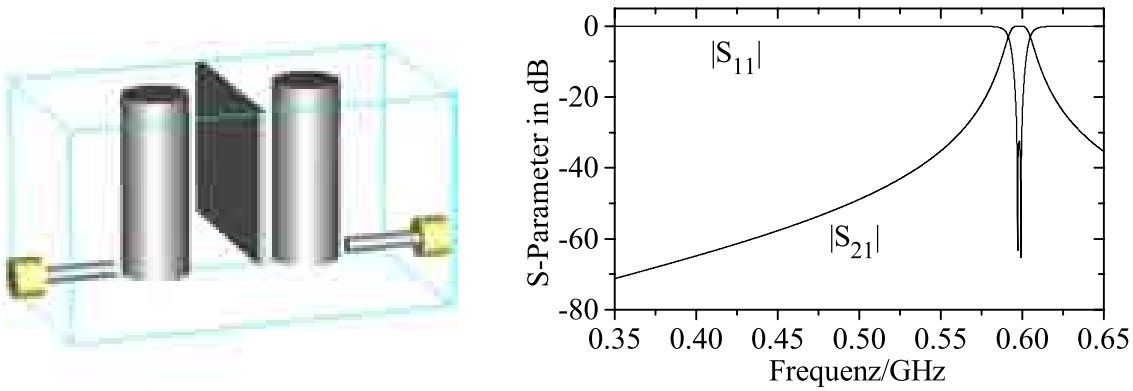


Abbildung 4.3: *Aufbau und Streuparameter des verwendeten Testbeispiels.*

In Abb. 4.4 werden drei Kurven gezeigt: Dies sind zum einen der Fehler der Impedanz δ_Z, gemittelt über 1000 Frequenzpunkte, die äquidistant im gewählten Frequenzbereich liegen, der gemittelte Fehler der beiden Eigenwerte δ_E sowie die gemittelte

Differenz der Eigenwerte δ_D im Vergleich zur vorherigen Berechnung. Die Auswertung erfolgt alle 10 Iterationen. Die Lösungsgenauigkeit des Eigenwertlösers beträgt 10^{-12}. Es zeigt sich, dass alle drei Kurven proportionales Verhalten aufweisen. Bei der Wahl von $\varepsilon_{\delta\text{eig}} = 10^{-8}$ hätte das Kriterium nach 380 Iterationen gegriffen und die Iteration gestoppt. Es zeigt sich, dass das Übertragungsverhalten eine gewisse Stagnation bei $7 \cdot 10^{-11}$ aufweist. Der zugrundeliegende Fehler entsteht bei der Approximation der Polstellen, ist aber für praktische Anwendungen bedeutungslos.

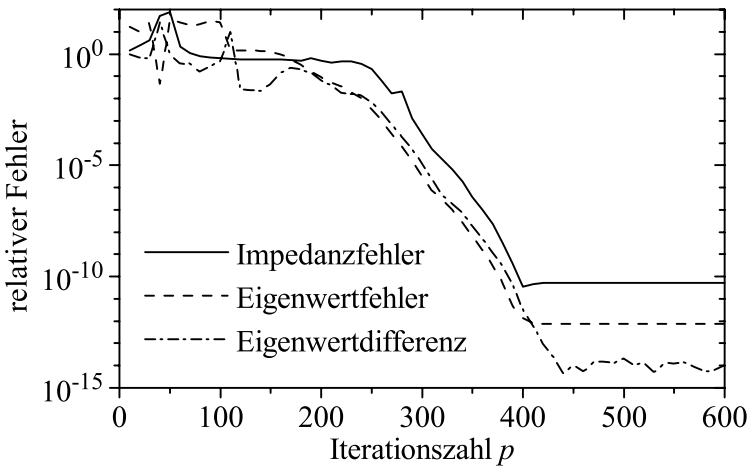


Abbildung 4.4: *Der Fehler der approximierten Impedanzfunktion δ_Z, der gemittelte Fehler aller im interessierenden Frequenzintervall liegenden Eigenwerte δ_E und die Differenz der Eigenwerte δ_D im Vergleich zum vorhergehenden Test einer partiellen Realisierung durch den Lanczos-Algorithmus.*

Das Abbruchkriterium erweist sich in der Praxis als sehr verlässlich und robust. Dennoch muss auf einige Eigenschaften hingewiesen werden, die das Verhalten erschweren können. Den entscheidenden Freiheitsgrad stellt die Wahl von $\varepsilon_{\delta eig}$ dar. Während bei nichtresonanten Strukturen bereits bei einer Fehlerschranke von 10^{-3} bis 10^{-4} eine sehr gute Übereinstimmung des Übertragungsverhaltens erkennbar ist, umgekehrt aber auch ein Fehler von 10^{-10} erreicht werden kann, ist die Situation bei hochresonanten Systemen mitunter komplizierter. Wird die Fehlerschranke zu großzügig gewählt, wird diese möglicherweise bereits von allen dominanten Eigenwerten erreicht, bevor tatsächlich *alle* im Frequenzband liegenden Eigenwerte erschienen sind. Dieser Effekt kann insbesondere auftreten, wenn eng beieinanderliegende, *geclusterte* Eigenwerte vorhanden sind, die erst verhältnismäßig spät „gefunden“ werden. Wird die Fehlerschranke zu streng angesetzt, zeigt sich im Gegenzug, dass vor Erreichen des geforderten Fehlers Geistereigenwerte die dominanten Eigenwerte verdoppeln. Das Kriterium kann in diesem Fall erst wieder greifen, wenn auch diese auskonvergiert sind. Dies führt nicht zum Versagen des Kriteriums, führt aber zu einem Modell von unnötiger Größe. Fehlerschranken zwischen 10^{-6} und 10^{-7} haben sich als guter Kompromiss für zahlreiche Testbeispiele bewährt.

Weitere wählbare Parameter sind die Differenz zwischen zwei Eigenwertüberprüfun-

gen Δp und die Iterationszahl der ersten Berechnung p_0. Je größer beide gewählt werden, je weniger Eigenwertberechnungen sind erforderlich, es steigt aber auch das Risiko, dass die ideale Modellgröße übersprungen wird. In beide Werte gehen sowohl die Systemgröße als auch die Portzahl als Einflussfaktoren ein. Da insbesondere die ersten Eigenwertberechnungen wenig numerischen Aufwand erfordern, sollte p_0 eher vorsichtig gewählt werden, bewährte Werte liegen zwischen 200 und 500 Iterationen. Da aufgrund des Blockalgorithmus zum Teil periodische Schwankungen in den Eigenwerten im Abstand von m erkennbar sind (siehe Abb. 4.5), sollte Δp ein ganzzahliges Vielfaches von m bilden, typischerweise mit Faktor 5 bis 20.

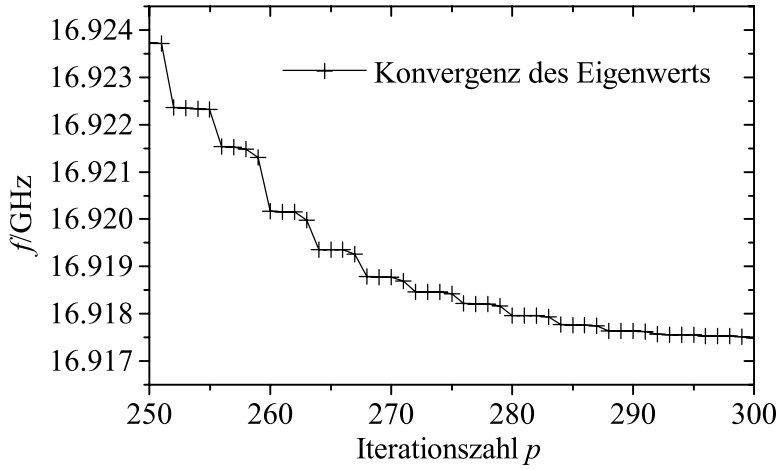


Abbildung 4.5: *Konvergenzverhalten eines Eigenwerts bei vier Startvektoren (Ports). Es zeigt sich ein periodisches Verhalten.*

Zur Berechnung der Eigenwerte aus dem Reduktionsprozess soll zum Abschluss noch auf einen Aspekt eingegangen werden. Während Krylov-Unterraum-Methoden zur Bestimmung von Eigenwerten üblicherweise mit Zufallsvektoren gestartet werden, von denen ausgegangen wird, das sie alle möglichen Raumrichtungen des Systems enthalten, werden die Algorithmen in Verbindung mit Modellreduktion mit den Matrizen $\mathbf{B}$ bzw. sowohl $\mathbf{B}$ als auch $\mathbf{C}$ gestartet. Damit werden die auffindbaren Eigenwerte von vornherein auf jene beschränkt, die von den Ports aus „sichtbar“ sind, die folglich von den Ports angeregt werden können und die umgekehrt auch in die Ports auskoppeln. Dies führt insbesondere dazu, dass divergenzbehaftete statische Moden von den divergenzfreien Portmoden nicht angeregt werden. Dieser Zusammenhang ist auch trotz endlicher Rechengenauigkeit gut erfüllt, wie in Abb. 4.6 für das obige Beispiel gezeigt. Während die Divergenz direkt nach der Multiplikation mit der Systemmatrix noch nahezu Null ist, wird durch die Orthonormierung des Lanczos-Vektors nach und nach eine Divergenz eingeführt. Diese steigt mit der Iterationszahl jedoch in der Folge an. Die Startdivergenz der Vektoren wird von dem auf 10^{-11} genau berechneten Portmode eingeführt. Da der Abbruch bereits bei 380 Iterationen erfolgt wäre, liegt der Divergenzanteil der Vektoren jedoch bei lediglich 10^{-7}. Auf Methoden wie die Einführung eines Grad-Div-Terms, die aus der klassischen Eigenwertberechnung bekannt sind und die die statischen Moden zu höheren Frequenzen verschieben, kann in Verbindung mit partiellen Realisierungen demnach

verzichtet werden. Eine solche Verschiebung wäre ohnehin nur unter der Bedingung sinnvoll, dass die statischen Moden nach der Verschiebung außerhalb des interessierenden Frequenzbandes liegen.

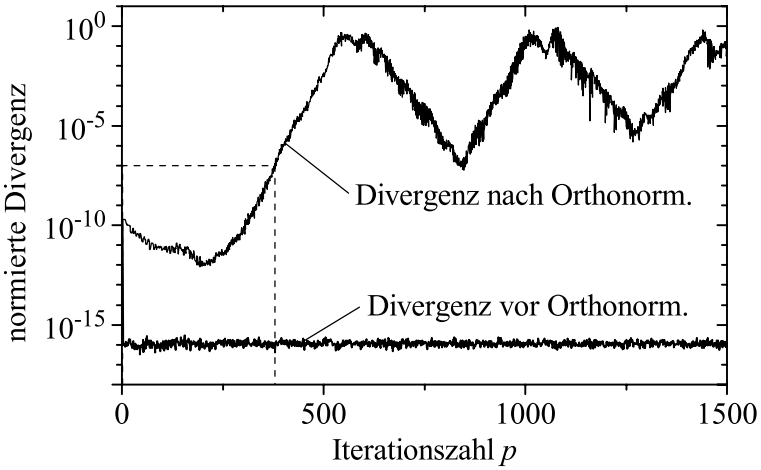


Abbildung 4.6: *Die Divergenz der Lanczos-Vektoren während der Iteration vor und nach dem Orthonormierungsschritt.*

4.3.1.3 Schwach verlustbehaftete Systeme

Mit der partiellen Realisierung durch Krylov-Unterraum-Verfahren wurde im vorigen Abschnitt ein sehr effizientes Verfahren zur Reduzierung der Modellordnung vorgestellt, welches allerdings nur auf eine einzelne Systemmatrix angewandt werden kann. Im Zusammenhang mit der Curl-Curl-Formulierung (Gl. 3.1.12) ist es jedoch wünschenswert, auch Verluste berücksichtigen zu können, also zugleich die Matrizen $\mathbf{A}_{CC}$ und $\mathbf{K}$ zu betrachten. Zwar ist es möglich, auf das lineare System zurückzugreifen, dies erhöht jedoch den Aufwand deutlich und die Passivität des reduzierten Modells ist nicht gewährleistet.

Eine Näherungslösung ergibt sich, wenn zur Erstellung des Unterraums die ebenfalls symmetrische aber komplexe Matrix $\mathbf{A}_{CC}^{(c)} = (\mathbf{A}_{CC} + s_0\mathbf{K})$ verwendet wird und die Matrizen $\mathbf{A}_{CC}$ und $\mathbf{K}$ im Anschluss separat auf das resultierende $\mathbf{V}_p$ projiziert werden. Für s_0 wird üblicherweise die Mittenfrequenz des interessierenden Spektrums multipliziert mit 2π, also ein rein imaginärer Wert, gewählt. Dieses Vorgehen ist vergleichbar zu dem Ansatz eines komplexen Dielektrikums zur Modellierung von Verlusten mit $\varepsilon^{(c)} = \varepsilon + \kappa/(j\omega_0)$. Als Folge wird das reduzierte Modell komplexwertig, was bei Verwendung als *Fast Frequency Sweep* nicht störend ist. Für die Erzeugung eines Ersatzschaltbilds ist es jedoch wünschenswert, das System reellwertig zu halten.

In vielen Fällen sind die Verluste jedoch rein parasitärer Natur und so gering, dass sie das elektromagnetische Feldbild innerhalb der Struktur im Vergleich zum verlustfreien Fall nicht nennenswert ändern. Unter diesen Umständen ist es daher meist ausreichend, den Projektionsunterraum wie zuvor allein mit der verlustfreien reellen

Systemmatrix $\mathbf{A}_{CC}$ zu bestimmen und im Anschluss die Verlustmatrix $\mathbf{K}$ ebenfalls symmetrisch darauf zu projizieren. Dieses Vorgehen entspricht der *Power-Loss-Methode* [2] in der klassischen Feldtheorie. Das reduzierte Modell ergibt sich zu

$$\mathbf{Z}_{pl}(s) = s\mathbf{B}_p^T \left(s^2\mathbf{I} + s\mathbf{V}_p^T\mathbf{K}\mathbf{V}_p + \mathbf{T}_p\right)^{-1} \mathbf{B}_p\,. \tag{4.3.12}$$

Es ist leicht zu erkennen, dass das System reell und symmetrisch bleibt und somit Stabilität und Passivität nach Abschnitt 3.2.3 erhalten werden. Zwar entsprechen sich durch dieses Vorgehen die Momente von Original- und reduziertem System nicht mehr exakt, erfahrungsgemäß ergeben sich aber sehr gute Übereinstimmungen für Verlustwinkel bis $\tan\delta = 0.1$, ein weit höherer Wert als typische Dielektrika aufweisen.

Nachteil des Verfahrens ist aber zweifelsohne, dass die Matrix $\mathbf{V}_p$ mit $n \times p$ Einträgen komplett im Speicher gehalten werden muss. Sollte dies nicht möglich sein, gibt es die folgenden Alternativen zur Berechnung von $\mathbf{K}_p = \mathbf{V}_p^T\mathbf{K}\mathbf{V}_p$:

- Die Matrix $\mathbf{V}_p$ wird abgesehen von den letzten $2m+1$ Spalten innerhalb des Lanczos Algorithmus nicht gleichzeitig benötigt. Die Vektoren können daher in periodischen Abständen auf Festplatte gespeichert und zur Projektion erneut geladen werden.

- Häufig sind nur einzelne Bereiche der Struktur verlustbehaftet, beispielsweise Dielektrika, während andere Materialien wie z. B. Luft als verlustfrei angenommen werden können. In diesem Fall hat die Diagonalmatrix $\mathbf{K}$ nur wenige Einträge ungleich Null und es genügt, die dünnerbesetzte Matrix $\mathbf{K}' = \mathbf{K}^{1/2}\mathbf{V}_p$ abzuspeichern und daraus $\mathbf{K}_p = \mathbf{K}'^T\mathbf{K}'$ zu berechnen.

- Die Matrix $\mathbf{K}_p$ zeigt, auch wenn sie nicht diagonaldominant ist, doch eine Betonung der Werte um die Diagonale. Bei simultaner Projektion auf eine feste Anzahl von Spalten in $\mathbf{V}_p$ werden genau diese inneren Werte berechnet. Der sich ergebende Fehler ist bei 50 bis 100 Spalten meist sehr gering, es ist allerdings schwierig, eine Abschätzung für ihn anzugeben. Typischerweise ist er geringer als bei Verwendung einer komplexen Systemmatrix, die die Leitfähigkeit bei der Mittenfrequenz fixiert.

Leider lassen sich die genannten Punkte nicht ohne Weiteres auf den Fall von Impedanzwänden übertragen. Zwar lässt sich ebenfalls nach Gl. 3.1.14 die Projektion

$$\mathbf{Z}_{p\mathrm{iw}}(s) = s\mathbf{B}_p^T \left(s^2\mathbf{I} + \frac{1}{\sqrt{s}}\mathbf{V}_p^T\mathbf{P}\mathbf{V}_p + \mathbf{T}_p\right)^{-1} \mathbf{B}_p\,. \tag{4.3.13}$$

bilden, allerdings weist die Matrix $\mathbf{P}$ nach Gl. 3.1.14 eine weit komplexere Bandstruktur als $\mathbf{K}$ auf. Sie ist nicht diagonal und auch die projizierte Matrix $\mathbf{P}_p = \mathbf{V}_p^T\mathbf{P}\mathbf{V}_p$ weist keine Diagonalbetonung auf. Es ist also die volle Matrix $\mathbf{V}_p$ zur Projektion zu verwenden.

4.3.1.4 Konvergenzbeschleunigung für partielle Realisierungen

Obwohl es einer der Hauptvorteile der partiellen Realisierung durch Krylov-Unterräume ist, dass der entsprechende Unterraum selbst nicht gespeichert werden muss, sondern die Projektion implizit erfolgt, sind doch bereits mehrfach Punkte angeklungen, in denen die Kenntnis der Matrix $\mathbf{V}_p$ sehr nützlich wäre. Diese sind, wie gerade beschrieben, die schwachen Verluste, insbesondere die Impedanzwandverluste, die symmetrische Projektion unsymmetrischer Systemmatrizen sowie die Berechnung von Feldlösungen, deren Lösungsvektor über Gl. 4.2.1 mit der für sich aussagelosen Lösung $\mathbf{y}$ des reduzierten Systems verkoppelt ist. Die sich stellende Frage lautet also, ob es möglich ist, das Verfahren in einer Form zu variieren, dass der Unterraum eine geringere Dimension erhält und auch für große Beispiele im Speicher gehalten werden kann.

Wegen der engen Zusammenhänge von Ordnungsreduktion, linearen Gleichungslösern und Eigenwertberechnung, jeweils basierend auf Krylov-Unterräumen, bietet es sich an, daraus bekannte Techniken auf die Ordnungsreduktion zu übertragen. So geht linearen Gleichungslösern meist eine so genannte *Vorkonditionierung* voraus, bei Eigenwertlösern werden Verfahren wie Beschleunigungspolynome oder interne Neustarts (engl. *Implicit Restarts*) verwendet. Diese drei Punkte sollen im Weiteren genauer untersucht werden.

Vorkonditionierung:

Ziel der Vorkonditionierung ist es, die Systemmatrix vor Start des Unterraum-Verfahrens so zu verändern, dass die Lösung für einen vorgegebenen Startvektor innerhalb von möglichst wenig Iterationsschritten konvergiert. Hierbei werden unter anderem sehr komplexe Algorithmen wie das Mehrgitter-Verfahren [55] angewandt, die das System durch Transformationen auf unterschiedlich feine Gitter mehrfach mit nur grober Genauigkeit lösen, bevor die endgültige Lösung auf dem feinsten Gitter erfolgt. Solch komplexe Verfahren greifen tief in die Struktur des Systems ein und lassen sich nicht ohne großen Aufwand in das hier beschriebene Projektionsverfahren integrieren, sie sollen daher im Rahmen dieser Arbeit nicht weiter verfolgt werden. Dasselbe gilt für die Klasse der *Successive Over-Relaxation, (SOR)*-Verfahren [45], die die Diagonale sowie obere und untere Dreiecksmatrix verwenden, um die Inverse von $\mathbf{A}$ nachzubilden. Während die genannten Matrizenteile für Gleichungslöser nur für eine Rückeinsetzung genutzt werden, müssten sie im Fall der Ordnungsreduktion tatsächlich invertiert werden, was für große Systeme nicht möglich ist. Zudem wirken diese Verfahren nur schmalbandig, was die breiterbandige Nutzung des reduzierten Modells nicht gewährleistet.

Ein verhältnismäßig einfaches Verfahren, die *Jakobi*-Vorkonditionierung [45], skaliert das System lediglich mit der Diagonale der Systemmatrix. Die Effizienz hängt von der Formulierung des Systems ab. Da die Systemmatrix des linearen verlustfreien Systems auf der Diagonalen nur Nulleinträge besitzt, ist eine derartige Vorkonditionierung in diesem Fall nicht möglich. Für Curl-Curl-Systeme wurden zwei Formulierungen eingeführt: $\mathbf{A}'_{CC} = \mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}$ oder $\mathbf{A}_{CC} = \mathbf{M}_\varepsilon^{-1/2}\mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}\mathbf{M}_\varepsilon^{-1/2}$, diese unterscheiden sich gerade durch die Diagonalmatrix $\mathbf{M}_\varepsilon$. Die Effizienz der Jakobi-Vorkonditionierung ist für beide Fälle in Tabelle 4.1 anhand der benötigten

Iterationszahl und Rechenzeit[3] zur Lösung eines Frequenzpunkts mit dem Lanczos-basierten BiCG-Verfahren für ein Beispiel mit ~ 30.000 Unbekannten dargestellt.

Formulierung	**ohne Jakobi VK**		**mit Jakobi VK**	
	Iter.	**CPU [s]**	**Iter.**	**CPU [s]**
$\mathbf{A}'_{CC} = \mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}$	> 10.000	> 1.500	3.346	883
$\mathbf{A}_{CC} = \mathbf{M}_\varepsilon^{-1/2}\mathbf{C}^T\mathbf{M}_\mu^{-1}\mathbf{C}\mathbf{M}_\varepsilon^{-1/2}$	730	119	3.328	880

Tabelle 4.1: *Vergleich des Aufwands zur Lösung eines Systems mit und ohne Jakobi-Vorkonditionierung für unterschiedliche Curl-Curl-Formulierungen. Betrachtet werden die Iterationszahl (Iter.) und die Rechenzeit (CPU).*

Es zeigt sich, dass die Vorkonditionierung nur im ersten Fall eine nennenswerte Verbesserung bringt, während sie im zweiten Fall eine Verschlechterung bewirkt. Anders formuliert bedeutet dies, dass die Matrix $\mathbf{M}_\varepsilon$ bereits einen besseren Vorkonditionierer darstellt als die Diagonale der Systemmatrix. Da üblicherweise die Formulierung $\mathbf{A}_{CC}$ gewählt wird, lässt sich durch diagonale Vorkonditionierung folglich keine Verkleinerung des Modells erzielen.

Polynombeschleunigung:

Einen in der Eigenwertberechnung erfolgreich verwendeten Ansatz bilden Beschleunigungspolynome wie beispielsweise Tschebyscheff-Polynome. Bei wiederholter Anwendung in Verbindung mit dem Neustart des Unterraum-Verfahrens wird von *Explicit Restarts* [45] gesprochen. Im Zusammenhang mit der Ordnungsreduktion sollen die Polynome jedoch nur einmalig auf die Startmatrix $\mathbf{B}$ angewendet werden.

Wird der Vektor $\mathbf{b}_m$ erneut entsprechend Gl. 4.2.10 als Überlagerung der Eigenvektoren dargestellt, führt die Anwendung eines Matrix-Polynoms $P_K(\mathbf{A})$ mit den Wurzeln r_k auf:

$$\mathbf{b}_{pol} = \left\{\prod_{k=1}^{K}(\mathbf{A} - r_k\mathbf{I})\right\}\mathbf{b}_m = \sum_{\eta=1}^{n}\alpha_\eta\left\{\prod_{k=1}^{K}(\lambda_\eta - r_k\mathbf{I})\right\}\mathbf{x}_\eta \tag{4.3.14a}$$

$$= \sum_{\eta=1}^{n}\alpha_\eta P_K(\lambda_\eta)\mathbf{x}_\eta. \tag{4.3.14b}$$

Mit der richtigen Wahl der Polynomnullstellen r_k lassen sich also die Anteile einzelner Eigenvektoren in $\mathbf{b}_{\text{pol}}$ betonen, während andere, meist die zu Eigenwerten weit außerhalb des interessierenden Bereichs gehörigen, stark unterdrückt werden können.

Eine geeignete Wahl bilden Tschebyscheff-Polynome [46] mit der Eigenschaft, dass sich der Funktionswert im Bereich der Nullstellen nur zwischen 1 und -1 bewegt, jenseits der Randnullstellen aber sehr steil ansteigt. Die Polynomnullstellen werden so gewählt, dass der unbeachtete Anteil des Spektrums stark gedämpft wird. Ein Beispiel für ein Tschebyscheff-Polynom ist in Abb. 4.7 gezeigt.

[3]Die Rechnungen erfolgten auf einem 731 MHz PC.

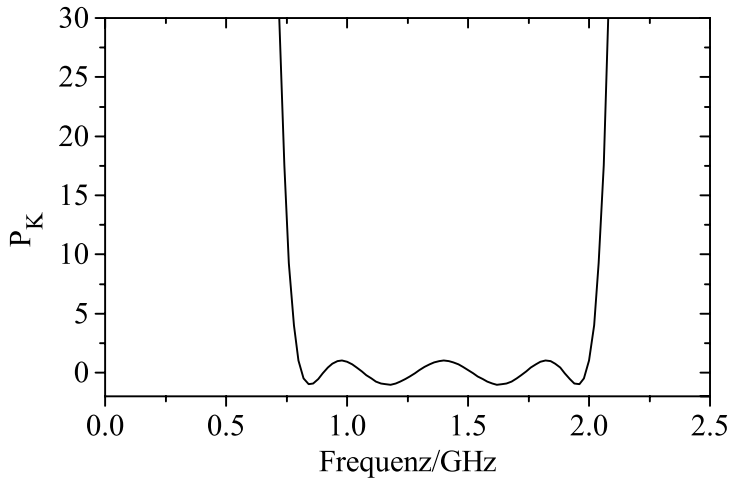


Abbildung 4.7: *Tschebyscheff-Polynom, das das zwischen 0,8 und 2 GHz liegende Spektrum dämpft.*

Wird der Lanczos Algorithmus mit den Vektoren $\mathbf{b}_{\mathrm{pol}}$ anstelle von $\mathbf{b}_m$ gestartet, konvergieren die Eigenwerte des reduzierten Systems deutlich schneller und das Abbruchkriterium wird bereits mit einer kleineren Anzahl von Unterraumvektoren erfüllt. Je größer die Ordnung K des Polynoms gewählt wird, je weniger Iterationen werden im Lanczos Algorithmus benötigt, was einen nahezu lückenlosen Übergang zur reinen Modalanalyse ermöglicht, die direkt die Eigenvektoren verwendet und im folgenden Abschnitt erläutert wird. Eine Abschätzung für die benötigte Anzahl von Polynomnullstellen ist in [68] gegeben.

Bei Betrachtung der Gesamtrechenzeit zur Erstellung der reduzierten Modelle zeigt sich, dass diese durch Beschleunigungspolynome unterschiedlicher Ordnung nahezu konstant bleibt. Die im Lanczos Verfahren eingesparte Zeit wird für die zusätzlichen Matrix-Vektor-Multiplikationen des Polynoms in etwa aufgewogen. Das Ziel, die Matrixgröße von $\mathbf{V}_p$ zu verringern, wird also ohne Vergrößerung der Rechenzeit problemlos erreicht.

Es ist zu beachten, dass bei Verwendung eines Beschleunigungspolynoms die Matrix $\mathbf{B}_p = \mathbf{V}_{pA}^T\mathbf{B}$ explizit berechnet werden muss, da durch die Polynomanwendung für die implizit entstehende Matrix des Lanczos-Algorithmus gilt: $\mathbf{B}_{pA} = \mathbf{V}_{pA}^T\mathbf{B}_{pol} \neq \mathbf{V}_{pA}^T\mathbf{B}$. Es soll zudem betont werden, dass es sich bei den resultierenden Modellen nicht mehr um partielle Realisierungen im engeren Sinne handelt, da keine Markov-Parameter mehr übereinstimmen. Die Approximationseigenschaften bleiben aber, zumindest bei gemäßigter Unterdrückung höherer Moden, gut erhalten. Werden die höheren Moden sehr stark gedämpft, wird das Fehlen ihrer Anteile jedoch im Spektrum durch einen Offset-Fehler in der Impedanz erkennbar. Dieser Fehler ist aus der Modalanalyse bekannt und kann durch einen Korrekturterm ausgeglichen werden. Das genaue Vorgehen wird im Rahmen der Modalanalyse im nächsten Abschnitt beschrieben.

Implicit Restarts:

Das Verfahren wird ebenfalls mit Erfolg bei der Eigenwertberechnung [54] angewendet und wurde in [58, 59] bereits in Verbindung mit Ordnungsreduktion gebracht.

Das Verfahren beruht zunächst auf dem Standard-Lanczos- oder Arnoldi-Algorithmus, der nach $k+p$ Iterationen gestoppt wird. Nach einer Eigenwertzerlegung der Matrix $\mathbf{T}_p$ werden die $k+p$ Eigenwerte s_i in k *gewünschte* (im interessierenden Frequenzbereich liegende) und p *unerwünschte* getrennt. Insbesondere für den unsymmetrischen Fall, in dem instabile Pole auftreten können, können diese zu den unerwünschten gerechnet und so beseitigt werden. Anschließend wird der Einfluss der ungewollten Eigenwerte bzw. -vektoren durch p Durchläufe der folgenden Befehle mit einer QR-Zerlegung eliminiert:

$$i = 1 \dots p: \qquad [\mathbf{Q}_i, \mathbf{R}_i] = qr(\mathbf{T}_p - s_i\mathbf{I}), \tag{4.3.15a}$$

$$\mathbf{T}_p = \mathbf{Q}_i^T \mathbf{T}_p \mathbf{Q}_i;\ \mathbf{V}_p = \mathbf{V}_p \mathbf{Q}_i, \tag{4.3.15b}$$

$$\hat{\mathbf{v}} = \hat{\mathbf{v}}^T \mathbf{Q}_i. \tag{4.3.15c}$$

Die Matrix $\mathbf{T}_p$ verändert hierdurch ihre Struktur nicht und kann im Folgenden auf den linken oberen $k \times k$-Block beschnitten werden. Entsprechend wird $\mathbf{V}_p$ auf die linken k Vektoren begrenzt. Der Lanczos-Algorithmus kann nun erneut für p Iterationen gestartet werden, bevor das obige Vorgehen erneut angewendet wird.

Das Verfahren kann ebenfalls als interne Anwendung eines Beschleunigungspolynoms interpretiert werden. Es ist daher auch möglich, andere Wurzeln des Polynoms zu wählen als die ungewollten Eigenwerte, beispielsweise erneut Nullstellen eines Tschebyscheff-Polynoms.

Praktische Tests haben jedoch gezeigt, dass das Verfahren für große Ausgangssysteme aufgrund der zahlreichen QR-Zerlegungen deutlich rechenaufwendiger ist als die oben beschriebene Polynombeschleunigung. Zudem sind die Parameter k und p für einen optimalen Ablauf schwer vorher bestimmbar und es zeigt sich, dass bereits eliminierte Anteile durch numerisches Rauschen immer wieder neu angeregt und später erneut eliminiert werden, was die Effizienz des Verfahrens einschränkt.

4.3.2 Korrigierte Modalanalyse

Werden zur Projektion anstelle der Krylov-Vektoren eine Anzahl von Eigenvektoren verwendet, spricht man von Modalanalyse [27, 69]. Die Eigenvektoren werden zuvor mit einem geeigneten Verfahren, meist ebenfalls einem Unterraumverfahren, explizit berechnet.

Nach Abschnitt 3.2.2.2 bildet der volle Raum der Eigenvektoren $\mathbf{x}_\eta$, $\eta = 1 \dots n$ im verlustlosen Fall eine orthogonalisierbare bzw. orthonormalisierbare Basis der Systemmatrix

$$\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T \quad \text{mit} \quad \mathbf{X}^{-1} = \mathbf{X}^T. \tag{4.3.16}$$

Für das verlustfreie Curl-Curl-System gilt für die Eigenwerte dabei

$$\mathbf{\Lambda} = \operatorname{diag}(\lambda_\eta) = \operatorname{diag}(-s_\eta^2). \tag{4.3.17}$$

Wird 4.3.16 anstelle von $\mathbf{A}$ in Gl. 3.1.12 eingesetzt, folgt für die Impedanz

$$\mathbf{Z}(s) = s\mathbf{B}^T \left(s^2\mathbf{I} + \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T\right)^{-1} \mathbf{B} \tag{4.3.18a}$$

$$= s\mathbf{B}^T\mathbf{X} \left(s^2\mathbf{I} + \mathbf{\Lambda}\right)^{-1} \mathbf{X}^T\mathbf{B} \tag{4.3.18b}$$

$$= s\mathbf{B}^T\mathbf{X} \operatorname{diag}\left(\frac{s}{s^2 - s_\eta^2}\right) \mathbf{X}^T\mathbf{B}. \tag{4.3.18c}$$

Mit dieser Formulierung lässt sich jeder Eintrag Z_{ij} von $\mathbf{Z}(s)$ auch als Summe über die Beiträge aller n Eigenvektoren schreiben

$$Z_{ij}(s) = \sum_{\eta=1}^{n} (\mathbf{x}_\eta^T\mathbf{b}_i)(\mathbf{x}_\eta^T\mathbf{b}_j)\frac{s}{s^2 - s_\eta^2} = \sum_{\eta=1}^{n} Z_{ij}^\eta(s). \tag{4.3.19}$$

Zur Ordnungsreduktion wird die obige Summe nach dem Term p abgebrochen. Mit dem Korrekturterm, der die übrigen Moden modelliert, ergibt sich

$$Z_{ij}(s) = \sum_{\eta=1}^{p} Z_{ij}^\eta(s) + Z_{ij}^{\text{corr}}(s) \quad \text{mit} \quad Z_{ij}^{\text{corr}}(s) \approx \sum_{\eta=p+1}^{n} Z_{ij}^\eta(s), \tag{4.3.20}$$

oder wieder in der gewohnten Projektionsschreibweise

$$\mathbf{Z}_p(s) = s\mathbf{B}^T\mathbf{X}_p \left(s^2\mathbf{I} + \mathbf{\Lambda}_p\right)^{-1} \mathbf{X}_p^T\mathbf{B} + \mathbf{Z}_{\text{corr}}(s). \tag{4.3.21}$$

Es ist offensichtlich, dass auch durch diese symmetrische Projektion Stabilität und Passivität im reduzierten Modell erhalten werden.

Die Güte der Approximation hängt stark von der Auswahl der Eigenwerte sowie vom Korrekturterm ab. Es ist offensichtlich, dass zumindest alle zu den im interessierenden Frequenzband liegenden Eigenwerten gehörigen Eigenvektoren verwendet werden sollten. Es zeigt sich folglich eine enge Analogie zum Abbruchkriterium der partiellen Realisierung (Abschnitt 4.3.1.2). Dennoch können auch außerhalb des Frequenzintervalls dominante Eigenwerte mit großem Einfluss auf das Ergebnis liegen. In [41] wird daher ein Dominanzmaß für Eigenwerte definiert. Dieses berücksichtigt letztlich die Größe der Residuen $(\mathbf{x}_\eta^T\mathbf{b}_i)(\mathbf{x}_\eta^T\mathbf{b}_j)$. Da aus systemtheoretischer Sicht jeder Eigenwert mit einem inneren Systemzustand verkoppelt ist, sind die Residuen ein Maß für die Steuer- und aufgrund der Symmetrie zugleich der Kontrollierbarkeit des zugehörigen Zustands. Eine Auswahl anhand des Residuums bewertet folglich die Zustände/Moden anhand ihrer Erreichbarkeit von den Ports. Nach demselben Prinzip arbeitet die Methode *Balanced Truncation*, auf die später noch ausführlicher eingegangen wird.

Um die Vorteile des kleinen Projektionsunterraums beizubehalten und somit die Zahl der Eigenvektoren nicht zu sehr durch Hinzunahme aller dominanten Eigenwerte ansteigen zu lassen, besteht ein alternativer Ansatz darin, nur eine Mindestanzahl

von Eigenvektoren, beispielsweise die im betrachteten Frequenzbereich liegenden, zu verwenden und den Fehler in der Approximation der Impedanz durch den Korrekturterm auszugleichen.

Eine sehr aufwendige Ableitung eines solchen Korrekturterms wurde in [69] vorgeschlagen, wo durch die Lösung eines komplementären Eigensystems mit ausgetauschten Randbedingungen an den Ports eine Approximation der Admittanz $\mathbf{Y}_p(s) = \mathbf{Z}_p^{-1}(s)$ berechnet wird. Aus dem Zusammenhang der Pole und Nullstellen von $\mathbf{Y}_p(s)$ und $\mathbf{Z}_p(s)$ ergibt sich schließlich eine sehr genaue Approximation des Korrekturterms $\mathbf{Z}_{\mathrm{corr}}(s)$.

Eine wesentlich einfachere Herleitung, eingeführt in [70] und ausführlich untersucht in [27], beruht auf der exakten Lösung des Systems an einem oder wenigen Frequenzpunkten. Für Moden, deren Eigenfrequenz weit vom interessierenden Frequenzspektrum liegt, gilt

$$\frac{s}{s^2 - s_\eta^2} \approx -\frac{s}{s_\eta^2} \quad \text{für} \quad |s| \ll |s_\eta|. \tag{4.3.22}$$

Folglich kann die Summe aller nichtbeachteten Moden durch eine lineare Funktion mit den exakten Lösungen als Stützwerten bestimmt werden. Der zusätzliche Aufwand für die exakte Gleichungslösung ist typischerweise gering, da sich durch die niederdimensionale Lösung $\mathbf{y}$ des reduzierten Systems Gl. 4.3.21 mit $\mathbf{x}_s = \mathbf{X}_p\mathbf{y}$ ein guter Startvektor für das Originalsystem angeben lässt.

Die Erfahrung zeigt, dass dieser Typ von Korrektur zu sehr genauen Approximationen der Übertragungsfunktion führt. Es sei erneut darauf hingewiesen, dass derselbe Korrekturterm auch in Verbindung mit polynombeschleunigten partiellen Realisierungen verwendet werden kann.

Der numerische Aufwand der korrigierten Modalanalyse ist mit dem der partiellen Realisierungen nur schwer vergleichbar, da er größtenteils von Typ und Parametersatz des verwendeten Eigenwertlösers abhängig ist. Die Erfahrung zeigt jedoch, dass er meist höher ist als für partielle Realisierungen.

Vorteil des Verfahrens ist aber zweifelsohne die geringe Größe des Projektionsraums. Werden nur die innerhalb des interessierenden Frequenzbereichs liegenden Eigenwerte berücksichtigt, kann die Modellgröße als minimal angesehen werden. Zudem ist die Berechnung von Feldlösungen nach Gl. 4.2.1 einfach durchführbar und analog zu Abschnitt 4.3.1.3 können schwache Verluste ebenfalls auf die Eigenvektoren projiziert werden, mit dem Vorteil, dass die Matrix $\mathbf{X}_p$ wesentlich leichter zu handhaben ist als die deutlich größere $\mathbf{V}_p$.

