Fehlerschätzer und Fehlerindikatoren für Modellordnungsreduktionsverfahren in der Finite-Elemente-Simulation elektromagnetischer Felder im Frequenzbereich

Dissertation
zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Naturwissenschaftlich-Technischen Fakultät
der Universität des Saarlandes

von

Yves Konkel

Saarbrücken

2018

Tag des Kolloquiums:	27. Juli 2018
Dekan:	Univ.-Prof. Dr. rer. nat. Guido Kickelbick
Berichterstatter:	Univ.-Prof. Dr. techn. Romanus Dyczij-Edlinger Univ.-Prof. Dr.-Ing. Michael Möller
Vorsitz:	Univ.-Prof. Dr. rer. nat. Rolf Pelster
Akad. Mitarbeiter:	Dr.-Ing. Joachim Schmitt

Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus anderen Quellen oder indirekt übernommenen Daten und Konzepte sind unter Angabe der Quelle gekennzeichnet.

Die Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem Verfahren zur Erlangung eines akademischen Grades vorgelegt.

Ich erkläre hiermit an Eides statt, dass die vorliegende Arbeit mit der elektronischen Version übereinstimmt. Ich erkläre darüber hinaus mit meiner Unterschrift, dass ich

- keine im Merkblatt "Hinweise zur Vermeidung von Plagiaten"der Naturwissenschaftlich-Technischen Fakultät beschriebene Form des Plagiats begangen habe,
- alle Methoden, Daten und Arbeitsabläufe wahrheitsgetreu dokumentiert habe und
- keine Daten manipuliert habe.

Aying, 25. April 2018

Kurzfassung

Die vorliegende Arbeit entwickelt auf Grundlage der Methode der finiten Elemente automatiserte Modellordnungsreduktionsverfahren, die eine schnelle und zuverlässige Charakterisierung passiver Mikrowellenstrukturen in Abhängigkeit der Frequenz ermöglichen. Es werden Strategien und Abbruchkriterien für Mehrpunkt-Verfahren untersucht, die sich auf Anregungsprobleme und Wellenleiterformulierungen anwenden lassen.

Das Hauptergebnis dieser Arbeit ist ein effizient auszuwertender a-posteriori-Fehlerschätzer für Einpunkt-Verfahren im Zusammenhang mit verlustlosen Systemen. Hierbei werden die Komponenten des ordnungsreduzierten Modells in zwei komplementäre Anteile aufgespalten: der erste Anteil ist aufgrund der Krylov-Unterraum-Iterationen hinreichend exakt bestimmt, der zweite Anteil kann mit Hilfe des bekannten Konvergenzverhaltens der Krylov-Verfahren abgeschätzt werden. Durch Wiederverwendung von Zwischenergebnissen und Anwendung schneller Auswerteverfahren kann der Rechenaufwand für den Fehlerschätzer im Verhältnis zum gesamten Modellordnungsprozess gering gehalten werden.

Anwendungsbeispiele belegen die Zuverlässigkeit und Effizienz der vorgestellten Ansätze.

Abstract

The present work develops automated model-order reduction techniques, which enable a fast and reliable characterization of passive frequency-dependent microwave structures. The investigated strategies and stopping criteria for multipoint methods are applicable to driven problems as well as waveguide formulations.

The main result of this work is given by an efficient a posteriori error estimator for single-point methods in combination with lossless systems. The fundamental idea of the presented approach lies in the separation of the components of the reduced model into two complementary parts: the first part is sufficiently approximated due to Krylov subspace iterations and seen to be exact. For the second part a lower bound of the error can be estimated using well-known convergence characteristics of Krylov subspace methods. Due to re-utilization of intermediate results and application of fast evaluation methods, the overall computation costs are low compared to the expense of the entire model order reduction process.

Real-world examples demonstrate the reliability and efficiency of the methods presented.

Inhaltsverzeichnis

1 Einleitung **1**

2 Modellierung passiver Mikrowellenstrukturen **5**

2.1 Passive Mikrowellenstrukturen 5

2.2 System der Maxwell-Gleichungen 7

2.3 Die vektorielle Helmholtz-Gleichung 8

2.4 Randbedingungen 9

2.4.1 Ideal elektrische Leiter 9

2.4.2 Ideal magnetische Leiter 10

2.4.3 Absorbierende Randbedingung 10

2.4.4 Tor-Randbedingung 11

2.5 Homogene Wellenleiter 11

2.5.1 Grundlegende Eigenschaften homogener Wellenleiter 11

2.5.2 Eigenwertproblem für homogene Wellenleiter 12

2.6 Starke Formulierung des Randwertproblems für Mikrowellenstrukturen 16

2.7 Schaltung von Mikrowellenstrukturen auf Systemebene 17

3 Finite-Elemente-Methode für Anregungsprobleme **21**

3.1 Funktionenräume und schwache Form 22

3.2 Finite-Elemente-Ansatz 25
3.2.1 Triangulierung des Feldgebiets 25
3.2.2 Finite-Elemente-Ansatzfunktionen 26
3.3 Transfinite-Elemente-Methode 27
3.4 Die FE-Formulierung als LTI-System 30
3.4.1 TE- und TM-Wellen im FE-System 32

4 Schnelle Frequenzgangberechnung linearer Systeme 35
4.1 Projektionsbasierte Modellordnungsreduktion 37
4.1.1 Mehrpunktverfahren 39
4.1.2 Einpunktverfahren 40
4.2 Einpunktverfahren für linear parametrierte Systeme 42
4.2.1 Asymptotic Waveform Evaluation 43
4.2.2 Krylov-Unterraum-Verfahren 44
4.2.3 Residuenberechnung im Arnoldi-Verfahren 48
4.2.4 Schnelle Auswertung von Ausgangsgröße und Residuum 49
4.2.5 Shift- und Invert-Vorkonditionierung 50
4.3 Systeme höherer Ordnung 51
4.4 Systeme in der Simulation elektrodynamischer Felder 52
4.5 Definition und Eigenschaften von ROM-Fehlern 54

5 Adaptive Strategien in der Mehrpunkt-Modellordnungsreduktion 57
5.1 Mehrpunkt-MOR in der Simulation elektrodynamischer Felder 58
5.1.1 Anregungsprobleme 59
5.1.2 Radial inhomogene Wellenleiter 60
5.2 Fehlerindikatoren 62

5.2.1 Residuen-basierte Indikatoren 62
5.2.2 Inkrementelle Indikatoren 64
5.3 Adaptive Strategien zur Bestimmung von Entwicklungspunkten . . . 65
5.3.1 Bisektionsmethode 66
5.3.2 Greedy-Methode 67
5.4 Numerische Untersuchungen 67
5.4.1 Vivaldi-Antenne 68
5.4.2 Bandpassfilter 72
5.4.3 Wellenleiter mit dielektrischem Einsatz 77
5.5 Fazit 82
5.5.1 Konvergenzraten 82
5.5.2 Vergleich von Bisektion und Greedy-Methode 82
5.5.3 Vergleich residuenbasierter und inkrementeller Fehlerindikatoren 82
5.5.4 Abbruchkriterien 83