4.3.3 Padé-Approximationen

Mit denselben Methoden und Algorithmen, die in den vorigen Abschnitten im Zusammenhang mit partiellen Realisierungen zur Ordnungsreduktion beschrieben wurden, lassen sich auch Taylorkoeffizienten bzw. Momente um vorgegebene Entwicklungsfrequenzen bestimmen. Dies führt auf klassische Padé-Approximationen [51].

Da der Entwicklungspunkt direkt in den interessierenden Frequenzbereich gelegt werden kann und er nicht wie bei partiellen Realisierungen im Unendlichen liegt, ist es zu erwarten, dass diese Modelle eine wesentlich geringere Ordnung haben als partielle Realisierungen. Neben den Vorteilen eines kleinen Modells ermöglicht die kleine Ordnung auch, dass die Projektionsmatrizen $\mathbf{V}_q$ und $\mathbf{W}_q$ weit einfacher im Speicher gehalten werden können, was beispielsweise die Berechnung von Feldlösungen wesentlich vereinfacht.

4.3.3.1 Padé-Approximationen durch Krylov-Unterräume

Die Entwicklungsfrequenz s_0 fließt durch folgende Umformung in das System ein, hier allgemein für das lineare System dargestellt:

$$\mathbf{Z}(s) = \mathbf{C}\left(\mathbf{I} + (s - s_0)\hat{\mathbf{A}}\right)^{-1}\hat{\mathbf{B}} \tag{4.3.23a}$$

$$\text{mit} \quad \hat{\mathbf{A}} = (\mathbf{A} + s_0\mathbf{I})^{-1}, \quad \hat{\mathbf{B}} = \hat{\mathbf{A}}\mathbf{B}. \tag{4.3.23b}$$

Ein entscheidender Nachteil dieser Vorgehensweise wird aus Gl. 4.3.23b unmittelbar deutlich: die Matrix $(\mathbf{A} + s_0\mathbf{I})$ muss invertiert werden. Da auf diese Weise die Bandstruktur von $\mathbf{A}$ verloren geht, ist dies aufgrund der Matrixdimension selbst für Systeme mittlerer Größe aus Gründen des benötigten Speichers nicht möglich, vom numerischen Aufwand der Inversion ganz abgesehen.

Alle Anwendungen von $\hat{\mathbf{A}}$ für Matrix-Vektor-Multiplikationen $\hat{\mathbf{A}}\mathbf{x} = \mathbf{y}$ müssen daher alternativ durch Lösen eines linearen Gleichungssystems Solve$\{(\mathbf{A} + s_0\mathbf{I})\mathbf{y} = \mathbf{x}\}$ behandelt werden. Sofern sich eine dünnbesetzte LU-Zerlegung [43] bilden lässt, kann dies verhältnismäßig einfach durch zwei Rückeinsetzungen in Dreiecksmatrizen erfolgen, für größere Systeme scheitert aber auch die LU-Zerlegung an Speichermangel, weswegen iterative Gleichungslöser, häufig auf Krylov-Unterraum-Verfahren basierend, verwendet werden müssen.

Wird das Bi-Lanczos-Verfahren nun mit den Matrizen $\hat{\mathbf{A}}$, $\hat{\mathbf{B}}$ und $\mathbf{C}$ gestartet, um die Krylov-Unterräume $\mathcal{K}(\hat{\mathbf{A}}, \hat{\mathbf{B}})$ und $\mathcal{K}(\hat{\mathbf{A}}^T, \mathbf{C}^T)$ aufzubauen, ergibt sich analog zu 4.3.2a ein reduziertes Modell der Form

$$\mathbf{Z}_p(s) = \mathbf{C}\mathbf{V}_q\left(\mathbf{W}_q^T\mathbf{V}_q + (s - s_0)\mathbf{W}_q^T\hat{\mathbf{A}}\mathbf{V}_q\right)^{-1}\mathbf{W}_q^T\hat{\mathbf{B}} \tag{4.3.24a}$$

$$= \mathbf{C}_q\left(\mathbf{I} + (s - s_0)\mathbf{T}_q\right)^{-1}\mathbf{B}_q. \tag{4.3.24b}$$

Um Verwechslungen zu vermeiden, werden die Krylov-Matrizen $\mathbf{W}$, $\mathbf{V}$ und $\mathbf{T}$, die aus dem invertierten System generiert wurden, im Weiteren mit dem Index q versehen, während diejenigen aus der direkten Anwendung weiterhin mit p indiziert werden.

Der Vorteil des Verfahrens zeigt sich unmittelbar, wenn 4.3.23a als geometrische Reihe geschrieben wird:

$$\mathbf{Z}(s) = \sum_{k=0}^{\infty}\mathbf{C}(-\hat{\mathbf{A}})^k\mathbf{B}\,(s - s_0)^k = \sum_{k=0}^{\infty}\mathbf{M}_k(s - s_0)^k\,. \tag{4.3.25}$$

Die Werte $\mathbf{M}_k = \mathbf{C}(-\hat{\mathbf{A}})^k\hat{\mathbf{B}}$ entsprechen Taylorkoeffizienten um die Kreisfrequenz s_0 und werden in Anlehnung an die Mechanik häufig auch als *Momente* bezeichnet. Entsprechend der Argumentation in Abschnitt 4.3.1 lässt sich direkt zeigen, dass auch bei dieser Projektion mit $m_q = \text{floor}(q/m)$ und $l_q = \text{floor}(q/l)$ bei m Eingangs- und l-Ausgangsports

$$\mathbf{M}_i = \mathbf{C}(-\hat{\mathbf{A}})^i\hat{\mathbf{B}} = \mathbf{C}_q(-\mathbf{T}_q)^i\mathbf{B}_q \quad \text{für} \quad 0 \leq i < m_q + l_q - 1 \tag{4.3.26}$$

gilt, und somit folglich die ersten $m_q + l_q$ Momente von Original- und reduziertem System übereinstimmen. Das reduzierte Modell 4.3.25 stellt demnach eine Padé-Approximation dar, bei der die maximal mögliche Anzahl von Taylorkoeffizienten um s_0 identisch ist.

Diese Zusammenhänge wurden erstmals in [61] vorgestellt, vor nahezu zehn Jahren wurden sie unabhängig voneinander in [58] und [63] erstmals im Rahmen der Netzwerkanalyse für größere Systeme erfolgreich angewandt. Insbesondere die Namensgebung *Padé Via Lanczos, PVL* aus [63] wurde sehr geläufig.

4.3.3.2 Ausnutzung von Symmetrien

Erneut wurde die obige Herleitung sehr allgemein gehalten und es wird üblicherweise $\mathbf{C} = \mathbf{B}^T$ und damit $m_q = l_q$ gelten. Auch im Fall symmetrischer Systemmatrizen $\mathbf{A}$ lässt sich zunächst jedoch nicht der symmetrische Lanczos-Algorithmus verwenden, da zwar $\hat{\mathbf{A}}^T = \hat{\mathbf{A}}$, aber auch $\mathbf{B} \neq \hat{\mathbf{B}}$ gilt.

Dennoch kann man sich die Symmetrie $\hat{\mathbf{A}}^T = \hat{\mathbf{A}}$ innerhalb des Bi-Lanczos-Algorithmus zunutze machen, da das Verhältnis der Startmatrizen über $\hat{\mathbf{B}} = \hat{\mathbf{A}}\mathbf{B}$ bekannt ist. Für zunächst reelle Extraktionsfrequenzen s_0 gilt damit für die links- und rechtsseitigen Krylov-Unterräume

$$\text{span}\{\mathbf{V}_q\} = \mathcal{K}(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \mathcal{K}(\hat{\mathbf{A}}, \hat{\mathbf{A}}\mathbf{B}) = \hat{\mathbf{A}}\,\mathcal{K}(\hat{\mathbf{A}}, \mathbf{B}) \qquad \text{und} \tag{4.3.27a}$$

$$\text{span}\{\mathbf{W}_q\} = \mathcal{K}(\hat{\mathbf{A}}^T, \mathbf{B}) = \mathcal{K}(\hat{\mathbf{A}}, \mathbf{B}). \tag{4.3.27b}$$

Es gilt folglich auch für die Krylov-Vektoren

$$\mathbf{w}_q = \hat{\mathbf{A}}^{-1}\mathbf{v}_q = (\mathbf{A} + s_0\mathbf{I})\mathbf{v}_q. \tag{4.3.28}$$

Für allgemeine komplexe s_0 bleibt $\mathbf{A}$ komplex symmetrisch und es gilt

$$\mathcal{K}(\hat{\mathbf{A}}^H, \mathbf{B}) = \left(\mathcal{K}(\hat{\mathbf{A}}^T, \mathbf{B})\right)^*. \tag{4.3.29}$$

Der Vektor $\mathbf{w}_q$ muss damit ebenfalls konjugiert werden:

$$\mathbf{w}_q = \left((\mathbf{A} + s_0\mathbf{I})\mathbf{v}_q\right)^*. \tag{4.3.30}$$

Insbesondere wenn die neuen Vektoren über ein numerisch aufwendiges Gleichungslösungsverfahren gewonnen werden müssen, genügt es, nur die Vektoren $\mathbf{v}_q$ auf diese

Weise zu bestimmen, während die Vektoren $\mathbf{w}_q$ durch eine einfache Matrix-Vektor-Multiplikation daraus generiert werden, was den numerischen Aufwand des Verfahrens in etwa halbiert.

Diese Einsparung kann nicht allgemein auf schiefsymmetrische Systeme, beispielsweise das verlustfreie lineare System, übertragen werden, da aus $\mathbf{A}^T = -\mathbf{A}$ nicht die Schiefsymmetrie von $\hat{\mathbf{A}}^T \neq -\hat{\mathbf{A}}$ folgt. Für den in der Praxis wichtigen Sonderfall einer reell schiefsymmetrischen Systemmatrix $\mathbf{A}$ und einer rein komplexen Entwicklungsfrequenz s_0 lässt sich jedoch zeigen, dass $\hat{\mathbf{A}} = -\hat{\mathbf{A}}^H$ gilt. Mit der Skalierungsrelation Gl. 4.2.6 kann $\mathcal{K}(\hat{\mathbf{A}}^H, \mathbf{B}) = \mathcal{K}(-\hat{\mathbf{A}}, \mathbf{B}) = \mathcal{K}(\hat{\mathbf{A}}, \mathbf{B})$ gezeigt werden, womit $\mathbf{w}_q$ auch nach Gl. 4.3.28 bestimmt werden kann.

Während die gerade beschriebene Vereinfachung für alle reellen schiefsymmetrischen Systeme gültig ist, kann für lineare FIT-Systeme ein Vorteil aus der speziellen Matrixstruktur gezogen werden. Dazu ist eine Blockstruktur, wie beispielsweise im linearen System Gl. 3.1.2, erforderlich. Insbesondere darf $\mathbf{B}$, wie dort beschrieben, nur im oberen Block Einträge haben, es muss folglich $\mathbf{B} = (\mathbf{R}',\ \mathbf{0})^T$ gelten. Unter diesen Umständen kann eine den Blockgrößen n_e und n_h entsprechende Matrix

$$\mathbf{E} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{pmatrix} \tag{4.3.31}$$

definiert werden, mit der sich das System folgendermaßen umstellen lässt

$$\mathbf{Z}(s) = \mathbf{B}^T \left(\mathbf{A}\mathbf{E} + s\mathbf{E}\right)^{-1} \mathbf{B}, \tag{4.3.32a}$$

$$= \mathbf{B}^T \left(\mathbf{I} + (s - s_0)\hat{\mathbf{A}}\right)^{-1} \hat{\mathbf{B}}, \tag{4.3.32b}$$

$$\text{mit} \quad \mathbf{E}\mathbf{B} = \mathbf{B},\ \ \hat{\mathbf{A}} = \bar{\mathbf{A}}\mathbf{E},\ \ \bar{\mathbf{A}} = (\mathbf{A}\mathbf{E} + s_0\mathbf{E})^{-1},\ \ \hat{\mathbf{B}} = \bar{\mathbf{A}}\mathbf{B}. \tag{4.3.32c}$$

Da nun sowohl $\mathbf{A}\mathbf{E}$ als auch $\mathbf{E}$ symmetrische Matrizen sind, ist auch die invertierte Systemmatrix $\hat{\mathbf{A}}$ für das lineare System symmetrisch, was selbst im Fall von Verlusten erfüllt bleibt. Für die zugehörigen Krylov-Unterräume gilt nun vergleichbar zu oben:

$$\operatorname{span}\{\mathbf{V}_q\} = \mathcal{K}(\hat{\mathbf{A}}, \hat{\mathbf{B}}) = \mathcal{K}(\bar{\mathbf{A}}\mathbf{E}, \bar{\mathbf{A}}\mathbf{B}) = \bar{\mathbf{A}}\,\mathcal{K}(\mathbf{E}\bar{\mathbf{A}}, \mathbf{B}) \quad \text{und} \tag{4.3.33a}$$

$$\operatorname{span}\{\mathbf{W}_q\} = \mathcal{K}(\hat{\mathbf{A}}^H, \mathbf{B}) = \left(\mathcal{K}(\hat{\mathbf{A}}^T, \mathbf{B})\right)^* = \left(\mathcal{K}(\mathbf{E}\bar{\mathbf{A}}, \mathbf{B})\right)^*. \tag{4.3.33b}$$

Damit gilt entsprechend Gl. 4.3.30 grundsätzlich für alle symmetrisierbaren linearen FIT-Systeme $\mathbf{w}_q = ((\mathbf{A}\mathbf{E} + s_0\mathbf{E})\mathbf{v}_q)^*$ und es genügt, nur ein Gleichungssystem pro Iterationsschritt zu lösen. Dieser Zusammenhang gilt auch für Matrizen der Netzwerkanalyse, das beschriebene Vorgehen wurde in ähnlicher Weise in [64] angewandt.

Es zeigt sich somit, dass, obwohl der symmetrische Lanczos-Algorithmus nicht direkt angewendet werden kann, in den üblichen Fällen dennoch eine einzige Systemlösung pro Iterationsschritt ausreichend ist. Dies gilt für reelle symmetrische oder schiefsymmetrische Systemmatrizen, also beispielsweise die Curl-Curl-Formulierung, aber auch für das verlustbehaftete lineare System.

4.3.3.3 Stabilität und Passivität

Da die Generierung einer Padé-Approximation durch Krylov-Unterräume stets auf einer unsymmetrischen Projektion auf die Matrizen $\mathbf{V}_q$ und $\mathbf{W}_q$ erfolgt, kann die Erhaltung von Stabilität und Passivität nicht grundsätzlich gewährleistet werden. Auch in den beschriebenen Sonderfällen, in denen der numerische Aufwand auf eine Gleichungslösung pro Iterationsschritt begrenzt wird, wird dennoch eine Matrix $\mathbf{W}_q \neq \mathbf{V}_q$ bestimmt und die Projektion erfolgt ebenfalls unsymmetrisch. Während die Möglichkeit von Instabilitäten die reine Verwendung des Modells für einen *Fast Frequency Sweep* nicht beeinträchtigt, muss zur Nutzung als Makromodell zumindest Stabilität, besser noch Passivität, garantiert werden können.

Die Erfahrung zeigt, dass meist nur einzelne Eigenwerte der reduzierten Systemmatrix $\mathbf{T}_q$ tatsächlich in die rechte „instabile“ Laplace-Halbebene wandern. In [58] wird der bereits beschriebene *Implicitly-Restarted-Arnoldi*-Algorithmus vorgeschlagen, um gezielt diese Eigenwerte in einem impliziten Neustart zu eliminieren und somit Stabilität zu erzielen. Durch die Herausnahme einzelner Eigenwerte stimmen die Momente zwar nicht mehr exakt mit denen des Originalsystems überein, es handelt sich demnach nicht mehr um eine Padé-Approximation im strengen Sinne, aber wenn nur wenige Eigenwerte betroffen sind, bleiben die Approximationseigenschaften dennoch sehr gut.

Ein von der Grundidee ähnliches Verfahren wird unter dem Namen *PVL*π in [71] vorgestellt. In einem Postprocessing-Schritt werden hierbei Nullstellen und Pole des reduzierten Systems berechnet und jene in der rechten Halbebene liegenden an der imaginären Achse gespiegelt. Die Anzahl der identischen Momente reduziert sich dabei genau um die Zahl der gespiegelten Pole und Nullstellen. Das Verfahren ist jedoch auf eindimensionale Übertragungsfunktionen beschränkt und daher für den typischeren Mehrportfall ungeeignet.

Wird entsprechend dem Vorgehen im Fall partieller Realisierungen nur die Matrix $\mathbf{V}_q$ genutzt oder anstelle des Bi-Lanczos- der Arnoldi-Algorithmus verwendet, führt dies auf eine symmetrische Projektion [74]

$$\mathbf{Z}_p(s) = \mathbf{B}^T\mathbf{V}_q \left(\mathbf{V}_q^T\mathbf{V}_q + (s-s_0)\mathbf{V}_q^T\hat{\mathbf{A}}\mathbf{V}_q\right)^{-1}\mathbf{V}_q^T\hat{\mathbf{B}}, \tag{4.3.34a}$$

$$= \mathbf{B}^T\mathbf{V}_q \left(\mathbf{I} + (s-s_0)\mathbf{H}_q\right)^{-1}\mathbf{B}_q. \tag{4.3.34b}$$

Allerdings werden auch hierdurch Stabilität und Passivität nicht garantiert, da das Grundsystem 4.3.23a selbst mit $\mathbf{B} \neq \hat{\mathbf{B}}$ nicht symmetrisch ist. Projiziert man jedoch alternativ das Originalsystem Gl. 3.1.2 auf obige Matrix $\mathbf{V}_q$, generiert aus $\hat{\mathbf{A}}$ und $\hat{\mathbf{B}}$, ergibt sich

$$\mathbf{Z}_p(s) = \mathbf{B}^T\mathbf{V}_q \left(s\mathbf{I} + \mathbf{V}_q^T\mathbf{A}\mathbf{V}_q\right)^{-1}\mathbf{V}_q^T\mathbf{B}, \tag{4.3.35a}$$

$$= \mathbf{B}_q'^T \left(s\mathbf{I} + \mathbf{A}_q\right)^{-1}\mathbf{B}_q'. \tag{4.3.35b}$$

Es ist direkt erkennbar, dass diese einseitige Projektion für reelle $s_0 \in \mathbb{R}$ Stabilität und Passivität erhält. Diese Formulierung wurde erstmals in [65] vorgestellt und der Name *Passive Reduced-Order Interconnect Macro-Modelling Algorithm, PRIMA*

geprägt. Es ist zu beachten, dass $\mathbf{B}'_q \neq \mathbf{V}_q^T\mathbf{B}$ gilt, folglich $\mathbf{B}$ tatsächlich auf $\mathbf{V}_q$ projiziert werden muss, und nicht das $\mathbf{B}_q$ aus dem Lanczos-/Arnoldi-Algorithmus verwendet werden kann.

Untersucht man die Approximationseigenschaften, zeigt sich für die Momente $\hat{\mathbf{M}}_i$ des reduzierten Modells nach 4.3.35 mit $\breve{\mathbf{A}} = (\mathbf{V}_q^T\mathbf{A}\mathbf{V}_q + s_0\mathbf{I})^{-1}$:

$$\hat{\mathbf{M}}_i = \mathbf{B}_q'^T\breve{\mathbf{A}}^{i+1}\mathbf{B}'_q \tag{4.3.36a}$$

$$\hat{\mathbf{M}}_i = \mathbf{M}_i \qquad \text{für} \qquad 0 \leq i < m_q - 1. \tag{4.3.36b}$$

Das Modell stimmt im Allgemeinen folglich nur in der Hälfte der möglichen Momente mit dem Originalsystem überein. Eine solche Approximation wird auch als Padé-Typ-Approximation bezeichnet. Ein ausführlicher Beweis der Aussagen 4.3.36 findet sich für die Entwicklungsfrequenz $s_0 = 0$ in [65] und allgemein in [72]. Der Preis für die garantierte Erhaltung von Stabilität und Passivität ist somit ein Modell, dessen Genauigkeit eingeschränkt ist. Allerdings handelt es sich bei der Anzahl der Momente nur um eine Abschätzung für die tatsächliche Approximationsgüte und in vielen Fällen zeigt sich, dass nicht wirklich die doppelte Modellgröße erforderlich ist, um dieselbe Genauigkeit zu erhalten.

Für den Fall der symmetrischen Curl-Curl-Systeme lässt sich darüberhinaus zeigen, dass für die Padé-Approximation mit $\hat{\mathbf{A}}^T = \hat{\mathbf{A}} = (\mathbf{A}_{CC} + s_0^2\mathbf{I})^{-1}$, $\hat{\mathbf{B}} = \hat{\mathbf{A}}\mathbf{B}$ und damit $\mathbf{W}_q = \hat{\mathbf{A}}^{-1}\hat{\mathbf{V}}_q$ gilt:

$$\mathbf{Z}_p(s) = \mathbf{B}^T\hat{\mathbf{V}}_q\left(\mathbf{W}_q^T\hat{\mathbf{V}}_q + (s^2 - s_0^2)\mathbf{W}_q^T\hat{\mathbf{A}}\hat{\mathbf{V}}_q\right)^{-1}\mathbf{W}_q^T\hat{\mathbf{B}}, \tag{4.3.37a}$$

$$= \mathbf{B}^T\hat{\mathbf{V}}_q\left(s^2\hat{\mathbf{V}}_q^T\hat{\mathbf{V}}_q + \hat{\mathbf{V}}_q^T\mathbf{A}_{CC}\hat{\mathbf{V}}_q\right)^{-1}\hat{\mathbf{V}}_q^T\mathbf{B}. \tag{4.3.37b}$$

$\hat{\mathbf{V}}_q$ wurde hier zur späteren Unterscheidung überdacht. Es bestätigt sich erneut, dass die Matrix $\mathbf{W}_q$ im Reduktionsprozess redundant ist. Vorallem zeigt sich aber, dass Formulierung Gl. 4.3.35 im Fall symmetrischer Systeme mit der Padé Approximation Gl. 4.3.37a übereinstimmt, also trotz einseitiger Projektion in einer Maximalzahl von Momenten identisch ist und, ein reelles s_0^2 vorausgesetzt, zudem Stabilität und Passivität gewährleistet. Wird durch eine QR-Zerlegung $\hat{\mathbf{V}}_q = \mathbf{V}_q\mathbf{U}$ orthonormiert, folgt mit $\mathbf{V}_q^T\mathbf{V}_q = \mathbf{I}$:

$$\mathbf{Z}_p(s) = \mathbf{B}^T\mathbf{V}_q\left(s^2\mathbf{I} + \mathbf{V}_q^T\mathbf{A}_{CC}\mathbf{V}_q\right)^{-1}\mathbf{V}_q^T\mathbf{B}. \tag{4.3.38}$$

Derselbe Unterraum $\mathbf{V}_q$ lässt sich auch direkt durch Verwendung des symmetrischen Lanczos- oder des Arnoldi-Algorithmus berechnen.

4.3.3.4 Mehrfache Entwicklungspunkte

Alle bisherigen Beschreibungen gingen stets von einem einzelnen Entwicklungspunkt aus. Gerade die Formulierung nach den Gln. 4.3.35 und 4.3.38 lassen jedoch eine sehr bequeme Erweiterung des Verfahrens auf mehrfache Entwicklungspunkte (engl. *multipoint Padé*) zu. In diesem Fall wird zur Projektion eine Matrix

$$\hat{\mathbf{V}}_q = (\mathbf{V}_1, \mathbf{V}_2, \ldots, \mathbf{V}_I) \tag{4.3.39}$$

verwendet, in der jede der I Untermatrizen $\mathbf{V}_i$ eine Anzahl von $r_i \leq q$ Unterraumvektoren um die I Entwicklungspunkte s_{0i} enthält. Die Unterraumdimension r_i kann dabei für jeden Punkt s_{0i} unterschiedlich gewählt werden. Im Grenzfall können also an I Entwicklungsfrequenzen jeweils ein Unterraumvektor bestimmt werden, der gerade der Systemlösung in diesem Punkt entspricht. Dies führt somit erneut auf rationale Interpolation. Um die Bedingung $\mathbf{V}_q^T\mathbf{V}_q = \mathbf{I}$ zu erhalten, muss $\hat{\mathbf{V}}_q$ erneut mit $\hat{\mathbf{V}}_q = \mathbf{V}_q\mathbf{U}$ orthonormiert werden, was bei typischen Unterraumgrößen aber keinen nennenswerten zusätzlichen Aufwand verursacht.

Auch für die klassische Padé Approximation, basierend auf Gl. 4.3.23a, existieren Erweiterungen für mehrere Entwicklungspunkte. Die verwendeten Algorithmen, die dem Lanczos- bzw. Arnoldi-Verfahren ähneln, werden als rationale Krylov-Algorithmen bezeichnet. Untersuchungen hierzu im Zusammenhang mit Ordnungsreduktion finden sich beispielsweise in [59].

Zum numerischen Aufwand ist grundsätzlich zu bedenken, dass jeder weitere Entwicklungspunkt erneut die Inversion der Systemmatrix erfordert. Erfolgt dies über eine LU-Zerlegung, müssen folglich I Zerlegungen berechnet werden, was den Aufwand für größere Systeme um den Faktor I anwachsen lässt. Erfolgt die Berechnung der Unterräume jedoch durch die Lösung von Gleichungssystemen, ist der Mehraufwand für mehrere Entwicklungspunkte marginal, da vorhandene Ergebnisse ohnehin nicht ausgenutzt werden können.

Erfahrungsgemäß lässt sich bei den interessierenden Frequenzbereichen elektrodynamischer Simulationen, die im Vergleich zu dem von den Eigenwerten abgedeckten Gesamtspektrum meist sehr schmalbandig sind, kein großer Gewinn in der Modellgröße durch mehrere Entwicklungspunkte erzielen. Es kann jedoch erreicht werden, dass der Fehler gleichmäßiger über den Frequenzbereich verteilt wird, wie in Abb. 4.8 zu erkennen. Es wird der Approximationsfehler in Abhängigkeit der Frequenz für eine einfache Padé Approximation achter Ordnung sowie einer *Multipoint*-Padé Approximation um drei Entwicklungsfrequenzen, die insgesamt ebenfalls achter Ordnung ist, dargestellt.

In vielen Fällen zeigt sich aber, dass auch eine einzelne Extraktionsfrequenz ausreichend ist.

4.3.3.5 Abbruchkriterium und Wahl der Entwicklungsfrequenz

Das Abbruchkriterium für Padé-Approximationen kann völlig analog zu dem in Abschnitt 4.3.1.2 für partielle Realisierungen vorgeschlagenen definiert werden. In Abb. 4.9 ist der gemittelte Fehler der Impedanzfunktion δ_Z, der Eigenwertfehler δ_E sowie die Eigenwertdifferenz δ_D relativ zum vorherigen Iterationsschritt für eine Padé-Approximation mit einem Entwicklungspunkt in der Mitte des interessierenden Frequenzintervalls dargestellt. Erneut zeigt sich ein proportionales Verhalten der drei Kurven. Auffällig ist jedoch, dass die benötigte Iterationszahl q zum Erreichen einer Fehlertoleranz von $\delta_D = 10^{-8}$ im Vergleich zur partiellen Realisierung wesentlich geringer ist, bereits acht Iterationen erweisen sich als ausreichend. Das

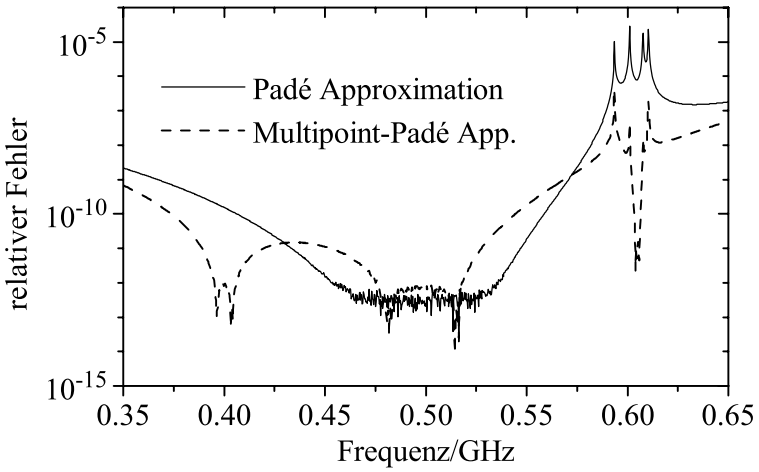


Abbildung 4.8: *Approximationsfehler zweier Padé-Approximationen achter Ordnung. Die erste bestimmt alle acht Momente um die Mittenfrequenz 0,5 GHz, die zweite ermittelt zwei Momente um 0,4 GHz sowie je drei um die Entwicklungsfrequenzen 0,5 und 0,6 GHz.*

Stufenverhalten der Kurven kann mit der Blockformulierung des Algorithmus erklärt werden, wobei dieses Beispiel $m = 2$ Ports besitzt.

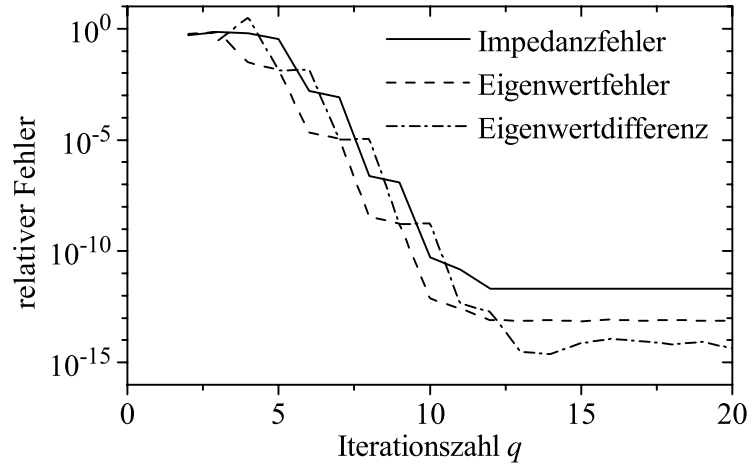


Abbildung 4.9: *Der Fehler der approximierten Impedanzfunktion δ_Z im Vergleich zum gemittelten Fehler aller im interessierenden Frequenzintervall liegenden Eigenwerte δ_E und deren Differenz δ_D für eine Padé-Approximation.*

Alternative Vorschläge für Abbruchkriterien [59, 73] schätzen den Fehler der Übertragungsfunktion ab, indem der Einfluss des $(2q + 1)$ten Moments, also des ersten nicht beachteten, untersucht wird. Die Iteration wird abgebrochen, wenn das Maximum dieses Schätzwerts unterhalb einer vorgegebenen Toleranz liegt.

Die Anzahl der benötigten Iterationen q zum Erreichen eines vorgegebenen Fehler hängt generell auch maßgeblich von Wahl der Entwicklungsfrequenz(en) ab. Es erscheint hierbei intuitiv sinnvoll, einen Entwicklungspunkt innerhalb des interes-

sierenden Frequenzbands zu wählen. Insbesondere für die Curl-Curl-Formulierung ist eine solche Wahl attraktiv, da sich für einen rein imaginären Punkt s_0 ein reelles s_0^2 als Verschiebung in 4.3.37a ergibt. In Verbindung mit einer symmetrischen Projektion ist dies Voraussetzung für die Erhaltung der Passivität.

Für die Wahl innerhalb des Frequenzintervalls bieten sich zwei grundsätzliche Möglichkeiten an:

- die Mitte des interessierenden Frequenzintervalls, $s_0 = j\pi(f_{\min} + f_{\max})$,
- der Schwerpunkt der K Polstellen s_k im betrachteten Frequenzbereich, falls diese bekannt sind, $s_0 = \frac{1}{K}\sum_k s_k$.

In Abb. 4.10 ist der Approximationsfehler der Padé-Approximation für Modelle fester Ordnung ($q = 6, 8, 10, 12$) in Abhängigkeit der rein imaginären Entwicklungsfrequenz aufgetragen. Die Reduzierung erfolgt in Curl-Curl-Formulierung. Zur Mittelung des Fehlers wurden nur die Werte innerhalb des interessierenden Frequenzbereichs von 0,35 - 0,65 GHz beachtet (in der Abbildung grau hinterlegt). Die beiden Polstellen liegen mit 0,6 und 0,61 GHz am Rand des Intervalls.

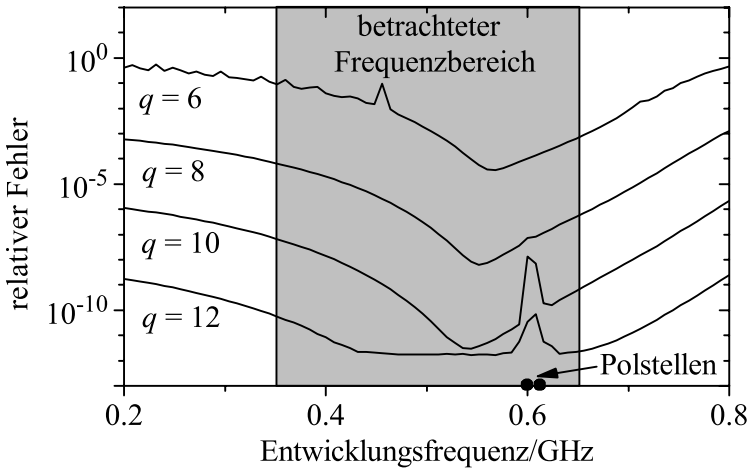


Abbildung 4.10: *Approximationsfehler einer Padé-Approximation in Curl-Curl-Formulierung mit festgelegten Modellgrößen q in Abhängigkeit einer imaginären Entwicklungsfrequenz.*

Es zeigt sich, dass sich ein minimaler Approximationsfehler tatsächlich in der Mitte des betrachteten Frequenzbereichs, leicht in Richtung der Polstellen verschoben, einstellt. Ein Entwicklungspunkt in zu unmittelbarer Nähe der Pole verschlechtert unter Umständen sogar den Gesamtfehler. Auch die Erfahrung mit anderen Beispielen zeigt, dass die Mittenfrequenz meist die verlässlichere Schätzung für den Entwicklungspunkt im Vergleich zum Schwerpunkt der Pole darstellt.

Soll das reduzierte Modell allein für einen *Fast Frequency Sweep* verwendet werden, können auch allgemeine komplexe Entwicklungspunkte verwendet werden. Bei einer linearen Formulierung muss zur Erhaltung der Passivität gar ein rein reeller Extraktionspunkt gewählt werden. In Abb. 4.11 wird für ein Filterbeispiel mit acht

Polstellen und einem betrachteten Intervall von 4–8 GHz eine Padé-Approximation in linearer Formulierung berechnet. Dargestellt ist die benötigte Modellgröße q in Abhängigkeit der komplexen Entwicklungsfrequenz, um einen Fehler von 10^{-6} zu erzielen.

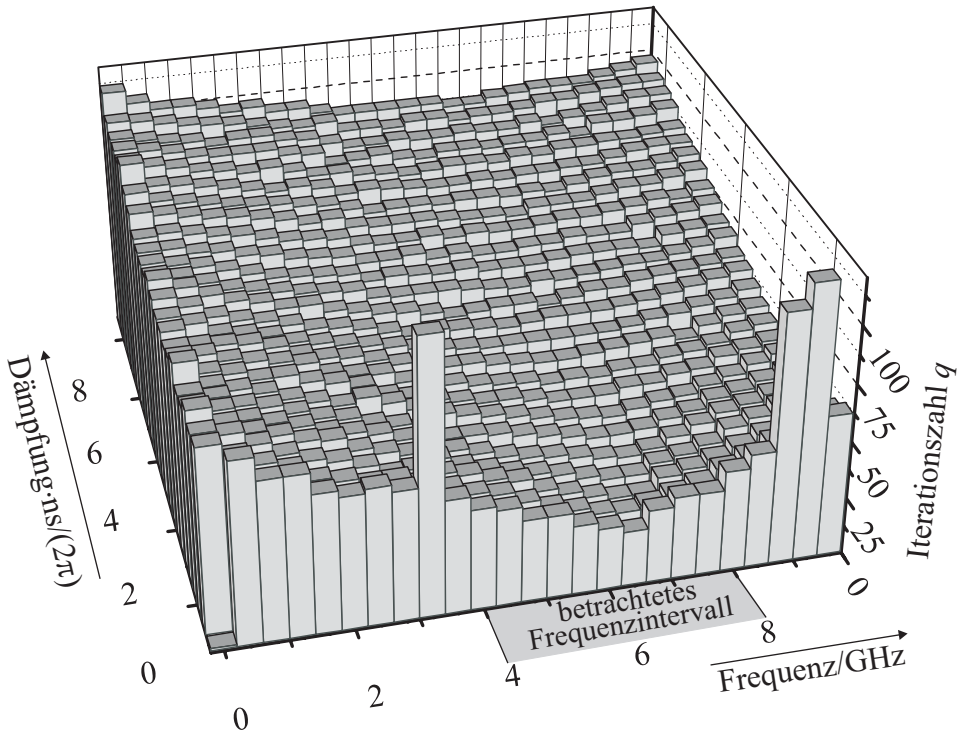


Abbildung 4.11: *Benötigte Anzahl von Iterationsschritten q, um einen Modellfehler von* 10^{-6} *zu erreichen, in Abhängigkeit eines komplexen Entwicklungspunkts. Es wurde die lineare Systemformulierung gewählt.*

Erneut zeigt sich die minimale Modellgröße etwa in der Mitte des betrachteten Frequenzintervalls mit $q = 24$. Im Vergleich dazu ist das kleinste passive Modell mit rein reeller Entwicklungsfrequenz bereits der Ordnung 80, also mehr als dreimal so groß. Dies betont erneut den Vorteil der Curl-Curl-Formulierung gegenüber dem linearen Fall zur Erzeugung passiver Modelle. Zudem zeigt sich, dass rein imaginäre Entwicklungsfrequenzen, wenn sie ungünstig in der Nähe von Polstellen liegen, die Modellgröße auch deutlich erhöhen können (beispielsweise bei 3.2 GHz). Dies kann umgangen werden, wenn man den Entwicklungspunkt leicht in den reellen Bereich verschiebt.

4.3.3.6 Padé-Approximationen für polynomiale Systeme

Die in den vorangegangen Abschnitten beschriebenen Verfahren, die Projektionen auf Krylov-Unterräume zur Reduzierung der Modellordnung nutzen, haben sich als numerisch sehr robust erwiesen. Die Momente der Originalfunktion sind in den Krylov-Unterräumen implizit enthalten und die reduzierten Modelle stimmen - abhängig von der jeweiligen Projektionstechnik - in q oder gar der Maximalzahl von

$2q$ Momenten mit denen des Originalsystems überein. In vielen Bereichen wurden daher explizite *moment-matching*-Techniken [61], wie die *Asymptotic Waveform-Evaluation, AWE* [62], weitgehend abgelöst. Diese beruhen auf der Idee, eine Anzahl von Taylormomenten der Originalfunktion zunächst explizit zu berechnen, um das reduzierte Modell im Anschluss durch Lösung zweier kleiner linearer Systeme (mit der Ordnung des reduzierten Modells) zu generieren. Die direkte Berechnung der Momente ist jedoch ein numerisch schlecht konditioniertes Problem, weswegen bei endlicher Rechengenauigkeit bei Modellordnungen größer etwa zehn die Hinzunahme neuer Momente praktisch keine neue Information birgt und der Reduktionsprozess stagniert.