6 A-posteriori-Fehlerschätzer in der Einpunkt-Modellordnungsreduktion 85

6.1 MOR für verlustlose elektrodynamische Systeme 87
6.2 Fehlerschätzer für die Impedanzmatrix 88
6.2.1 Bewertung der ROM-Eigenwerte 91
6.3 Numerischer Aufwand bei der Fehlerschätzung 92
6.4 Fehlerschätzer für MIMO-Systeme 95
6.5 Asymptotischer Fehlerschätzer für Streuparameter 95
6.6 Numerische Beispiele 97
6.6.1 Bandpassfilter 97

6.6.2 Filter mit dielektrischen Resonatoren 99

7 Zussammenfassung **105**

Abbildungsverzeichnis **107**

Kapitel 1

Einleitung

Elektromagnetische Wellenphänomene sind in technischen Anwendungen von immer größerer Bedeutung. Neben den klassischen Anwendungsfeldern in der Hochfrequenztechnik, wie beispielsweise der Auslegung von Antennen und deren Speisenetzwerke, erlangt die Untersuchungen der Welleneigenschaft elektromagnetischer Felder zunehmend auch in anderen Bereichen an Bedeutung. So wird beispielsweise bei der Entwicklung integrierter Schaltkreise und bei Untersuchungen im Zusammenhang mit elektromagnetischer Verträglichkeit die Analyse der Wellenphänome unverzichtbar. Die dabei betrachteten Strukturen besitzen üblicherweise komplizierte Geometrien und setzen sich aus einer Vielzahl von Einzelkomponenten zusammen, weshalb in der Regel keine analytischen Ansätze verfolgt werden können. Stattdessen muss die Behandlung mit numerischen Verfahren erfolgen.

Bei den numerischen Methoden sind Integralgleichungs- und Differenzialgleichungsverfahren zu unterscheiden. Erstere kommen insbesondere bei Anwendungen zum Einsatz, bei denen die Abstrahlungscharakteristik flächenhafter Quellen zu bestimmen ist. Sobald jedoch inhomogene Materialien auftreten wird eine Volumendiskretisierung erforderlich. Das Bestimmen der Lösung des zugehörigen linearen Gleichungssystems ist aufgrund der vollbesetzten Matrix daher schon für einfache Strukturen sehr rechenintensiv. Zur Ausdünnung der Matrizen und damit zur Steigerung der Effizienz von Integralgleichungsverfahren existieren jedoch wirkungsvolle Verfahren, wie die fast multipole Methode [Rok85], adaptive cross-approximation [Beb00] und $\mathcal{H}$-Matrizen [Hac99].

Differenzialgleichungsverfahren hingegen diskretisieren nicht Quellen sondern Felder und sind somit grundsätzlich nur auf beschränkte Gebiete anwendbar. Abstrahlung in den Freiraum lässt sich durch Kopplung mit Integralgleichungsverfahren oder durch Näherungen wie absorbierende Randbedingungen [EM77], [BT80], infinite Elemente [Bet77] oder perfectly matched layers [SKLL95] berücksichtigen. Differenzialgleichungsverfahren werden in der Regel als Teilbereichs- oder Kolloka-

tionsmethoden ausgeführt, die auf schwachbesetzte Matrizen führen. Innerhalb der Differenzialgleichungsverfahren kann zwischen den Zeitbereichs- und den Frequenzbereichsverfahren unterschieden werden.

Im Zeitbereich kommen oftmals die Methode der Finiten Differenzen (FD) [Yee66], [TH05] oder die Finite Integrationstechnik (FIT) [Wei96] zum Einsatz, weil diese explizite Zeitschrittverfahren erlauben und daher keine Matrixfaktorisierungen benötigen.

Im Frequenzbereich hingegen ist die Methode der finiten Elemente (FE) [SF73] weit verbreitet. Diese beruht im Gegensatz zu den meisten Zeitbereichsverfahren auf unstrukturierten Gebietsdiskretisierungen und bietet daher wesentlich höhere Flexibilität in der Modellierung von Geometrie und Materialeigenschaften. Zudem lassen sich durch Ansatzfunktionen höherer Ordnung bessere asymptotische Konvergenzraten erzielen als mit FD-Verfahren. Durch die Einführung diskontinuierlicher Galerkinverfahren [RH73] hat sich die FE-Methode auch im Zeitbereich zu einem überaus leistungsfähigen Lösungsansatz entwickelt.

In dieser Arbeit werden lineare zeitinvariante Materialeigenschaften vorausgesetzt. Dies erlaubt eine Beschreibung der Problemstellung im Frequenzbereich und damit die Verwendung finiter Elemente. Die FE-Diskretisierung der im Weiteren betrachteten Anregungsprobleme führt auf frequenzabhängige lineare Gleichungssysteme mit hunderttausend bis einigen Millionen Unbekannten. Die Lösung solcher Systeme in einem einzigen Frequenzpunkt ist mit entsprechend ausgestatteten Computern und Verwendung effizienter Algorithmen in akzeptabler Zeit zu bewerkstelligen. Hierbei können direkte [Met], [Par] oder (semi-)iterative [HFDE03], [HFDE04] Verfahren eingesetzt werden. Vielfach ist jedoch in Anwendungen nicht das elektromagnetische Feld für einen Frequenzpunkt, sondern vielmehr das breitbandige Systemverhalten von Interesse. Zur Bestimmung des Frequenzgangs für eine feste Bandbreite muss das Systemverhalten an hinreichend vielen Frequenzstützstellen ausgewertet werden. Nur so können auch schmalbandige Effekte dargestellt werden. Das Lösen des zuvor genannten Gleichungssystems für alle einzelnen Frequenzstützstellen ist bereits für einfache Strukturen mit einem hohen numerischen Aufwand verbunden. Im Rahmen von Optimierungen, wird dieser Ansatz zusätzlich unattraktiv, da hierbei die Zahl der Systemauswertungen leicht im Bereich von mehreren Zehntausend liegt.

Hier bieten Verfahren der Modellordnungsreduktion elegante Möglichkeiten, diese Einschränkungen zu vermeiden. Ein wichtiger Vertreter im Bereich der Modellordnungsreduktion ist durch das balancierte Abschneiden [Moo81] gegeben, welches auf Singulärwertzerlegungen der Operatoren basiert. Die Vorteile dieses Ansatzes liegen in der Existenz globaler Fehlerschranken [Glo84] und der gesicherten Stabilität des ordnungsreduzierten Modells. Allerdings sind beim balancierten Abschneiden vollbesetzte Lyapunov-Gleichungen zu lösen, weshalb schon für Systemdimensionen von wenigen Tausend Freiheitsgraden der Rechenaufwand sehr hoch ist. Ansätze, die die Anwendung des balancierten Abschneidens auch auf Systeme höherer Ordnung

erlauben, werden beispielsweise in [LW02] und [Ben04] vorgestellt.