Explizite *moment-matching*-Techniken haben aber ihrerseits den Vorteil, dass sie auf beliebige Systeme angewendet werden können, solange deren Taylormomente bekannt sind, während Krylov-Unterraumtechniken auf lineare Systeme wie Gl. 3.1.2 beschränkt sind. Systeme höheren Grades müssen, wie in den Abschnitten 3.1.2 und 3.1.3 beschrieben, durch Hinzunahme von Freiheitsgraden zunächst in ein lineares System transformiert oder durch ein Kleinsignalmodell um eine vorgegebene Betriebsfrequenz linearisiert werden, was einer lokalen Näherung entspricht. Auch die Projektion geringer Verluste nach Abschnitt 4.3.1.3 ist als Näherungslösung anzusehen, die aufgrund dieser Limitierung der Krylov-Unterraumverfahren nötig wird.

Ein Verfahren zur Reduzierung von Systemen zweiten Grades, wurde unter dem Namen *ENOR* [75] vorgeschlagen. Zur Berechnung ordnungsreduzierter Modelle allgemeiner polynomialer Systeme nach Gl. 3.1.16 soll im Folgenden jedoch als Alternative zu Krylov-Unterraumverfahren mit *Galerkin Asymptotic Waveform-Evaluation, GAWE* eine Weiterentwicklung des AWE-Verfahrens vorgestellt werden. Hierbei wird der Algorithmus, der in seiner ursprünglichen Fassung nur für Einportsysteme vorgeschlagen wurde, auf eine Blockformulierung erweitert, die auch die Reduktion von Mehrportsystemen zulässt. Die Herleitung wird an dieser Stelle nur verkürzt wiedergegeben, eine detailliertere Beschreibung findet sich in [98].

Ausgangspunkt ist eine allgemeine polynomiale Zustandsraumbeschreibung des Systems nach Gl. 3.1.16. Um eine vereinfachte Herleitung der Blockformulierung zu ermöglichen, wird das System zunächst ohne den Vektor $\mathbf{i}$ betrachtet. Damit ergibt sich mit der $n \times m$-dimensionalen Zustandsmatrix $\mathbf{X}$:

$$\sum_{k=0}^{N_A} \left(s^k \mathbf{A}_k\right) \mathbf{X} = \sum_{k=0}^{N_B} \left(s^k \mathbf{B}_k\right). \tag{4.3.40}$$

Gl. 3.1.16 ergibt sich daraus wiederum über $\mathbf{x} = \mathbf{X}\mathbf{i}$. Entsprechend Gl. 4.3.23 lässt sich ein Entwicklungspunkt s_0 einführen

$$\sum_{k=0}^{N_A} \left((s-s_0)^k \hat{\mathbf{A}}_k\right) \mathbf{X} = \sum_{k=0}^{N_R} \left((s-s_0)^k \hat{\mathbf{B}}_k\right) \mathbf{i} \tag{4.3.41a}$$

$$\text{mit} \quad \hat{\mathbf{A}}_k = \sum_{i=k}^{N_A} \binom{n}{k} \hat{\mathbf{A}}_i, \quad \hat{\mathbf{B}}_k = \sum_{i=k}^{N_B} \binom{n}{k} \hat{\mathbf{B}}_i. \tag{4.3.41b}$$

Wird auch die frequenzabhängige Zustandsmatrix als Taylorreihe geschrieben

$$\mathbf{X} = \mathbf{X}(s) = \sum_{k=0}^{\infty} (s - s_0)^k \hat{\mathbf{V}}_k, \tag{4.3.42}$$

ergibt sich durch Einsetzen in 4.3.41 und Koeffizientenvergleich für die Vektoren $\hat{\mathbf{V}}_k$ die rekursive Vorschrift:

$$\hat{\mathbf{V}}_0 = \hat{\mathbf{A}}_0^{-1} \hat{\mathbf{B}}_0 \tag{4.3.43a}$$

$$\hat{\mathbf{V}}_1 = \hat{\mathbf{A}}_0^{-1} (\hat{\mathbf{B}}_1 - \hat{\mathbf{A}}_1 \mathbf{V}_0) \tag{4.3.43b}$$

$$\hat{\mathbf{V}}_2 = \hat{\mathbf{A}}_0^{-1} (\hat{\mathbf{B}}_2 - \hat{\mathbf{A}}_1 \mathbf{V}_1 - \hat{\mathbf{A}}_2 \mathbf{V}_0) \tag{4.3.43c}$$

$$\vdots$$

$$\hat{\mathbf{V}}_k = \hat{\mathbf{A}}_0^{-1} \left(\hat{\mathbf{B}}_k - \sum_{i=1}^{\min(N_A,k)} \hat{\mathbf{A}}_i \mathbf{V}_{k-i} \right). \tag{4.3.43d}$$

Somit folgt für das Residuum

$$\mathbf{r}_k = \sum_{k=0}^{N_B} (s - s_0)^k \hat{\mathbf{B}}_k - \sum_{k=0}^{N_A} \left((s - s_0)^k \hat{\mathbf{A}}_k \right) \sum_{k=0}^{K} (s - s_0)^k \hat{\mathbf{V}}_k, \tag{4.3.44}$$

dass es nach K Schritten senkrecht auf dem Unterraum der Taylorkoeffizienten steht

$$\mathbf{r}_K \perp [\hat{\mathbf{V}}_0, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{K-1}], \tag{4.3.45}$$

was dem Verfahren den Namen *Galerkin*-AWE einbringt. Entwickelt man zudem $\mathbf{r}_K$ ebenfalls in eine Taylorreihe, zeigt sich dass, die ersten K Koeffizienten aufgrund der obigen Herleitung Null sind, das Modell im Hinblick auf die Approximation des Residuums also optimal ist. Dies gilt auch, wenn der Unterraum mit Hilfe einer QR-Zerlegung orthonormiert wird [76], was numerische Vorteile bringt

$$\hat{\mathbf{V}} = [\hat{\mathbf{V}}_0, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{K-1}] = \mathbf{V}_q \mathbf{U}. \tag{4.3.46}$$

Im Gegensatz zum klassischen AWE wird nun aber keine Padé-Approximation berechnet, sondern eine Projektion wie in Abschnitt 4.3.3.3 angewandt. Dies führt auf das reduzierte System in Impedanzdarstellung:

$$\mathbf{Z}'_p(s) = \mathbf{L}\mathbf{V}_p \left(\sum_{k=0}^{N_A} (s - s_0)^k \mathbf{V}_p^H \hat{\mathbf{A}}_k \mathbf{V}_p \right)^{-1} \mathbf{V}_p^H \sum_{k=0}^{N_R} (s - s_0)^k \hat{\mathbf{B}}_k. \tag{4.3.47}$$

Dieses stimmt folglich ebenfalls in K Momenten mit dem Originalsystem überein und ist somit eine Padé-Typ-Approximation, wie auch die symmetrische Projektion in 4.3.34. Entsprechend dem dortigen Vorgehen lässt sich auch das Originalsystem auf den Unterraum $\mathbf{V}_q$ projizieren:

$$\mathbf{Z}_p(s) = \mathbf{L}\mathbf{V}_p \left(\sum_{k=0}^{N_A} s^k \mathbf{V}_p^H \mathbf{A}_k \mathbf{V}_p \right)^{-1} \mathbf{V}_p^H \sum_{k=0}^{N_R} s^k \mathbf{B}_k. \tag{4.3.48}$$

Es zeigt sich direkt, dass Stabilität und Passivität unter der Voraussetzung eines reellen s_0 aufgrund der symmetrischen Projektion erneut erhalten werden. Im Gegensatz zum Arnoldi-reduzierten linearen System zeigt die Erfahrung zudem, dass bei genügend hoher Approximationsgüte auch bei imaginären Entwicklungsfrequenzen, mit dem Vorteil kleinerer Modelle, die Passivität meist erhalten bleibt. Dies kann jedoch nicht ohne Weiteres allgemein gezeigt werden und hängt vermutlich von der Lage der Pole ab, die bei verlustbehafteten Strukturen meist nahe der imaginären Achse liegen.

Bedauerlicherweise weist das hier beschriebene Verfahren dieselben Schwächen wie auch das Standardverfahren AWE auf. So werden die Unterraumvektoren $\hat{\mathbf{V}}$ zunehmend linear abhängig, da sie zu dem dominanten Eigenvektor von $\hat{\mathbf{A}}_0^{-1}\hat{\mathbf{A}}_1$ tendieren. Im Mehrportfall $m > 1$ können zusätzliche lineare Abhängigkeiten innerhalb der Blöcke $\hat{\mathbf{V}}_k$ auftreten, die nicht ohne weiteres während der Berechnung des Unterraums deflationiert werden können. Die Orthonormierung am Ende des Verfahrens hat meist auch nur geringen Einfluss auf das Konvergenzverhalten, da bereits der Unterraum $\hat{\mathbf{V}}$ nicht mehr den vollen Spaltenrang aufweist.

Das nahe liegende Vorgehen, die Vektoren $\mathbf{V}_k$ direkt bei ihrer Erzeugung gegen die vorhandenen Vektoren zu orthogonalisieren, sollte vermieden werden, da hierdurch im Allgemeinen nicht mehr die Momente übereinstimmen. Für den linearen Fall $N_A = 1$ und $N_K = 1$ ist dies natürlich möglich und führt auf den bereits beschriebenen Fall des Arnoldi-Algorithmus. Auch in anderen Fällen führt die Orthogonalisierung häufig auf gute Ergebnisse, jedoch kann das Verfahren auch unvermittelt fehlschlagen und verliert somit folglich seine Zuverlässigkeit. In [76] wird daher ein Verfahren vorgeschlagen, das mehrere Modelle mit jeweils niedriger Ordnung um verschiedene Entwicklungspunkte vorschlägt. Dies entspricht der Idee, die unter dem Namen *Complex Frequency Hopping, CFH* [77] bereits im Zusammenhang mit AWE vorgeschlagen wurde.

Ein anderer Ansatz, die Konditionierung zu verbessern, ist in [78] unter dem Namen *Well Conditioned AWE, WCAWE* beschrieben und wurde ebenfalls in [98] auf den Mehrportfall erweitert. Es soll im Rahmen dieser Arbeit nur kurz zusammengefasst werden.

Die Grundidee basiert darauf, dass die Projektion so lange eine Padé-Typ-Approximation bildet, wie die Matrix $\mathbf{U}$ in der Zerlegung 4.3.46 obere Dreiecksgestalt aufweist, die Orthonormalität der Matrix $\mathbf{V}_q$ ist dabei nicht zwingend erforderlich. Die Dreiecksgestalt von $\mathbf{U}$ kann jedoch am einfachsten durch eine QR-Zerlegung erreicht werden. Der Einfluss der Matrix $\mathbf{U}$ kann nun sukzessive durch Korrekturterme in der Berechnung der Momente 4.3.43 berücksichtigt werden. Ist $\mathbf{U}$ invertierbar, folgt nach [78] mit

$$\mathbf{Q}_i(n,k) = \prod_{t=i}^{k} \mathbf{U}(t : n-k+t-1, t : n-k+t-1)^{-1} \tag{4.3.49}$$

für die Berechnung der Momente:

$$\hat{\mathbf{V}}_0 = \hat{\mathbf{A}}_0^{-1}\hat{\mathbf{B}}_0 \tag{4.3.50a}$$

$$\hat{\mathbf{V}}_1 = \hat{\mathbf{A}}_0^{-1}(\hat{\mathbf{B}}_1\mathbf{E}_1^T\mathbf{Q}_1(2,1)\mathbf{E}_1 - \hat{\mathbf{A}}_1\mathbf{V}_0) \tag{4.3.50b}$$

$$\vdots$$

$$\hat{\mathbf{V}}_k = \hat{\mathbf{A}}_0^{-1}\left(\sum_{i=1}^{\min(N_R,k)} \hat{\mathbf{B}}_k\mathbf{E}_1^T\mathbf{Q}_1(k,i)\mathbf{E}_{k-i} - \hat{\mathbf{A}}_1\mathbf{V}_{k-1} - \sum_{i=2}^{\min(N_A,k)} \hat{\mathbf{A}}_i\mathbf{V}_q\mathbf{Q}_2(k,i)\mathbf{E}_{k-i}\right). \tag{4.3.50c}$$

Die Indizierung in 4.3.49 und 4.3.50 ist blockweise zu verstehen, wobei jeder Eintrag die Größe $m \times m$ aufweist. Dies gilt auch für die verallgemeinerten Einheitsvektoren $\mathbf{E}_n$, deren $n$ter Block eine $m \times m$ Einheitsmatrix darstellt, während alle anderen Einträge Null sind. Die Projektion erfolgt schließlich analog zu Gl. 4.3.47 bzw. Gl. 4.3.48 auf den orthonormierten Unterraum $\mathbf{V}_q$.

Auch wenn mit WCAWE die numerischen Instabilitäten des AWE-Verfahrens behoben werden konnten, verbleibt dennoch der hohe numerische Aufwand, der vergleichbar zu Padé-Approximationen, basierend auf Krylov-Unterräumen, ist. Wird in der Herleitung von GAWE oder WCAWE $\sigma = s - s_0$ durch $\sigma = 1/s$ ersetzt, führt dies entsprechend partieller Realisierungen auf ein reduziertes System, dessen Momente um Unendlich übereinstimmen. Anstelle der Matrix $\mathbf{A}_0$ muss in 4.3.43 bzw. 4.3.50 die Matrix $\mathbf{A}_{N_A}$ invertiert werden, was trivial ist, da diese stets die Einheitsmatrix ist. Der numerische Aufwand ist folglich drastisch reduziert. Der Algorithmus wurde in [98] implementiert, führte aber leider aufgrund systematisch auftretender linearer Abhängigkeiten der Unterraumvektoren nicht zu brauchbaren Ergebnissen.

4.3.4 Two-Step-Lanczos

Im bisherigen Verlauf des Kapitels wurden eine Reihe von Projektionsverfahren auf unterschiedliche Unterräume vorgestellt. Sie lassen sich grob in zwei Gruppen einteilen:

- Verfahren, die eine kurze Rechenzeit aufweisen, aber auf verhältnismäßig große Unterräume und damit auf große Modelle führen: partielle Realisierung, beschleunigte partielle Realisierung.
- Verfahren, die sehr kleine Modelle generieren, aber hohen numerischen Aufwand erfordern: Modalanalyse, Padé Approximationen, WCAWE.

Für einen wirklich „schnellen“ *Fast Frequency Sweep* ist der numerische Aufwand einer Padé-Approximation häufig zu hoch, umgekehrt sind partielle Realisierungen meist zwar in Sekunden oder zumindest wenigen Minuten berechnet, sie sind jedoch noch zu groß, um einen effektiven Durchlauf mit sehr vielen Frequenzpunkten zu berechnen. Soll das Modell als Makromodell bzw. Ersatzschaltbild verwendet werden, sollte die Modellordnung ohnehin so gering wie möglich sein.

Eine deutliche Effizienzsteigerung ergibt sich aus der Kopplung der beiden Krylov-basierten Methoden: In einem ersten Schritt wird das Originalsystem durch partielle Realisierung auf ein Modell mittlerer Größe reduziert. Dieses hat dann eine Größe, bei der problemlos eine dünnbesetzte LU-Zerlegung erfolgen kann, was die sehr schnelle Berechnung einer Padé-Approximation im zweiten Schritt ermöglicht. Die Gesamtrechenzeit ist damit im Vergleich zu einer puren partiellen Realisierung nur minimal vergrößert, während die Modellgröße identisch mit der einer direkten Padé-Approximation ist. Da die Modellberechnung in beiden Schritten durch den Lanczos-Algorithmus erfolgt, wurde dieses Verfahren unter dem Namen *Two-Step Lanczos, TSL* erstmals in [102] vorgeschlagen. Bei Verwendung eines eigenwertbasierten Stopp-Kriteriums in beiden Schritten führt dies auf einen ausgewogenen Approximationsfehler im Verlauf des Verfahrens, wobei der Algorithmus vollständig automatisiert abläuft.

Es sollte jedoch erneut beachtet werden, dass die Effizienz des ersten Schritts nur gegeben ist, wenn die Systemmatrix in Gl. 3.1.2 oder Gl. 3.1.12 einfach erzeugt werden kann, was sich maßgeblich auf die einfache Invertierbarkeit der Materialmatrizen zurückführen lässt. Dies ist folglich bei FIT durch diagonale Matrizen gegeben, lässt sich aber nicht ohne Weiteres auf andere Simulationsverfahren wie beispielsweise FE anwenden. Bei FE ist eine partielle Realisierung je nach Aufwand für die Inversion der Materialmatrizen unter Umständen numerisch aufwendiger als die direkte Berechnung einer Padé Approximation.

Der zweite Schritt von TSL ist trotz der Namensgebung nicht grundsätzlich auf den Lanczos-Algorithmus festgelegt. Je nach Anwendungsfall kann auch der Arnoldi-Algorithmus oder eine Modalanalyse Vorteile bieten, wobei Rechenzeiten bei der Größe des vorreduzierten Modells hierbei nahezu keine Rolle mehr spielen.

Auch wenn das TSL-Verfahren ursprünglich für den klassischen Fall linearer Systeme entwickelt wurde, zeigt es seine besondere Stärke bei Anwendung auf Curl-Curl-Systeme [106, 107]. So ist, wie in Abschnitt 4.3.1.1 beschrieben, der Aufwand für die partielle Realisierung, der den dominanten Anteil im Gesamtaufwand darstellt, bei Curl-Curl-Systemen im Vergleich zum linearen Fall etwa halbiert. Die sich ergebende Modellgröße ist ebenfalls in etwa halbiert, was auch zusätzlich die LU-Zerlegung im zweiten Schritt beschleunigt.

Zudem werden Stabilität und Passivität im Curl-Curl-Fall in beiden Schritten bei gleichzeitiger Übereinstimmung einer maximalen Zahl von Momenten mit dem Originalsystem garantiert. Im Gegensatz zum linearen Fall gilt dies insbesondere auch, wenn im zweiten Schritt eine optimale imaginäre Entwicklungsfrequenz gewählt wird, was im Gegenzug eine minimale Modellgröße zur Folge hat. Dies ermöglicht garantiert passive Makromodelle mit sehr niedriger Ordnung und entsprechend Ersatzschaltbilder mit wenigen Elementen.

Schwach verlustbehaftete Systeme lassen sich entsprechend den beschriebenen Verfahren durch Projektion der Verlustmatrizen $\mathbf{K}$ oder $\mathbf{P}$ auf die jeweiligen Unterräume $\mathbf{V}_p$ und $\mathbf{V}_q$ einfach integrieren. Dies führt auf

$$\mathbf{Z}_{TSL}(s) = \mathbf{B}_q^T \left(s^2\mathbf{I} + s\mathbf{V}_q^T\mathbf{V}_p^T\mathbf{K}\mathbf{V}_p\mathbf{V}_q + \frac{1}{\sqrt{s}}\mathbf{V}_q^T\mathbf{V}_p^T\mathbf{P}\mathbf{V}_q\mathbf{V}_p + \mathbf{A}_q \right)^{-1} \mathbf{B}_q. \quad (4.3.51)$$

Zur Berechnung der reduzierten Verlustmatrizen stehen die Verfahren aus den Abschnitten 4.3.1.3 und 4.3.1.4 zur Verfügung. Ein Gesamtüberblick über TSL ist in Abb. 4.12 gegeben. Der numerische Aufwand bzw. die Komplexität des Verfahrens lässt sich experimentell bestimmen und sei hier exemplarisch erneut für das verlustfreie Zweiport-Filterbeispiel mit zwei Resonanzen im interessierenden Frequenzbereich dargestellt. Die Diskretisierung des Filters ist unterschiedlich fein gewählt, so dass die Gitterpunktzahl von 1800 bis zu 2,5 Millionen Punkten variiert, was bei Curl-Curl-Formulierung bis zu 7,5 Millionen Unbekannte im Originalsystem bedeutet. Untersucht werden zum Einen die benötigte Zahl von Iterationen der partiellen Realisierung im ersten Schritt, um das Eigenwert-Fehlerkriterium mit einer Genauigkeit von 10^{-6} zu erfüllen, zum anderen wird die Gesamtrechendauer beider Schritte verglichen. Die Rechnungen erfolgten auf einem PC mit 731 MHz Prozessortaktung. Das Ergebnis ist in doppelt logarithmischer Darstellung in Abb. 4.13 gezeigt.

Es zeigt sich, dass die Iterationszahl im ersten Schritt bei hoher Zahl von Gitterpunkten eine Ordnung von etwa $\mathcal{O}(n^{0.45})$ aufweist. Für niedrige Gitterpunktzahlen unter 20.000, Werte, die in der Praxis für dieses Beispiel durchaus ausreichen würden, ist die Steigung der Ausgleichsgeraden sogar noch geringer und beträgt etwa 1/3. Die Komplexität der Gesamtrechenzeit ergibt sich zu $\mathcal{O}(n^{4/3})$, ein Wert, der dem bekannter Verfahren wie dem CG-Algorithmus oder Zeitbereichsrechnungen nach der Leapfrog-Methode entspricht.

Aufgrund der speziellen Vorteile bei der Berechnung von resonanten Systemen, wurde der TSL-Algorithmus in [79] bereits erfolgreich zur Optimierung von Filterstrukturen eingesetzt.

Abschließend sollen noch die unterschiedlichen Ansätze verglichen werden, verlustbehaftete Strukturen mit TSL zu berechnen. Diese sind:

- partielle Realisierung / unsymmetrische Padé-Approximation des linearen Systems,
- symmetrische Projektion in beiden Schritten (Padé-Typ-Approximation), ebenfalls bei Betrachtung des linearen Systems und
- Projektion der Verluste auf die Unterräume, die sich durch die Lösung des verlustfreien Curl-Curl-Systems ergeben.

Die ersten beiden Varianten beschreiben die Verluste im Rahmen der Approximationsgenauigkeit exakt, wobei die zweite Variante in weniger Momenten mit dem Original übereinstimmt, dafür aber Passivität erhält. Die dritte Lösung ist eine Näherungslösung unter der Annahme, dass die Verluste wenig Einfluss auf das Feldbild im Resonator haben. Im verlustfreien Fall sind alle drei Varianten äquivalent.

Es wird erneut das Testfilter betrachtet und angenommen, dass dieses nun anstelle von Luft mit einem verlustbehafteten Dielektrikum gefüllt ist. Die Verlustwinkel werden als $\tan\delta = 0{,}001,\ 0{,}01$ und $0{,}1$ angenommen. Die Transmission des Filters wird hierdurch zum Teil erheblich verringert, wie in Abb. 4.14 dargestellt. Der Wert von $\tan\delta > 0,01$ wird nur zum Test des Verfahrens betrachtet, realistische Materialien haben üblicherweise kleinere Werte.

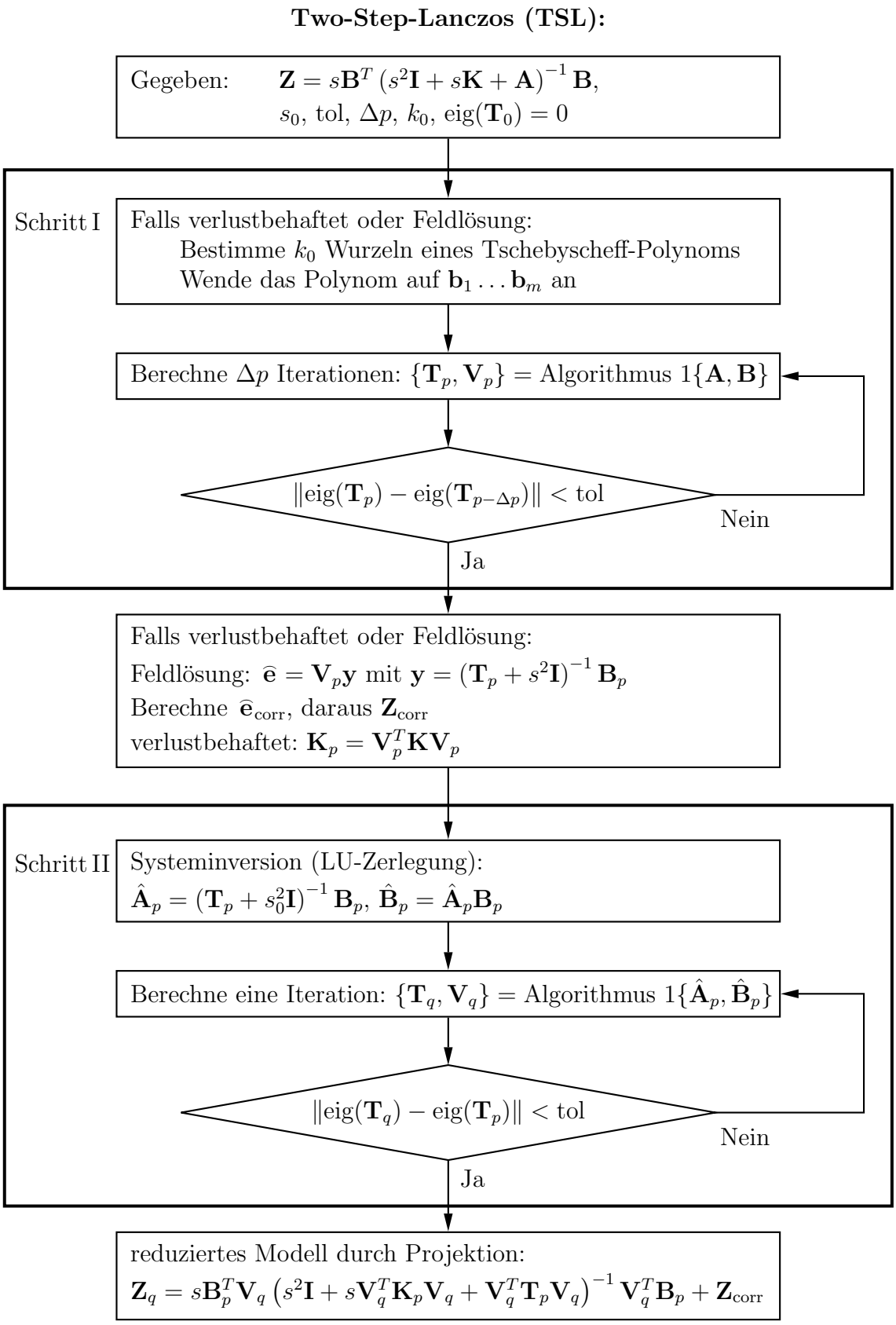


Abbildung 4.12: *Blockdiagramm des Two-Step-Lanczos-Algorithmus.*

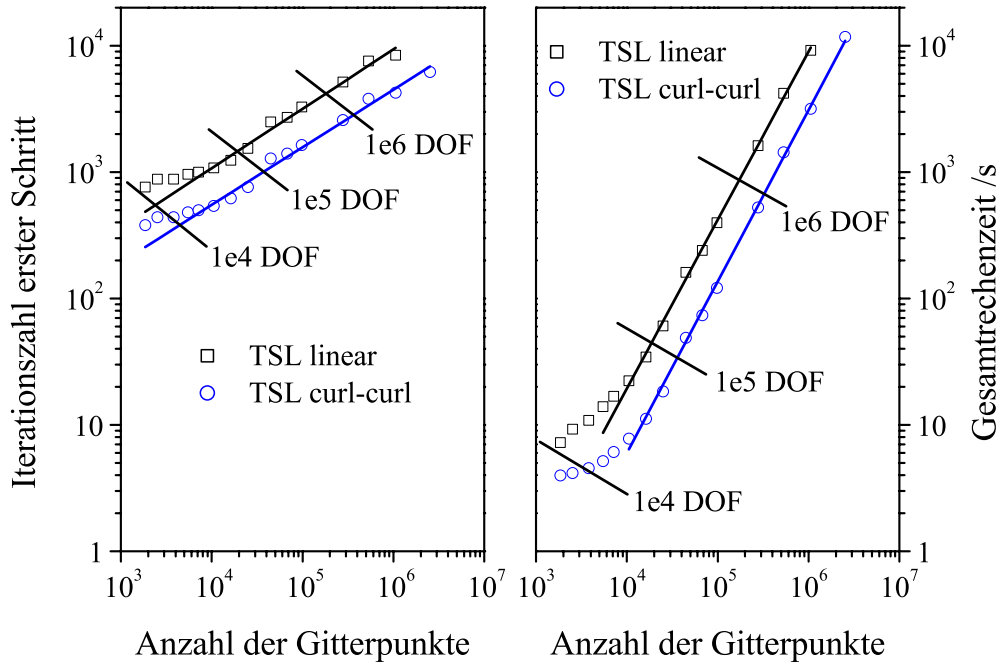


Abbildung 4.13: *Komplexität des TSL-Algorithmus: Iterationszahl des ersten Schritts und Gesamtrechenzeit sowohl für die lineare als auch die Curl-Curl-Formulierung. Die größten Modelle haben 7,5 und 6,3 Millionen Unbekannte für den linearen bzw. Curl-Curl-Fall, der Rechenzeitgewinn zwischen beiden Formulierungen entspricht etwa einen Faktor von drei.*

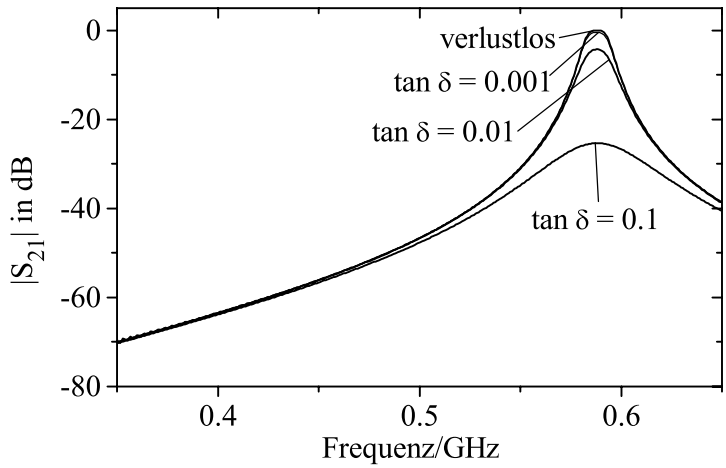


Abbildung 4.14: *Transmission des Testfilters bei zunehmender Leitfähigkeit des Füllmaterials. Die Kurven des verlustfreien Filters und bei* $\tan\delta = 0.001$ *überdecken sich nahezu.*

Zur Berechnung werden für den ersten Schritt im linearen Fall $p = 850$ und im Curl-Curl-Fall $p = 400$ Iterationen benutzt, im zweiten Schritt jeweils $q = 12$. Diese Werte sind so gewählt, dass der Approximationsfehler des verlustfreien Falls keinen nennenswerten Einfluss auf den Vergleich hat.

Der über den gesamten betrachteten Frequenzbereich gemittelte relative Fehler zur direkt bestimmten Lösung des vollen Systems ist für die unterschiedlichen Verfahren in Tabelle 4.2 aufgeführt.

$\tan\delta$	0	0.001	0.01	0.1
Padé-Approx. linear	3.29e-010	1.37e-010	7.38e-011	7.37e-011
sym. Projektion linear	3.29e-010	7.76e-009	3.13e-008	7.25e-008
Projektion Curl-Curl	3.45e-011	1.67e-008	4.17e-007	1.11e-006

Tabelle 4.2: *Gemittelter Approximationsfehler für unterschiedliche Verfahren zur Behandlung verlustbehafteter Systeme.*

Für den Fall einer unsymmetrischen Projektion in beiden Schritten zeigt sich deutlich, dass der Approximationsfehler nahezu unabhängig von der Größe der Verluste ist. Der Fehler nimmt sogar leicht ab, was damit zu erklären ist, dass die Polstellen flacher werden und die Referenzlösung in der Nähe von Singularitäten stets ungenau ist. Wird die symmetrische Projektion in beiden Schritten verwendet, steigt der Approximationsfehler mit den Verlusten an. Dies lässt sich erklären, da bei Verlusten zwar weniger Momente übereinstimmen, beide Verfahren im verlustfreien Fall jedoch ineinander übergehen. Der stärkste Anstieg des Approximationsfehlers ist für den Curl-Curl-Fall zu registrieren. Auch dies ist verständlich, da die Verluste im Prozess der Reduzierung überhaupt nicht betrachtet werden, sondern nur nachträglich berücksichtigt werden. Diese Annahme ist nur bis zu gewissen Leitfähigkeiten zulässig.

Betrachtet man die sich einstellenden Approximationsfehler absolut, zeigt sich, dass alle Varianten gute Näherungen bilden. Selbst ein Fehler von 10^{-6} ist in nahezu allen praktischen Fällen durchaus ausreichend. In Kombination mit der Rechenzeitersparnis erscheint die Curl-Curl-Variante bei parasitären Verlusten daher als die effizienteste. Zur Behandlung allgemeiner Leitfähigkeiten muss jedoch zu einer der linearen Varianten übergegangen werden.

4.3.5 Weitere Verfahren

Im Folgenden sollen noch drei weitere Verfahren kurz angesprochen werden, die enge Zusammenhängezu den bereits beschriebenen Methoden aufweisen.

4.3.5.1 Laguerre-Approximationen

Betrachtet man das allgemeine lineare oder linearisierte System nach Gl. 3.1.2 und führt die folgende Koordinatentransformation

$$u = \frac{s - \alpha}{s + \alpha} \tag{4.3.52}$$

mit $\alpha = 2\pi f_{\max}$ durch, führt dies auf eine Impedanzfunktion der Form

$$\mathbf{Z}(u) = \mathbf{B}^T \left((\alpha\mathbf{I} + \mathbf{A}) + u(\alpha\mathbf{I} - \mathbf{A})\right)^{-1} \mathbf{B}. \tag{4.3.53}$$

Wird der Lanczos-Algorithmus nun entsprechend einer Padé-Approximation mit den Matrizen

$$\hat{\mathbf{A}} = (\alpha\mathbf{I} + \mathbf{A})^{-1}(\alpha\mathbf{I} - \mathbf{A}) \quad \text{und} \quad \hat{\mathbf{B}} = (\alpha\mathbf{I} + \mathbf{A})^{-1}\mathbf{B} \tag{4.3.54}$$

gestartet und eine symmetrische Projektion nach Gl. 4.3.35 durchgeführt, entspricht dies einer Approximation der Impulsantwort des Systems in Laguerre-Polynomen. Eine detaillierte Beschreibung der Eigenschaften dieser Approximation findet sich in [80].

4.3.5.2 Rationale Interpolation

Werden zur Projektion des Systems bereits berechnete Feldlösungen verwendet, entspricht das reduzierte Modell einer rationalen Interpolation. Zugleich kann diese Approximation auch als eine Multipoint-Padé-Approximation mit je einem Moment pro Entwicklungspunkt betrachtet werden. Wie bereits erwähnt, hängt die Qualität einer solchen Näherung allerdings stark von der Wahl der Stützpunkte ab und ist daher schwer steuerbar.

Das Verfahren lässt sich jedoch sehr vorteilhaft in einem klassischen *Frequency Sweep* zur Berechnung des Übertragungsverhaltens im Frequenzbereich nutzen. Hierbei wird das System zunächst auf eine feste Anzahl der letzten berechneten Feldlösungen projiziert. Das reduzierte System wird gelöst und das Ergebnis dient im Anschluss als Startlösung für einen klassischen iterativen Solver zur Lösung des Originalsystems. Da die Startlösung aufgrund der angewandten Approximation bereits nahe an der tatsächlichen Lösung liegt, wird die Konvergenz des Solvers stark beschleunigt. Da sowohl die Feldlösungen als auch der iterative Solver eng mit Krylov-Unterräumen verbunden sind, besitzt dieses Verfahren wiederum eine enge Verwandschaft zu partiellen Realisierungen. Für genauere Beschreibungen siehe [81].

4.3.5.3 Balanced Truncation

Ein in der Regelungstechnik sehr weit verbreitetes Verfahren zur Reduzierung der Ordnung sind so genannte *Balanced Truncations*, gelegentlich auch als optimale Hankel-Norm-Reduzierung bezeichnet. Sie beruhen auf der Grundidee [67], dass jedes System durch eine Ähnlichkeitstransformation in eine Form gebracht werden kann, in der jeder Zustand gleichermaßen beobachtbar und steuerbar ist. Dieses System wird als ausgewogen (engl. *balanced*) bezeichnet. Zustände, die in dieser Darstellung nur schwer erreichbar sind, können schließlich eliminiert werden.

Ausgangspunkt des Verfahrens ist die Betrachtung der Gramschen Steuerbarkeitsmatrix $\mathbf{W}_C$ und die entsprechende Beobachtbarkeitsmatrix $\mathbf{W}_O$ nach Abschnitt 3.2.4. Um diese zu berechnen, muss die Lyapunov-Gleichung gelöst werden [35].

Werden beide Matrizen mit $\mathbf{W}_C = \mathbf{X}\mathbf{X}^T$ und $\mathbf{W}_O = \mathbf{Y}\mathbf{Y}^T$ Cholesky-zerlegt und anschließend eine Singulärwertzerlegung [43] von $\mathbf{X}^T\mathbf{Y} = \mathbf{U}_L\mathbf{\Sigma}\mathbf{U}_R^T$ durchgeführt, folgt für die Projektionsmatrix:

$$\mathbf{V}_{\mathrm{bt}} = \mathbf{X}\mathbf{U}_L\mathbf{\Sigma}^{-1/2} = (\mathbf{\Sigma}^{-1/2}\mathbf{U}_R^T\mathbf{Y}^T)^{-1}. \tag{4.3.55}$$

Wird das System auf $\mathbf{V}_{\mathrm{bt}}$ projiziert, gilt für die Gramschen Steuerbarkeits- und Beobachtbarkeitsmatrizen des resultierende Systems:

$$\bar{\mathbf{W}}_C = \bar{\mathbf{W}}_O = \mathbf{\Sigma}. \tag{4.3.56}$$

Die Werte σ_k der Diagonalmatrix $\mathbf{\Sigma}$ werden als Hankel-Singulärwerte bezeichnet und sind in dieser Formulierung ein direktes Maß für die Erreichbarkeit des zugehörigen Zustands. Ein reduziertes Modell kann folglich erzeugt werden, indem die Spalten der Matrix $\mathbf{V}_{\mathrm{bt}}$, die zu kleinen Hankel-Singulärwerten gehören, eliminiert werden.

Der Erfolg dieses Verfahrens beruht auf der Tatsache, dass sich für das gesamte Spektrum ein Maximalfehler der Übertragungsfunktion aus den unberücksichtigten Hankel-Singulärwerten σ_k angeben lässt:

$$\max_{\omega}\|\mathbf{Z}(\omega) - \mathbf{Z}_{\mathrm{bt}}(\omega)\| \leq 2\sum \sigma_k. \tag{4.3.57}$$

Die Frage der benötigten Modellgröße ist also wesentlich einfacher zu beantworten als bei den bisher beschriebenen Verfahren. Umgekehrt lässt sich in dieser Methode, anders als bei Padé-Approximationen, jedoch kein Bereich angeben, in der der Approximationsfehler besonders gering ist.

Da sowohl die Lösung der Lyapunov-Gleichung auch auch die Singulärwertzerlegung mit kubischem Aufwand über der Systemgröße ansteigt, gilt dieses Verfahren im Zusammenhang mit elektrodynamischen Simulationen als praktisch nicht einsetzbar. In Verbindung mit TSL ist es jedoch möglich, *Balanced Truncation* auf das vorreduzierte System nach dem ersten oder dem zweiten Schritt anzuwenden, wobei das obige Fehlerkriterium in diesem Fall natürlich nur noch relativ zur bereits gemachten Approximation gilt. Dies kann insbesondere Vorteile bringen, wenn Padé-Approximationen aufgrund zu hoher Portzahlen ihre Effizienz verlieren.

Kapitel 5

Spektralschätzung aus Zeitbereichsdaten

Alle im letzten Kapitel beschriebenen Verfahren zur Modellreduzierung verwenden als Ausgangsbasis direkt das FIT-diskretisierte Modell der Struktur, die eigentliche Lösung erfolgt erst auf der reduzierten Ebene. Während dies im Fall resonanter Strukturen bei Berechnung einer partiellen Realisierung oder Anwendung des TSL-Verfahrens sehr effizient sein kann, kann die Berechnung einer Padé Approximation oder einer Modalanalyse im allgemeinen Fall, z. B. in Verbindung mit offenen Rändern, auch aufwändig werden.

Da mit dem expliziten Zeitbereichsverfahren ein sehr effizientes Lösungsverfahren für FIT-Systeme existiert, besteht ein alternativer Ansatz darin, zunächst das vollständige System mit einer geeigneten Anregung im Zeitbereich zu lösen und im Anschluss das gewünschte reduzierte Modell oder das Übertragungsverhalten aus den Zeitbereichsdaten zu gewinnen. Dieses Vorgehen ist immer dann sinnvoll, wenn die Lösung des Systems weniger Zeit in Anspruch nimmt, als die Berechnung eines reduzierten Modells.

Im folgenden Kapitel soll zunächst das explizite Zeitintegrationsverfahren für FIT-Systeme kurz erläutert werden. Zur Spektralschätzung aus den Zeitsignalen werden im Anschluss parameterbasierte Verfahren angewendet, die für resonante Strukturen weit bessere Eigenschaften aufweisen als die auf einer diskreten Fouriertransformation (DFT) basierenden. Zum Aufbau eines Modells aus Zeitbereichsdaten hat sich der 4SID-Algorithmus bewährt, der am Ende des Kapitels kurz vorgestellt werden soll.

5.1 FIT-Simulationen im Zeitbereich

Die allgemeine Lösung des linearen FIT-Systems nach Gl. 3.1.1 im Zeitbereich ergibt sich durch die Beziehung Gl. 3.1.5 in Abschnitt 3.1.1. Soll das System im Zeitbereich numerisch gelöst werden, erfordert dies neben der ohnehin bereits erfolgten räumlichen Diskretisierung auch die Diskretisierung der Zeitachse. Dies erfolgt für

HF-Anwendungen häufig durch die Abtastung der Signale mit einer festen Zeitschrittweite: $t^{(g)} = g \cdot \Delta t,\ g \in \mathbb{N}$. Der kontinuierliche Ableitungsoperator $\frac{d}{dt}$ wird außerdem durch den zentralen Differenzenquotienten ersetzt:

$$\frac{d}{dt} f(g\Delta t) = \frac{f((g+\frac{1}{2})\Delta t) - f((g-\frac{1}{2})\Delta t)}{\Delta t} + \mathcal{O}(\Delta t^2). \tag{5.1.1}$$

Die Berechnung der Ableitung aus um einen halben Zeitschritt versetzten Funktionswerten führt hierbei auf eine quadratische Fehlerordnung. Im Folgenden werden die elektrischen Größen $\widehat{\mathbf{e}}$ und $\widehat{\widehat{\mathbf{d}}}$ zu den *vollen* Zeitschritten $g\Delta t$ und die magnetischen Vektoren $\widehat{\mathbf{h}}$ und $\widehat{\widehat{\mathbf{b}}}$ zu den *halben* Zeitschritten $(g+\frac{1}{2})\Delta t$ abgetastet.

Unter Berücksichtigung der Materialmatrizen und durch Einsetzen des Differenzenquotienten ergibt sich für den verlustfreien Fall ein explizites Rekursionsschema, das so genannte *Leapfrog*-Schema. In einer Matrixschreibweise nach [13] lautet es:

$$\mathbf{x}^{(g+1)} = \mathbf{A}_t \mathbf{x}^{(g)} + \mathbf{q}^{(g)} \tag{5.1.2}$$

mit der System- bzw. Zeitschrittmatrix $\mathbf{A}_t$ und den Vektoren $\mathbf{x}^{(g)}$ und $\mathbf{q}^{(g)}$:

$$\mathbf{A}_t = \begin{pmatrix} \mathbf{I} - \Delta t^2 \mathbf{M}_\varepsilon^{-1} \mathbf{C}^T \mathbf{M}_\mu^{-1} \mathbf{C} & \Delta t \mathbf{M}_\varepsilon^{-1} \mathbf{C}^T \\ -\Delta t \mathbf{M}_\mu^{-1} \mathbf{C} & \mathbf{I} \end{pmatrix}, \tag{5.1.3a}$$

$$\mathbf{x}^{(g)} = \begin{pmatrix} \widehat{\mathbf{e}}^{(g)} \\ \widehat{\mathbf{h}}^{(g+1/2)} \end{pmatrix}, \quad \mathbf{q}^{(g)} = \begin{pmatrix} \Delta t \mathbf{M}_\varepsilon^{-1} \widehat{\widehat{\mathbf{j}}}_e^{(g+1/2)} \\ \mathbf{0} \end{pmatrix}. \tag{5.1.3b}$$

Die Berechnung eines neuen Zeitwerts besteht also im wesentlichen aus einer Matrix-Vektor-Multiplikation. Zudem muss stets nur ein einziger Feldvektor im Speicher gehalten werden, was die Methode sehr effizient macht. Das beschriebene Verfahren ist auf kartesischen Gittern mathematisch äquivalent zu der *Finite Difference Time Domain, FDTD* Methode, die erstmals 1966 in [16] vorgeschlagen wurde.

Eine bedeutende Einschränkung des Verfahrens stellt die Bedingung dar, dass die Zeitinteration nach Gl. 5.1.2 nur stabil ist, wenn alle Eigenwerte $|\lambda(\mathbf{A}_t)| \leq 1$ sind. Diese Forderung lässt sich mit dem größten Eigenwert der Systemmatrix des linearen Systems gemäß Gl. 3.1.1 nach [13] zu

$$\Delta t_{\max} = \frac{2}{|\lambda_{\max}(\mathbf{A}_l)|} \tag{5.1.4}$$

umformen. Eine Abschätzung für den maximalen Zeitschritt gibt zudem das *Courant-Friedrichs-Levi*-Kriterium

$$\Delta t_{\max} = \min_k \sqrt{\frac{\varepsilon_k \mu_k}{\frac{1}{\Delta u_k^2} + \frac{1}{\Delta v_k^2} + \frac{1}{\Delta w_k^2}}}. \tag{5.1.5}$$

Diese Beschränkung limitiert das beschriebene Verfahren de facto auf Hochfrequenzanwendungen und selbst dort ist die vom Stabilitätskriterium bzw. zum Erreichen einer gewissen Genauigkeit nach Gl. 5.1.1 vorgegebene Abtastrate häufig weit höher, als dies aus informationstheoretischer Sicht notwendig wäre.

Weitere wichtige Eigenschaften des Verfahrens wie Energie- und Ladungserhaltung sowie die Erweiterung für verlustbehaftete Materialien werden ausführlich in [13, 20] beschrieben.

5.2 Filterbasierte Spektralschätzung

Einschwingvorgänge oder Pulsberechnungen in der Zeitbereichsreflektometrie stellen typische Zeitbereichsanwendungen dar. Aufgrund der Effizienz des Verfahrens gegenüber Frequenzbereichsrechnungen werden häufig jedoch auch typische Frequenzbereichsgrößen wie Übertragungsfunktionen mittels Zeitbereichsrechnungen gewonnen. Hierzu wird die Struktur mit einem breitbandigen, typischerweise gaußmodulierten, Puls angeregt. Im Anschluss werden die entsprechenden resultierenden Zeitsignale durch die diskrete Fouriertransformation (DFT) in den Frequenzbereich überführt.

Da bei der Berechnung der DFT implizit die periodische Fortsetzung des Signals vorausgesetzt wird, liefert die Transformation nur verlässliche Ergebnisse, wenn das Signal bereits auf nahezu Null abgeklungen ist. Sinkt die Signalamplitude nur sehr langsam ab, wie beispielsweise bei resonanten Strukturen, kann das Verfahren in Verbindung mit dem vorgegebenen maximalen Zeitschritt $\Delta t_{\max}$ seine Effizienz verlieren.

Eine Alternative zur Transformation eines Zeitsignals in den Frequenzbereich liefern spezielle Signalverarbeitungsverfahren, die den Signalverlauf im Zeitbereich mit Hilfe der Impulsantwort digitaler Filter nachbilden. Aus den optimierten Filterkoeffizienten dieser so genannten *linearen Prädiktoren* kann ebenfalls das Spektrum berechnet werden. Sobald das Filter mit genügender Genauigkeit gefunden ist, hat der weitere Verlauf des Zeitsignals keine weitere Bedeutung, weswegen kein Abschneidefehler auftritt. Die Problemstellung wird auch in Abb. 5.1 dargestellt.

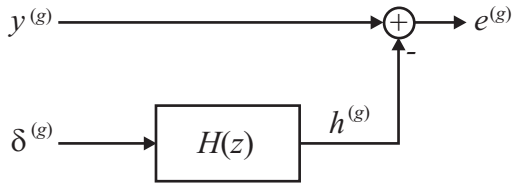


Abbildung 5.1: *Signalmodellierung mit Hilfe der Impulsantwort $h^{(g)}$ des Filters H(z).*

Da alle Signale nur zeitdiskret vorliegen, wird die Übertragungsfunktion $H(z)$ im Bildbereich der Z-Transformation [34] angegeben. Zur Laplace-Transformation gilt der Zusammenhang $z = e^{s\Delta t}$. Die diskrete Diracfolge wird durch $\delta^{(g)}$ gekennzeichnet. Das Filteroptimierungsproblem lautet nun mit der zu schätzenden Signalfolge $y^{(g)}$, der Impulsantwort $h^{(g)}$ des Filters $H(z)$ und dem Prädiktionsfehler $e^{(g)} = h^{(g)} - y^{(g)}$

$$\sum_g |e^{(g)}|^2 = \sum_g |h^{(g)} - y^{(g)}|^2 \to \min. \tag{5.2.1}$$

Nach dem Parsevalschen Theorem [34] ist diese Bedingung gleichbedeutend mit der Minimierung des Frequenzbereichsfehlers $E(e^{j\omega\Delta t})$:

$$\int_{-\pi/\Delta t}^{\pi/\Delta t} |E(e^{j\omega\Delta t})|^2 \, d\omega = \int_{-\pi/\Delta t}^{\pi/\Delta t} |H(e^{j\omega\Delta t}) - Y(e^{j\omega\Delta t})|^2 \, d\omega \to \min. \tag{5.2.2}$$

5.2.1 ARMA-Modelle

Eine verhältnismäßig einfache Möglichkeit, ein solches Filter zu wählen, stellen so genannte *autoregressive* Filter (AR-Filter) mit der Übertragungsfunktion

$$H(z) = \frac{b_0}{1 + a_1 z + a_2 z^2 + \ldots + a_K z^K} \tag{5.2.3}$$

dar. Das Optimierungsproblem entspricht in diesem Fall der Lösung eines meist überbestimmten Gleichungssystems der Dimension $G \times K$. Hierbei ist G die Länge der zur Optimierung herangezogenen Zahlenfolge, K die Filterordnung.

Bei der Auswahl der Signalfolge sollten einige Punkte beachtet werden:

- Das Signal einer FIT-Zeitbereichsrechnung ist aufgrund des vorgegebenen Zeitschritts im Vergleich zum Shannonschen Abtasttheorem meist deutlich überabgetastet. Zur effizienten Filterbestimmung ist es daher ratsam, zunächst eine Reduzierung (engl. *downsampling*) der Folge vorzunehmen. Besteht die Gefahr, dass sich die periodisch fortgesetzten Spektren überlagern (engl. *aliasing*), muss zudem vorher eine digitale Filterung des Signals vorgenommen werden.
- Zu Beginn der Zeitfolge enthält das Spektrum häufig noch eine sehr große Anzahl von Schwingungen, die schnell abklingen. Wird ein späterer Ausschnitt der Folge untersucht, enthält dieser nur noch die Anteile von resonanten Schwingungen mit höherer Güte. Zu deren Beschreibung ist meist ein Filter kleinerer Ordnung ausreichend. Der nicht betrachtete vordere Teil der Folge kann später mit Hilfe einer DFT zum geschätzten Spektrum superponiert werden.

Ein speziell auf die Anforderungen der Zeitbereichssimulationen in FIT angepasstes Verfahren zur Berechnung eines AR-Filters wird detailliert in [20] beschrieben.

Ein AR-Filter ist aufgrund seiner Struktur gut geeignet, Polstellen des Systems aufzufinden. Sollen jedoch Nullstellen des Spektrums genau abgebildet werden, erfordert dies eine sehr hohe Filterordnung oder führt gar zum Scheitern des Verfahrens. Es erscheint also sinnvoll, ein Filter zu wählen, das neben einem Nennerpolynom auch ein Zählerpolynom besitzt

$$H(z) = \frac{b_0 + b_1 z + b_2 z^2 + \ldots + b_{(K-1)} z^{(K-1)}}{1 + a_1 z + a_2 z^2 + \ldots + a_K z^K} = \frac{B(z)}{A(z)}. \tag{5.2.4}$$

Ein solches Filter wird auch als *Autoregressives Moving-Average, ARMA*-Modell bezeichnet. Stimmt die Länge der Signalfolge genau mit der Anzahl der bestimmbaren Filterkoeffizienten überein, führt dies auf eine rationale Interpolation des Zeitsignals. Soll nun allerdings der Fehler nach den Gln. 5.2.1 bzw. 5.2.2 für einen längeren Signalausschnitt minimiert werden, erfordert dies eine komplexe nichtlineare Optimierung. In der Praxis weicht man daher auf eine indirekte Modellierung aus. Aus einem Teil der Signalfolge wird zunächst mit der oben beschriebenen Methode das Nennerpolynom bestimmt und dieses mit der Signalfolge gefaltet. Die Koeffizienten des Zählerpolynoms werden schließlich auf das Ergebnis der Faltung optimiert,

wie symbolisch in Abb. 5.2 dargestellt. Anstelle des Fehler in Gl. 5.2.1 wird damit allerdings der folgende Fehler minimiert:

$$\sum_g |e_i^{(g)}|^2 = \sum_g |y^{(g)} * a^{(g)} - b^{(g)}|^2 \rightarrow \min, \tag{5.2.5}$$

bzw.

$$\int_{-\pi/\Delta t}^{\pi/\Delta t} |E_i(e^{j\omega\Delta t})|^2 \, d\omega = \int_{-\pi/\Delta t}^{\pi/\Delta t} |Y(e^{j\omega\Delta t})A(e^{j\omega\Delta t}) - B(e^{j\omega\Delta t})|^2 \, d\omega \rightarrow \min. \tag{5.2.6}$$

Der Zusammenhang zum obigen Fehler lautet damit im Z-Bereich

$$E(z) = \frac{E_i(z)}{A(z)}. \tag{5.2.7}$$

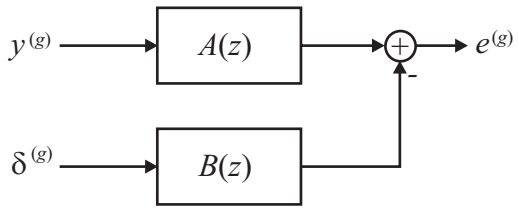


Abbildung 5.2: *Signalmodellierung beim Prony-Verfahren.*

Eine Methode, die auf der indirekten Modellierung beruht, ist das so genannte *Prony*-Verfahren. Es benötigt die Lösung zweier linearer Systeme. Eine genaue Darstellung findet sich ebenfalls in [20].

5.2.2 Iterativer Prony und Verfahren nach Steiglitz-McBride

Um anstelle des indirekten Fehlers $E_i(z)$ den tatsächlichen Fehler $E(z)$ zu minimieren ohne auf nichtlineare Optimierung zurückgreifen zu müssen, finden zwei iterative Verfahren Anwendung [97]. Diese sind das iterative Prony-Verfahren nach [82, 83] und das Verfahren nach Steiglitz und McBride [84]. Beide gehen von einem Prony-Modell als Startlösung aus.

Das iterative Prony-Verfahren wandelt das Startmodell unter der Voraussetzng nur einfacher Pole zunächst in eine Pol-Residuen-Darstellung

$$H(z) = \frac{B(z)}{A(z)} = \sum_{k=1}^{K} \frac{c_k z}{z - z_k}, \tag{5.2.8}$$

oder im Zeitbereich

$$h^{(g)} = \sum_{k=1}^{K} c_k z_k^g \tag{5.2.9}$$

um. Im Folgenden wird zwar die getrennte Behandlung der Nennerkoeffizienten z_k und der Zählerkoeffizienten c_k wie beim klassischen Prony-Verfahren beibehalten, durch die abwechselnde Optimierung der Koeffizienten nähert sich der Fehler des Verfahrens jedoch iterativ dem Fehler $E(z)$ an. Ein Iterationsdurchlauf besteht stets aus den beiden Schritten:

- Optimierung der Polstellen z_k bei festgesetzten Residuen nach dem Marquardt-Algorithmus [82],
- Berechnung der neuen Residuen c_k durch ein Least-Square-Verfahren ähnlich dem beim klassischen Prony.

Das Verfahren nach Steiglitz-McBride nutzt ebenfalls den Nenner des Prony-Filters, diesmal um die Anregungssignale $h^{(g)}$ und $y^{(g)}$ vorzufiltern, weswegen das Verfahren gelegentlich auch als *Iteratives Prefiltering* bezeichnet wird. Das Vorgehen wird in Abb. 5.3 verdeutlicht.

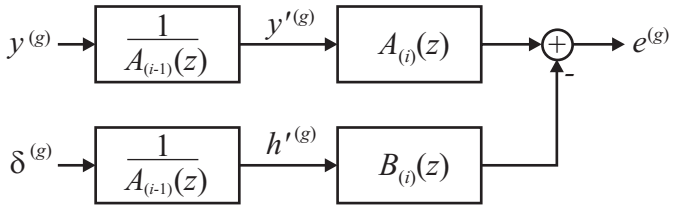


Abbildung 5.3: *Signalmodellierung beim Steiglitz-McBride-Verfahren.*

Als Fehler ergibt sich in der Z-Ebene:

$$E_{\mathrm{smb}}(z) = \frac{B_{(i)}(z)}{A_{(i-1)}(z)} - Y(z)\frac{A_{(i)}(z)}{A_{(i-1)}(z)}. \tag{5.2.10}$$

Konvergiert das Verfahren, so gilt $A_{(i)}(z) = A_{(i-1)}(z)$ und E_{smb} geht damit in den gesuchten direkten Fehler

$$E(z) = \frac{B(z)}{A(z)} - Y(z) \tag{5.2.11}$$

über.

Eine weitere Möglichkeit des Verfahrens nach Steiglitz-McBride ergibt sich aus der Tatsache, dass zur Fehleroptimierung nicht die Impulsantwort des Filters $b_{(i)}^{(g)}$, sondern das gefaltete Signal $h'^{(g)} * b_{(i)}^{(g)}$ verwendet wird. Folglich kann auch von vornherein anstelle von $\delta^{(g)}$, wie beim klassischen Prony, ein beliebiges Anregungssignal verwendet werden. Wenn zur Schätzung direkt das Ausgangs- und das Eingangssignal berücksichtigt werden, bedeutend dies, dass das Filter direkt das Übertragungsverhalten, beispielsweise den Streuparameter, modelliert. Allerdings müssen in diesem Fall alle Zeitsignale von Anbeginn verwendet werden, was die Modellgröße im Vergleich zur Schätzung eines späteren Signalausschnitts meist vergrößert.

5.2.3 Ein Beispiel

Neben der Schätzung des Übertragungsverhaltens resonanter Systeme bieten die beschriebenen Signalverarbeitungsmethoden vor allem auch sehr genaue Abschätzungen für die Güten der einzelnen Resonanzen. Die Güte ergibt sich hierbei wie bereits in Abschnitt 3.3.3 erwähnt aus dem komplexen Pol nach

$$Q_k = \frac{\mathrm{Im}\{s_k\}}{2\mathrm{Re}\{s_k\}} = \frac{\mathrm{Im}\{\ln(z_k)/\Delta t\}}{2\mathrm{Re}\{\ln(z_k)/\Delta t\}}. \tag{5.2.12}$$

Als Beispiel soll ein neunzelliger Resonator nach Abb. 5.4 aus der Teilchenbeschleunigerphysik dienen. Er ist Bestandteil des TESLA-Projekts [85]. Während für die Beschleunigungsmoden in dem Resonator hohe Güten erwünscht sind, treten in der Struktur auch so genannte *gefangene* Moden mit sehr hoher Güte auf, die das passierende Elektronenpaket empfindlich stören können. Die Analyse solcher hochresonanter Güten erweist sich als aufwendig. Ein Überblick über Verfahren, die meist auf einer Modalanalyse beruhen, ist in [33] gegeben. An dieser Stelle sollen die genannten Spektralschätzungsmethoden verwendet werden, um die Güten aus einem Zeitsignal zu bestimmen, das während einer transienten Simulation im Inneren der Struktur aufgezeichnet wurde. Aus Anwendungssicht ist für die zu erwartenden extrem hohen Güten dabei eine Genauigkeit von einigen Prozent durchaus ausreichend.

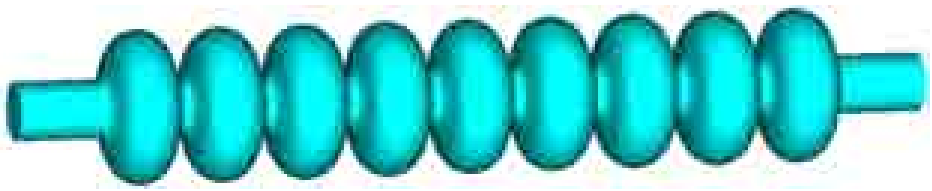


Abbildung 5.4: *Neunzelliges TESLA-Beschleunigungsmodul.*

Zunächst soll nur eine einzelne Resonatorzelle betrachtet werden, da die darin auftretenden moderaten Güten einen Vergleich der Verfahren zulassen. Es wird ein Signalausschnitt von 0.25–0.42 μs mit 2000 Werten verwendet, dessen Abtastrate auf den 1,3-fachen Wert des Shannontheorems reduziert wurde. Die Ergebnisse im Vergleich zu einer Modalberechnung sind in Tabelle 5.1 zusammengefasst. Die Ordnung aller Verfahren beträgt 70, das iterative Prony-Verfahren verwendet 10 Iterationen, das Steiglitz-McBride-Verfahren 5. Der iterative Prony wurde aus [83] verwendet. Während die niedrigen Güten von allen Verfahren gut geschätzt werden,

Frequenz in GHz	Güten				
	Modal	AR	Prony	It. Prony	St.-McBr.
3,3364	1,58e3	1,54e3	1,54e3	1,54e3	1,54e3
4,1347	7,28e5	6,01e5	5,88e5	7,10e5	7,13e5
4,1530	2,02e7	4,34e5	3,76e5	5,84e5	2,27e7
4,3858	2,76e5	2,84e5	2,88e5	6,91e5	2,63e5

Tabelle 5.1: *Güteschätzung aus dem Zeitsignal eines TESLA-Einzellers im Vergleich zum Modalwert.*

wird die scharfe Resonanz bei 4.15 GHz nur vom Steiglitz-McBride-Verfahren genau ermittelt. Zur Analyse eines Zeitsignals aus der vollen neunzelligen Struktur wird daher nur dieses Verfahren herangezogen. Es werden zwei Signalausschnitte mit je 4000 Werten betrachtet, eines von 3,5–4,05 μs, das andere etwas später von 4,45–5 μs. Unabhängig von Signalausschnitt und Modellordnung werden die Frequenzen der technisch relevanten gefangenen Moden um 3,05 GHz stets vollständig und sehr genau gefunden. Die entsprechende Resonanz ist auch im Einzeller zu finden, weist dort jedoch nur eine Güte von etwa 60 auf, die erst durch die Kopplung der neun Zellen in der vollen Struktur deutlich vergrößert wird. Die geschätzten Güten sind in Tabelle 5.2 dargestellt. Aus Gründen der numerischen Stabilität wurden nur je

zwei Iterationen des Verfahrens durchgeführt.

Frequenz in GHz	Modal	Ausschnitt I: 3,5 - 4,05 μs			Ausschnitt II: 4,45 - 5 μs		
		Or. 100	Or. 125	Or. 150	Or. 100	Or. 125	Or. 150
3,0583	5,96e6	5,65e5	1,08e6	4,46e5	2,58e6	5,79e6	7,48e6
3,0615	3,47e5	1,23e5	2,23e5	1,20e5	2,93e5	4,38e5	5,02e5
3,0689	3,53e5	3,33e5	3,39e5	3,49e5	3,37e5	3,39e5	3,39e5

Tabelle 5.2: *Güteschätzung aus dem Zeitsignal eines TESLA-Neunzellers mit dem Verfahren nach Steiglitz-McBride für zwei Signalausschnitte und unterschiedliche Modellordnung (Or.) Als Referenz dient ein Modalwert aus [27].*

Es zeigt sich, dass der Mode bei 3,0689 GHz immer sehr exakt gefunden wird. Die anderen Güten mit geringerer Signalamplitude schwanken stärker und werden erst im hinteren Signalausschnitt zuverlässiger gefunden. Es sei jedoch erwähnt, dass auch die Referenzgüten der Modalanalyse bei unterschiedlichen Gitterauflösungen in vergleichbarem Rahmen schwanken. Zusammenfassend bietet die Signalschätzung nach Steiglitz-McBride damit auch für sehr hohe Güten eine sehr verlässliche Schätzung der Resonanzfrequenz sowie eine brauchbare Abschätzung der Güte.

5.3 4SID

Zum Abschluss dieses Zeitbereichskapitels soll noch kurz darauf hingewiesen werden, dass neben den beschriebenen Methoden zur Schätzung einzelner Signale auch Verfahren existieren, die aus ein- oder mehrdimensionalen Zeitsignalen eine Zustandsraumdarstellung mit den Matrizen **A**, **B**, **C** und **D** generieren. Diese Verfahren werden als *Subspace-based State-Space System Identifikation* oder kurz *4SID* bezeichnet.

Hierzu werden die diskreten Eingangs- sowie die Ausgangszeitsignale geeignet in einem linearen System zusammengefasst. Aus diesem werden durch Anwendung einer zur Systemmatrix orthogonalen Matrix und einer Singulärwertzerlegung die Zustandsraummatrizen gewonnen. Eine genaue Beschreibung von 4SID in Verbindung mit FIT-Zeitbereichsrechnungen findet sich in [86]. In exakter Arithmetik wäre die benötigte Systemgröße durch die Singulärwertzerlegung festgelegt, bei endlicher Rechengenauigkeit erfolgt ein Abbruch ähnlich der *Balanced Truncation* über die Größe der Singulärwerte.

Dieses Verfahren kann folglich ebenfalls dazu dienen, Modelle reduzierter Ordnung aus einer FIT-Diskretisierung zu erzeugen, nicht direkt aus den Systemmatrizen wie im vorigen Kapitel, sondern über den Umweg einer Zeitbereichsrechnung. Dies kann insbesondere dann Vorteile bringen, wenn dispersive Materialien oder PML-Berandungen verwendet werden. In diesem Fall werden die FIT-Systeme, wie in Abschnitt 3.1.3 beschrieben, durch Linearisierung sehr groß und unsymmetrisch, was die Reduzierung durch Projektion erschwert. Gerade in diesen Fällen führt eine Zeitbereichsrechnung über den Leapfrog-Algorithmus häufig auf eine sehr schnelle Lösung, die dann zur Modellgenerierung durch 4SID verwendet werden kann. Bedauerlicherweise wird jedoch die Passivität von 4SID nicht generell erhalten, was einen deutlichen Nachteil des Verfahrens gegenüber Projektionsverfahren darstellt.

Kapitel 6

Generierung von Ersatzschaltbildern

In den vorigen Kapiteln wurden zahlreiche Verfahren beschrieben, die aus einer FIT-Diskretisierung ein Modell mit geringer Ordnung generieren. Soll das resultierende Modell als Makromodell verwendet werden, ist insbesondere darauf zu achten, dass die Passivität im Reduktionsprozess erhalten bleibt. Wird das Makromodell darüberhinaus mit einem klassischen Netzwerk verknüpft und soll eine gemeinsame Simulation durchgeführt werden, muss dieses in den Netzwerksimulator eingebracht werden. Während moderne Simulatoren oft einen direkten Import der abstrakten Zustandsraumdarstellung zulassen, ist es in anderen Fällen wünschenswert, ein Ersatzschaltbild aus den konzentrierten Elementen R, L, C sowie idealen Übertragern oder gesteuerten Quellen anzugeben.

In einer Einführung sollen zunächst die Gemeinsamkeiten und Unterschiede der Netzwerk- und Feldtheorie sowie der FIT-Diskretisierung beschrieben werden. Im Weiteren werden verschiedene Möglichkeiten angegeben, wie aus dem ordnungsreduzierten FIT-Modell ein Netzwerk generiert werden kann. Diese beruhen auf einer Interpretation des Zustandsraums als Knotenanalysemodell, nutzen topologische Eigenschaften des Modells oder basieren auf einer Pol-Residuen-Zerlegung.

6.1 Einführung

Die Feldtheorie basierend auf den Maxwellschen Gleichungen nach Abschnitt 2.1 bildet eine fundamentale und allgemeine Beschreibung aller elektromagnetischen Phänomene in Gebieten mit inhomogener Materialverteilung. Die unbekannten Feld- und Flussgrößen hängen in diesen partiellen Differentialgleichungen sowohl von den drei Raumrichtungen als auch von der Zeit ab.

Eine Beschreibung mit höherem Abstraktionsgrad liefert die Kirchhoffsche oder auch Netzwerktheorie, die auf der Definition von Spannungen, Strömen sowie *konzentrierten Elementen* beruht. Spannungen werden hierbei als Wegintegral über das elektrische Feld und Ströme als Flächenintegral der Stromdichte innerhalb elektrischer

Leiter aufgefasst. Die Ausdehnung der konzentrierten Elemente wie Widerstand, Spule oder Kondensator wird als gegen Null gehend angenommen und ihr Verhalten durch die konstitutiven Gleichungen wie das Ohmsche Gesetz beschrieben. Der räumlichen Verteilung bzw. Verschaltung der verschiedenen Elemente kommt folglich eine rein topologische Eigenschaft zu, alle auftretenden Parameter hängen allein von der Zeit ab.

Die Netzwerktheorie ist demnach eine Näherung der Maxwellschen Gleichungen, die Gültigkeit hat, wenn alle verwendeten Elemente tatsächlich klein im Verhältnis zur betrachteten Wellenlänge sind, diese durch gut leitende Drähte verbunden und Signallaufzeiten vernachlässigbar sind. Auch wenn diese Approximation zunächst sehr grob erscheint, ist sie für viele klassische elektrische Schaltungen mit ausreichender Genauigkeit erfüllt. Klarer Vorteil der Kirchhoffschen Theorie ist, dass hiermit weit komplexere Netzwerke berechnet werden können, als dies mit der Feldtheorie je möglich wäre.

Bedeutende Phänomene wie Pulslaufzeiten und -dispersion, Nebensprechen und Abstrahlung werden hierbei jedoch nicht modelliert. Sind einzelne Bereiche der Schaltung von solchen Effekten betroffen, was bei zunehmenden Betriebsfrequenzen immer wahrscheinlicher wird, ist eine mathematische Kopplung der beiden Theorien möglich [88]. In vielen Fällen wird es jedoch vorgezogen, diese Phänomene durch Ersatzschaltbilder zu modellieren und diese in das Netzwerk einzubeziehen.

In mathematischer Beschreibung führt die Kirchhoffsche Theorie auf ein System gewöhnlicher Differentialgleichungen mit endlicher Anzahl von Zustandsvariablen sowie einfacher algebraischer Gleichungen.

Durch die Diskretisierung der Maxwellschen Gleichungen durch die FI Methode wird ebenfalls eine endliche Anzahl von Zustandsvariablen definiert, die Spannungs- oder Stromcharakter haben. Die Materialbeziehungen entsprechen konstitutiven Relationen und die Matrizen $\mathbf{C}$ und $\widetilde{\mathbf{C}}$ haben rein topologische Eigenschaften. Formell bedeutet die Diskretisierung des Modells also die Approximation einer feldbehafteten Struktur durch ein konzentriertes Modell. Besonders deutlich wird dies, wenn das lineare FIT-System nach Gl. 3.1.1 mit der Systemstruktur verglichen wird, die sich bei der Netzwerksimulation nach der modifizierten Knotenanalyse [89] ergibt: Beide weisen exakt die gleiche Blockstruktur auf. Für den Sonderfall zweidimensional diskretisierter Elemente kann das FIT-Modell direkt als RLC-Schaltkreis interpretiert werden. Diese Eigenschaft wurde beispielsweise in [90] zur Analyse von zweidimensionalen Spannungsversorgungsnetzwerken genutzt. Im allgemeinen dreidimensionalen Fall ist die Angabe eines RLC-Modells jedoch nicht ohne Weiteres möglich, da sich die Topologie der Curl-Matrizen aufgrund der versetzten Gitter nicht auf ein klassisches Netzwerk übertragen lässt. Werden hingegen auch Gyratoren, die in der Realität nicht vorkommen, jedoch als abstraktes Hilfsmittel definiert und mit Hilfe von gesteuerten Quellen realisierbar sind, zugelassen, lässt sich auch in diesem Fall ein Netzwerk angeben [91].

Die direkte Interpretation der FIT-Matrizen als Ersatzschaltbild einer Struktur kommt allerdings üblicherweise nicht in Frage, da die Modellgröße und folglich die Anzahl der benötigten Elemente viel zu groß wäre. Ein vielversprechender Ansatz

ergibt sich aber, wenn das FIT-Modell zunächst durch eine der im vorigen Kapitel beschriebenen passivitätserhaltenden Techniken in der Ordnung reduziert und im Anschluss in ein Ersatzschaltbild transformiert wird. Da durch die Ordnungsreduktion die oben genannte Blockstruktur verloren geht, kann zur Realisierung anstelle der modifizierten Knotenanalyse auch die klassische Knotenanalyse herangezogen werden.

6.2 Interpretation als Knotenanalysemodell

Die klassische Knotenanalyse [92] bietet eine systematische Möglichkeit, alle in einem Netzwerk auftretenden Spannungen zu berechnen, solange der Schaltkreis weder unabhängige Spannungsquellen noch stromgesteuerte Elemente enthält. Zunächst wird ein Knoten als Bezugsknoten gewählt, wobei die Spannungen aller anderen relativ zum Bezugsknoten als Unbekannte betrachtet werden. Wird nun die Kirchhoffsche Knotenregel

$$\sum_{k=1}^{K} i_k = 0, \tag{6.2.1}$$

hier für einen Knoten mit K Zweigen angegeben, für jeden der q übrigen Knoten des Netzwerks abgesehen vom Bezugspunkt aufgestellt, ergibt sich ein lineares Gleichungssystem der Struktur

$$\mathbf{G}\mathbf{u}_z = \mathbf{i}_z. \tag{6.2.2}$$

Der q-dimensionale Unbekanntenvektor $\mathbf{u}_z$ enthält alle q unbekannten Knotenpotenziale. Der Anregungsvektor $\mathbf{i}_z$ enthält die mit dem Knoten q verbundenen unabhängigen Stromquellen. Der Wert ist positiv, wenn der Strom in den Knoten, negativ, falls er aus dem Knoten fließt. Die Matrix $\mathbf{G}$ ist quadratisch und repräsentiert die Leitwerte bzw. im allgemeinen Fall die Admittanzen $i_k = G(u_j - u_k)$ der Elemente des Netzwerks. Hierbei enthalten die Diagonaleinträge die Summe aller mit dem Knoten j verbundenen Admittanzen, es gilt $G_{jj} = \sum G_j$. Im symmetrischen Fall enthalten die übrigen Elemente $G_{jk} = G_{kj}$ den negativen Wert der Admittanz, die die Knoten j und k miteinander verbindet. Spannungsgesteuerte Stromquellen zwischen dem Knoten j und dem Bezugsknoten mit der Steuergröße u_k führen auf unsymmetrische reelle Einträge $G_{jk}(\neq G_{kj})$.

Eine einfache RLC-Ersatzanordnung für ein (reduziertes) FIT-System lässt sich damit finden, wenn dessen Zustandsraumdarstellung in die aus Gl. 6.2.2 abgeleitete Form überführt werden kann:

$$\left(\frac{1}{s}\mathbf{G}_L + \mathbf{G}_R + s\mathbf{G}_C + \mathbf{G}_{CS}\right)\mathbf{u}_z = \mathbf{B}_q\mathbf{i}. \tag{6.2.3}$$

Die Matrizen $\mathbf{G}_L$, $\mathbf{G}_R$ und $\mathbf{G}_C$ müssen reell symmetrisch sein und beschreiben das induktive, resistive und kapazitive Verhalten des Ersatznetzwerks. Sollen alle Netzwerkelemente positiv sein, muss die Matrix diagonaldominant sein, wobei alle Nebendiagonalelemente negativ sind. Eine Matrix mit dieser Eigenschaft wird auch als

L-Matrix bezeichnet. Da das Ersatzschaltbild aber ohnehin von der Struktur her keinen physikalischen Bezug aufweist und gängige Netzwerksimulatoren wie SPICE [93] problemlos auch negative Bauteile akzeptieren, ist die zweite Bedingung meist von geringer Bedeutung, zumal die Passivität des Modells durch den Reduktionsprozess gewährleistet ist. Die Matrix $\mathbf{G}_{CS}$ ist schließlich ebenfalls reell, aber nicht symmetrisch und enthält keine Einträge auf der Diagonalen. Sie beschreibt alle spannungsgesteuerten Stromquellen. Auch die Koppelmatrix $\mathbf{B}_q$ mit $\mathbf{i}_z = \mathbf{B}_q\mathbf{i}$ darf nur reelle Werte enthalten und hat die Dimension Anzahl Netzwerkknoten × Anzahl Ports. Der Stromfluss in die Ports, wird entsprechend der Zustandsraumdarstellung Gl. 3.1.2 durch den Vektor $\mathbf{i}$ repräsentiert.