Einen anderen Ansatz stellen die Projektionsverfahren dar. Bei diesen werden ordnungsreduzierte Modelle generiert, bei denen die Übertragungsfunktion und gegebenenfalls auch deren Ableitungen mit dem Originalmodell in einer definierten Anzahl von Stützstellen übereinstimmen. Anstatt einer Singulärwertzerlegung sind lediglich Matrix-Vektor-Multiplikationen oder Vorwärts-/Rückwärtseinsetzungen faktorisierter Matrizen zu berechnen. Da die beteiligten Matrizen in vielen Anwendungen dünnbesetzt sind, können diese Verfahren auch für sehr hochdimensionale Systeme effizient eingesetzt werden. In Abhängigkeit der Anzahl der Stützstellen werden Einpunkt- und Mehrpunktverfahren unterschieden. Die ersten Ansätze dieser Art [PR90], [SCNZ94] sind Einpunktverfahren, die den Momentenabgleich explizit vollziehen. Neuere Methoden [FF95] basieren auf Projektion und nutzen Krylov-Unterraumverfahren, um numerische Stabilität zu gewährleisten. Eine weitere Steigerung der Robustheit lässt sich durch die Kombination von Einpunkt- und Mehrpunktverfahren erzielen [Gri97]. Die frühen Ansätze der momentenabgleichenden Verfahren setzen eine lineare Parametrisierung der Systemmatrix voraus. Die Verfahren gemäß [SLL03b], [BS05a], [SL06] erlauben auch eine quadratische oder allgemein polynomiale Abhängigkeit der Systemmatrix von einem Parameter. Ist die Systemmatrix von mehreren Parametern in unterschiedlichen Potenzen abhängig, können *multivariate* Verfahren [Far07] eingesetzt werden.

Aus der Berechnung ordnungsreduzierter Modelle der zweiten Kategorie resultieren keine unmittelbaren Fehlerschranken. Als sehr effizient haben sich inkrementelle Abbruchkriterien [SLL02], [RRM09], [SFDE09] bzw. heuristische Fehlerschätzer [BSSY99] erwiesen. In Kapitel 5 werden entsprechende Kriterien im Zusammenhang mit Mehrpunktverfahren diskutiert, unterschiedliche Strategien gegenübergestellt und die Leistungsfähigkeit im Bereich der FE-Simulation elektromagnetischer Felder analysiert.

Inkrementelle Abbruchkriterien und auch heuristische Fehlerschätzer sind zwar gute Indikatoren, um die Qualität des ordnungsreduzierten Modells einschätzen zu können, eine garantierte Fehlerschranke liefern sie aber nicht. In [CHMR10] und [PS10] werden *a-posteriori-Fehlerschätzer* vorgeschlagen, die dieser Einschränkung nicht unterliegen. Allerdings ist die Bestimmung der Fehlerschranke mit diesen Verfahren sehr rechenintensiv und dominiert schon bei wenig komplexen Anwendungen den gesamten MOR-Prozess. In Kapitel 6 wird daher ein neuer Fehlerschätzer für linear parametrierte Systeme vorgestellt. Diese Systemklasse stellt in technischen Anwendungen einen wichtigen Spezialfall dar, beispielsweise bei der Modellierung verlustloser elektromagnetischer Strukturen. Der vorgestellte a-posteriori-Fehlerschätzer liefert eine beweisbare Fehlerschranke und ist numerisch effizient zu berechnen. Sowohl Speicheranforderung als auch Gesamtrechenzeiten sind im Vergleich zum gesamten MOR-Prozesses als gering zu bewerten. Numerische Beispiele belegen die Zuverlässigkeit und Effizienz des Verfahrens.

Kapitel 2

Modellierung passiver Mikrowellenstrukturen

2.1 Passive Mikrowellenstrukturen

Die Hochfrequenztechnik beschreibt elektromagnetische Phänomene, bei denen die auftretenden Wellenlängen ähnlich der Abmessungen der untersuchten Strukturen sind. Die Welleneigenschaft der elektromagnetischen Feldgrößen kann unter diesen Gegebenheiten nicht vernachlässigt werden. Ein typisches Anwendungsfeld ist durch die Mikrowellentechnik gegeben. Aufgrund des Verhältnisses von Wellenlänge und Bauteilgröße sowie der zu übertragenden Leistungen stellen hier Hohlleiterstrukturen geeignete Schaltungselemente dar. Es existieren unterschiedliche Angaben zur Abgrenzung des Mikrowellenfrequenzbereichs. Für technische Anwendungen wird üblicherweise der Bereich von 300 MHz bis 300 GHz genannt [Poz05, S. 1]. Die im Folgenden behandelten Untersuchungen beschränken sich ausschließlich auf *passive* Mikrowellenstrukturen, welche schematisch durch das in Abbildung 2.1 dargestellte Modell beschrieben werden können. Hierbei weist das beschränkte Feldgebiet

$$\Omega \subset \mathbb{R}^d, d \in \{2, 3\}, \tag{2.1}$$

Bereiche örtlich variierender Materialeigenschaften auf, die durch die Materialtensoren der *magnetischen Permeabilität* μ, der *elektrischen Permittivität* ε und der *elektrischen Leitfähigkeit* σ beschrieben werden. Darüber hinaus ist der geschlossene Rand

$$\Gamma = \Gamma_{\mathrm{WG}} \cup \Gamma_{\mathrm{R}} \cup \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}} \tag{2.2}$$