Ein Netzwerk lässt sich nun wie folgt finden: Jeder Zeile in Gl. 6.2.3 entspricht ein Knoten j des Netzwerks. Nebendiagonalelemente der ersten drei Matrizen beschreiben je ein konzentriertes Bauteil, beispielsweise $-G_{L,jk} = -G_{L,kj}$ eine Spule zwischen den Knoten j und k mit dem Wert $G_{L,jk}$. Der Diagonaleintrag plus der Summe aller übrigen Elemente dieser Zeile, z. B. $G_{L,j} = G_{L,jj} + \sum_k G_{L,jk}$ entsprechen einem Bauteil zwischen Knoten j und Bezugspunkt. Die Matrix $\mathbf{G}_{CS}$ wird wie oben beschrieben durch gesteuerte Quellen realisiert. Jeder Eintrag von $B_{q\,jk}$ koppelt den Einfluss des Ports k in Knoten j und wird durch eine stromgesteuerte Stromquelle zwischen Knoten j und dem Bezugsknoten mit der Steuergröße i_k realisiert.

Bisher unbeachtet blieb die zweite Gleichung einer typischen Zustandsdarstellung: $\mathbf{u} = \mathbf{B}_q^T\mathbf{u}_z$. Deren Einträge koppeln das Potential jedes Knotens gewichtet mit $B_{q,jk}$ als Spannung in den Port. Für jeden Port muss somit ein weiterer Knoten definiert werden. Zwischen diesem Knoten und dem Bezugspunkt werden die Einträge von $\mathbf{B}_q$ durch in Reihe geschaltete spannungsgesteuerte Spannungsquellen mit den Wichtungsfaktoren $B_{q,jk}$ realisiert. Beispiele für diese Beschreibung folgen später anhand wichtiger Anwendungsfälle.

Zunächst stellt sich die Frage, wie das reduzierte Modell in Zustandsraumdarstellung in die Form 6.2.3 überführt werden kann. Da in grober Näherung jeder Matrixeintrag in einer der Matrizen einem Schaltungselement entspricht, ist zudem wünschenswert, dass die Matrixgleichung 6.2.3 so dünn wie möglich besetzt ist, um ein Ersatzschaltbild mit einer möglichst geringen Zahl von Bauteilen zu erhalten.

6.2.1 Lineare Systeme

Betrachtet man das Resultat einer Padé-Approximation bzw. des zweiten Schritts von TSL für den allgemeinen Fall verlustbehafteter linearer Systeme, ergibt sich für das reduzierte Modell nach Gl. 4.3.24

$$(\mathbf{I} + s\mathbf{T}_q - s_0\mathbf{T}_q)\,\mathbf{y} = \mathbf{B}_q\,\mathbf{i} \tag{6.2.4a}$$

$$\mathbf{u} = \mathbf{C}_q\,\mathbf{y}. \tag{6.2.4b}$$

Zwar sind alle Matrizen reell (ein reelles s_0 angenommen), aber die Matrix $\mathbf{T}_q$ ist nicht symmetrisch, $s\mathbf{T}_q$ demnach nach 6.2.3 nicht realisierbar.

Dieser Umstand kann behoben werden, wenn die Transformation 4.3.23 zunächst rückgängig gemacht wird

$$\left(s\mathbf{I} + \hat{\mathbf{T}}_q\right)\mathbf{y} = \hat{\mathbf{B}}_q\,\mathbf{i} \tag{6.2.5a}$$

$$\mathbf{u} = \mathbf{C}_q\,\mathbf{y} \tag{6.2.5b}$$

$$\text{mit} \quad \hat{\mathbf{T}}_q = (\mathbf{T}_q - s_0\mathbf{I})^{-1}, \quad \hat{\mathbf{B}}_q = \hat{\mathbf{T}}_q\mathbf{B}_q. \tag{6.2.5c}$$

Die erforderliche Matrixinversion kann bei der geringen Ordnung des reduzierten Modells problemlos direkt durchgeführt werden. Die Schaltung kann nun realisiert werden, $s\mathbf{I}$ entspricht Kondensatoren mit der Kapazität 1 Farad, die Diagonale der Matrix $\hat{T}_{q,kk}$ entspricht Widerständen zwischen den Knoten k und dem Bezugspunkt, die übrigen Einträge werden durch gesteuerte Quellen wie oben beschrieben modelliert. Um realistischere Werte für die Widerstände und Kapazitäten zu bekommen, ist es auch möglich, die Systemgleichung in 6.2.5a mit einem Faktor α zu multiplizieren. Die resultierende Schaltung ist in Abb. 6.1 dargestellt. Die Matrizen $\hat{\mathbf{T}}_q$ und $\hat{\mathbf{B}}_q$ sind voll besetzt, während $\mathbf{C}_q$ nur im oberen Dreieck Einträge hat, was auf ein resultierendes Ersatzmodell mit $q^2+mq+0,5\,(m^2+m)$ Elementen führt. Aus Gründen der Übersichtlichkeit ist von den $q-1$ spannungsgesteuerten Stromquellen, die aus $\hat{\mathbf{T}}_q$ resultieren, nur eine pro Knoten abgebildet.

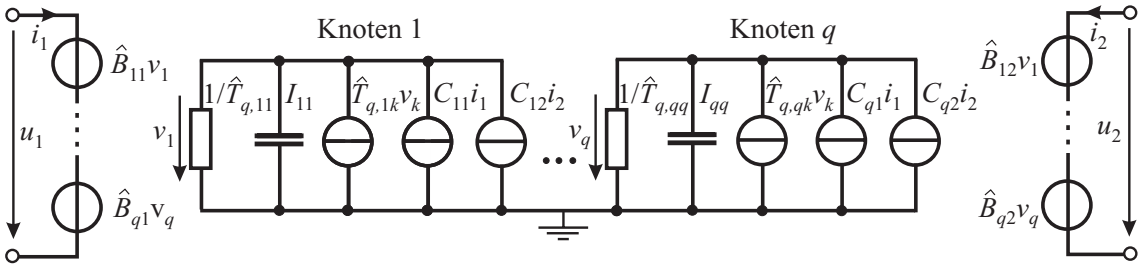


Abbildung 6.1: *Realisierung eines unsymmetrischen linearen Systems durch gesteuerte Quellen bei Interpretation als Knotenanalysemodell. Die Diagonalelemente von* $\hat{\mathbf{T}}_q$ *und* $\mathbf{I}$ *entsprechen Widerständen bzw. Kapazitäten.*

Eine deutliche Einsparung an Elementen ergibt sich, wenn das System zunächst in eine dünnbesetzte Form gebracht wird. Dies ist am einfachsten möglich, indem ein weiterer – im TSL-Fall der dritte – Lanczosschritt mit $\mathbf{T}_r = \mathbf{W}_r^T\hat{\mathbf{T}}_q\mathbf{V}_r$, $\mathbf{B}_r = \mathbf{W}_r^T\hat{\mathbf{B}}_q$, $\mathbf{C}_r = \mathbf{C}_q\mathbf{V}_r$ und $\mathbf{W}_r^T\mathbf{V}_r = \mathbf{I}$ durchgeführt wird:

$$(s\mathbf{I} + \mathbf{T}_r)\,\mathbf{y}' = \mathbf{B}_r\,\mathbf{i} \tag{6.2.6a}$$

$$\mathbf{u} = \mathbf{C}_r\,\mathbf{y}'. \tag{6.2.6b}$$

Die Matrix $\mathbf{T}_r$ hat nun erneut Bandstruktur und die Matrizen $\mathbf{B}_r$ und $\mathbf{C}_r$ nur Einträge im rechten oberen Dreieck.

6.2.2 Curl-Curl Systeme

Ersatzschaltbilder für reduzierte Curl-Curl-Systeme können entsprechend dem oben beschriebenen Vorgehen bei linearen Systemen generiert werden. Zunächst sollen

hierbei rein verlustfreie Strukturen betrachtet werden.

6.2.2.1 Verlustfreie Curl-Curl Systeme

Auch Curl-Curl-Systeme sollten zunächst in eine dünnbesetzte Form gebracht werden. Dies kann ebenfalls nach Gl. 6.2.6a erfolgen. Wurde das System durch eine einseitige Projektion nach Gl. 4.3.35 reduziert, ist auch das dünnbesetzte System unmittelbar symmetrisch, d.h., $\mathbf{T}_q$ ist symmetrisch und es gilt $\mathbf{C}_q^T = \mathbf{B}_q$. Das System lautet:

$$\left(s\mathbf{I} + \frac{1}{s}\mathbf{T}_r\right)\mathbf{y}' = \mathbf{B}_r\,\mathbf{i} \tag{6.2.7a}$$

$$\mathbf{u} = \mathbf{B}_r^T\,\mathbf{y}'. \tag{6.2.7b}$$

Wurde das System durch eine unsymmetrische Padé Approximation reduziert, ist es stets möglich, es ebenfalls in eine symmetrische Form zu transformieren. Diese Tatsache ergibt sich daraus, dass beide Arten der Reduktion für verlustfreie Curl-Curl-Systeme nach Abschnitt 4.3.3.3 mathematisch äquivalent sind.

Nach [94] kann die Symmetrisierung durch eine Diagonalmatrix $\mathbf{F}$ erfolgen. Die Diagonaleinträge F_{jj} der Matrix $\mathbf{F}$ lassen sich beispielsweise wie folgt finden:

$$a = B_{r,\,11}/C_{r,\,11} \tag{6.2.8a}$$

$$F_{jj} = a\,\frac{T_{r,\,(j-1)\,j}}{T_{r,\,j\,(j-1)}}. \tag{6.2.8b}$$

Mit $\hat{\mathbf{T}}_r = \mathbf{F}\mathbf{T}_r$ und $\hat{\mathbf{B}}_r = \mathbf{F}\mathbf{B}_r = \mathbf{C}_r^T$ ergibt sich schließlich das symmetrische System

$$\left(s\mathbf{F} + \frac{1}{s}\hat{\mathbf{T}}_r\right)\mathbf{y}' = \hat{\mathbf{B}}_r\,\mathbf{i} \tag{6.2.9a}$$

$$\mathbf{u} = \hat{\mathbf{B}}_r^T\,\mathbf{y}'. \tag{6.2.9b}$$

Aufgrund numerischer Ungenauigkeiten ist die Symmetrie der Systemmatrix typischerweise nur näherungsweise erfüllt, meist auf sechs bis acht Nachkommastellen genau. Die exakte Symmetrie kann durch $\hat{\mathbf{T}}_r' = \frac{1}{2}(\hat{\mathbf{T}}_r + \hat{\mathbf{T}}_r^T)$ gewährleistet werden.

Das System kann nun als Netzwerk erneut allein durch Kapazitäten, Induktivitäten sowie gesteuerte Quellen zur Ein- und Auskoppelung der Ports aufgebaut werden. Aufgrund der Symmetrie der Matrix $\hat{\mathbf{T}}_r$ kann diese ohne weitere gesteuerte Quellen realisiert werden. Das Netzwerk ist in Abb. 6.2 für $m = 2$ und $q = 6$ dargestellt. Es enthält $(m+2)q + 0,5\,(m^2 + m)$ Elemente. Die Werte der Induktivitäten und Kapazitäten lassen sich, wie oben beschrieben, einfach aus $\hat{\mathbf{T}}_r$ und $\mathbf{F}$ bestimmen. Es bleibt zu beachten, dass einige der Induktivitäten negative Werte annehmen können, das Modell folglich nur bedingt physikalisch ist, es zur Kopplung mit anderen Netzwerken aber uneingeschränkt verwendet werden kann.

Betrachtet man das Ersatzschaltbild eines idealen Übertragers in Abb. 6.3 mit dem Übertragungsverhalten

$$\frac{u_1}{u_2} = \ddot{u} = \frac{i_2}{i_1}, \tag{6.2.10}$$

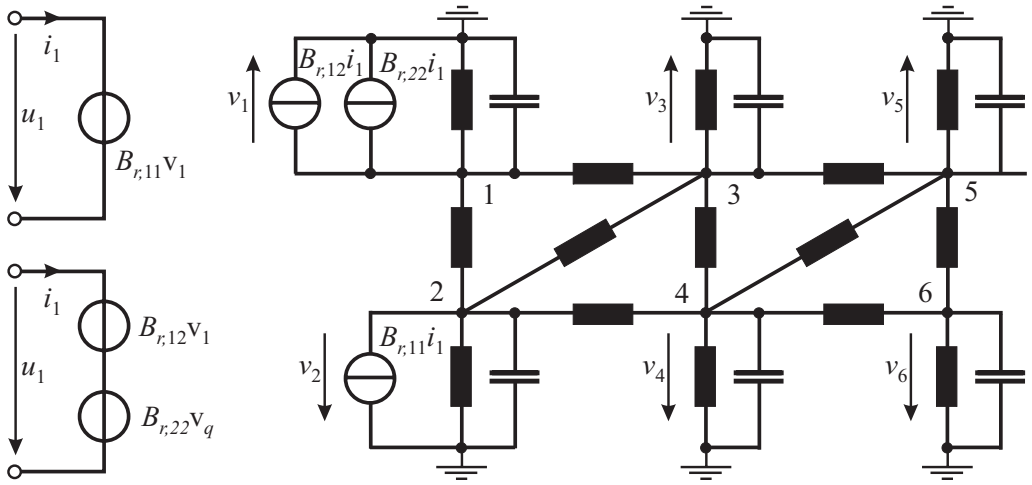


Abbildung 6.2: *Realisierung des symmetrisierten Curl-Curl-Systems durch Induktivitäten und Kapazitäten. Gesteuerte Quellen dienen zur Ein- und Auskoppelung der Portsignale.*

zeigt sich, dass jedes Paar einer spannungsgesteuerten Spannungsquelle und einer stromgesteuerten Stromquelle auch als idealer Übertrager interpretiert werden kann. Für die Verwendung innerhalb eines Netzwerksimulators bieten gesteuerte Quellen jedoch numerische Vorteile und werden im Ersatzschaltbild daher bevorzugt.

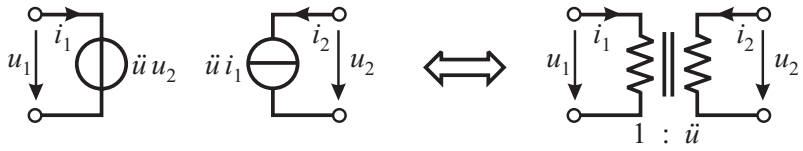


Abbildung 6.3: *Ersatzschaltbild eines idealen Übertragers.*

In [94] wird ein Verfahren vorgestellt, das die Matrix $\hat{\mathbf{B}}_r$ auf eine rein topologische Matrix mit ausschließlich 1 und -1 als Einträgen reduziert. Auf diese Weise wird kein separates Netzwerk für die Ports benötigt. Zudem werden Alternativen vorgeschlagen, wie negative Elemente umgangen werden können.

Da durch Projektion reduzierte Curl-Curl-Systeme wie ebenfalls in Abschnitt 4.3.3.3 gezeigt stets Stabilität und Passivität erhalten, die Matrix $\hat{\mathbf{T}}_r$ folglich positiv semidefinit ist, kann ein dünnbesetztes symmetrisches System auch mit Hilfe einer Eigenwertzerlegung einfach erzeugt werden. Mit $\mathbf{\Lambda} = \mathbf{X}^{-1}\mathbf{T}_r\mathbf{X}$, $\mathbf{X}^T\mathbf{X} = \mathbf{I}$ und $\mathbf{B}_e = \mathbf{X}^T\mathbf{B}_r$ ergibt sich aus 6.2.7a das symmetrische System:

$$\left(s\mathbf{I} + \frac{1}{s}\mathbf{\Lambda}\right)\mathbf{y}'' = \mathbf{B}_e\,\mathbf{i} \tag{6.2.11a}$$

$$\mathbf{u} = \mathbf{B}_e^T\,\mathbf{y}''. \tag{6.2.11b}$$

Das System besteht somit aus zwei diagonalen Matrizen $\mathbf{I}$ und $\mathbf{\Lambda}$ mit nur positiven (bzw. Null-) Einträgen sowie der Ein- und Auskoppelmatrix $\mathbf{B}_e$. Die Realisierung bei Interpretation als Knotenanalysemodell führt auf je eine Kapazität und

eine Induktivität in Form eines Parallelschwingkreises zwischen jedem Knoten und dem Bezugsknoten (siehe Abb. 6.4). Die Werte können erneut durch einen Faktor a skaliert werden. Zur Ein- und Auskoppelung der Ports werden wieder gesteuerte Quellen verwendet. Ersetzt man die gesteuerten Quellen nach Abb. 6.3 durch ideale Übertrager, zeigt sich durch die positiven Werte aller Elemente, dass die Schaltung tatsächlich physikalisch und passiv ist. Ein auf diese Weise erzeugtes Ersatzschaltbild für verlustfreie Systeme, wie beispielsweise Resonatoren, ist folglich mit rein physikalischen Elementen realisierbar und mehr als eine rein mathematische Ersatzanordnung. Die Anzahl der Elemente ist mit $(2m + 2)q$ allerdings meist höher als die $(m+2)q+0,5\,(m^2+m)$ die sich durch die Transformation des Systems nach Gl. 6.2.9a ergeben.

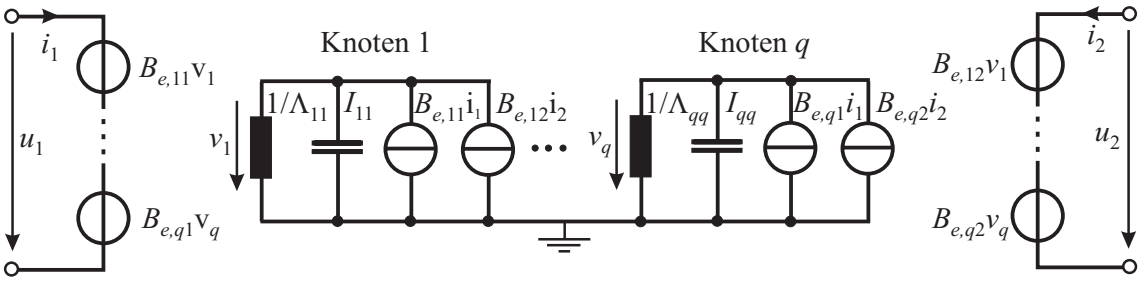


Abbildung 6.4: *Ersatzschaltbild einer verlustlosen Struktur mit $m = 2$ Ports, die in Curl-Curl-Formulierung reduziert wurde.*

6.2.2.2 Verlustbehaftete Curl-Curl Systeme

Beide für verlustfreie Curl-Curl-Systeme vorgestellten Varianten lassen sich auch auf verlustbehaftete Systeme erweitern. Zunächst sollen rein ohmsche Verluste durch geringe Materialleitfähigkeiten betrachtet werden. In diesem Fall wird Gl. 6.2.9a um eine weitere vollbesetzte Matrix $\mathbf{K}_{r\mathbf{V}} = \mathbf{V}_r^T \mathbf{K}_q \mathbf{V}_r$ erweitert. Entsprechend gilt bei Eigenwertzerlegung nach Gl. 6.2.11a $\mathbf{K}_{r\mathbf{X}} = \mathbf{X}_r^T \mathbf{K}_q \mathbf{X}_r$. Das System ergibt sich damit für den zweiten Fall

$$\left(s\mathbf{I} + \mathbf{K}_{r\mathbf{X}} + \frac{1}{s}\mathbf{\Lambda}\right)\mathbf{y}'' = \mathbf{B}_e\,\mathbf{i} \tag{6.2.12a}$$

$$\mathbf{u} = \mathbf{B}_e^T\,\mathbf{y}''. \tag{6.2.12b}$$

wobei der erstere Fall aus Gl. 6.2.9a völlig analog formuliert werden kann.

Die Einträge der Matrizen $\mathbf{K}_{r\mathbf{V}}$ bzw. $\mathbf{K}_{r\mathbf{X}}$ entsprechen der Widerstandsmatrix in Gl. 6.2.3 und können wie bereits beschrieben als Leitwerte G realisiert werden. Die Anzahl der Elemente erhöht sich damit auf $(2m + 2,5)q + 0,5q^2$. Zudem geht der direkte physikalische Bezug erneut verloren, da einzelne Widerstände negative Werte annehmen können. Die Passivität bleibt davon, wie bereits mehrfach betont, jedoch unbeeinträchtigt. Für den Fall der Eigenwertzerlegung ist das resultierende Ersatzschalbild in Abb. 6.5 dargestellt.

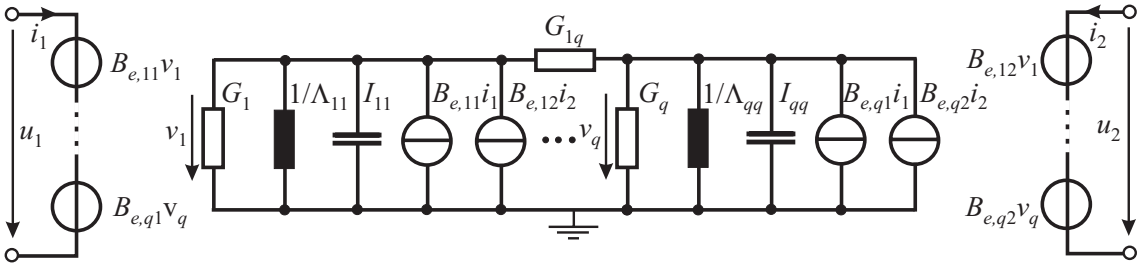


Abbildung 6.5: *Ersatzschaltbild einer verlustbehafteten Struktur mit $m = 2$ Ports, die in Curl-Curl-Formulierung reduziert wurde.*

Aufgrund der in Abschnitt 4.3.1.3 bereits festgestellten Dominanz der Diagonalwerte in den Matrizen $\mathbf{K}_{r\mathbf{V}}$ bzw. $\mathbf{K}_{r\mathbf{X}}$ können die Nebendiagonalelemente häufig auch vernachlässigt werden, ohne dass ein nennenswerter Fehler entsteht. Dies reduziert die Anzahl der benötigten Elemente auf $(2m+3)q$, zudem sind in den meisten Fällen alle Elemente erneut rein positiv.

Werden auch Verluste aufgrund des Impedanzwandmodells berücksichtigt, entspricht das Vorgehen von der Grundidee dem der ohmschen Verluste. Allerdings lässt sich die $-1/(s\sqrt{s})$-Abhängigkeit nicht direkt einer Bauteilgattung zuordnen. Die Funktion muss vielmehr zunächst durch eine rationale Funktion approximiert werden. Erster Gedanke ist sicherlich, auch in diesem Fall eine Padé-Approximation zu nutzen. Es zeigt sich aber, dass die Padé-Approximationen von $-1/(s\sqrt{s})$ in nahezu allen Fällen auf instabile Modelle führen. Um dies zu umgehen, kann auf nichtlineare Optimierung wie in [95] zurückgegriffen werden. Da die Verluste aber üblicherweise sehr klein sind, ist es meist ausreichend, auf noch einfachere Ersatzanordnungen zurückzugreifen. Betrachtet man Real- und Imaginärteil von $-1/(j\omega\sqrt{j\omega})$, zeigt sich, dass beide identisch sind, für die Frequenz Null bei plus Unendlich beginnen und für $\omega \to \infty$ gegen Null tendieren. Dieses Verhalten kann am ehesten durch eine negative Induktivität, die den Imaginärteil nachbildet, sowie einen Widerstand, der über den Realteil gemittelt ist, approximiert werden. Auch wenn die negativen Induktivitäten sehr klein sind, ist die Passivität hierdurch jedoch nicht mehr gewährleistet. Soll diese sicher gestellt werden, kann die negative Induktivität durch eine positive Kapazität ersetzt werden, allerdings mit größerem Fehler. Häufig erweist es sich sogar als ausreichend, nur den Realteil zu beachten und durch einen Widerstand anzunähern.

6.3 Pol-Residuen-Darstellung

Ersatzschaltbilder aus der Interpretation des Zustandsraums als Knotenanalysemodell führen auf allgemeine Weise zu verhältnismäßig kleinen, aber sehr abstrakten Netzwerken. Ein alternativer Ansatz besteht darin, die Impedanzmatrix zunächst in

eine Pol-Residuen-Darstellung zu bringen. Für jedes Element Z_{jl} gilt damit:

$$Z_{jl} = \sum_{k=1}^{q} \frac{r_k}{s - s_k}. \tag{6.3.1}$$

Hierbei treten zwei Typen von Polstellen auf, rein reelle und zueinander konjugiert komplexe Polpaare. Beide Typen lassen sich durch ein einfaches Ersatzschaltbild nach Abb. 6.6 realisieren.

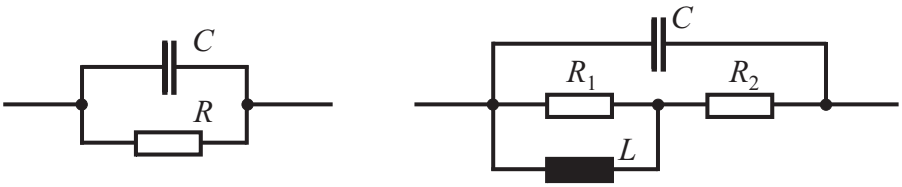


Abbildung 6.6: *Realisierung eines einzelnen reellen (links) und zweier konjugiert komplexer (rechts) Pol-Residuen-Terme als Impedanz.*

Im Fall einer reellen Polstelle lassen sich die Werte für R und C direkt zuordnen:

$$C = \frac{1}{r_k}, \qquad R = -\frac{r_k}{s_k}. \tag{6.3.2}$$

Im Fall konjugiert komplexer Polpaare führt die Zuordnung auf etwas komplizierte Ausdrücke

$$C = \frac{1}{2\mathrm{Re}\{r_k\}}, \qquad R_2 = \frac{2\mathrm{Re}\{s_k^* r_k\}}{|s_k|^2}, \tag{6.3.3a}$$

$$R_1 = \frac{1 + C^2 R_2^2 |s_k|^2 + 2\mathrm{Re}\{s_k\} C R_2}{-2\mathrm{Re}\{s_k\} C - C^2 R_2^2 |s_k|^2}, \qquad L = \frac{R_1}{C|s_k|^2 (R_1 + R_2)}. \tag{6.3.3b}$$

Im Einportfall kann die Impedanz Z_{11} einfach durch Reihenschaltung der einzelnen Teilimpedanzen verwirklicht werden. Im Mehrportfall ist wieder eine Verkoppelung der Einzelimpedanzen Z_{jl} über ideale Übertrager bzw. gesteuerte Quellen erforderlich. Eine sehr allgemeine und zugleich einfache Variante wird in [98] beschrieben. Diese benötigt allerdings $m^2(2q + 1)$ Schaltungselemente. Insbesondere die quadratische Abhängigkeit von m kann hier bei einer großen Portzahl zu sehr großen Ersatzschaltbildern führen. Eine alternative Variante, die mit weniger Elementen auskommt, wird in [96] beschrieben.

Kapitel 7

Anwendungsbeispiele

Die im bisherigen Verlauf dieser Arbeit beschriebenen Methoden zur Modellreduzierung sowie zur schnellen Berechnung des Frequenzverhaltens sollen an einer Reihe von typischen Anwendungsbeispielen auf ihre Eigenschaften und insbesondere ihre Praxistauglichkeit getestet werden. Bei den ersten beiden Beispielen handelt es sich um resonante Filterstrukturen, einmal bei mittlerer Systemgröße und komplizierter Polstruktur, zum anderen ein Wellenleiterfilter mit einer großen Zahl von Unbekannten und annähernd Tschebyscheff-verteilten Polstellen. Das dritte Beispiel einer Patchantenne stellt einen typischen Fall eines Systems mit offenen Rändern dar, während das letzte Beispiel eine Interconnect-Struktur einer integrierten Schaltung mit einer großen Anzahl von 50 diskreten Ports beschreibt.

7.1 Langer-Filter

Als erstes Beispiel zur Untersuchung und zum Vergleich der im Rahmen dieser Arbeit entwickelten und vorgestellten Verfahren dient ein dielektrisches Filter, das so genannte *Langer-Filter*. Das Filter wird bereits seit vielen Jahren zum Test von Simulationsverfahren und -programmen zur Berechnung elektromagnetischer Probleme verwendet, im Zusammenhang mit FIT beispielsweise in [13, 27] und hat somit quasi den Status eines Referenzbeispiels bekommen.

Die Filterstruktur ist symmetrisch aufgebaut und wird durch zwei koaxiale Ports angeregt. Das Resonanzverhalten wird durch zwei dielektrische Ringe gesteuert, die mit $\varepsilon_r = 38$ eine hohe Permittivität aufweisen. Der Aufbau des Filters ist in Abb. 7.1 gezeigt, eine genaue Bemaßung findet sich in [13]. Das Filter dient als scharfes Bandpassfilter bei ca. 4,6 GHz. Aus numerischer Sichtweise ist es aber insbesondere auch interessant, die Resonanzen oberhalb von 6,5 GHz mit in die Betrachtungen einzubeziehen.

Aufgrund der Struktursymmetrie ist es ausreichend, nur die Hälfte des Filters zu diskretisieren, wobei eine magnetische Randbedingung in der Symmetrieebene benutzt wird. Wird die kleinste auftretende Wellenlänge mit zehn Gitterlinien pro Wellenlänge abgetastet, führt dies auf 14.265 Gitterzellen.

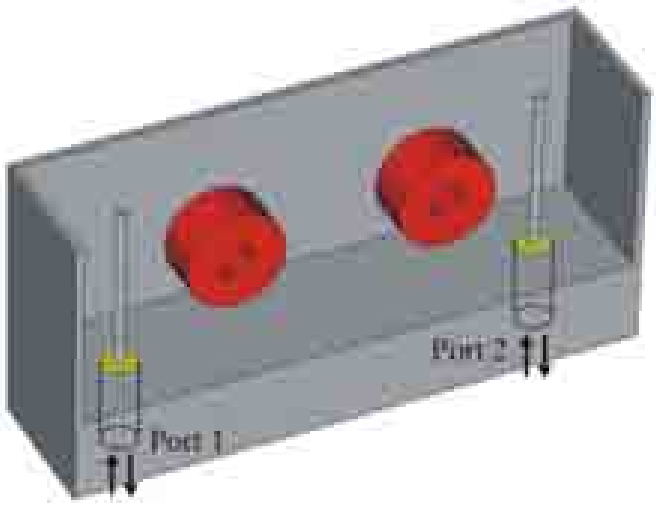


Abbildung 7.1: *Aufbau des analysierten dielektrischen Filters. Ein Teil des Gehäuses wurde weggelassen, damit die innere Struktur sichtbar wird.*

Mit rund 45.000 Unbekannten im Curl-Curl-Fall oder entsprechend rund 90.000 im linearen Fall ist das Filter von der Systemgröße eher im mittleren Bereich angesiedelt, mit 8 Eigenfrequenzen im interessierenden Frequenzbereich von 4–8 GHz zählt es jedoch vom Übertragungsverhalten zu den anspruchsvolleren resonanten Beispielen, zumal die Eigenwerte zum Teil hohe Güten aufweisen und eng beieinander liegen.

Der Transmissonsfaktor $|S_{21}|$ als Ergebnis einer Zeitbereichsrechnung ist in Abb. 7.2 gezeigt. Die Berechnung dieser als Referenz betrachteten Kurve erfolgte mit dem Programm CST MICROWAVE STUDIO® (MWS) [30]. Alle übrigen Verfahren nutzen ebenfalls MWS zur Modellierung und zur Diskretisierung der Struktur. Die Materialmatrizen werden aus MWS exportiert und für die eigenen Programme als Grundlage genutzt. Alle Vergleiche erfolgen jeweils mit dem exakt gleichen räumlichen Modell.

Als weitere Ergebnisse werden in Abb. 7.2 die Lösung einer partiellen Realisierung mit $p = 600$ Iterationen, einer TSL-Rechnung in Curl-Curl-Formulierung mit einem Abbruchkriterium von 10^{-6} in beiden Schritten, sowie eine SPICE-Frequenzbereichsrechnung des aus dem TSL-System gewonnenen Ersatzschaltbildes dargestellt. Die TSL- und die SPICE-Kurve liegen hierbei ununterscheidbar auf der Referenzkurve.

Die Implementierung des ersten Schritts von TSL verwendet hierbei nicht direkt die Curl-Curl-Matrix, sondern führt die Operatoren $\mathbf{C}$ und $\mathbf{C}^T$ sukzessive aus. Das Abbruchkriterium des TSL-Durchlaufs wird im ersten Schritt erstmals nach 600 Iterationen aufgerufen und in der Folge alle 40 Schritte wiederholt. Der automatische Stopp erfolgt schließlich nach 1000 Iterationen. Im zweiten Schritt führt das Abbruchkriterium auf 22 Iterationen. Die Gesamtrechenzeit beträgt auf einem 731 MHz PC nur 42 Sekunden, was weniger als 5 % der vergleichbaren Rechenzeit im Zeitbereich in Verbindung mir einem AR-Filter ist. Die zusätzliche Rechenzeit, die das Abbruchkriterium im ersten Schritt benötigt, macht etwa 7 % der Gesamtzeit aus.

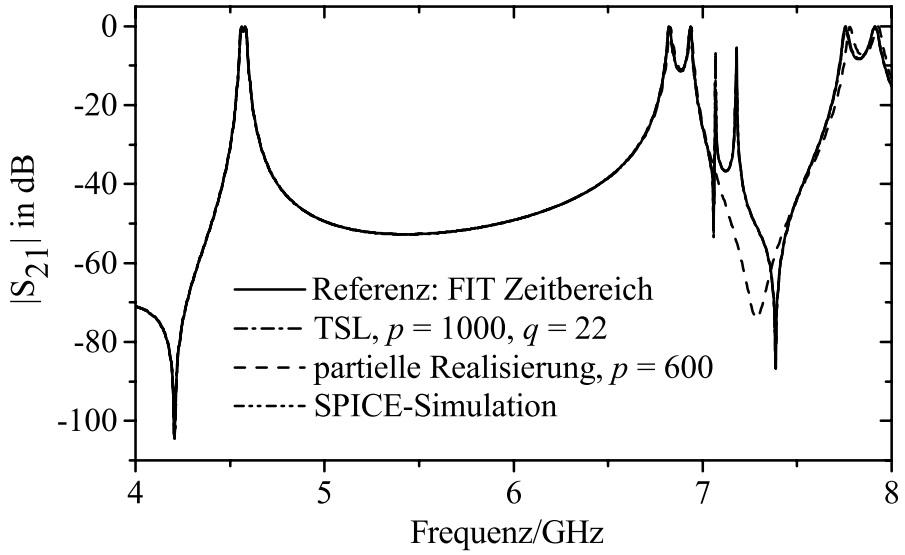


Abbildung 7.2: *Die Frequenzabhängigkeit des S_{21}-Parameters. Die Kurven der Referenz, das TSL-Ergebnis und die SPICE-Simulation liegen ununterscheidbar übereinander. Die partielle Realisierung mit p=600 Iterationen wurde vor Erfüllung des Abbruchkriteriums abgebrochen.*

TSL stellt damit insbesondere im Vergleich zu den direkten Padé-Approximationen wie beispielsweise PVL einen deutlichen Rechenzeitgewinn bei gleicher Modellgüte dar. So benötigt PVL im günstigsten Fall 613 s. Da eine LU-Zerlegung aufgrund der Systemgröße nicht möglich ist, werden innerhalb von PVL hierbei iterative Löser eingesetzt: Im linearen Fall wird das *BiConjugate-Gradient*-Verfahren (BiCG) [43], im symmetrischen Curl-Curl-Fall das *Conjugate-Gradient*-Verfahren (CoCG) verwendet. Die geforderte Genauigkeit des Solverresiduums wird dabei auf 10^{-6} gesetzt, ein Wert der ausreichend ist, um Folgefehler in der weiteren Iteration zu vermeiden. Ein Rechenzeitvergleich diverser Verfahren wird tabellarisch am Ende dieses Abschnitts gegeben.

Wird die partielle Realisierung bzw. der erste TSL-Schritt bereits vorzeitig nach 600 Iterationen abgebrochen, zeigt sich, dass die stark separierten Polstellen bereits gut approximiert sind, während die dicht beieinander liegenden zwischen 7 und 8 GHz noch gar nicht oder nur ungenau gefunden sind. Wird das System mit klassischen iterativen Lösern im Frequenzbereich berechnet, stellt sich heraus, dass bis zu etwa 7 GHZ um 600 Solver-Iterationen ausreichend sind, während über 7 GHz um die 1000 benötigt werden. Dies betont erneut den engen Zusammenhang zwischen partiellen Realisierungen und klassischer Gleichungslösung.

Zur Untersuchung der Konvergenz von TSL sind in Abb. 7.3 die Abweichungen verschiedener TSL-Modelle zu unterschiedlichen Referenzlösungen dargestellt. Wird ein sehr großes TSL-System mit 2000 Iterationen im ersten und 50 Iterationen im zweiten Schritt betrachtet und mit einer Frequenzbereichslösung des vollen Systems verglichen, zeigt sich, dass der Fehler in weiten Bereichen unter 10^{-9} liegt, was der Genauigkeit des iterativen Lösers bei der Berechnung der Referenzlösung entspricht.

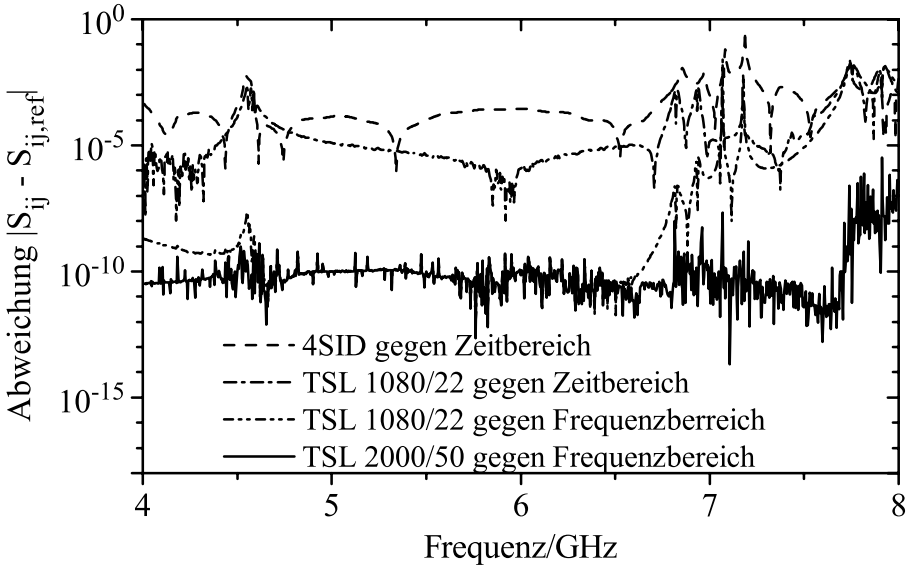


Abbildung 7.3: *Abweichung des über TSL berechneten Transmissionsfaktors S_{21} relativ zu Frequenz- und Zeitbereichslösungen des vollen Systems. Als Vergleich ist der Fehler eines nach dem 4SID-Verfahren aus Zeitbereichsdaten generierten Modells dargestellt.*

Lediglich am rechten Rand des betrachteten Spektrums steigt der Fehler leicht auf 10^{-6} an. Wird ein kleineres TSL-Modell mit 1080 und 22 Iterationen im ersten und zweiten Schritt gewählt, bleibt der Fehler bis zu 6,5 GHz ebenfalls im Bereich von 10^{-9}, steigt oberhalb diese Frequenz aber auf Werte von bis zu 10^{-2} an. Dies entspricht der Erfahrung, dass der Lanczos-Algorithmus die äußeren, d. h. kleinen Eigenwerte zuerst genau findet, während die oberen noch weniger gut approximiert sind. In der Praxis erweist sich aber auch diese Abweichung als tolerierbar, mit dem Vorteil der kürzeren Rechenzeit und der Tatsache, dass die endgültige Modellgröße nur von Ordnung 22 anstelle von 50 ist.

Vergleicht man dass Modell TSL 1080/22 mit einer Zeitbereichsrechnung aus MWS, liegt die minimale Abweichung bei 10^{-6}. Diese Abweichung ist kein Fehler von TSL, sondern ergibt sich durch den Abschneidefehler sowie die Genauigkeit der Portberechnung in MWS. Auch die größere Abweichung an den Polen erklärt sich durch die Restenergie in den Resonanzen zum Zeitpunkt des Abbruchs der Zeitbereichsrechnung. Der Zeitdispersionsfehler [13] der Referenzlösung wurde an dieser Stelle herausgerechnet. Es zeigt sich dabei, dass die Abweichung der TSL-Kurve zur Zeitbereichsreferenz im Allgemeinen kleiner als der allgemein tolerierte Fehler durch die Zeitdispersion ist. Die letzte Kurve vergleicht die TSL-Systeme mit einem nach dem 4SID-Verfahren aus Zeitbereichsdaten erstellen Modell. Dies wird später in diesem Abschnitt vorgestellt.

Der Vergleich des TSL-Modells mit dem FIT-System kann auch im Zeitbereich erfolgen. Hierzu wird ein trapezförmiger digitaler Puls betrachtet, der eine Anstiegszeit von 0,1 ns aufweist, den Wert Eins für 0,3 ns hält und innerhalb von ebenfalls 0,1 ns wieder auf Null absinkt (siehe Abb. 7.4). Als Referenz dient erneut eine MWS-

Rechnung. Die Vergleichsrechnung des TSL-Modells erfolgt mit dem generierten Ersatzschaltbild in SPICE. Dieses wurde als Netzwerk mit 22 LC-Schwingkreisen und Signaleinkopplung durch ideale Übertrager realisiert. Verwendet man gesteuerte Quellen anstelle der Übertrager in SPICE, führt dies auf ein Ersatzschaltbild mit 132 linearen Elementen. Da SPICE Strom- und Spannungsgrößen, MWS jedoch Wellenamplituden verwendet, werden die Wellengrößen a und b der SPICE-Rechnung aus den berechneten Größen u und i nach Gl. 2.2.38 rückgerechnet, so dass auch a(t) der SPICE-Rechnung nicht mit der Originalanregung übereinstimmt. Die Ergebnisse beider Verfahren zeigen nach Abb. 7.4 große Übereinstimmung. Die Rechenzeit in SPICE beträgt 5,6 s im Vergleich zu 23 s in MWS. Allerdings müssen hierbei zur SPICE-Zeit noch die 42 s für die TSL-Modellgenerierung gerechnet werden.

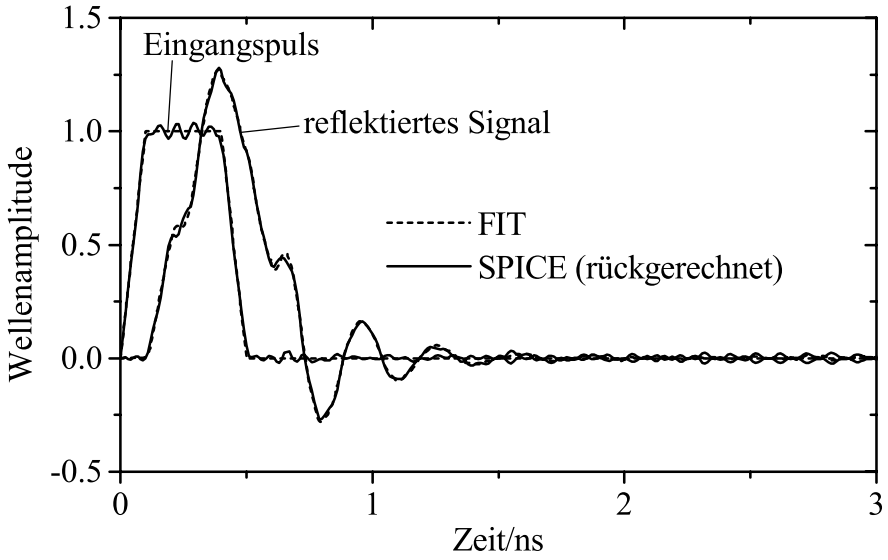


Abbildung 7.4: *Simulation im Zeitbereich: Der einfallende Digitalpuls sowie das reflektierte Signal, wobei die FIT-Kurven gestrichelt und das Ergebnis der SPICE-Simulation durchgezogen dargestellt ist. Die hochfrequenten Wellen, die das Signal überlagern stammen von einem Pol des Ersatzschaltbilds, der außerhalb des betrachteten Frequenzbereichs liegt. Die Abweichung im Anregungspuls erklärt sich durch die Berechnung der Wellengrößen aus Strom und Spannung.*

Die hochfrequenten Anteile der SPICE-Signale stammen von einer Polstelle, die außerhalb des interessierenden Frequenzbereichs liegt und vom Abbruchkriterium daher nicht berücksichtigt wurde. Wird der entsprechende Pol des reduzierten Systems entfernt, entfällt die hochfrequente Überlagerung, dafür ist die Übereinstimmung im Frequenzbereich weniger gut als in Abb. 7.2 dargestellt.

Wird zunächst ein Tschebyscheff-Polynom auf die Startvektoren des Originalsystems angewendet, lässt sich die Größe des Unterraums im ersten TSL-Schritt deutlich reduzieren. Im hier betrachteten Fall wird ein Beschleunigungspolynom mit der Ordnung 450 verwendet, was die Zahl der benötigten Iterationen im ersten Schritt bei Anwendung des Abbruchkriteriums auf nur 330 begrenzt. Die Größe der Matrix $\mathbf{V}_p$ hat damit bei doppelter Genauigkeit eine Größe von etwa 110 MB und kann

in modernen PCs bequem im Speicher gehalten werden. Die Anzahl der Matrix-Vektor-Multiplikationen wird durch Verwendung des Polynoms zwar von 1000 im Standard-TSL auf 1230 erhöht, die Rechenzeit ist jedoch dennoch leicht reduziert, da entsprechend weniger Orthogonalisierungsschritte und Aufrufe des Abbruchkriteriums erforderlich sind. Wie in Abb. 7.5 gezeigt, weicht die Kurve der beschleunigten partiellen Realisierung jedoch deutlich von der Referenz ab, was einen Korrekturschritt notwendig macht, der weitere 28 s in Anspruch nimmt. Die beschleunigte und korrigierte partielle Realisierung zeigt wieder eine sehr gute Übereinstimmung mit der Referenzkurve. Die Gesamtrechenzeit beträgt in diesem Fall 65 s, was gegenüber dem einfachen TSL etwas erhöht, im Vergleich zu bestehenden Verfahren aber noch immer etwa eine Größenordnung schneller ist.

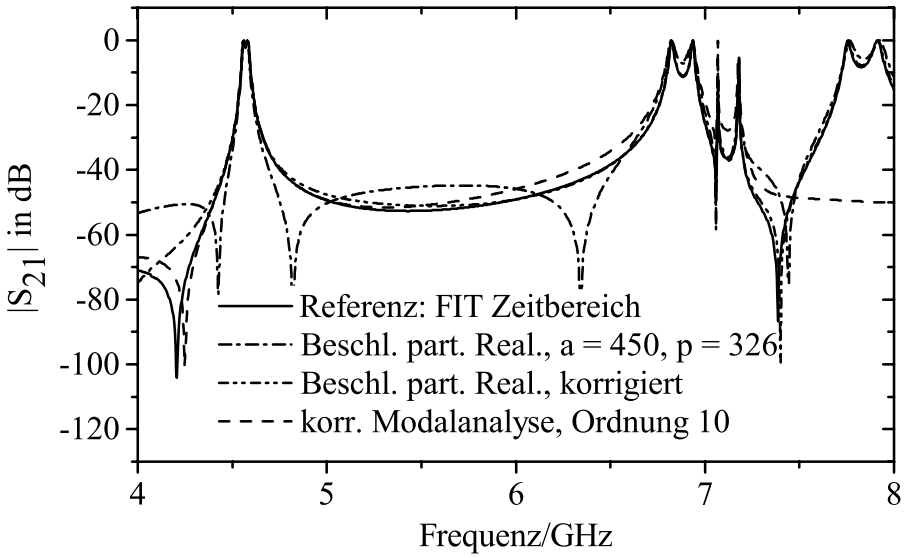


Abbildung 7.5: *Der S_{21}-Parameter für einen beschleunigten ersten Schritt. Die unkorrigierte Variante zeigt einen deutlichen Fehler gegenüber der Referenzkurve, während die korrigierte Kurve eine gute Übereinstimmung aufweist. Eine korrigierte Modalanalyse mit 10 Moden zeigt ebenfalls bis 7 GHz gute Approximationseigenschaften.*

Als Vergleich wird in Abb. 7.5 auch eine korrigierte Modalanalyse mit nur 10 Moden dargestellt. Die Kurve stimmt bis 7 GHz ebenfalls gut mit der Zeitbereichskurve überein, die darüber liegenden Moden wurden nicht berücksichtigt. Die Rechenzeit beträgt 181 s, sie ist jedoch stark abhängig vom verwendeten Eigenwertlöser.

Für alle bisherigen Rechnungen wurde das Filter als verlustlos angenommen. Zum Test des verlustbehafteten TSL-Algorithmus (TSLlossy) wird nun angenommen, dass die dielektrischen Ringe eine Leitfähigkeit von 0,2 S/m aufweisen, was einem Verlustwinkel von $\tan\delta \approx 0,016$ entspricht. Dieser Wert stellt eine bewusste Übertreibung dar, um die Eigenschaften des Verfahrens zu testen, realistische Materialverluste in Dielektrika sind deutlich geringer. Die Berechnung erfolgt, indem die Verlustmatrix $\mathbf{K}$ auf die gespeicherten Unterräume des verlustfreien TSL projiziert wird. Um eine Überlagerung von Fehlern an dieser Stelle zu vermeiden, wird auf

die Anwendung eines Beschleunigungspolynoms verzichtet, zumal die 1000 Vektoren von $\mathbf{V}_p$ im ersten Schritt noch im Speicher gehalten werden konnten. Die Rechenzeit vergrößert sich durch den Projektionsschritt leicht auf 48 s.

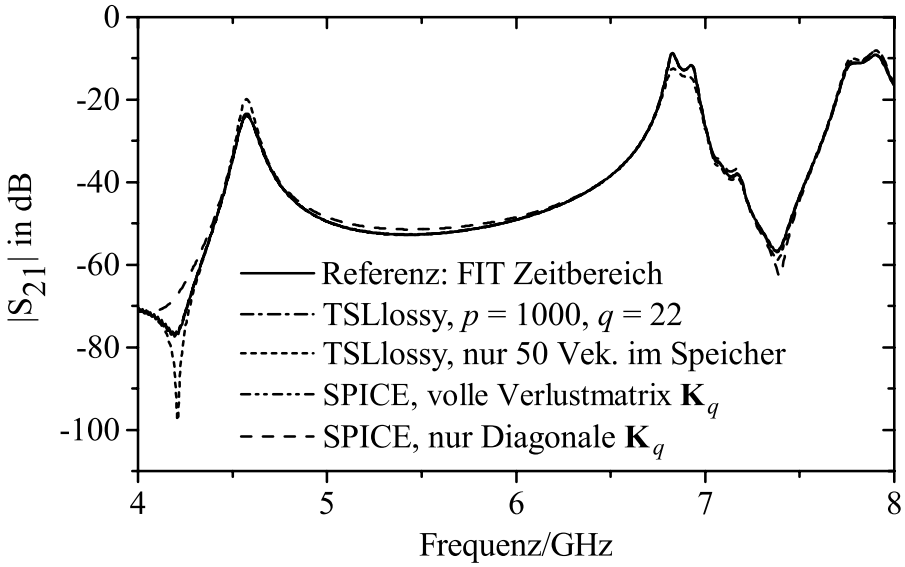


Abbildung 7.6: *Die S_{21}-Übertragungsfunktion im verlustbehafteten Fall. Die Kurven der Referenzlösung, TSLlossy und die der SPICE-Simulation unterscheiden sich nahezu nicht.*

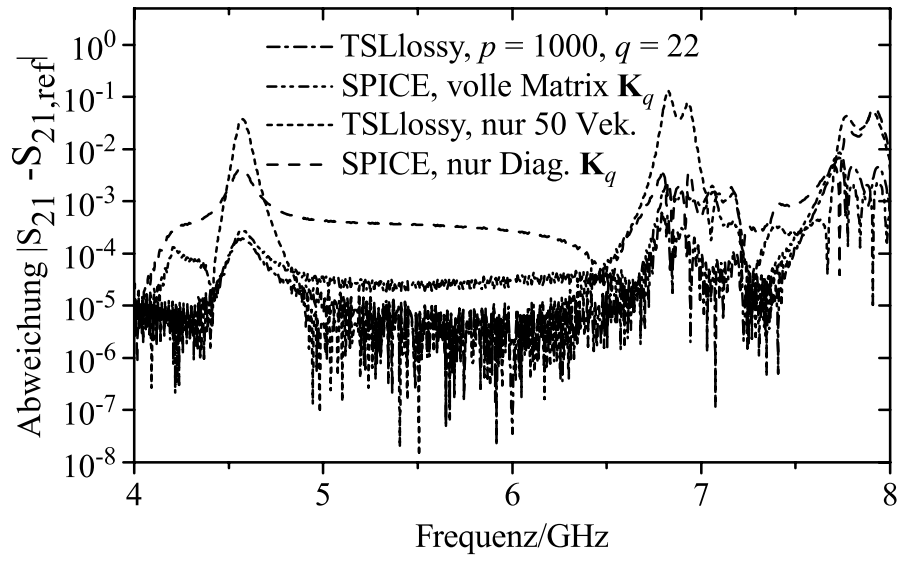


Abbildung 7.7: *Die Abweichung der Transmission des verlustbehafteten Filters von der Referenzlösung. Die TSL-Projektion und die SPICE-Lösung zeigen eine sehr geringe Abweichung, die Fehler bei nur 50 gespeicherten Vektoren oder alleiniger Betrachtung der Diagonale der reduzierten Verlustmatrix liegen bei maximal* 10^{-1}.

Erneut zeigt TSL sehr gute Approximationseigenschaften, die Kurven sind, wie in Abb. 7.6 zu erkennen, nicht unterscheidbar. Die Abweichung des Transmissionsfaktors zur Zeitbereichslösung liegt nach Abb. 7.7 an den Polstellen bei etwa 10^{-3}, in den übrigen Frequenzbereichen unter 10^{-4}. Sie liegt damit im selben Rahmen wie die Abweichung des verlustlosen Systems relativ zur Zeitbereichsreferenz. Etwas größere, aber noch immer akzeptable Abweichungen von maximal 10^{-1} an den Polstellen ergeben sich, wenn im ersten Schritt anstelle der 1000 Vektoren nur die letzten 50 im Speicher gehalten werden.

Wird ein Ersatzschaltbild der verlustbehafteten Struktur generiert, hat dieses, wenn alle Einträge der Verlustmatrix berücksichtigt werden, 385 Elemente. Ein SPICE-Frequenzdurchlauf benötigt damit 6,6 s, wobei die resultierende Kurve auf der Referenz liegt. Werden nur die Diagonaleinträge der Verlustmatrix berücksichtigt, sind die benötigten 154 Netzwerkelemente erneut rein positiv und die Simulationszeit verkürzt sich auf 1,4 s, allerdings wiederum bei kleinen Abweichungen in der Kurve.

Neben dielektrischen Materialien mit geringen Leitfähigkeiten stellen auch die hohen aber endlichen Leitfähigkeiten des Resonatorgehäuses eine Quelle für Verluste dar. Diese werden mit Impedanzrandbedingungen unter Annahme einer geringen magnetischen Leitfähigkeit modelliert. Erneut wird mit einer Leitfähigkeit des Metalls von 10^4 S/m ein übertriebener Wert angenommen, Kupfer hat beispielsweise $5,7 \cdot 10^7$ S/m. Analog zu den oben beschriebenen Verlusten erfolgt die Einbeziehung in das Modell durch Projektion der Impedanzrandverluste, repräsentiert durch die Matrix $\mathbf{P}$, auf die Unterräume des verlustfreien Systems (TSLimpw). Wie in Abb. 7.8 zu sehen, besteht auch hier eine große Übereinstimmung zu der Referenzkurve aus MWS, der Fehler liegt außerhalb der Polstellen wieder im Bereich von 10^{-4}, an den Polstellen bei 10^{-2} (siehe Abb. 7.9). Die Abweichung erklärt sich zum Teil durch die unterschiedlichen Ansätze, die zur Modellierung der Eindringtiefe verwendet werden. Das TSL-Modell beruht auf den in dieser Arbeit vorgeschlagenen Impedanzwänden, die eine geringe magnetische Leitfähigkeit nutzen, die MWS-Referenzlösung nutzt eine hohe elektrische Leitfähigkeit. Selbst wenn die im Allgemeinen nicht passive $(-1/(s\sqrt{s}))$-Abhängigkeit der TSL-Impedanzmatrix einfach durch den konstanten Wert des Realteils bei der Mittenfrequenz ersetzt wird, weichen die Kurven nur gering voneinander ab, der resultierende Fehler vergrößert sich um etwa eine Größenordnung und liegt an den Polstellen bei maximal 10^{-1}. Das System ist somit wieder passiv und kann wie ein gewöhnliches verlustbehaftetes Modell zu einem Netzwerk transformiert werden.

Als Alternative zu TSL bzw. ordnungsreduzierten Modellen, die auf Projektion beruhen, wird auch ein aus Zeitbereichsdaten generiertes Modell betrachtet. Dieses wird mit dem kurz angesprochenen 4SID-Verfahren erzeugt, die Implementierung hierzu wurde aus [87] entnommen. Für den verlustfreien Fall zeigt sich in Abb. 7.10, dass dieses lineare Modell der Ordnung 68 innerhalb des betrachteten Frequenzintervalls ebenfalls eine gute Approximation darstellt, auch wenn die Übereinstimmung nicht an die von TSL heranreicht, wie aus dem Vergleich in Abb. 7.3 deutlich wird. Es ist jedoch auch offensichtlich, dass das resultierende Modell nicht passiv ist, da der Transmissionsfaktor schon kurz außerhalb des interessierenden und zur Optimierung herangezogenen Bereichs von 4–8 GHz den Wert Eins deutlich übersteigt.

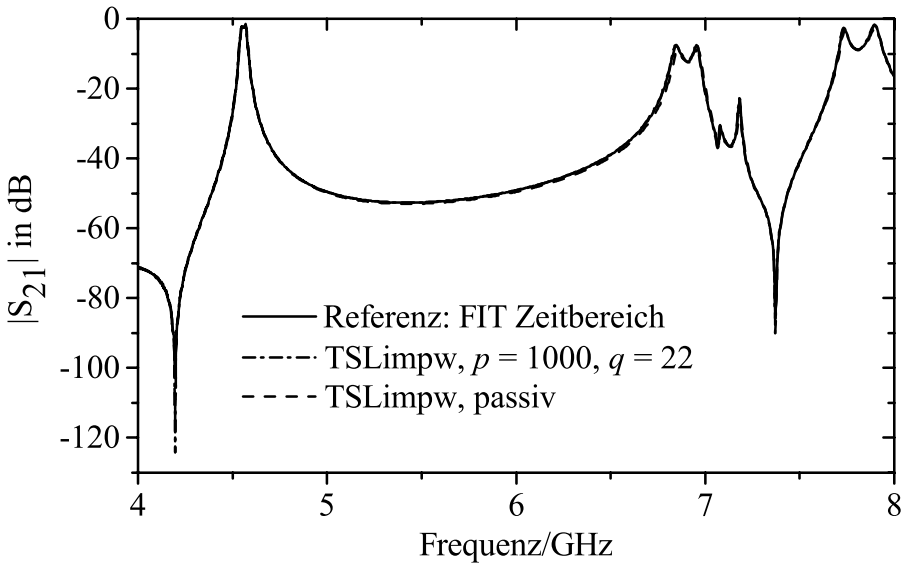


Abbildung 7.8: *Die Transmission S_{21} bei Annahme verlustbehafteter Impedanzwände. Die Kurven der Referenz und von TSLimpw überdecken sich, die der passiven Näherung weicht leicht ab.*

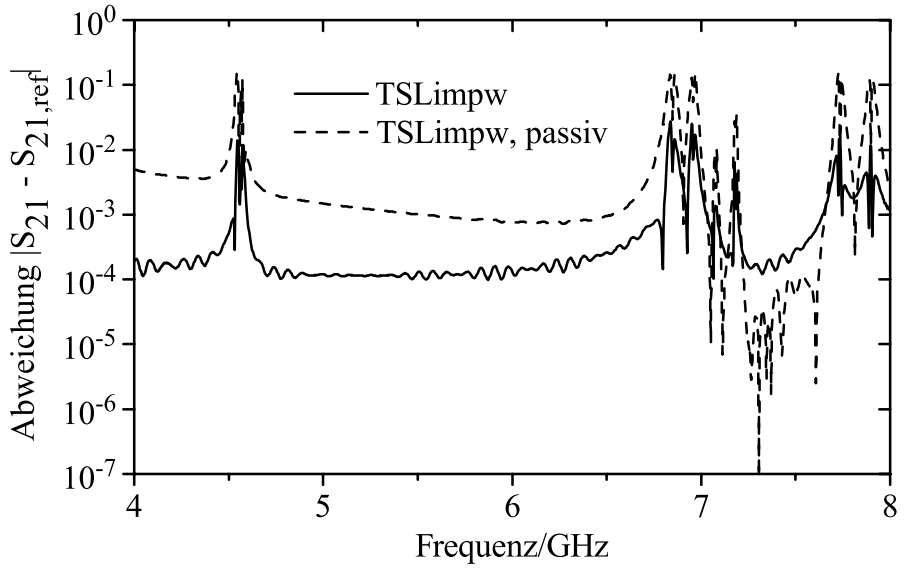


Abbildung 7.9: *Die Abweichung des TSL-reduzierten Impedanzmodells, das auf magnetischen Verlusten basiert, von der Referenzlösung, die die Eindringtiefe durch eine äquivalente elektrische Leitfähigkeit modelliert. Die passive Näherung vergrößert den Fehler um etwa eine Größenordnung.*

Wird nicht direkt ein Makromodell oder ein Ersatzschaltbild benötigt, sondern nur die Übertragungsfunktion der Streuparameter, lassen sich auch mit Signalverarbei-

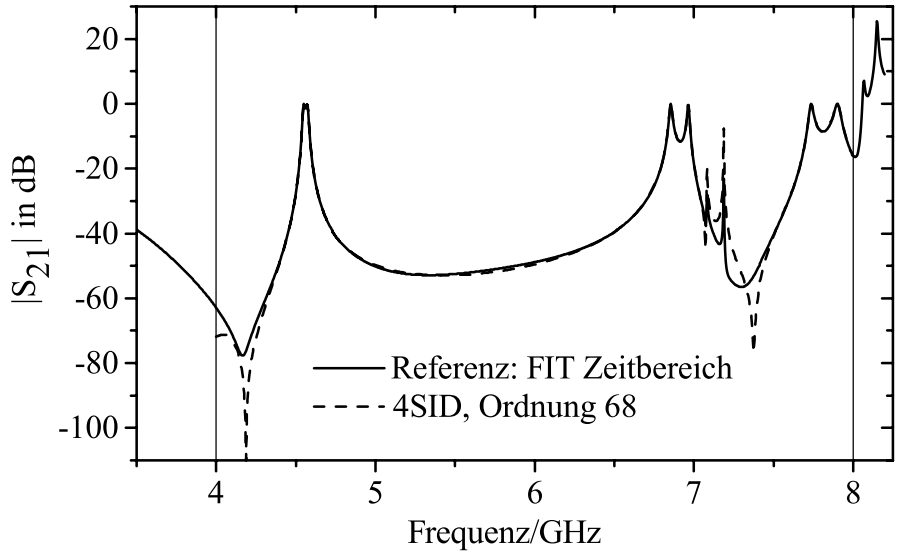


Abbildung 7.10: *Der Transmissionsfaktor eines durch das 4SID-Verfahren erzeugten Makromodells für das verlustlose Langer-Filter.*

tungsmethoden deutliche Rechenzeitgewinne gegenüber bestehenden Verfahren erzielen. Ausgangspunkt ist das Zeitsignal einer transienten Simulation. Da die Struktur mit einem gaußförmigen Puls angeregt wird, der oberhalb von 10 GHz keine erkennbaren Frequenzanteile hat, wird die Abtastrate ohne weitere Filterung so weit reduziert, dass das Abtasttheorem gerade 10 GHz korrekt wiedergibt. Wird die Zeitbereichsrechnung abgebrochen, wenn 11 ns des transienten Signals berechnet sind, führt dies auf 220 Abtastwerte. Werden zudem die ersten 40 Werte, die den Eingangspuls enthalten, übersprungen, findet das Steiglitz-McBride-Verfahren mit der Ordnung 70 und fünf Iterationen bereits ein Modell, das abgesehen von den Polstellen einen mittleren Fehler von nur $5 \cdot 10^{-3}$ aufweist. Als Referenz dient die DFT eines Zeitsignals von 1,8 μs mit also mehr als der 150-fachen Anzahl von Signalwerten. Die Rechenzeit für die Berechnung der Übertragungsfunktion beträgt mit diesem Verfahren 91 s, 90 s für die transiente Rechnung und 1 s für die Filterbestimmung.

Das iterative Prony-Verfahren findet mit demselben Zeitsignal bei Verwendung von 10 Iterationen ebenfalls bereits ein befriedigendes Modell, wobei allerdings die resonanten Pole 7,07 und 7,18 GHz zunächst nur ungenau modelliert werden. Verwendet man 300 Zeitwerte, was 15 ns des Signals entspricht, sind die Schätzungen von Steiglitz-McBride und iterativem Prony etwa äquivalent. Die Berechnung dauert in diesem Fall 120 s. Auch bei längeren Zeitsignalen stagniert der Fehler dieser Methoden bei etwa 10^{-3}, was vermutlich auf die nicht ausreichende Genauigkeit der Referenzlösung zurückzuführen ist.

Werden schließlich die Güten verglichen, die sich aus den Filtermodellen sowie aus TSL nach dem in Abschnitt 3.3.3 beschriebenen Vorgehen ergeben, zeigt sich auch hier eine gute Übereinstimmung. In Tabelle 7.1 werden die beiden Resonanzen des Passbands bei 4,55 GHz sowie die beiden hohen Güten um 7,1 GHz betrachtet. Insbesondere TSL und Steiglitz-McBride weisen sehr ähnliche Werte für die Güten auf.

Wird eine längere Folge des Zeitsignals betrachtet, nähert sich auch der iterative Prony diesen Werten an. Es sei nochmals darauf hingewiesen, dass zur Berechnung der Güten nur 42 bzw. 120 s Rechenzeit erforderlich waren, was im Vergleich zu bestehenden Schätzverfahren eine deutliche Verkürzung darstellt.

Frequenz in GHz	Güten		
	TSL	Steiglitz-McBride	iterativer Prony
4,56	327	332	330
4,58	300	305	306
7,07	26.568	24.132	41.594
7,18	9.100	8.472	7.081

Tabelle 7.1: *Vergleich der Güteberechnung.*

Zum Abschluss der Betrachtung dieses Beispiels sollen alle verwendeten Verfahren in Tabelle 7.2 noch einmal zusammengefasst werden. Als Maß für die Komplexität der Algorithmen wird jeweils die benötigte Rechenzeit sowie die Anzahl der verwendeten Matrix-Vektor-Multiplikationen angegeben. Die Matrix-Vektor-Multiplikationen beziehen sich zwar auf unterschiedliche Systeme, bilden aber dennoch einen vergleichbaren Richtwert. Es zeigt sich, dass der im Rahmen dieser Arbeit vorgeschlagene TSL-Algorithmus alle bestehenden Verfahren an Effizienz deutlich übersteigt. Zur Spektralschätzung stellt das Verfahren nach Steiglitz-McBride ebenfalls eine nennenswerte Verbesserung gegenüber dem bisherigen AR-Filter dar.

Verfahren	**Rechenzeit[s]**	**M.-V.-Multip.**
FIT Zeitbereich	9.000	1.500.000
FIT Zeitbereich + AR-Filter	1.047	59.803
FIT Zeitb. + iterativer Prony	117	12.802
FIT Zeitb. + Steiglitz-McBride	**91**	**9.754**
PVL linear, BiCG, Ordnung 22	6.901	69.207
PVL Curl-Curl, COCG, Ord. 22	613	18.146
TSL linear, Ordnung 2020/22	137	2020
TSL Curl-Curl, Ord. 1000/22	**42**	**1000**
TSL Curl-Curl, Ord. 330/22, acc. 450	37	1.230
TSL Curl-Curl, Ord. 330/22, acc. + korr.	65	1.924
TSLlossy, Ord. 1000/22	48	1.000
TSLimpw, Ord. 1000/22	48	1.000
korr. Modalanalyse, 16 Moden	344	10.049
korr. Modalanalyse, 10 Moden	181	18.146

Tabelle 7.2: *Vergleich des Aufwands verschiedener Verfahren zur Reduzierung der Ordnung des Langer-Filters. Betrachtet werden die Rechenzeit sowie die Anzahl der Matrix-Vektor-Multiplikationen.*

7.2 6-Kreis Hohlleiter-Filter

Das zweite untersuchte Beispiel stellt ebenfalls eine Filterstruktur dar und ist in Abb. 7.11 dargestellt. Es handelt sich um ein Wellenleiterfilter, das an beiden Enden mit dem 2D-Feldbild des Grundmodes angeregt wird. Die Filterwirkung beruht auf der Kopplung von sechs gekoppelten Hohlleiterresonatoren, wobei die Übertragungspole bzw. -nullstellen näherungsweise Tschebyscheff-verteilt sind. Das Filter stellt damit einen scharfen Bandpass zwischen 7,1 und 7,3 GHz dar. Hohlleiterfilter haben den grundsätzlichen Vorteil, dass sie sehr verlustarm sind und auch für hohe Leistungen eingesetzt werden können.

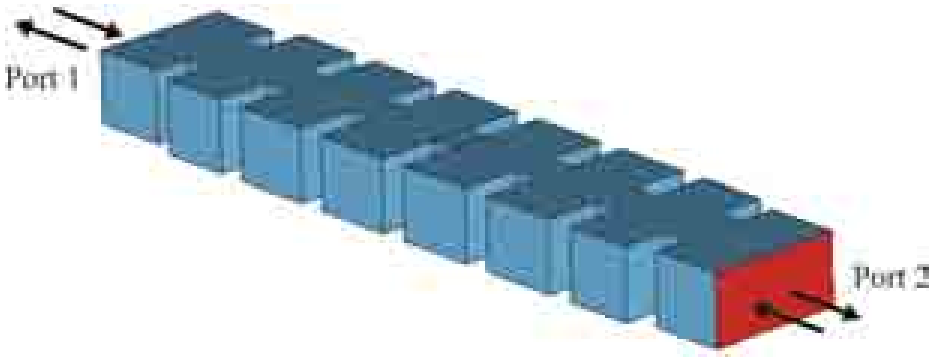


Abbildung 7.11: *Aufbau des betrachteten Hohlleiterfilters.*

Um die Flanken des Passbands mit genügender Genauigkeit bestimmen zu können, sind 40 Gitterlinien pro Wellenlänge erforderlich. Bei Ausnutzung einer Symmetrieebene der Struktur ergibt sich eine Anzahl von 159.720 Gitterpunkten. Dies führt auf ein System von 479.160 Unbekannten in Curl-Curl-Formulierung und 958.320 im Falle eines linearen Zustandsraums. Die Systemgröße ist im Vergleich zum Langer-Filter folglich deutlich vergrößert. Die Normierung der Streuparameter erfolgt bei der Berechnung so, dass für jeden Frequenzpunkt die jeweilige Wellenimpedanz als Referenz dient, der Bezugswiderstand ist damit also variabel.

Die Anwendung des TSL-Algorithmus auf das Curl-Curl-System erweist sich erneut als den übrigen Verfahren deutlich überlegen. Die Übertragungsfunktion der Streuparameter lässt sich an 1000 Frequenzpunkten auf einem 731 MHz PC in weniger als fünf Minuten berechnen. Dem stehen mehr als 70 min gegenüber, die das Zeitbereichsverfahren in Verbindung mit einem AR-Filter benötigt und knapp 15 min für die korrigierte Modalanalyse. Wie in Abb. 7.12 zu erkennen, liegen die Kurven der Streuparameter in logarithmischer Darstellung nahezu ununterscheidbar übereinander. TSL benötigt bei Anwendung des automatischen Stoppkriteriums 760 Iterationen im ersten Schritt und 10 im zweiten. Das Abbruchkriterium wurde hierbei erstmals nach 500 Iterationen angewandt und anschließend alle 20 Iterationen wiederholt, wobei etwa 1 % der Rechenzeit für das Kriterium aufgewendet wurde.

Die folgende Abb. 7.13 zeigt den absoluten Fehler der Streuparameter im Vergleich zu einer Frequenzbereichsrechnung mit MWS. Auch hier ist die maximale Abweichung der Streuparameter kleiner als 10^{-4}.

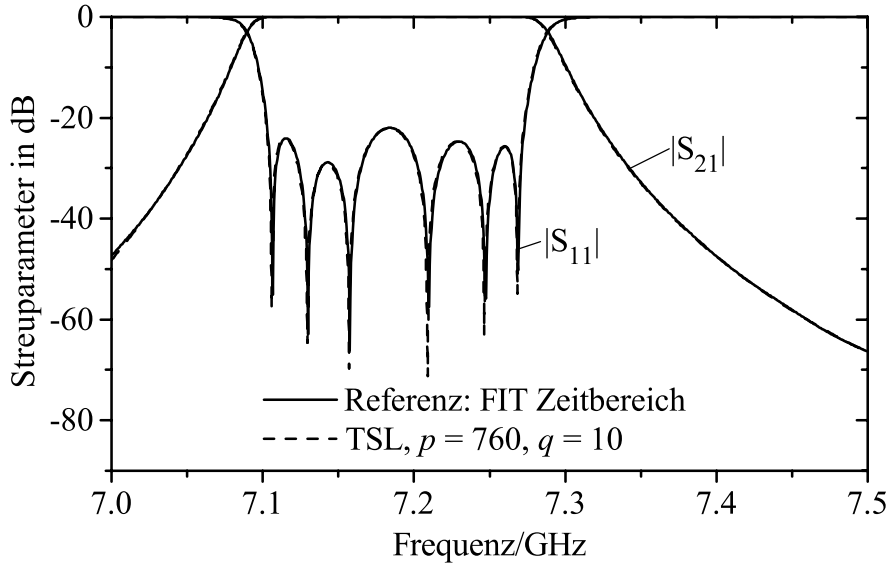


Abbildung 7.12: *Der Frequenzverlauf der Streuparameter des Hohlleiterfilters. Die Referenzkurven und die Ergebnisse von TSL mit* $p = 760$ *und* $q = 10$ *sind nahezu ununterscheidbar.*

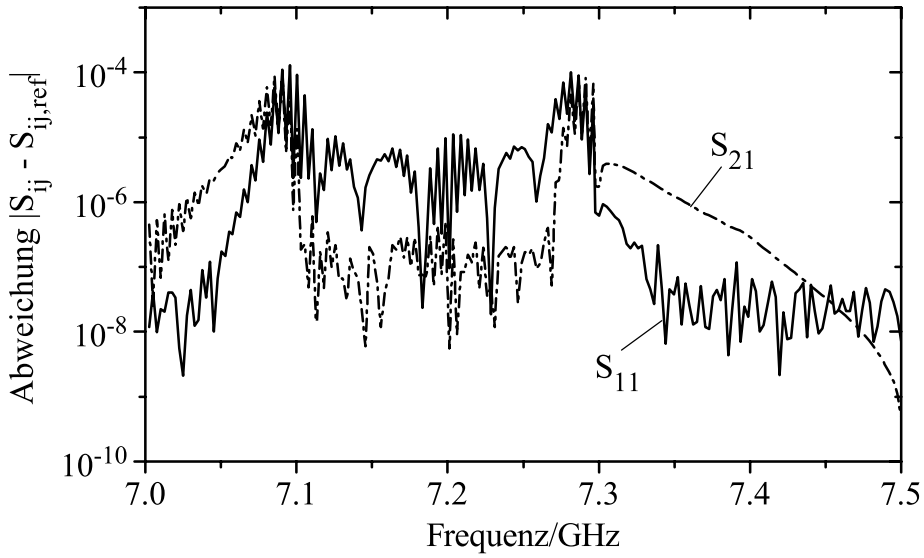


Abbildung 7.13: *Die Abweichung der TSL-Streuparameter des Hohlleiterfilters im Vergleich zur Lösung des nicht reduzierten Systems.*

Im Fall einer Modalanalyse werden acht Moden verwendet. Dies sind die sechs Pole, die innerhalb des interessierenden Frequenzbereichs liegen, sowie zwei darunter liegende. Aus Abb. 7.14 wird deutlich, dass die einfache Modalanalyse nur schwache Ähnlichkeiten mit der Referenzkurve aufweist, während die Kurve nach der Korrektur eine deutlich bessere Approximation darstellt. Dies betont erneut die Bedeutung der Korrektur in der Modalanalyse. Die Rechenzeit der Modalanalyse beträgt 1.027

Sekunden, wovon 27 % für den Korrekturschritt benötigt wurden.