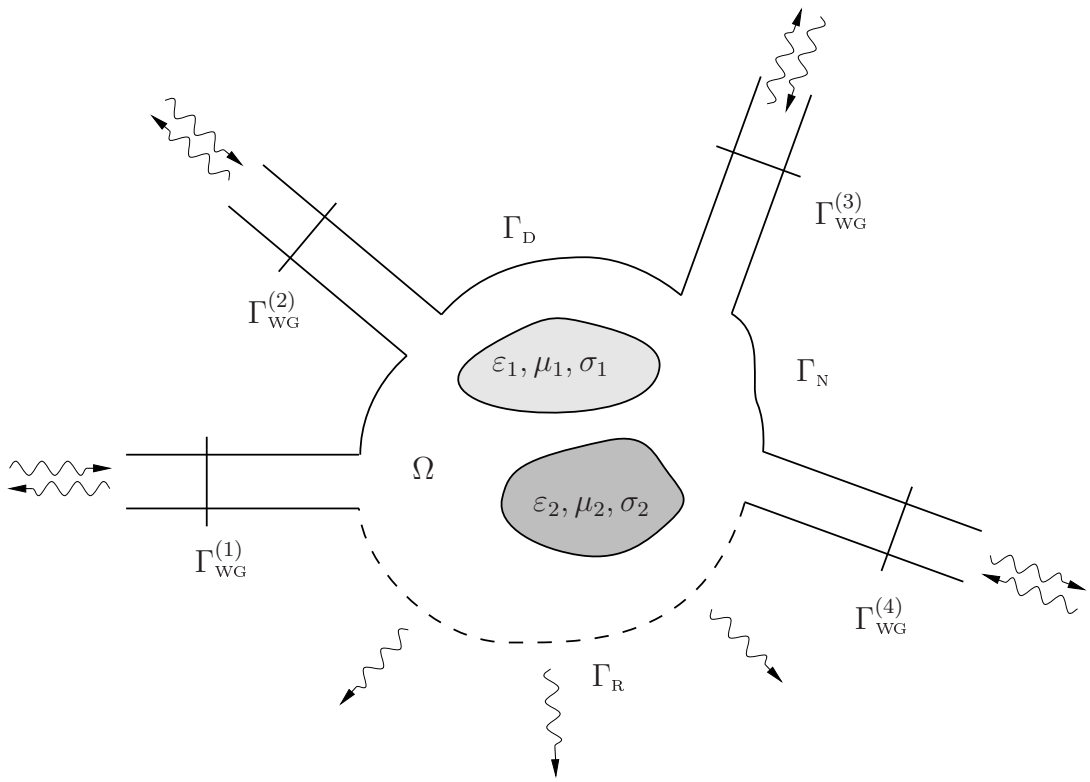


Abbildung 2.1: Modales Mehrtor mit Randbedingungen.

von Ω als Vereinigung disjunkter Teilmengen gegeben. Die Randsegmente repräsentieren hierbei die folgenden Eigenschaften:

Γ_{WG}	... Tor bzw. Wellenleiter,
Γ_{R}	... Abstrahlung in den Freiraum,
Γ_{D}	... idealer elektrischer Leiter
Γ_{N}	... idealer magnetischer Leiter.

Die Randsegmente selbst können wiederum durch eine Vereinigung nicht zusammenhängender Teilmengen gegeben sein, also

$$\Gamma_x = \bigcup_{i \in \{1,2,\ldots,N^x\}} \Gamma_x^{(i)}, \quad x \in \{\mathrm{WG}, \mathrm{R}, \mathrm{D}, \mathrm{N}\},\ N^x \in \mathbb{N}. \tag{2.3}$$

Elektromagnetische Felder in Ω werden über Quellen im Feldgebiet und über Feldbelegungen auf dem Rand Γ angeregt.

In den nachfolgenden Abschnitten wird ausgehend von dem System der Maxwell-Gleichungen ein Randwertproblem hergeleitet, welches den Ausgangspunkt für die rechnergestützte Feldsimulation darstellt.

2.2 System der Maxwell-Gleichungen

Bei klassischen Betrachtungen werden die elektromagnetischen Felder in einem Gebiet Ω durch das System der Maxwell-Gleichungen beschrieben. Im Zeitbereich lauten diese in differenzieller Form

$$\nabla \times \boldsymbol{\mathcal{E}} = -\frac{\partial \boldsymbol{\mathcal{B}}}{\partial t}, \tag{2.4a}$$

$$\nabla \times \boldsymbol{\mathcal{H}} = \boldsymbol{\mathcal{J}} + \frac{\partial \boldsymbol{\mathcal{D}}}{\partial t}, \tag{2.4b}$$

$$\nabla \cdot \boldsymbol{\mathcal{D}} = \varrho, \tag{2.4c}$$

$$\nabla \cdot \boldsymbol{\mathcal{B}} = 0, \tag{2.4d}$$

wobei $\boldsymbol{\mathcal{E}}$ die *elektrische Feldstärke*, $\boldsymbol{\mathcal{H}}$ die *magnetische Erregung*, $\boldsymbol{\mathcal{B}}$ die *magnetische Flussdichte* und $\boldsymbol{\mathcal{D}}$ die *elektrische Verschiebungsdichte* bezeichnen. Die *elektrische Stromdichte*

$$\boldsymbol{\mathcal{J}} = \boldsymbol{\mathcal{J}}_l + \boldsymbol{\mathcal{J}}_e \tag{2.5}$$

lässt sich in die *Leitungsstromdichte* $\boldsymbol{\mathcal{J}}_l$ und die *eingeprägte Stromdichte* $\boldsymbol{\mathcal{J}}_e$ unterteilen, wobei letztere – wie auch die *Raumladungsdichte* ϱ – zu den Quellgrößen zu zählen ist. Darüber hinaus kann auf makroskopischer Ebene der Einfluss unterschiedlicher Materialien mittels der Konstitutivgleichungen ausgedrückt werden. Unter der Voraussetzung, dass kein magnetisches Gedächtnis vorliegt und sich die Materialeigenschaften linear verhalten, gilt

$$\boldsymbol{\mathcal{D}} = \varepsilon \boldsymbol{\mathcal{E}}, \tag{2.6a}$$

$$\boldsymbol{\mathcal{B}} = \mu \boldsymbol{\mathcal{H}}, \tag{2.6b}$$

$$\boldsymbol{\mathcal{J}}_l = \sigma \boldsymbol{\mathcal{E}}. \tag{2.6c}$$

Hierin sind μ, ε und σ als ortsabhängige Materialtensoren zu verstehen, die die Feldgrößen $\boldsymbol{\mathcal{E}}$ und $\boldsymbol{\mathcal{H}}$ auf die entsprechenden Flussgrößen $\boldsymbol{\mathcal{D}}$, $\boldsymbol{\mathcal{B}}$ und $\boldsymbol{\mathcal{J}}$ abbilden. Im Vakuum vereinfachen sich die Materialtensoren ε und μ zu skalaren, orts- und richtungsunabhängigen Größen

$$\varepsilon_0, \mu_0 \in \mathbb{R}, \tag{2.7}$$

was die Darstellung

$$\varepsilon = \varepsilon_0 \varepsilon_r \qquad \text{und} \tag{2.8a}$$

$$\mu = \mu_0 \mu_r \tag{2.8b}$$

mit den *relativen* Materialtensoren ε_r bzw. μ_r erlaubt. Für die weiteren Betrachtungen werden isotrope Materialien angenommen, bei denen die Materialeigenschaften zwar orts- aber nicht richtungsabhängig sind. Damit vereinfachen sich die Materialtensoren zu skalarwertigen Funktionen. Die in dieser Arbeit vorgestellten Verfahren sind ebenfalls uneingeschränkt anwendbar, wenn für die vorkommenden Materialien gilt:

- ε_r und μ_r sind durch symmetrisch positiv definite Tensoren gegeben.
- σ ist durch einen symmetrisch positiv semi-definiten Tensor beschrieben.