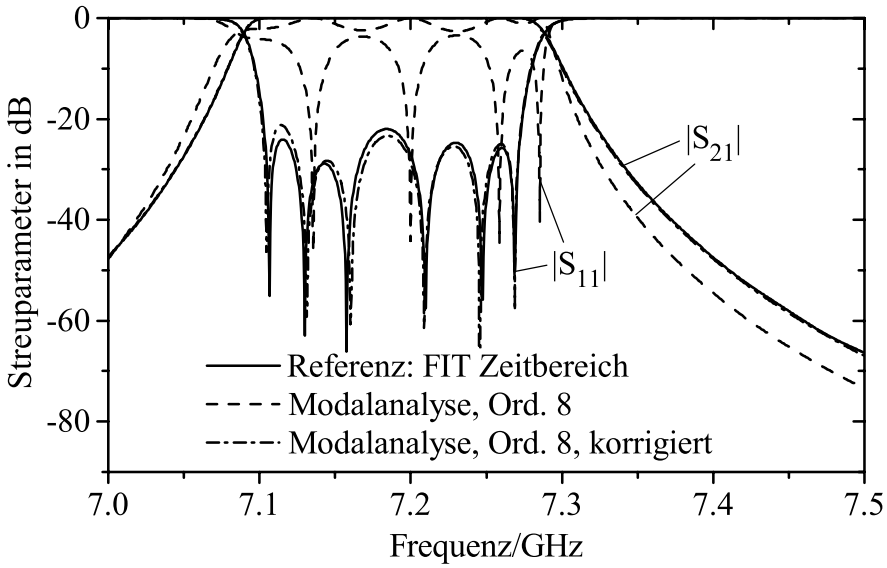


Abbildung 7.14: *Derselbe Frequenzverlauf bei Anwendung der einfachen sowie der korrigierten Modalanalyse mit 8 Moden. Nur die korrigierte Kurve zeigt gute Übereinstimmung mit der Referenzkurve. Es zeigt sich die Bedeutung der Korrektur für diese Methode.*

Wird das Spektrum aus einer Zeitbereichsrechnung gewonnen, lässt sich die Länge des benötigten Zeitsignals mit den beschriebenen Signalverarbeitungsansätzen wieder deutlich reduzieren. In der als Referenz betrachteten transienten Simulation in Verbindung mit einem AR-Filter wird ein Signal von 28 ns Länge oder 59.801 Zeitschritten benötigt. Mit dem Verfahren nach Steiglitz-McBride reicht bei gleicher Approximationsqualität bereits ein Signal von etwa 14 ns mit folglich nur etwa der Hälfte der Zeitschritte (29.866). Dies halbiert in etwa auch die Rechenzeit von 4.657 s auf 2.238 s, ein Wert der allerdings dennoch deutlich über dem vergleichbaren Aufwand für TSL oder eine Modalanalyse liegt. Einen Überblick gibt die Tabelle 7.3.

Verfahren	**Rechenzeit[s]**	**Matrix-Vektor-Mult.**
FIT Zeitbereich + AR-Filter	4.657	59.801
FIT Zeitbereich + Steiglitz-McBride	2.238	29.866
TSL linear, Ordnung 1.560/10	881	1.560
TSL Curl-Curl, Ordnung 760/10	**287**	**760**
korr. Modalanalyse, 8 Moden	1.027	3.245

Tabelle 7.3: *Vergleich des Aufwands verschiedener Verfahren zur Reduzierung der Ordnung des Hohlleiterfilters.*

7.3 Patchantenne

Als typisches Beispiel für eine Struktur, die offene Randbedingungen erfordert, soll im Weiteren eine Patchantenne untersucht werden, die in Abb. 7.15 gezeigt ist. Antennen dieser Art zeichnen sich dadurch aus, dass sie sehr Platz sparend sind und einfach sowie günstig hergestellt werden können, weswegen sie vielfache Verwendung finden. Die hier betrachtete Antenne besteht aus einer dünnen, als unendlich leitfähig angenommenen, rechteckigen Metallschicht mit den Maßen $12,45 \times 16$ mm, die auf einem 0,8 mm dicken dielektrischen Substrat aus Teflon mit $\varepsilon = 2,2$ aufgebracht ist. Die Unterseite des Substrats ist ebenfalls metallisiert, was im diskreten Modell durch eine elektrische Randbedingung realisiert ist. Oberhalb der Antenne befinden sich 3 mm Luft, der weitere Außenraum wird durch eine vierschichtige PML-Randbedingung nachgebildet. Die Speisung der Antenne erfolgt über eine 50 Ω Mikrostreifenleitung, die wiederum durch einen Wellenleiterport angeregt wird.

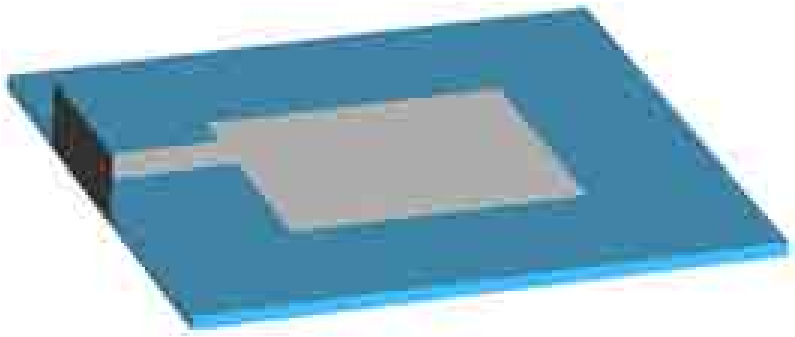


Abbildung 7.15: *Aufbau der untersuchten Patchantenne. Die Fläche, an der die Signaleinspeisung durch einen Wellenleiterport erfolgt, ist links im Bild gekennzeichnet.*

Die Antenne wurde für eine Sendefrequenz von 7,5 GHz optimiert, es ist also zu erwarten, dass der Reflexionskoeffizient für diese Frequenz minimal wird. Untersucht werden soll die Antenne in einem Frequenzbereich von 6–9 GHz. Da bei ausreichender Modellgröße die Kurven aller Verfahren die Referenzkurve überdecken, wird diese einmal in Abb. 7.16 dargestellt. Die späteren Untersuchungen vergleichen nur die erforderliche Systemgröße, um die Referenz auf einen vorgegebenen Fehler zu approximieren.

Wird die kleinste auftretende Wellenlänge mit 15 Gitterlinien abgetastet, führt dies auf etwa 4.200 Gitterzellen innerhalb der Struktur sowie dieselbe Anzahl von Zellen im absorbierenden PML-Material. Zur Beschreibung der Antenne und der absorbierenden Randbedingung als System existieren nun drei Möglichkeiten:

- Wird ein komplexes dämpfendes Material entsprechend Gl. 2.2.26 angenommen, kann wie in den vorigen Beispielen ein lineares oder ein Curl-Curl-System aufgestellt werden. Allerdings werden die Matrizen komplex, weswegen das Modell nicht kausal und folglich nicht für Zeitbereichssimulationen geeignet ist. Das Curl-Curl-System hat nach Eliminierung aller Nullspalten und -zeilen eine Ordnung von 17.744.

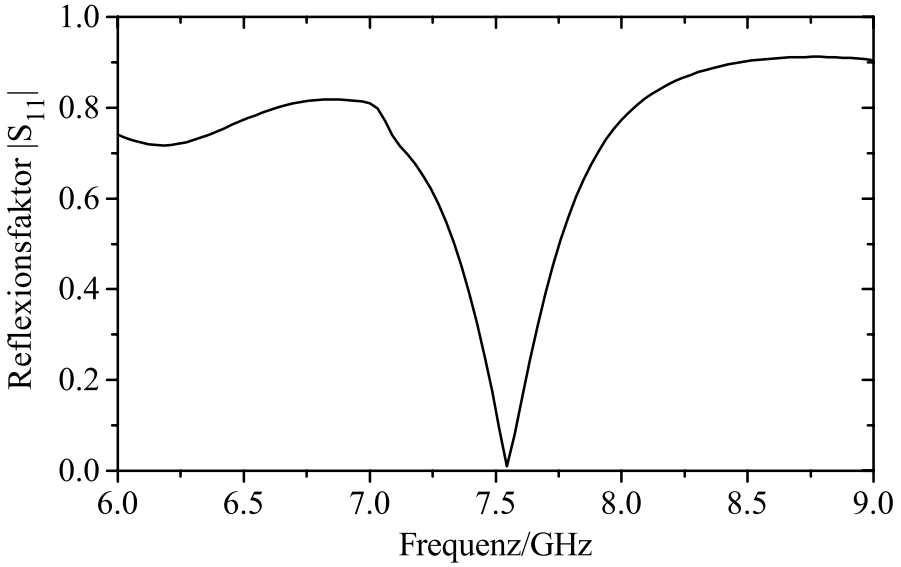


Abbildung 7.16: *Reflexionsfaktor der Patchantenne. Bei 7,5 GHz liegt die Resonanz der Antenne, d.h., es erfolgt die maximale Abstrahlung.*

- Wird ein frequenzabhängiges komplexes Material nach der Beziehung Gl. 2.2.27 verwendet, entsteht ein System zweiten Grades nach Gl. 3.1.23b. Dieses hat nach der Eliminierung 36.064 Unbekannte und kann mit Hilfe des WCAWE-Algorithmus in der Ordnung reduziert werden.

- Das System zweiten Grades lässt sich wiederum durch Hinzunahme von Variablen durch Gl. 3.1.15 in ein in der Frequenz lineares System transformieren und entsprechend mit den bekannten Verfahren reduzieren. Das lineare System hat etwa 47.432 Unbekannte. Setzt man diese in Relation zu den 4.200 Gitterzellen, zeigt sich, dass die Systemgröße bei Verwendung von PML-Randbedingungen massiv anwächst.

Bedauerlicherweise zeigt sich, dass sich die Kondition der Matrizen im Vergleich zum Fall der resonanten Systeme durch das PML-Material merklich verschlechtert und partielle Realisierungen in einem frühen Stadium stagnieren. Der effiziente TSL-Algorithmus kann daher in diesem Fall nicht angewendet werden. Zur Untersuchung werden daher die Lanczos-basierte Padé-Approximation (PVL), die Arnoldi-basierte symmetrische Padé-Typ-Approximation und das WCAWE-Verfahren verwendet.

Da die Matrizen für eine LU-Zerlegung bereits zu groß sind, wird das iterative BiCG-Verfahren zur Lösung der linearen Systeme genutzt. Eine Vorkonditionierung wird hierbei nicht verwendet, da sich eine unvollständige LU-Zerlegung als unwirksam herausgestellt hat und andere Verfahren für die entsprechenden Matrixstrukturen nicht ohne weiteres verfügbar sind. Als Entwicklungspunkt wird die Mitte des interessierenden Frequenzbandes bei 7,5 GHz gewählt, die geforderte Solvergenauigkeit ist 10^{-6}. In Tabelle 7.4 wird die jeweilige Modellordnung verglichen, die benötigt wird, um einen mittleren Fehler von 10^{-4} bzw. 10^{-6} zur Referenz einzuhalten.

Fehler	Padé Curl-Curl	Padé linear	Padé-Typ linear	WCAWE
10^{-4}	8	7	12	12
10^{-6}	11	11	19	19

Tabelle 7.4: *Größe des mit unterschiedlichen Verfahren berechneten reduzierten Modells, um den vorgegebenen Fehler zu erreichen.*

Es zeigt sich klar, dass die Größe des reduzierten Modells vom Typ der Approximation, nicht aber von der gewählten Formulierung abhängt. So haben die Padé-Approximationen der Curl-Curl und der linearen Formulierung etwa die gleiche Ordnung, ebenso die symmetrischen Projektionen mit Arnoldi im linearen Fall und mit WCAWE im System zweiten Grades. Interessiert die Modellgröße, produzieren also die reinen Padé-Approximationen die kleinsten Modelle. Vergleicht man den numerischen Aufwand, benötigt aber jede Iteration der Padé-Approximation zwei Systemlösungen pro Iteration, die symmetrischen Verfahren nur eine. Der Gesamtaufwand ist demnach bei den Padé-Typ-Approximationen – dieselbe Systemformulierung vorausgesetzt – geringer.

Rechenzeiten haben an dieser Stelle nur geringe Aussagekraft, da nicht optimierte Gleichungslöser verwendet werden. Um aber einen Richtwert zu geben: Um einen Fehler von 10^{-4} zu erreichen, benötigt die Padé-Approximation 348 s im Curl-Curl- und 3.368 s im linearen Fall. Die symmetrische Projektion braucht 2.955 s, WCAWE 6.865 s. Alle Rechnungen erfolgten auf einem 1,7 GHz PC. Da die Zeitbereichsimulation desselben Modells gerade 22 s benötigt, ist ein *Fast Frequency Sweep* folglich keinesfalls sinnvoll bzw. möglich. Ein reduziertes Modell nach dem 4SID-Verfahren lässt sich aus den Zeitsignalen ebenfalls innerhalb von Sekunden berechnen, das Verfahren ist für Strukturen mit offener Berandung also deutlich effizienter. Die Genauigkeit des 4SID-Modells erreicht mit 10^{-2} bei der Ordnung 15 allerdings nicht ganz die Werte der projizierten Modelle.

Keines der generierten Modelle erhält die Passivität des Systems. Da die Matrix $\mathbf{A}_0$ nach Gl. 3.1.23a keine positiv reelle Matrix darstellt, kann selbst eine reell symmetrische Projektion die Passivität des Systems nicht gewährleisten. Die Frage nach einer passiven Näherung mit geringer Ordnung für PML-berandete Strukturen bleibt folglich ein ungelöstes Problem.

7.4 Chip-Interconnect-Modell

Als abschließendes Beispiel soll das Gehäuse eines integrierten Schaltkreises (engl. *Integrated Circuit, IC*) untersucht werden. Es handelt sich um das P-TSSOP24-Gehäuse der Infineon Technologies AG. Die vollständige Struktur ist links in Abb. 7.17 gezeigt; die eigentlich relevante Zuleitungsstruktur, die so genannte *Interconnect*-Struktur, mit den *Pins* und dem Siliziumplättchen, das die logische Schaltung beinhaltet, ist auf der rechten Seite abgebildet. Das Gehäuse besitzt 24 Zuleitungen und kann diverse Schaltungen aufnehmen. Ausgelegt ist es für den Mobilfunkbereich mit Betriebsfrequenzen von 900 MHz bzw. 1,8 GHz.

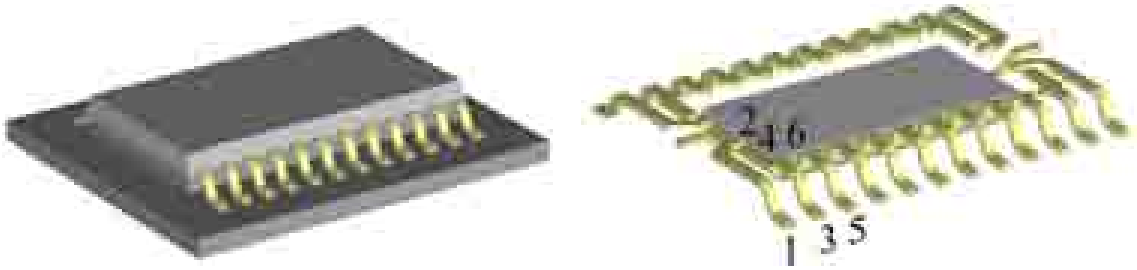


Abbildung 7.17: *Das untersuchte Chipgehäuse mit 24 Zuleitungen. Links ist die simulierte Gesamtstruktur einschließlich des Substrats dargestellt, rechts die zum Teil im Inneren liegenden metallischen Zuleitungen und das Siliziumplättchen. Die Nummerierung markiert die Ports, deren Streuparameter in der folgenden Abbildung dargestellt sind.*

Im diskreten Modell wird das Silizium als Material, nicht aber als logische Schaltung betrachtet. Schwerpunkt des Interesses sind die Feldeffekte, wie Laufzeiten und Nebensprechen, die sich in und zwischen den Zuleitungen auswirken. Jeder der Pins wird daher am Anfang und am Ende mit je einem diskreten Port zur geerdeten Grundmetallisierung des Stubstrats verbunden und darüber angeregt. Der Innenwiderstand dieser Ports beträgt 50 Ω. Je ein Port verbindet auch das Siliziumplättchen und eine Abschirmplatte unterhalb des Siliziums mit der Erdung. Die Gesamtstruktur weist somit 50 Ports auf, eine Anzahl, die für jede Form der Makromodellierung eine große Herausforderung darstellt. Da in der Digitaltechnik Pulse mit Frequenzanteilen bis zur zehnten Oberschwingung relevant sein können, wird die Struktur in einem Bereich von 0–25 GHz simuliert. Wird die Gitterauflösung so gewählt, dass die kürzeste Wellenlänge mit 10 Gitterlinien abgetastet wird, führt dies auf eine Gitterpunktzahl von 22.704.

Da die Abstrahlung der Zuleitungen in den freien Raum sehr gering ist, wurde das Rechengebiet, abgesehen von der Metallisierung des Substrats, mit magnetischen Rändern umgeben. Die Lösung ist hierbei nahezu identisch im Vergleich zu einer mit PML-Berandungen berechneten. Die Probleme der Reduzierung der Ordnung, die anhand des vorigen Beispiels beschrieben wurden, können somit vermieden werden.

Die Reduzierung des Modells erfolgt durch TSL in der Curl-Curl-Formulierung, die rund 68.000 Unbekannte enthält. Die Koppelmatrix $\mathbf{B}$ enthält nun 50 Vektoren. Um ein Taylormoment des Originalsystems zu bestimmen sind also 50 Lanczos-Schritte erforderlich. Die Curl-Operatoren werden erneut sukzessive ausgeführt. Als Abbruchkriterium wird in beiden Schritten ein Wert von 10^{-4} verwendet. Im ersten Schritt wird es erstmals nach 2000 und in der Folge alle 50 Iterationen durchgeführt. Der erste Schritt bricht damit nach 5400 Iterationen und 3.623 s Rechenzeit ab. Der zweite Schritt berechnet 137 Iterationen und benötigt dafür 92 s. Die Gesamtrechenzeit liegt also knapp über einer Stunde. Ein Durchlauf einer Zeitbereichsmulation benötigt im Vergleich 268 s. Werden jedoch alle 50 Ports nacheinander angeregt, um tatsächlich die volle Streumatrix zu bestimmen, führt dies mit 13.400 s auf mehr als die dreifache Rechenzeit. Wenn die vollständige Übertragungsmatrix benötigt wird,

wie dies beispielsweise für ein Makromodell der Fall ist, erweist sich TSL wieder überlegen gegenüber 4SID oder anderen Verfahren, die auf Zeitsignalen basieren.

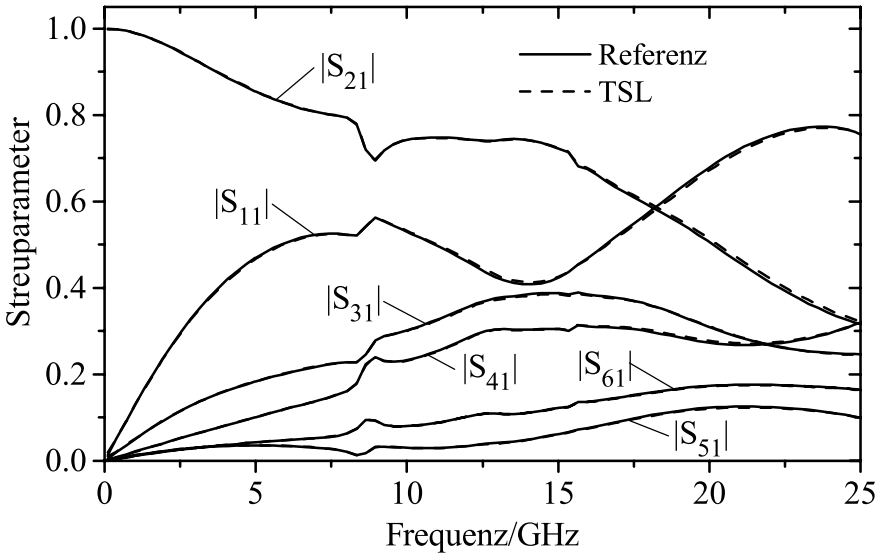


Abbildung 7.18: *Der Verlauf ausgewählter S-Parameter des Chipmodells bei Anregung von Port 1. Verglichen werden die Referenzlösung und das TSL-Ergebnis.*

Wie in Abb. 7.18 dargestellt, zeigen die TSL-Streuparameter grundsätzlich eine sehr gute Übereinstimmung mir der MWS-Referenz. Aus den insgesamt von dem Modell repräsentierten $50 \times 50 = 2.500$ Übertragungsfunktionen werden hierbei natürlich nur ein kleiner Teil abgebildet. Betrachtet werden Transmission und Reflexion bei Anregung von Port 1 sowie der Einfluss auf die beiden benachbarten Zuleitungen. Die zugehörige Portnummerierung ist ebenfalls in Abb. 7.17 rechts zu erkennen.

Aus den TSL-Matrizen lässt sich erneut ein garantiert passives Ersatzschaltbild mit 822 Elementen generieren. Dieselbe Chipstruktur wurde in [101] auch genutzt, um einen alternativen Ansatz zu testen, ein Ersatzschaltbild zu erstellen. Hierbei wird jede Zuleitung einzeln durch ein RLC-Leitungsmodell nachgebildet. Das Nebensprechen wird durch Koppelkapazitäten und zusätzliche Kopplungen zwischen den Induktivitäten nachgebildet. Die Struktur des Ersatzschaltbildes ist also vorgegeben, die Koeffizienten des Modells werden lediglich aus den Streuparametern für eine Frequenz optimiert. Die Qualität des Modells kann zusätzlich über die Anzahl der in Reihe geschalteten Leitungsglieder gesteuert werden, was als Kaskadierung bezeichnet wird, genauere Details in [101]. In Abb. 7.19 werden die besten in dieser Veröffentlichung erreichten Fehler mit denen des TSL-Verfahrens verglichen. Da die Fehler der mit TSL berechneten Streuparameter alle im gleichen Rahmen liegen, wird auf eine einzelne Beschriftung in der Abbildung verzichtet.

Zum Vergleich werden zwei Leitungsmodelle herangezogen. Das erste wurde bei 10 GHz optimiert und es wird nur eine Kaskade verwendet, was auf einen verhältnismäßig konstanten Fehler von 0,05 bis zu Frequenzen von 10 GHz führt. Das andere bei 1 GHz optimierte Modell in drei Kaskaden erreicht einen sehr geringen Fehler bis etwa 5 GHz, steigt dann aber stark an. Zur Optimierung wurden nur drei Zuleitungen

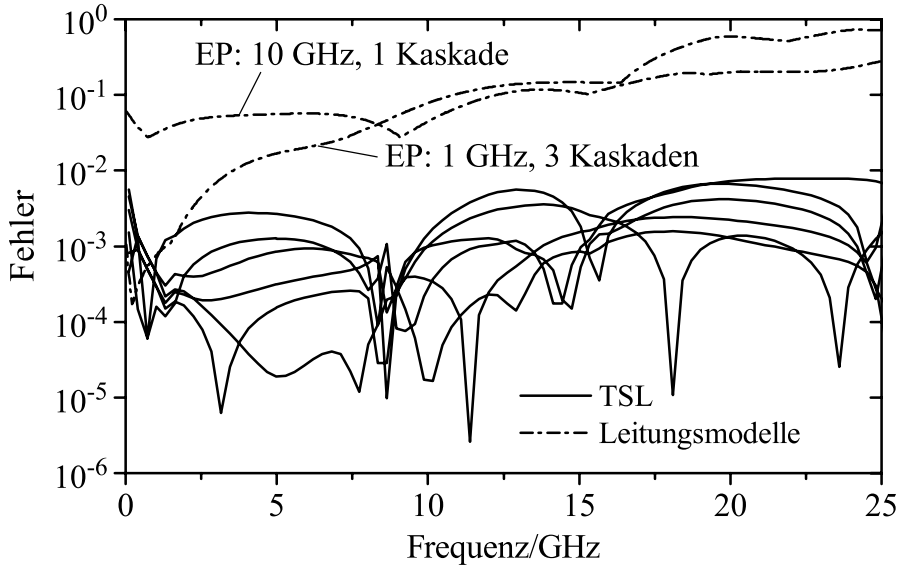


Abbildung 7.19: *Der Fehler der mit TSL berechneten Streuparameter im Vergleich zu einem Leitungsmodell, das an unterschiedlichen Entwicklungspunkten (EP) optimiert wurde. Da alle TSL-Fehler vergleichbar sind, wird auf separate Kennzeichnung verzichtet.*

modelliert. Die Zahl der Elemente beträgt für das erste Netzwerk mit einer Kaskade 34 Elemente, für das zweite mit drei Kaskaden 88. Erweitert man das Verfahren allerdings auf alle 50 Ports wie im TSL-Fall, steigt die Zahl der Koppelkoeffizienten quadratisch an. Auch der Aufwand für die Optimierung wächst deutlich.

Der Fehler von TSL liegt im gesamten betrachteten Frequenzbereich bei unter 10^{-2}, das resultierende Modell ist dem Leitungsmodell folglich sowohl von der Genauigkeit als auch vom benötigten Aufwand deutlich überlegen.

Kapitel 8

Zusammenfassung und Ausblick

Zielsetzung dieser Arbeit war es, ein Modell einer elektromagnetischen Struktur zu generieren, das mit möglichst geringer Ordnung das Übertragungsverhalten des Bauteils in einem gewissen Frequenzbereich nachbildet. Ausgangspunkt hierfür ist eine Diskretisierung der Struktur nach der Methode der Finiten Integration.

Zur Generierung des Modells existieren dabei zwei grundsätzlich unterschiedliche Ansätze: Entweder wird das diskretisierte Modell zunächst berechnet und aus den Ergebnissen ein Makromodell extrahiert oder das Makromodell wird direkt aus der mathematischen Beschreibung der diskretisierten Struktur gewonnen. Die Berechnung des reduzierten Modells direkt aus dem Originalsystem hat insbesondere den Vorteil, dass während des Reduktionsprozesses auch eine Lösung des Modells bereitgestellt wird. So erfolgt die Lösung des Systems als *Fast Frequency Sweep* auf der reduzierten Basis anstelle der des großen Originalsystems. Diese Arbeit konzentriert sich daher auf den zweiten Fall, zieht aber Ergebnisse aus dem ersten Ansatz zum Vergleich heran. Auch die Untersuchung der Spektralschätzverfahren fällt in die erste Kategorie.

Die Methode der Finiten Integration (FIT) stellt eine konsistente Übertragung der kontinuierlichen Maxwellschen Gleichungen in den Gitterraum dar. Neben einer allgemeinen Einführung wird die Behandlung des Randes näher betrachtet. Hierbei wird insbesondere eine neue Formulierung der Impedanzrandbedingung für sehr gute Leiter vorgeschlagen, die auf geringen magnetischen Verlusten basiert, damit eine deutliche Verbesserung der Kondition des Systems bewirkt und eine vereinfachte Behandlung bei der späteren Reduzierung der Ordnung ermöglicht. Zur Anregung wird anstelle des klassischen offenen Wellenleiterrands eine Formulierung vorgestellt, die ein geschlossenes System mittels eines äquivalenten Stroms anregt. Dies führt auf eine Impedanzdarstellung des Systems in klassischer oder erweiterter Zustandsraumdarstellung. Mit den Methoden der Systemtheorie können daraus wichtige Systemeigenschaften wie Kausalität, Stabilität und Passivität einfach untersucht werden. Es zeigt sich hierbei, dass mit FIT diskretisierte Systeme in nahezu allen Fällen diese Eigenschaften erhalten, sofern sie von der vorgegebenen Struktur erfüllt sind. Ausnahmen bilden PML-berandete Strukturen, deren Passivität nicht allgemein nachgewiesen werden kann, sowie der vorgeschlagene Impedanzrand, wobei die Passivität

im zweiten Fall durch eine kleine Näherung wieder garantiert werden kann. Zwischen den aufgestellten Impedanzmodellen und den in der Praxis meist relevanteren Streumatrizen besteht ein enger Zusammenhang.

Die untersuchten Systeme bilden die Grundlage für die Reduzierung der Modellordnung durch einen allgemeinen Projektionsansatz. Unter der Voraussetzung, dass alle Systemmatrizen positiv reelle Matrizen darstellen, wird bei einer symmetrischen Projektion auf einen reellen Unterraum die Passivität des Originalsystems auch im reduzierten Modell erhalten. Dies ist für die meisten vorgestellten Formulierungen erfüllt, Ausnahmen bilden aber linearisierte Systeme sowie Systeme PML-berandeter Strukturen. Wird hingegen eine unsymmetrische Projektion auf zwei Unterräume angewendet, kann im Allgemeinen eine verbesserte Approximationsgüte erzielt werden.

Eine entscheidende Rolle als Projektionsmatrizen spielen Krylov-Unterräume. Werden sie direkt aus dem System bestimmt, ergibt sich durch Projektion eine partielle Realisierung, werden sie aus dem invertierten System berechnet, folgt durch Projektion eine Padé-Approximation. Die bisher wenig beachteten partiellen Realisierungen führen zwar generell auf größere Modelle, sind vom Rechenaufwand einer Padé-Approximation aber häufig weit überlegen. Wendet man zur Bestimmung einer partiellen Realisierung aus der Eigenwertberechnung bekannte Techniken wie Beschleunigungspolynome an, lässt sich zudem ein nahezu fließender Übergang zur Modalanalyse finden. Somit wird es möglich, den Projektionsunterraum im Speicher zu halten, um z. B. Feldlösungen zu berechnen. Werden die höheren Moden zu stark unterdrückt, führt dies auf einen Fehler, der durch eine einfache aber effiziente Korrektur behoben werden kann.

Die Vorteile der beiden genannten Verfahren werden schließlich durch den im Rahmen dieser Arbeit entwickelten Two-Step-Lanczos-Algorithmus (TSL) verbunden. In einem ersten Schritt wird eine partielle Realisierung berechnet. Diese hat bereits eine Größe, in der sie leicht LU-zerlegt werden kann, was die nachfolgende Berechnung einer Padé-Approximation auf effiziente Weise ermöglicht. Das resultierende Modell hat folglich die Dimension einer Padé-Approximation, während der Rechenaufwand den einer partiellen Realisierung nur leicht übersteigt. Zudem wird ein eigenwertbasiertes Abbruchkriterium vorgeschlagen, womit sich ein vergleichbarer Fehler in beiden Schritten ergibt. Der Algorithmus kann somit vollständig automatisch ablaufen. Besondere Effizienz zeigt das Verfahren, wenn es in Verbindung mit einer Curl-Curl-Formulierung verwendet wird. Das System bleibt auch bei Verwendung von imaginären Entwicklungspunkten, die sich als besonders geeignet erweisen, reell und symmetrisch. Das reduzierte Modell erhält folglich bei maximaler Approximationsgenauigkeit seine Passivität. Das Verfahren kann einfach erweitert werden, um geringe parasitäre Verluste zu berücksichtigen. Diese werden auf die Unterräume des verlustfreien Systems projiziert. Durch diese Näherung stimmen zwar die Momente mit denen des Originalsystems nicht mehr exakt überein, die Vorteile der Methode werden aber in vollem Umfang erhalten.

Anhand mehrerer realistischer Filterbeispiele wird die Effizienz des TSL-Algorithmus eindrucksvoll belegt. Der Rechenzeitgewinn gegenüber klassischen Zeitbereichsverfahren in Verbindung mit einem AR-Filter liegt bei einem Faktor zwischen 10

und 25. Das Verfahren wurde für bis zu 7,5 Millionen Unbekannte getestet und erweist sich als sehr robust. Auch die Erweiterung auf Verluste führt zu keinen erkennbaren Genauigkeitseinbußen. Auch im kritischen Anwendungsfall einer *Interconnect*-Struktur mit 50 Ports erwies sich TSL als konkurrenzfähig im Vergleich zu Zeitbereichsverfahren.

Die aus TSL resultierenden Makromodelle lassen sich auch unmittelbar in ein Ersatzschaltbild wandeln und somit in einem Netzwerksimulator verwenden. Insbesondere aus einem verlustlosen TSL-Modell lässt sich ein physikalisches Netzwerk allein aus LC-Schwingkreisen und wahlweise idealen Übertragern oder linearen gesteuerten Quellen realisieren. Für andere Formulierungen kommt das Netzwerk eher einem mathematischen Hilfsmittel gleich, welches auch negative Elemente beinhalten kann. Ist die Passivität des zu Grunde liegenden Modells gewährleistet, bleibt sie auch im Netzwerk erhalten.

Wie das Beispiel einer Patchantenne zeigt, lässt sich der Projektionsansatz auch auf stark verlustbehaftete Systeme, z. B. in Verbindung mit einem PML-Rand, anwenden. Allerdings zeigt sich, dass der erste Schritt von TSL in vielen Fällen früh stagniert. Reduzierte Modelle müssen in diesem Fall direkt über Padé-basierte Approximationen generiert werden. Allerdings bietet gerade in diesem Fall der Zeitbereich eine effiziente Alternative und Makromodelle lassen sich über andere Verfahren wie 4SID aus den Zeitbereichsdaten gewinnen.

Sofern nur das Übertragungsverhalten der Struktur gesucht ist, können auch für resonante Systeme Methoden benutzt werden, die auf Zeitsignalen aufbauen. Ein eigener Abschnitt widmet sich Verfahren, die aus dem Zeitsignal die Koeffizienten eines digitalen Filters optimieren und daraus das Spektrum berechnen. Wenn das Zeitsignal noch nicht vollständig abgeklungen ist, weist dieses Vorgehen deutliche Vorteile gegenüber einer DFT auf. Neben den bereits häufig eingesetzten AR- und Prony-Filtern bewähren sich hierbei zwei iterative Algorithmen, das iterative Prony-Verfahren und das Verfahren nach Steiglitz-McBride. Insbesondere das zweite zeigt sich als sehr zuverlässig zur Schätzung von zum Teil sehr hohen Güten im Bereich von 10^6, wie am Beispiel eines neunzelligen TESLA-Resonators gezeigt wird.

Als Fazit bleibt festzustellen, dass der TSL-Algorithmus ein sehr robustes und vielseitiges Verfahren für verlustlose oder schwach verlustbehaftete resonante Strukturen bildet. Insbesondere zur Berechnung von Filtern ist TSL zur Zeit, was die Rechenzeiten angeht, konkurrenzlos. Für den klassischen Anwendungsfall, in dem das Port-Übertragungsverhalten gesucht ist, kann TSL immer eingesetzt und somit neben klassischen Zeit- und Frequenzbereichslösern als drittes alternatives Verfahren zur Berechnung angesehen werden. Darüberhinaus eröffnet TSL neue Möglichkeiten zur Feld-Netzwerk-Kopplung.

Als Ausblick sind weitere Anwendungsfälle und Erweiterungen für TSL denkbar. So kann das Verfahren zur schnellen Optimierung von Filterstrukturen eingesetzt werden. Auch die Verwendung des TSL-Makromodells zur Berechnung einer Teilstruktur innerhalb des Rechengebiets – alternativ zu bzw. in Verbindung mit Untergittern – könnte untersucht werden. Neben den Streuparametern oder Güten lassen sich auch weitere sekundäre Größen wie beispielsweise effektive Materialeigenschaf-

ten bei Meta-Materialien aus dem TSL-Modell gewinnen. Auch zusätzliche verallgemeinerte Ports, z.B. für Strahlimpedanzen, sind denkbar. Es erscheint weiterhin vielversprechend, TSL mit anderen Verfahren zur Ordnungsreduzierung, wie z.B. einer *Balanced Truncation*, zu verknüpfen. Die Reduzierung der Ordnung in Verbindung mit PML-berandeten Strukturen erweist sich bisher als ineffizient, was vor allem an der langsamen Konvergenz der verwendeten iterativen Gleichungslöser liegt. Hier könnte eine effektive Vorkonditionierung vermutlich Abhilfe schaffen. Letztlich wäre eine Formulierung oder zumindest eine Näherung, die die Passivität einer PML-berandeten Struktur im reduzierten Modell gewährleistet, sehr wünschenswert.

Anhang A

Der Bi-Lanczos-Algorithmus

Der folgende Bi-Lanczos-Algorithmus führt eine unsymmetrische (Block-)Tridiagonalisierung einer allgemeinen quadratischem Matrix $\mathbf{A}$ durch. Er verwendet eine Bandformulierung, die unterschiedliche Dimensionen der Startmatrizen $\mathbf{B}$ und $\mathbf{C}$ zulässt. Der Algorithmus orientiert sich an den in [64, 53] vorgeschlagenen Versionen, führt jedoch eine veränderte Orthogonalisierung durch. Auch die Behandlung komplexer Matrizen ist erweitert. Deflationierung und *Look-Ahead* werden an dieser Stelle nicht betrachtet, für eine ausführliche Darstellung siehe [53].

Abweichend zu den Angaben in Abschnitt 4.2.2.3 sind die Matrizen $\mathbf{V}$ und $\mathbf{W}$ orthogonal, aber nicht orthonormal definiert, es gilt $\mathbf{W}^T\mathbf{V} = \mathbf{\Delta} = \mathrm{diag}(\delta_i)$. Dies erleichtert die Erweiterung auf Deflationierung und *Look-Ahead*. Als Folge werden innerhalb des Algorithmus zwei Lanczos-Matrizen $\mathbf{T}$ und $\widetilde{\mathbf{T}}$ sowie $\mathbf{\Delta}$ genutzt. Die zur Reduzierung der Ordnung benötigten Matrizen $\mathbf{A}_p$, $\mathbf{B}_p$ und $\mathbf{C}_p$ können nach Ablauf des Algorithmus leicht aus $\mathbf{T}$, $\widetilde{\mathbf{T}}$ und $\mathbf{\Delta}$ extrahiert werden.