Im Rahmen der vorliegenden Arbeit werden die Materialbeziehungen (2.6) als linear und zeitinvariant angenommen, was die Beschreibung der Maxwell-Gleichungen im Frequenzbereich erlaubt. Die Zeitabhängigkeit der Feldgrößen ist dann in der Form $e^{j\omega t}$ mit der imaginären Einheit $j := \sqrt{-1}$ und der *Kreisfrequenz* ω darstellbar. Die Zeitableitungen lassen sich in diesem Fall gemäß der Vorschrift

$$\frac{\partial}{\partial t} \to j\omega \tag{2.9}$$

algebraisieren, so dass die Maxwell-Gleichungen im Frequenzbereich die Form

$$\nabla \times \boldsymbol{E} = -j\omega \boldsymbol{B}, \tag{2.10a}$$
$$\nabla \times \boldsymbol{H} = \boldsymbol{J} + j\omega \boldsymbol{D}, \tag{2.10b}$$
$$\nabla \cdot \boldsymbol{D} = \rho, \tag{2.10c}$$
$$\nabla \cdot \boldsymbol{B} = 0 \tag{2.10d}$$

annehmen. Hierin stellen $\boldsymbol{E}, \boldsymbol{H}, \boldsymbol{D}, \boldsymbol{B}, \boldsymbol{J}, \rho$ die den Feldgrößen $\boldsymbol{\mathcal{E}}, \boldsymbol{\mathcal{H}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{D}}, \boldsymbol{\mathcal{J}}, \varrho$ entsprechenden Phasoren dar. Für die Phasoren gelten die Materialbeziehungen (2.6) in unveränderter Weise:

$$\boldsymbol{D} = \varepsilon \boldsymbol{E}, \tag{2.11a}$$
$$\boldsymbol{B} = \mu \boldsymbol{H}, \tag{2.11b}$$
$$\boldsymbol{J}_l = \sigma \boldsymbol{E}. \tag{2.11c}$$

Die Algebraisierung der Zeitableitung ist insbesondere bei der numerischen Behandlung der Maxwell-Gleichungen im Zusammenhang mit der Finite-Elemente-Methode von elementarer Bedeutung. Die folgenden Herleitungen basieren daher ausschließlich auf der Darstellung im Frequenzbereich.

2.3 Die vektorielle Helmholtz-Gleichung

Das System der Maxwell-Gleichungen lässt unter Einbeziehung der Konstitutivgleichungen die Eliminierung einer Feldgröße zu, so dass ein Randwertproblem in nur einer unbekannten Feldgröße formuliert werden kann. Hierzu wird (2.11b) in (2.10a) eingesetzt und anschließend der Operator $\nabla \times \mu^{-1}$ angewendet. Dies führt auf die *vektorielle Helmholtz-Gleichung* in der Unbekannten $\boldsymbol{E}$,

$$\nabla \times \mu^{-1} \nabla \times \boldsymbol{E} + j\omega\sigma \boldsymbol{E} - \omega^2 \varepsilon \boldsymbol{E} = -j\omega \boldsymbol{J}_e. \tag{2.12}$$

Das System partieller Differenzialgleichungen erster Ordnung wird damit in *eine* partielle Differenzialgleichung zweiter Ordnung überführt. Völlig analog kann auch eine Formulierung in der magnetischen Erregung $\boldsymbol{H}$ als unbekannte Größe beschrieben werden. Für die weiteren Betrachtungen wird aus Gründen der Vereinfachung eine Parametrisierung mit der Freiraumwellenzahl

$$k_0 = \omega\sqrt{\varepsilon_0\mu_0} \tag{2.13}$$

gewählt. Wird außerdem der *Freiraumwellenwiderstand*

$$\eta_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}} \tag{2.14}$$

eingeführt, kann die vektorielle Helmholtz-Gleichung mit (2.8) in der Form

$$\boxed{\nabla \times \mu_r^{-1}\nabla \times \boldsymbol{E} + jk_0\eta_0\sigma\boldsymbol{E} - k_0^2\varepsilon_r\boldsymbol{E} = -jk_0\eta_0\boldsymbol{J}_e} \tag{2.15}$$

geschrieben werden. Die in dieser Arbeit betrachteten elektromagnetischen Problemstellungen werden für die numerische Behandlung stets auf die $\boldsymbol{E}$-Feld-Formulierung zurückgeführt, die auf der Differenzialgleichung (2.15) basiert.

2.4 Randbedingungen

Entsprechend der Darstellung in Abbildung 2.1 liegen bei der betrachteten Problemstellung unterschiedliche Randbedingungen vor. Deren physikalischer Hintergrund und die entsprechenden mathematischen Modellierungen werden nachfolgend beschrieben.

2.4.1 Ideal elektrische Leiter

In Körpern ideal elektrischer Leitfähigkeit verschwinden alle elektromagnetischen Felder. Entsprechend verschwindet auch an der Grenzfläche zu idealen Leitern die Tangentialkomponente von $\boldsymbol{E}$. Diese Eigenschaft lässt sich mathematisch als *homogene Dirichlet-Randbedingung* formulieren,

$$\hat{\boldsymbol{n}} \times (\boldsymbol{E} \times \hat{\boldsymbol{n}}) = 0 \qquad \text{auf } \Gamma_{\mathrm{D}}. \tag{2.16}$$

Hierin bezeichnet $\hat{\boldsymbol{n}}$ den aus dem Gebiet Ω zeigenden Einheitsnormalenvektor.

2.4.2 Ideal magnetische Leiter

Ideal magnetisch leitende Randbedingungen werden zur Modellierung von Symmetrien eingesetzt, wenn a priori feststeht, dass an der gewählten Symmetrieebene die Tangentialkomponente der magnetischen Erregung $\boldsymbol{H}$ verschwindet:

$$\hat{\boldsymbol{n}} \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) \times \hat{\boldsymbol{n}} = -jk_0\eta_0 \hat{\boldsymbol{n}} \times (\boldsymbol{H} \times \hat{\boldsymbol{n}}) = 0 \qquad \text{auf } \Gamma_{\text{N}}. \tag{2.17}$$

Entsprechend ist die Grenzfläche zu einem idealen magnetischen Leiter als *homogene Neumann-Randbedingung* zu verstehen.