Der Algorithmus lautet im Einzelnen:

gegeben: $\mathbf{A}$, $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_m)$, $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_l)$, p

for $i = 1 : p + \max(m, l)$ — *Erzeuge Krylov- und Hilfsvektoren*

  if $i \leq m$ — *Neuer Vektor* $\mathbf{v}$

    $\mathbf{v}_i = \mathbf{b}_i$

  else

    $\mathbf{v}_i = \mathbf{A}_i \mathbf{v}_{i-m}$

  end

  $j_0 = \max(1, i - 2m)$ — *Biorthogonalisierung* $\mathbf{v}$ *zu* $\mathbf{w}$

  for $j = j_0 : i - 1$

    $t_{j,i} = \dfrac{\mathbf{w}_j^H \mathbf{v}_i}{\delta_j}$

    $\mathbf{v}_i = \mathbf{v}_i - t_{j,i} \mathbf{v}_j$

  end

$t_{i,i} = \|\mathbf{v}_i\|_2$
$\mathbf{v}_i = \mathbf{v}_i / t_{i,i}$

if $i \leq l$ — *Neuer Vektor* $\mathbf{w}$
  $\mathbf{w}_i = \mathbf{c}_i$
else
  $\mathbf{w}_i = \mathbf{A}_i^H \mathbf{v}_{i-l}$
end

$j_0 = \max(1, i - 2l)$ — *Biorthogonalisierung* $\mathbf{w}$ *zu* $\mathbf{v}$
for $j = j_0 : i - 1$
  $\tilde{t}_{j,i} = \dfrac{\mathbf{v}_j^H \mathbf{w}_i}{\delta_j^*}$
  $\mathbf{w}_i = \mathbf{w}_i - \tilde{t}_{j,i} \mathbf{w}_j$
end
$\tilde{t}_{i,i} = \|\mathbf{w}_i\|_2$
$\mathbf{w}_i = \mathbf{w}_i / \tilde{t}_{i,i}$

$\delta_i = \mathbf{w}_i^H \mathbf{v}_i$
end

$\mathbf{A}_p = [t_{1..p,m+1..m+p}]$ — *Matrizen des reduzierten Systems*
$\mathbf{B}_p = [t_{1..p,1..m}]$
$\mathbf{C}_p = [\widetilde{t}_{1..p,1..l}]^T \boldsymbol{\Delta}$

$\mathbf{V}_p = [\mathbf{v}_{1..p}]$ — $\mathbf{V}_p$ *und* $\mathbf{W}_p$ *werden üblicherweise nicht im*
$\mathbf{W}_p = [\mathbf{w}_{1..p}] \boldsymbol{\Delta}^{-\mathbf{1}}$ *Speicher gehalten!*

Symbolverzeichnis

Allgemeine Symbole und Schreibweisen

$\mathbb{N}$, $\mathbb{R}$	Menge der natürlichen bzw. reellen Zahlen
floor(x)	Nächstkleinere ganze Zahl
$\vec{X}$, $\mathbf{x}$	Vektordarstellungen
$\mathbf{M}$, $\mathbf{D}$, $\mathbf{I}$	Matrix, Diagonalmatrix, Einheitsmatrix
$\mathbf{U}$	Obere Dreiecksmatrix
$\mathbf{T}$, $\mathbf{H}$	(Block-)Tridiagonalmatrix, obere (Block-)Hessenberg-Matrix
x_u	Skalare Komponente eines Vektors $\mathbf{x}$
$\mathbf{M}_{kl}$	Element der Matrix $\mathbf{M}$
$\mathbf{x}^T, \mathbf{M}^T$	Transponierte eines Vektors bzw. einer Matrix
diag(x_k)	Diagonalmatrix mit den Einträgen x_k
eig($\mathbf{M}$)	Eigenwerte der Matrix $\mathbf{M}$
span($\mathbf{M}$)	Von den Vektoren der Matrix $\mathbf{M}$ aufgespannter Unterraum
rank($\mathbf{M}$)	Rang der Matrix $\mathbf{M}$
j	Imaginäre Einheit
x^*	Konjugiert komplexer Wert (auch Vektoren und Matrizen)
$\mathbf{x}^H, \mathbf{M}^H$	Konjugiert komplexe Transponierte eines Vektors bzw. einer Matrix
$\mathrm{Re}\{x\}, \mathrm{Im}\{x\}$	Real- und Imaginärteil einer komplexen Größe
∂A, ∂V	Berandung einer Fläche, eines Volumens
$\mathcal{O}$	LANDAU-Symbol zur Behandlung von Fehlerordnungen

Klassische Feldtheorie

$\vec{E}$	Elektrische Feldstärke
$\vec{D}$	Dielektrische Verschiebung
$\vec{H}$	Magnetische Feldstärke
$\vec{B}$	Magnetische Flussdichte
ρ	Elektrische Raumladungsdichte
$\vec{J}$, $\vec{J}_l$, $\vec{J}_k$, $\vec{J}_e$	Elektrische Stromdichte (gesamt, Leitungsstromdichte, Konvektionsstromdichte, eingeprägte Stromdichte)
k	Komplexe Wellenzahl
ε, $\overleftrightarrow{\varepsilon}$	Skalare bzw. tensorielle Permittivität
μ, $\overleftrightarrow{\mu}$	Skalare bzw. tensorielle Permeabilität
$\overleftrightarrow{\chi}_e$ $\overleftrightarrow{\chi}_m$	Tensorielle elektrische bzw. magnetische Suszeptibilität
$\vec{P}$, $\vec{M}$	Polarisation, Magnetisierung
κ	elektrische Leitfähigkeit

t	Zeit
λ	Wellenlänge
f, ω, s	Frequenz, Kreisfrequenz, Laplace-Variable

Methode der Finiten Integration

G, $\widetilde{G}$	Primäres bzw. duales Gitter
V_n, $\widetilde{V}_n$	Volumen einer primären bzw. dualen Gitterzelle
A_n, $\widetilde{A}_n$	Elementarfläche des primären bzw. dualen Gitters
L_n, $\widetilde{L}_n$	Elementarlänge des primären bzw. dualen Gitters
P_n, $\widetilde{P}_n$	Primärer bzw. dualer Gitterpunkt
I, J, K	Anzahl der Gitterebenen in u-, v- und w-Richtung
N_p	Anzahl der Gitterpunkte
Δ, $\Delta_{\{u,v,w\}}$	Gitterschrittweiten
Δt, Δt_{max}	Zeitschrittweite (allgemein, maximal)
$\widehat{\mathbf{e}}$, $\widehat{e}$	Gittervektor, Komponente der diskreten elektrischen Spannung
$\widehat{\widetilde{\mathbf{d}}}$, $\widehat{\widetilde{d}}$	Gittervektor, Komponente der diskreten elektrischen Flüsse
$\widehat{\widetilde{\mathbf{j}}}$, $\widehat{\widetilde{j}}$	Gittervektor, Komponente der diskreten elektrischen Ströme
$\widehat{\widetilde{\mathbf{j}}}_e$, $\widehat{\widetilde{j}}_e$	Gittervektor, Komponente der eingeprägten elektrischen Ströme
$\mathbf{q}$, q	Gittervektor, Komponente der diskreten elektrischen Ladungen
$\widehat{\mathbf{h}}$, $\widehat{h}$	Gittervektor, Komponente der diskreten magnetischen Spannung
$\widehat{\widetilde{\mathbf{b}}}$, $\widehat{\widetilde{b}}$	Gittervektor, Komponente der diskreten magnetischen Flüsse
$\mathbf{P}_{\{u,v,w\}}$	Diskrete partielle Differentiationsoperatoren
$\mathbf{C}$, $\widetilde{\mathbf{C}}$	Rotationsoperator auf dem primären bzw. dualen Gitter
$\mathbf{S}$, $\widetilde{\mathbf{S}}$	Divergenzoperator auf dem primären bzw. dualen Gitter
$-\widetilde{\mathbf{S}}^T$, $-\mathbf{S}^T$	Gradientenoperator auf dem primären bzw. dualen Gitter
$\mathbf{M}_\varepsilon$	Verknüpfungsmatrix zwischen elektrischen Flüssen und Spannungen
$\mathbf{M}_\mu$	Verknüpfungsmatrix zwischen magnetischen Flüssen und Spannungen
$\mathbf{M}_\kappa$	Verknüpfungsmatrix zwischen Leitungsströmen und elektrischen Spannungen
$\mathbf{K}_e$	Matrix zur Beschreibung der klassischen Impedanzrandbedingung
$\mathbf{K}_m$	Matrix zur Beschreibung der Impedanzrandbedingung mit magnetischer Leitfähigkeit
$c_{\{u,v,w\}}$	Komplexer Dämpfungsterm des PML-Mediums
$\Lambda_{\{u,v,w\}}$	Komplexer Materialtensor des PML-Mediums
$\widehat{\mathbf{e}}_{t,m}$, $\widehat{\mathbf{h}}_{t,m}$	Transversales Feldbild des Modes m
$\mathbf{N}_{\{u,v,w\}}$	Topologische Matrix zur Bildung des diskretes Kreuzprodukts
$\mathbf{b}_m$	Anregungsvektor des Ports m

Systemdarstellungen

u, $\mathbf{u}$	Spannung, Vektor der Portspannungen
i, $\mathbf{i}$	Strom, Vektor der Portströme
$\mathbf{H}$	Allgemeine Übertragungsfunktion
$\mathbf{Z}$	Impedanzmatrix
$\mathbf{Z}_L$	Diagonalmatrix der Leitungsimpedanzen
a, b	Wellenamplitude der ein- bzw. auslaufenden Welle
$\mathbf{a}$, $\mathbf{b}$	Vektoren der Wellenamplituden aller Ports
$\mathbf{S}$	Streumatrix
$\mathbf{x}$, $\mathbf{y}$	Zustandsvektoren
$\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$, $\mathbf{D}$	Matrizen der klassischen Zustandsraumdarstellung
$\mathbf{A}_l$	Systemmatrix eines linearen Systems
$\mathbf{A}_{CC}$, $\mathbf{A}'_{CC}$	Systemmatrix eines Curl-Curl-Systems mit bzw. ohne Berücksichtigung von $\mathbf{M}_\varepsilon$
$\mathbf{R}$	Matrix der Portanregungsvektoren
$\mathbf{K}$	Matrix der Verluste einer Curl-Curl-Darstellung
$\mathbf{P}$	Matrix der Impedanzwand-Verluste einer Curl-Curl-Darstellung
$\mathbf{W}_C$, $\mathbf{W}_O$	Gramsche Steuerbarkeits-, Beobachtbarkeitsmatrix
$\bar{\mathbf{Z}}$	Normierte Impedanz
$\mathbf{D}_\beta$	Diagonale Gitterkorrekturmatrix der Ports
$\delta\bar{Z}$, $\Delta\bar{Z}$	Relativer, absoluter Fehler der normierten Impedanz
$\delta\bar{S}$, $\Delta\bar{S}$	Relativer, absoluter Fehler der Streuparameter
Q_k	Güte der Resonanz k

Reduzierung der Ordnung

$\mathcal{K}(\mathbf{A}, \mathbf{b})$	Krylov-Unterraum der Matrix $\mathbf{A}$ und des Vektors $\mathbf{b}$
$P(\mathbf{A})$	Polynom einer Matrix $\mathbf{A}$
$\mathbf{V}$, $\mathbf{W}$	Projektionsmatrizen
$\hat{\mathbf{v}}$, $\hat{\mathbf{V}}$	Hilfsvektor, Hilfsvektorenmatrix
p	Dimensionen der Projektionsmatrizen einer partiellen Realisierung/erster Schritt TSL
q	Dimensionen der Projektionsmatrizen einer Padé-Approximation/ zweiter Schritt TSL
$\mathbf{T}_p$, $\mathbf{B}_p$, $\mathbf{C}_p$	Resultierende Matrizen des Lanczos-Algorithmus
$\mathbf{A}_p$, $\mathbf{B}_p$, $\mathbf{C}_p$	Zustandsraumdarstellung des projizierten Systems einer partiellen Realisierung
$\mathbf{A}_q$, $\mathbf{B}_q$, $\mathbf{C}_q$	Zustandsraumdarstellung des projizierten Systems einer Padé-Approximation
δ_Z, δ_E, δ_D	Impedanzfehler, Eigenwertfehler, Eigenwertdifferenz
ε_Z, ε_{eig}, $\varepsilon_{\delta\text{eig}}$	Fehlertoleranz der Impedanz, des Eigenwerts, der Eigenwertdifferenz
r_k	Wurzeln eines Tschebyscheff-Beschleunigungspolynoms
$\mathbf{X}$	Matrix der Eigenvektoren
$\mathbf{\Lambda}$	Diagonalmatrix der Eigenwerte
$\mathbf{Z}_{\text{corr}}$	Impedanz-Korrekturmatrix

Spektralschätzung

$y^{(g)}$	Abgetastete Signalfolge zum Zeitpunkt $g\Delta t$
z	z-Bereichsvariable
$H(z)$	Übertragungsfunktion eines digitalen Filters
$h^{(g)}$	Impulsantwort des Filters $H(z)$
$A(z)$, $B(z)$	Polynome in z, Nenner und Zähler von $H(z)$
$\delta^{(g)}$	Diskrete Dirakfolge
$e^{(g)}$, $E(z)$	Fehler im diskreten Zeit- und z-Bereich

Netzwerktheorie

$\mathbf{G}$	Admittanzmatrix in der Knotenanalyse
$\mathbf{G}_L$, $\mathbf{G}_R$, $\mathbf{G}_C$	Symmetrische Matrizen mit den Beiträgen der Induktivitäten, Widerstände und Kapazitäten zu $\mathbf{G}$
$\mathbf{G}_{CR}$	Unsymmetrischer Anteil von $\mathbf{G}$
$\mathbf{F}$	Diagonalmatrix zur Symmetrisierung eines Curl-Curl-Systems
$\ddot{u}$	Übersetzungsverhältnis eines idealen Übertragers
$\mathbf{K}_{r\mathbf{V}}$, $\mathbf{K}_{r\mathbf{X}}$	Verlustmatrizen nach drittem Lanczosschritt bzw. Eigenwertzerlegung

Abkürzungsverzeichnis

4SID	Subspace-based State-Space System Identification
BEM	Boundary Elements Method
AWE	Asymptotic Waveform Evaluation
DFT	Diskrete Fouriertransformation
FDTD	Finite Difference Time Domain
FE, FEM	Finite Element (Method)
FIT	Finite Integration Technique
GAWE	Galerkin Asymptotic Waveform Evaluation
MWS	MICROWAVE STUDIO®
PML	Perfectly Matched Layers
PRIMA	Passive Reduced Interconnect Modelling Algorithm
PVL	Padé Via Lanczos
SPICE	Simulation Program with Integrated Circuit Emphasis
TESLA	Tera Electron Volt Superconducting Linear Accelerator
TSL	Two-Step Lanczos
WCAWE	Well-Conditioned Asymptotic Waveform Evaluation

Literaturverzeichnis

Klassische Feldtheorie

[1] J. C. Maxwell: *A Treatise on Electricity and Magnetism*, 2 Bände, Oxford University Press, London, 1. Auflage, 1873.

[2] J. D. Jackson: *Klassische Elektrodynamik*, Walter de Gruyter Verlag, Berlin, New York, 1983.

[3] K. Simonyi: *Theoretische Elektrotechnik*, J. A. Barth, Deutscher Verlag der Wissenschaften, Leipzig, Berlin, 10. Auflage, 1993.

[4] R. E. Collin: *Field Theory of Guided Waves*, McGraw-Hill, New York, 1960.

[5] G. Matthaei, L. Young, E.M.T. Jones: *Micro Wave Filters, Impedance-Matching Networks, and Coupling Structures*, Artech House, 1980.

Numerische Feldberechnung und Methode der Finiten Integration

[6] P. K. Banerjee, R. Butterfield: *Boundary Element Methods in Engineering Science*, McGraw-Hill Book Company (UK) Limited, London, New York, St Louis u.a., 1981.

[7] O. C. Zienkiewitz: *The Finite Element Method*, McGraw-Hill Book Company Maidenhead, 1977.

[8] T. Weiland: *Eine Methode zur Lösung der Maxwellschen Gleichungen für sechskomponentige Felder auf diskreter Basis*, Archiv für Elektronik und Übertragungstechnik (AEÜ), Band 31, S. 116-120, 1977.

[9] T. Weiland: *Elektromagnetisches CAD - Rechnergestützte Methoden zur Berechnung von Feldern*, Skriptum zur Vorlesung, Technische Universität Darmstadt, 2001.

[10] T. Weiland: *Time Domain Electromagnetic Field Computation with Finite Difference Methods*, International Journal of Numerical Modelling, Band 9, S. 295-319, 1996.

[11] S. Gutschling: *Zeitbereichsverfahren zur Simulation elektromagnetischer Felder in dispersiven Materialien*, Dissertation D17, Technische Universität Darmstadt, 1998.

[12] H. Krüger: *Zur numerischen Berechnung transienter elektromagnetischer Felder in gyrotropen Materialien*, Dissertation D17, Technische Universität Darmstadt, 2000.

[13] M. Dohlus: *Ein Beitrag zur numerischen Berechnung elektromagnetischer Felder im Zeitbereich*, Dissertation D17, Technische Universität Darmstadt, 1992.

[14] R. Schuhmann: *Die nichtorthoganale Finite-Integrations-Methode zur Simulation elektromagnetischer Felder*, Dissertation D17, Technische Universität Darmstadt, 1999.

[15] T. Weiland, R. Schuhmann, R. B. Greegor, C. G. Parazzoli, A. M. Vetter, D. R. Smith, D. C. Vier, S. Schultz: *Ab Initio Numerical Simulation of Left-Handed Metamaterials*, Journal of Applied Physics, Band 90, Nr.10, S. 5419-5424, 2001.

[16] K. S. Yee: *Numerical Solution of Initial Boundary Value Problems Involving Maxwell's Equations in Isotropic Media*, IEEE Transactions on Antennas and Propagation, Band 14, Nr. 3, S. 302-307, 1966.

[17] M. Hilgner: *Zur Generierung problemangepaßter Gitter zur Berechnung elektromagnetischer Felder mit der Methode der Finiten Integration*, Dissertation D17, Technische Universität Darmstadt, 2000.

[18] D. Reinecke: *Zur Simulation dünner, perfekt elektrisch leitender Schichten durch die Methode der Finiten Integration*, Dissertation D17, Technische Universität Darmstadt, 2002.

[19] H. Spachmann: *Die Methode der Finiten Integration höherer Ordnung zur numerischen Berechnung elektromagnetischer Felder*, Dissertation D17, Technische Universität Darmstadt, 2003.

[20] P. Thoma: *Zur numerischen Lösung der Maxwellschen Gleichungen im Zeitbereich*, Dissertation D17, Technische Universität Darmstadt, 1997.

[21] G. Mur: *Absorbing Boundary Conditions for the Finite-Difference Approximation of the Time-Domain Electromagnetic Field Equations*, IEEE Transactions on Electromagnetic Compatibility, Band EMC-23, S. 377-382, 1981.

[22] J.-P. Bérenger: *A Perfectly Matched Layer for the Absorption of Electromagnetic Waves*, Journal of Computational Physics, Band 114, S. 185-200, 1994.

[23] Z. S. Sacks, D. M. Kingsland, R. Lee, J. F. Lee: *A Perfectly Matched Anisotropic Absorber for Use as an Absorbing Boundary Condition*, IEEE Transactions on Antennas and Propagation, Band 43, Nr. 12, S. 1460-1463, 1995.

[24] L. Zhao, A. C. Cangellaris: *GT-PML: Generalized Theory of Perfectly Matched Layers and Its Application to the Reflectionless Truncation of Finite-Difference Time-Domain Grids*, IEEE Transactions on Microwave Theory and Techniques, Band 44, Nr. 12, S. 2555-2562, 1996.

[25] W. C. Chew, J. M. Jin, E. Michielssen: *Fast and Efficient Algorithms in Computational Electromagnetics*, Artech House, Boston, London, 2001.

[26] P. Hammes: *Zur numerischen Berechnung von Streumatrizen im Hochfrequenzbereich*, Dissertation D17, Technische Universität Darmstadt, 1999.

[27] B. Trapp: *Zur numerischen Berechnung hochfrequenter elektromagnetischer Felder auf der Basis von Eigenlösungen*, Dissertation D17, Technische Universität Darmstadt, 2002.

[28] T. Weiland: *A Numerical Method for the Solution of the Eigenwave Problem of Longitudinally Homogenous Waveguides*, Archiv für Elektronik und Übertragungstechnik (AEÜ), Band 31, S. 308-314, 1977.

[29] F. Hănţilă, D. Ioan: *Voltage-Current Relation of Circuit Elements with Field Effects*, Proccedings of the 6th International IGTE Symposium, Graz, Österreich, S. 41-46, 1994.

[30] CST GmbH: *MICROWAVE STUDIO*, Bad Nauheimer Strasse 19, 64289 Darmstadt.

[31] T. Weiland: *On the Unique Numerical Solution of Maxwellian Eigenvalueproblems in Three Dimensions*, Particle Accelerators, Band 17, S. 227-242, 1985.

[32] P. Thoma, T. Weiland: *Numerical Stability of Finite Difference Time Domain Methods*, IEEE Transactions on Magnetics, Band 34, Nr. 5, S. 2740-2743, 1998.

[33] R. Schuhmann, T. Weiland: *Rigorous Analysis of Trapped Modes in Accelerating Cavities*, Physical Review Special Topics - Accelerators and Beams, Band 3, 122002, http://publish.aps.org/abstract/PRSTAB/v3/e122002, 2000.

Systemtheorie

[34] R. Unbehauen: *Systemtheorie*, Oldenbourg Verlag, München, 1990.

[35] C. T. Chen: *Linear System Theory and Design*, Oxford University Press, New York, 1999.

[36] G. Newstead: *General Circuit Theory*, The Broadwater Press Ltd, Welwyn Garden City, 1959.

[37] R. A. Rohrer, H. Nosrati: *Passivity Considerations in Stability Studies of Numerical Integration Algorithms*, IEEE Transactions on Circuits and Systems, Band CAS-28, S. 857866, 1981.

[38] B. D. O. Anderson, S. Vongpanidlerd: *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, New Jersey, 1973.

[39] R. W. Newcomb: *Linear Multiport Synthesis*, McGraw-Hill, New York, 1966.

[40] F. W. Fairman: *Linear Control Theory: The State Space Approach*, John Wiley, West Sussex, England, 1998.

[41] O. Föllinger: *Regelungstechnik*, Hüthig Verlag, Heidelberg, 1990.

Mathematische Grundlagen

[42] R. Zurmühl: *Matrizen*, Spinger Verlag, Berlin, Göttingen, Heidelberg, 1958.

[43] G. H. Golub, C. F. Van Loan: *Matrix computations*, The Johns Hopkins University Press, Baltimore, 3rd edition, 1996.

[44] L. N. Trefethen, D. Bau: *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[45] Y. Saad: *Numerical Methods for Large Eigenvalue Problems: Theory and Algorithms*, John Wiley, New York, 1992.

[46] T. J. Rivlin: *Chebychev Polynomials*, Wiley, New York, 1976.

[47] D. Arnold, R. Falk, R. Winther: *Multigrid in H(Div) and H(Curl)*, Numerische Mathematik, Band 85, S. 197-218, 2000.

[48] C. Lanczos: *An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators*, J. Research of the National Bureau of Standards, Band 45(4), S. 255-282, 1950.

[49] M. H. Gutknecht, Z. Strakos: *Accuracy of Two Three-Term and Three Two-Term Recurrences for Krylov Space Solvers*, SIAM Journal on Matrix Analysis and Applications, Band 22, Nr 1, S. 213-229, 2000.

[50] W. E. Arnoldi: *The Principle of Minimized Iteration in the Solution of the Matrix Eigenvalue Problem*, Quarterly of Applied Mathematics, Band 9, S. 17-29, 1951.

[51] G. A. Baker, P. Graves-Morris: *Padé Approximants*, 2 Bände, Addison-Wesley Publishing Company, Inc., Massachusetts, 1981.

[52] J. K. Cullum, R. A. Willoughby: *Lanczos algorithms for large symmetric eigenvalue computations, Volume 1: Theory*, Birkhäuser, Boston, 1985.

[53] J. I. Aliaga, D. L. Boley, R. W. Freund, V. Hernandez: *A Lanczos-type method for multiple starting vectors*, Mathematics of Computation, Band 69, S. 1577-1601, 2000.

[54] D. C. Sorensen: *Implicit Application of Polynomial Filters in a k-Step Arnoldi Method*, SIAM Journal on Matrix Analysis and Applications, Band 13, S. 357-385, 1992.

[55] P. Wesseling: *An Introduction to Multigrid Methods*, John Wiley & Sons Chichester, New York, u.a., 1991.

Reduzierung der Modellordnung

[56] L. Zhao, A. C. Cangellaris: *Reduced-Order Modeling of Elektromagnetic Field Interactions in Unbounded Domains Truncated by Perfectly Matched Layers*, Microwave and Optical Technology Letters, Band 17, Nr. 1, S. 62-66, 1998.

[57] W. B. Gragg, A. Lindquist: *On the Partial Realization Problem*, Linear Algebra and its Applications, Band 50, S. 227-319, 1983.

[58] E. J. Grimme, D. C. Sorensen, P. Van Dooren: *Model Reduction of State Space Systems via an Implicitly Restarted Lanczos Method*, Technical Report CRPC-TR94458, Rice Univ., Houston, TX 77251-1892, USA, 1994.

[59] E. J. Grimme: *Krylov Projection Methods for Model Reduction*, PhD-Thesis, University of Illinois, 1977.

[60] T. Zhou, S. L. Dvorak, J. L. Prince: *Application of the Padé via Lanczos (PVL) algorithm to electromagnetic systems with expansion at infinity*, Proc. of 2000 Electronic Components and Technology Conference, S. 1515-1520, 2000.

[61] Y. Shamash: *Linear System Reduction Using Padé Approximation to Allow Retention of Dominant Modes*, International Journal of Control , Band 21, S. 257-272, 1975.

[62] L. T. Pillage, R. A. Rohrer: *Asymptotic Waveform Evaluation for Timing Analysis*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Band 9, Nr. 4, S. 352-366, 1990.

[63] P. Feldmann, R. W. Freund: *Efficient Linear Circuit Analysis by Padé Approximation via the Lanczos Process*, Proc. of EURO-DAC 94, EURO-VHDL 94, IEEE Computer Society Press, S. 170-175, 1994.

[64] R. W. Freund: *Krylov-Subspace Methods for Reduced-Order Modeling in Circuit Simulation*, Journal of Computational and Applied Mathematics, Band 123, S. 395-421, 2000.

[65] A. Odabasioglu, M. Celik, L. T. Pileggi: *PRIMA: Passive Reduced-Order Interconnect Macromodeling Algorithm*, IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Band 17, Nr. 8, S. 645-654, 1998.

[66] A. C. Cangellaris, L. Zhao: *Rapid FDTD Simulation without Time Stepping*, Microwave and Guided Wave Letters, Band 9, Nr.1, S. 4-6, 1999.

[67] B. C. Moore: *Principal Component Analysis in Linear Systems: Controllability, Observability and Model Reduction*, IEEE Transactions on Automatic Control, Band 26, S. 17-32, 1981.

[68] D. Schmitt, R. Schuhmann, T. Weiland: *The Complex Subspace Iteration for the Computation of Eigenmodes in Lossy Cavities*, International Journal of Numerical Modelling, Band 8, S. 385-398, 1995.

[69] M. Dohlus, R. Schuhmann, T. Weiland: *Calculation of Frequency Domain Parameters Using 3D Eigensolutions*, International Journal of Numerical Modelling, Band 12, S. 41-68, 1999.

[70] R. Schuhmann, T. Weiland: *Recent Advances in Finite Integration Technique for High Frequency Applications*, in: W. H. A. Schilders, E. J. W. ter Maten, S. H. M. J. Houben (eds.): Scientific Computing in Electrical Engineering (Mathematics in Industry, The European Consortium for Mathematics in Industry, Vol. 4), Springer, Berlin, ISBN 3-540-21372-4, S. 46-57, 2004.

[71] Z. Bai, P. Feldmann, R. W. Freund: *Stable and Passive Reduced-Order Models Based on Partial Padé Approximation Via the Lanczos Process*, Bell Laboratories, Numerical Analysis Manuscript 97-3-10, http://cm.bell-labs.com/cm/cs/doc/97 3-10.ps.gz, 1997.

[72] R. W. Freund, P. Feldmann: *Reduced-Order Modeling of Large Linear Passive Multi-Terminal Circuits Using Matrix-Padé Approximation*, Proceedings of the Design Automation and Test in Europe Conference, IEEE Computer Society Press, S. 530-537, 1998.

[73] Z. Bai, Q. Ye: *Error Estimation of The Padé Approximation of Transfer Functions Via The Lanczos Process*, Electronic Transactions on Numerical Analysis. Band 7, S. 1-17, 1998.

[74] I.M. Elfadel, D.D. Ling: *A Block Rational Arnoldi Algorithm for Multipoint Passive Model-Order Reduction of Multiport RLC Networks*, IEEE Computer-Aided Design, Digest of Technical Papers, 66-71, 1997.

[75] B. N. Sheehan *ENOR: Model Order Reduction of RLC Circuits Using Nodal Equations for Efficient Factorization*, Proceedings of the 36th ACM/IEEE Conference on Design Automation, S. 17-21, 1999.

[76] R. D. Slone, R. Lee, J.-F. Lee: *Multipoint Galerkin asymptotic waveform evaluation for model order reduction of frequency domain FEM electromagnetic radiation problems*, IEEE Transactions on Antennas and Propagation, Band 49, Nr. 10, S. 1504-1513, 2001.

[77] E. Chiprout, M. S. Nakhla: *Analysis of Interconnect Networks Using Complex Frequency Hopping (CFH)*, IEEE Transactions on Computer-Aided Design of Circuits and Systems, Band 14, Nr. 2, S. 186-200, 1995.

[78] R. D. Slone, R. Lee, J. F. Lee: *Well-Conditioned Asymptotic Waveform Evaluation for Finite Elements*, IEEE Transactions on Antennas and Propagation, Band 51, Nr. 9, S. 2442-2447, 2003.

[79] K. Krohne, R. Vahldieck: *A Fast Filter Optimization Scheme Based on Model Order Reduction*, IEEE MTT-S International Microwave Symposium Digest, IEEE Microwave Theory and Techniques Society, Philadelphia, PA, USA, S. 21-24, 2003.

[80] L. Knockaert, D. De Zutter: *Passive Reduced Order Multiport Modeling: The Padé-Laguerre, Krylov-Arnoldi-SVD Connection*, Archiv für Elektronik und Übertragungstechnik (AEÜ), Band 53, Nr. 5, S. 254-260, 1999.

[81] M. Clemens, M. Wilke, R. Schuhmann, T. Weiland: *Subspace Projection Extrapolation Scheme for Transient Field Simulations*, Proceedings of the Conference on the Computation of Magnetic Fields (COMPUMAG), Saratoga Springs, USA, S. 10-11, 2003.

Signalverarbeitung

[82] C. W. Therrien, C. H. Velasco: *An Iterative Prony Method for ARMA Signal Modelling*, IEEE Transactions on Signal Processing, Band 43, Nr. 1, S. 358-361, 1995.

[83] M. Elsayed: *Iterative Prony-Verfahren*, Studienarbeit an der Universität Erlangen, 1997.

[84] K. Steiglitz, L. E. McBride: *A Technique for the Identification of Linear Systems*, IEEE Transactions on Automatoic Control, Band AC-10, S. 461-464, 1965.

[85] F. Richard, j. R. Schneider, D. Trines, A. Wagner: *TESLA – Technical Design Report, Part I – Executive Summary*, TESLA Report 2001-23, DESY, 2001.

[86] I. Munteanu, D. Ioan: *A Survey on Parameter Extraction Techniques for Coupling Electromagnetic Devices to Electric Circuits*, U. v.Rienen, M. Günther, D. Hecht (eds.): Scientific Computing in Electrical Engineering (Lecture Notes in Computational Sciences and Engineering, Vol. 18), Springer, Berlin, S. 337-357, 2001.

[87] F. G. Canavero, S. Grivet-Talocia, H. Krüger, I. A. Maio, I. S. Stievano, P. Thoma: *On the Modeling of Packages from Their Time Responses*, VI Workshop on Signal Propagation on Interconnects (SPI), Castelvecchio Pascoli (LU), Italien, S. 69-72, 2002.

Netzwerksimulation

[88] M. Witting: *Simulation elektrischer Netzwerke unter Berücksichtigung ihrer elektromagnetischen Umgebung*, Dissertation D17, Technische Universität Darmstadt, 1997.

[89] C. W. Ho, A. E. Ruehli, P. A. Brennan: *The Modified Nodal Approach to Network Analysis*, IEEE Transactions on Circuits and Systems, Band CAD-25, S. 504-509, 1975.

[90] A. Paech: *Transiente Simulation von Ausbreitungsvorgängen auf On-Chip Spannungsversorgungen mit frequenzabhängigen Parametern*, Diplomarbeit an der TU-Darmstadt, Fachbereich 18, 2002.

[91] I. J. Craddock, C. J. Railton, J. P. McGeehan: *Derivation and application of a passive equivalent circuit for the finite difference time domain algorithm*, IEEE Microwave and Guided Wave Letters, Band 6, Nr. 1, S. 40-42, 1996.

[92] H. Clausert, G. Wiesemann: *Grundgebiete der Elektrotechnik*, 2 Bände, Oldenbourg Verlag München, Wien, 1988.

[93] L.W. Nagel: *SPICE: A Computer Program to Simulate Semiconductor Circuits*, Ph.D. Thesis, Univ. of California, Berkeley, 1975.

[94] I. Munteanu, D. Ioan: *Equivalent Circuits for Reduced-Order Models of Electromagnetic Devices*, Technical report no. 21/2001, Numerical Methods Laboratory, 'Politehnica' University of Bucharest, 2001.

[95] T. Wittig: *Implementierung eines Filtersyntheseverfahrens zur Weiterverarbeitung numerischer Simulationsergebnisse*, Diplomarbeit an der TU-Darmstadt, Fachbereich 18, 1998.

[96] T. Mangold, P. Russer: *Full-Wave Modelling and Automatic Equivalent-Circuit Generation of Millimeter-Wave Planar and Multilayer Structures*, IEEE Transactions on Microwave Theory and Techniques, Band 47, S. 851-858, 1999.

Betreute Diplomarbeiten im Rahmen dieser Arbeit

[97] B. Trabert: *Spektralschätzung resonanter Zeitsignale in der numerischen Feldsimulation*, Diplomarbeit an der TU-Darmstadt, Fachbereich 18, 2001.

[98] K. Krohne: *Verfahren zur Reduzierung der Modellordnung FIT diskretisierter elektromagnetischer Strukturen*, Diplomarbeit an der TU-Darmstadt, Fachbereich 18, 2002.

Konferenz- und Zeitschriftenbeiträge im Rahmen dieser Arbeit

[99] I. Munteanu, T. Wittig, T. Weiland, D. Ioan: *FIT/PVL Circuit Parameter Extraction for General Electromagnetic Devices*, IEEE Transactions on Magnetics, Band 36, Nr. 4, S. 1421-1425, 2000.

[100] T. Wittig, I. Munteanu, D. Ioan, T. Weiland: *Reduction of Electromagnetic Circuit Elements Based on FIT Discretization*, Proceedings of the 2000 USNC/URSI National Radio Science Meeting, Salt Lake City, USA, S. 271, 2000.

[101] T. Wittig, T. Weiland, F. Hirtenfelder, W. Eurskens: *Efficient Parameter Extraction of High-Speed IC-Interconnects Based on 3D Field Simulations Using FIT*, Proceedings of the 14th International Zurich Symposium and Technical Exhibition on Electromagnetic Compatibility (EMC 2001), Zürich, Schweiz, S. 281-286, 2001.

[102] T. Wittig, I. Munteanu, R. Schuhmann, T. Weiland: *Model Order Reduction with a Two-Step Lanczos Algorithm*, IEEE Transactions on Magnetics, Band 38, Nr. 2, S. 673-676, 2002.

[103] T. Wittig, I. Munteanu, R. Schuhmann, T. Weiland: *Efficient Parameter Extraction Based on 3D Field Simulation*, Proceedings of the Fourth International Workshop on Computational Electromagnetics in the Time Domain (CEM-TD), Nottingham, S. 209-214, 2001.

[104] M. Clemens, T. Wittig, T. Weiland: *Numerische Feldsimulation mit der Methode der Finite Integration - Chancen und Grenzen*, tm - Technisches Messen, Sonderausgabe zu „EMV“, S. 90-102, 2002.

[105] R. Schuhmann, T. Wittig, I. Munteanu, B. Trapp, T. Weiland: *Time and Frequency Domain Simulations of Highly Resonant RF-Devices*, Proceedings of the Int. Conference on Electromagnetics in Advanced Applications (ICEAA 01), Turin, S. 125-128, 2001.

[106] T. Wittig, I. Munteanu, R. Schuhmann, T. Weiland: *Model Order Reduction and Equivalent Network Generation for a FIT Curl-Curl Formulation*, Proceedings of the 18th Annual Review of Progress in Applied Computational Electromagnetics, ACES 2002, Monterey, USA, S. 265-272, 2002.

[107] T. Wittig, I. Munteanu, R. Schuhmann, T. Weiland: *Model Order Reduction and Equivalent Circuit Extraction for FIT Discretized Electromagnetic Systems*, International Journal of Numerical Modelling, Band 15, S. 517-533, 2002.

[108] T. Wittig, R. Schuhmann, T. Weiland: *Ordnungsreduzierte Modelle auf Basis von FIT-Diskretisierungen*, Tagungsband der Kleinheubacher Tagung, Miltenberg, Deutschland, S. 28, 2003.

[109] T. Wittig, R. Schuhmann, T. Weiland: *Efficient Model Order Reduction of FIT Discretized Electromagnetic Structures for Field-Circuit Coupled Problems*, Proceedings of the Workshop on Optimization and Coupled Problems in Electromagnetism, Neapel, Italien, 2003.

[110] T. Wittig, R. Schuhmann, T. Weiland: *Fast Full Wave Simulation of Interconnects based on FIT and Model Order Reduction*, Proceedings of the Progress in Electromagnetics Research Symposium (PIERS), Pisa, Italien, S. 4, 2004.

[111] T. Wittig, R. Schuhmann, T. Weiland: *Efficient Model Order Reduction Based on a Two-Step Lanczos Approach*, Digest of the International Microwave Symposium (IEEE MTT-S), Forth Worth, USA, 2004.

[112] T. Wittig, R. Schuhmann, T. Weiland: *Model Order Reduction for Large Systems in Computational Electromagnetics*, Linear Algebra and Its Applications, Sonderausgabe „Order Reduction of Large-Scale Systems“, im Druck, 2004.

Danksagung

An dieser Stelle möchte ich mich bei all denen bedanken, die in den vergangenen Jahren zum Gelingen der vorliegenden Arbeit beigetragen haben.
Insbesondere gilt mein Dank

- Herrn Prof. Dr.-Ing. Thomas Weiland für die wissenschaftliche Betreuung und die ausgezeichneten Arbeitsmöglichkeiten am Institut,
- Herrn Prof. Dr. Wil Schilders für die freundliche Übernahme des Korreferates und sein Interesse an meiner Arbeit,
- Frau Prof. Dr.-Ing. Irina Munteanu und Herrn Dr.-Ing. Rolf Schuhmann für die kompetente und freundschaftliche Betreuung, ihnen sowie den Herren Dipl.-Ing. Denis Sievers und Dipl.-Ing. Robert Oestreich für die gewissenhafte Durchsicht des Manuskripts und die vielen wertvollen Anmerkungen und Diskussionen,
- den Herren Dipl.-Ing. Franz Hirtenfelder, Dipl.-Ing. Wigand Koch und Dr. techn. Stefan Reitzinger der CST GmbH für die freundliche Überlassung von Anwendungsbeispielen und zahlreiche hilfreiche Hinweise zur Arbeit,
- allen meinen ehemaligen und jetzigen Kolleginnen und Kollegen am Institut für Theorie Elektromagnetischer Felder der TU Darmstadt für ihre Hilfsbereitschaft und gute Zusammenarbeit, durch die sie mir eine schöne Promotionszeit in einem hervorragenden und freundschaftlichen Arbeitsklima ermöglicht haben,
- der Gesellschaft für Schwerionenforschung (GSI) Darmstadt für die finanzielle Unterstützung meiner Promotion,
- meiner Familie und meinen Freunden, die mich jederzeit unterstützt und ermuntert haben.

Wissenschaftlicher Werdegang

Tilmann Wittig
geb. am 12.05.1972

1978-1982	Grundschulzeit an der Hirschbachschule in Reinheim-Georgenhausen
1982-1984	Förderstufe an der Dr.-Kurt-Schumacher-Schule in Reinheim
1984-1991	Besuch des Gymnasiums Albert-Einstein-Schule in Groß-Bieberau, Abschluss mit Abitur
1991-1992	Zivildienst bei der Reinheimer Bürgergemeinschaft für Behinderte
1992-1994	Grundstudium der Elektrotechnik an der Technischen Universität Darmstadt
1996-1997	Studienaustausch an die Heriot-Watt University Edinburgh, Schottland zur Anfertigung der Studienarbeit
1994-1998	Hauptstudium mit Vertiefungsfach Nachrichtentechnik an der Technischen Universität Darmstadt, Abschluss mit Diplom (Dipl.-Ing.)
1998-2003	Wissenschaftlicher Mitarbeiter am Institut für Theorie Elektromagnetischer Felder des Fachbereichs Elektrotechnik und Informationstechnik der Technischen Universität Darmstadt