2.4.3 Absorbierende Randbedingung

Differenzialgleichungsverfahren, wie die in dieser Arbeit behandelte FE-Methode, basieren auf der Diskretisierung des Feldgebiets Ω, also einer beschränkten zusammenhängenden Teilmenge in $\mathbb{R}^d$. Tritt Abstrahlung in den Freiraum auf, so ist diese beispielsweise durch Kopplung mit Integralgleichungsverfahren zu bestimmen [HEL94], [ZVL06]. Für den Fall des Modellproblems in Abbildung 2.1 sind die Feldverläufe außerhalb von Ω nicht von Interesse. Hier ist die Abstrahlung nur im Sinne einer Verlustleistung zu betrachten. Daher ist es ausreichend künstliche Randbedingungen zu modellieren, die auftreffende elektromagnetische Wellen absorbieren. In der vorliegenden Arbeit werden daher *absorbierenden Randbedingungen erster Ordnung* (ABC: absorbing boundary condition) zur Beschreibung von Abstrahlungen in den Freiraum herangezogen [Pet88]. Hierbei werden im Sinne einer *homogenen Robin-Randbedingung*

$$\begin{aligned} &\hat{\boldsymbol{n}} \times (\boldsymbol{E} \times \hat{\boldsymbol{n}}) - \eta_0 (-\frac{1}{jk_0\eta_0} \mu_r^{-1} \nabla \times \boldsymbol{E}) \times \hat{\boldsymbol{n}} \\ &\qquad \overset{\mu_r=1}{=} \hat{\boldsymbol{n}} \times (\boldsymbol{E} \times \hat{\boldsymbol{n}}) - \eta_0 \boldsymbol{H} \times \hat{\boldsymbol{n}} = 0 \qquad \text{auf } \Gamma_{\text{R}} \end{aligned} \tag{2.18}$$

die Tangentialkomponenten von $\boldsymbol{E}$ und $\boldsymbol{H}$ über den *Feldwellenwiderstand des Freiraums* η_0 in Bezug gesetzt und somit senkrecht einfallende ebene Wellen ideal absorbiert.

Mit absorbierenden Randbedingungen kann völlig analog auch Abstrahlung in andere, nichtleitende Medien mit $\epsilon_r, \mu_r \neq 1$ modelliert werden. Hierbei ist lediglich der Wellenwiderstand des Mediums $\eta := \eta_0 \sqrt{\frac{\mu_r}{\epsilon_r}}$ mitzuführen. Für weitere Ansätze zur Modellierung von Abstrahlung in den Freiraum wird auf die entsprechende Literatur verwiesen: [WK89], [Ber94], [SKLL95].

2.4.4 Tor-Randbedingung

Erfolgt die Anregung einer Struktur mittels modaler Tore über den Rand des Feldgebiets, beispielsweise über Wellenleiter Γ_{WG}, so kann dies in Form einer *inhomogenen Neumann-Randbedingungen* modelliert werden. Hierbei wird die Anregung durch Vorgabe der am Rand tangential gerichteten magnetischen Erregung $\bar{\boldsymbol{H}}_t$ gemäß

$$\hat{\boldsymbol{n}} \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) \times \hat{\boldsymbol{n}} = -jk_0\eta_0 \bar{\boldsymbol{H}}_t \qquad \text{auf } \Gamma_{\mathrm{WG}} \tag{2.19}$$

eingeprägt. Die besonderen Eigenschaften modaler Tore und die somit vorzugebenden Felder $\bar{\boldsymbol{H}}_t$ werden in Abschnitt 2.5 behandelt.

2.5 Homogene Wellenleiter

2.5.1 Grundlegende Eigenschaften homogener Wellenleiter

Bei dem Modellproblem gemäß Abbildung 2.1 erfolgt die Anregung der Struktur über zylindrische Wellenleiter, wobei die Randsegmente $\Gamma_{\mathrm{WG}}^{(i)}, i = 1, 2, \ldots, N^{\mathrm{port}}$ gerade die Querschnittflächen der einmündenden Wellenleiter darstellen. Charakterisierend für zylindrische Wellenleiter ist die Symmetrie entlang einer ausgezeichneten Richtung. Für die folgenden Betrachtungen sei dies entsprechend Abbildung 2.2 die z-Richtung. Die nachfolgend betrachteten Wellenleiter weisen in der transversalen Ebene homogene Materialeigenschaften auf und werden daher als *homogene Wellenleiter* bezeichnet. Die Berücksichtigung der allgemeineren Klasse *axial* homogener Wellenleiter, bei denen die Materialeigenschaften nur in z-Richtung konstant sind, in radialer Richtung jedoch variieren dürfen, ist in der FE-Methode und auch im Kontext der Modellordnungsreduktion [FHDE04] [SFDE08] möglich.

Im Rahmen der vorliegenden Arbeit werden die Wellenleiter zusätzlich als verlustlos angenommen, d. h.

$$\sigma = 0 \qquad \text{in } \Omega_{\mathrm{WG}}, \tag{2.20}$$

$$\sigma \to \infty \qquad \text{in metallischen Leitern}, \tag{2.21}$$

$$\varepsilon_r, \mu_r \in \mathbb{R} \qquad \text{in } \Omega_{\mathrm{WG}}. \tag{2.22}$$

Ein Wellenleiter mit den beschriebenen Eigenschaften ist schematisch in Abbildung 2.2 dargestellt. Die Randbedingungen Γ_{D} und Γ_{N} repräsentieren in Analogie zu Abbildung 2.1 auch hier den elektrisch ideal leitenden bzw. den magnetisch ideal leitenden Rand, so dass die in Abschnitt 2.4 beschriebenen mathematischen Modelle auch hier gültig sind. Wie im übrigen Feldgebiet müssen die Felder auch im Wellenleiter den Maxwell-Gleichungen bzw. der Helmholtz-Gleichung (2.12) genügen, jedoch lassen die speziellen Eigenschaften, Symmetrie in der Geometrie und Homogenität der Materialien, vereinfachte Formulierungen zu.

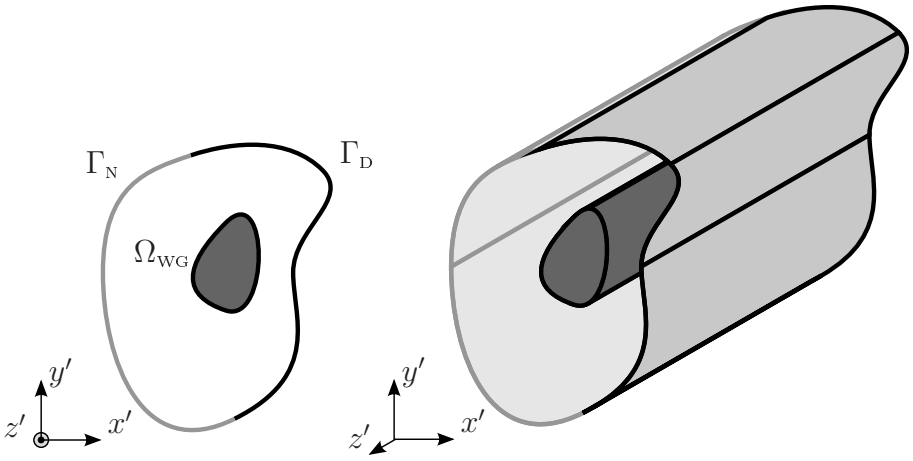


Abbildung 2.2: Axial homogener, zylindrischer Wellenleiter.

2.5.2 Eigenwertproblem für homogene Wellenleiter

Aufgrund der Homogenität der Materialien und nicht vorhandener Quellen im Feldgebiet Ω_{WG} vereinfachen sich (2.12) und (2.10c) zu

$$\nabla \times \nabla \times \boldsymbol{E} - k_0^2 \varepsilon_r \mu_r \boldsymbol{E} = 0 \tag{2.23}$$

beziehungsweise

$$\nabla \cdot \boldsymbol{D} = \nabla \cdot \varepsilon \boldsymbol{E} = \varepsilon \nabla \cdot \boldsymbol{E} = 0. \tag{2.24}$$

Mit der Identität

$$\nabla \times \nabla \times \boldsymbol{E} = \nabla(\nabla \cdot \boldsymbol{E}) - (\nabla \cdot \nabla)\boldsymbol{E} \tag{2.25}$$

gilt somit für homogene Wellenleiter

$$-\Delta \boldsymbol{E} - k^2 \boldsymbol{E} = 0, \tag{2.26}$$

wobei

$$\Delta = \nabla \cdot \nabla \tag{2.27}$$

den auf Komponentenebene wirkenden Laplace-Operator und

$$k := \omega \sqrt{\varepsilon \mu} \tag{2.28}$$

die Wellenzahl im vorliegenden Medium bezeichnet. Die Symmetrie in z-Richtung erlaubt für die elektrische Feldstärke einen Produktansatz der Form

$$\boldsymbol{E}(\boldsymbol{t}, z) = \boldsymbol{e}(\boldsymbol{t}) e^{-\gamma z} \tag{2.29}$$

und analog für die magnetische Erregung den Ansatz

$$\boldsymbol{H}(\boldsymbol{t}, z) = \boldsymbol{h}(\boldsymbol{t}) e^{-\gamma z}. \tag{2.30}$$

Hierin steht $\boldsymbol{t}$ für die transversalen Komponenten des Koordinatensystems, d. h. $\boldsymbol{t} = \binom{x}{y}$ im Falle von Abbildung 2.2. Die komplexe Zahl

$$\gamma = \alpha + j\beta \text{ mit } \alpha, \beta \geq 0, \tag{2.31}$$

stellt den *Ausbreitungskoeffizienten* dar, welcher den Feldverlauf in z-Richtung beschreibt. Der Realteil α ist der *Dämpfungskoeffizient* und der Imaginärteil β der *Phasenkoeffizient*. Für $\beta > 0$ pflanzt sich eine Welle in positive z-Richtung fort, und ein nicht verschwindender Realteil $\alpha > 0$ bewirkt ein exponentielles Abklingen der Welle entlang der positiven z-Richtung.

Die folgenden Untersuchungen zur Herleitung eines geeignet gestellten Wellenleiterproblems beschränken sich auf bezüglich z vorwärts laufende Wellen. Vertauschen des Vorzeichens im Exponent von (2.29) führt demnach auf Wellen, die sich entgegen der z-Richtung fortpflanzen. Das Eigenwertproblem zu rückwärts laufenden Wellen unterscheidet sich daher nur in den entsprechenden Vorzeichen.

Für die Wellenform in (2.29) erlaubt der Produktansatz eine Algebraisierung der Ableitung bezüglich der z-Richtung in der Form

$$\frac{\partial}{\partial z} \to -j\gamma. \tag{2.32}$$

In die Maxwell-Gleichungen eingesetzt lässt sich somit zeigen, dass die transversalen Komponenten

$$\boldsymbol{e}_t := \hat{\boldsymbol{z}} \times (\boldsymbol{e} \times \hat{\boldsymbol{z}}), \qquad \boldsymbol{h}_t := \hat{\boldsymbol{z}} \times (\boldsymbol{h} \times \hat{\boldsymbol{z}}) \tag{2.33}$$

allein aus den z-Komponenten

$$e_z := \boldsymbol{e} \cdot \hat{\boldsymbol{z}}, \qquad h_z := \boldsymbol{h} \cdot \hat{\boldsymbol{z}} \tag{2.34}$$

der Faktoren $\boldsymbol{e}(\boldsymbol{t})$ und $\boldsymbol{h}(\boldsymbol{t})$ bestimmt werden können [Sim79, S. 738]. Liegt ein mehrfach zusammenhängendes Wellenleitergebiet Ω_{WG} vor, treten außerdem Wellenformen auf, deren z-Komponenten verschwinden. Die somit rein transversal gerichteten Feldkomponenten lassen sich aus einem elektrostatischen Randwertproblem ermitteln [Poz05, S. 94].

In homogenen Wellenleitern können demnach die folgenden Wellenformen unterschieden werden:

- *Transversal elektrische (TE) Wellen* mit

$$e_z = 0 \text{ und } h_z \neq 0, \tag{2.35}$$

- *Transversal magnetische (TM) Wellen* mit

$$e_z \neq 0 \text{ und } h_z = 0, \text{ sowie} \tag{2.36}$$

- *Transversal elektromagnetische (TEM) Wellen* mit

$$e_z = 0 \text{ und } h_z = 0. \tag{2.37}$$

Zur Bestimmung von h_z und e_z und somit der TE- und TM-Wellen ist das im Folgenden beschriebene skalare ebene Eigenwertproblem zu lösen.

Problem 2.5.1. Sei $\Omega_{\text{WG}} \in \mathbb{R}^2$ ein beschränktes Gebiet mit homogenen Materialeigenschaften ε und μ. Gesucht sind Eigenfunktionen $u : \Omega_{\text{WG}} \to \mathbb{C}$ und Eigenwerte $k_c^2 \in \mathbb{R}$, die das Eigenwertproblem

$$-\Delta_t u - k_c^2 u = 0 \qquad \text{in } \Omega_{\text{WG}}, \tag{2.38a}$$

$$u = 0 \qquad \text{auf } \Gamma_u, \tag{2.38b}$$

$$\frac{\partial u}{\partial n} = 0 \qquad \text{auf } \Gamma \backslash \Gamma_u. \tag{2.38c}$$

lösen. Hierbei ist

$$\Delta_t = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \tag{2.39}$$

als in der Transversalebene wirkender Laplaceoperator und u als Platzhalter für entweder $e_z(\boldsymbol{t})$ oder $h_z(\boldsymbol{t})$ zu verstehen. Für den Fall $u = e_z$ entspricht Γ_u einer ideal elektrisch und $\Gamma \backslash \Gamma_u$ einer ideal magnetisch leitfähigen Randbedingung. Für $u = h_z$ sind die Zuordnungen der Randbedingungen gerade umgekehrt.

Das Problem für e_z führt dann auf TM-Wellen, während mit h_z TE-Wellen gefunden werden. Die Eigenwerte k_c^2 sind über die Dispersionsgleichung

$$k_c^2 := \gamma^2 + k^2 \tag{2.40}$$

direkt mit dem Ausbreitungskoeffizienten γ der Wellenform verknüpft. Dem Wert k_c kann somit die physikalische Bedeutung der *Grenzwellenzahl* zugewiesen werden. Für Arbeitswellenzahlen $k < k_c$ gilt $\mathbb{R} \supset \gamma = \alpha > 0$, so dass die zugehörige Wellenform exponentiell gedämpft wird und nicht entlang des Wellenleiters propagiert. Ist hingegen $k > k_c$ und somit $\gamma = j\beta$ rein imaginär, breitet sich die Welle ungedämpft aus.

Auf die numerische Behandlung des Eigenwertproblems 2.5.1 wird in der vorliegenden Arbeit nicht weiter eingegangen. Für effiziente Algorithmen zur Berechnung der Wellenformen wird auf [FHDE04] verwiesen.

Mit den bisher eingeführten Bezeichnungen lassen sich Felder in Wellenleitern in der Form

$$\boldsymbol{E}_v(\boldsymbol{t}, z) = \boldsymbol{e}_v(\boldsymbol{t}) e^{-\gamma z} = [\boldsymbol{e}_t(\boldsymbol{t}) + e_z(\boldsymbol{t})\hat{\boldsymbol{z}}] e^{-\gamma z}, \tag{2.41a}$$

$$\boldsymbol{E}_r(\boldsymbol{t}, z) = \boldsymbol{e}_r(\boldsymbol{t}) e^{+\gamma z} = [\boldsymbol{e}_t(\boldsymbol{t}) - e_z(\boldsymbol{t})\hat{\boldsymbol{z}}] e^{+\gamma z}, \tag{2.41b}$$

$$\boldsymbol{H}_v(\boldsymbol{t}, z) = \boldsymbol{h}_v(\boldsymbol{t}) e^{-\gamma z} = [\boldsymbol{h}_t(\boldsymbol{t}) + h_z(\boldsymbol{t})\hat{\boldsymbol{z}}] e^{-\gamma z}, \tag{2.41c}$$

$$\boldsymbol{H}_r(\boldsymbol{t}, z) = \boldsymbol{h}_r(\boldsymbol{t}) e^{+\gamma z} = [-\boldsymbol{h}_t(\boldsymbol{t}) + h_z(\boldsymbol{t})\hat{\boldsymbol{z}}] e^{+\gamma z}, \tag{2.41d}$$

darstellen, wobei der Index v jene Wellenformen beschreibt, die sich bezüglich z vorwärts, d. h. in positiver Richtung ausbreiten, während r die bezüglich der z-Richtung rückwärts laufenden Wellen kennzeichnet. Gemäß der Eigenschaft (2.35) verschwindet die Komponente $e_z(\boldsymbol{t})$ im Falle einer TE-Welle. Entsprechend wird für TM-Wellen wegen (2.36) die Komponente $h_z(\boldsymbol{t})$ zu Null.

Die hier aufgezeigten Eigenschaften der Wellenformen führen auf zwei weitere Ergebnisse von zentraler Bedeutung.

Satz 2.5.2. *Seien $(\boldsymbol{E}^m, \boldsymbol{H}^m, \gamma_m)$ und $(\boldsymbol{E}^n, \boldsymbol{H}^n, \gamma_n)$ Eigenformen in einem homogenen Wellenleiter mit beschränktem Querschnitt. Dann erfüllen die zugehörigen Transversalkomponenten $\boldsymbol{e}_t^m$ und $\boldsymbol{h}_t^n$ die Orthogonalitätsrelation*

$$\int_{\Omega_{WG}} (\boldsymbol{e}_t^m \times \boldsymbol{h}_t^n) \cdot \hat{\boldsymbol{z}} \, \mathrm{d}\Omega = 0, \text{ falls } m \neq n \text{ und } \gamma_m \neq \gamma_n. \tag{2.42}$$

Beweis. Siehe [Col91, S. 333 ff.]. □

Die Form (2.42) nimmt für den Fall $m = n$ Werte ungleich Null an, so dass stets eine Normierung der Eigenformen der Art

$$\int_{\Omega_{\mathrm{WG}}} (\boldsymbol{e}_t^m \times \boldsymbol{h}_t^n) \cdot \hat{\boldsymbol{z}} \, \mathrm{d}\Omega = \delta_{mn} \tag{2.43}$$

vorgenommen werden kann. Hierin bedeutet δ_{mn} das Kronecker-Symbol. Darüber hinaus gilt außerdem der Satz zur Darstellbarkeit:

Satz 2.5.3. *Wellenformen eines homogenen Wellenleiters mit beschränktem Querschnitt bilden ein vollständiges orthonormales Funktionensystem. Damit sind Felder im Wellenleiter stets als Überlagerungen der Wellenformen gemäß*

$$\boldsymbol{E} = \sum_{k=1}^{\infty} \left(a_k \boldsymbol{E}_v^k + b_k \boldsymbol{E}_r^k \right), \tag{2.44a}$$

$$\boldsymbol{H} = \sum_{k=1}^{\infty} \left(a_k \boldsymbol{H}_v^k + b_k \boldsymbol{H}_r^k \right) \tag{2.44b}$$

darstellbar, wobei a_k die komplexe Amplitude einer in positiver z-Richtung und b_k die Amplitude einer in negativer z-Richtung propagierenden Wellenform bezeichnet.

Beweis. Siehe [Col91, S. 359 f.]. □

Ist das Wellenleitergebiet mehrfach zusammenhängend, d. h. es liegen $N > 1$ Elektroden vor, so sind zusätzlich $N - 1$ TEM-Wellen ausbreitungsfähig.

Das Verhalten elektromagnetischer Strukturen wird in der Regel nur in einem beschränkten Frequenzbereich

$$I_\omega = [\omega_{min}, \omega_{max}] := \{\omega \in \mathbb{R} : \omega_{min} \leq \omega \leq \omega_{max}\} \tag{2.45}$$

bzw. im entsprechenden Wellenzahlintervall

$$I_k = [k_{min}, k_{max}] \tag{2.46}$$

untersucht. Zur Darstellung der Felder im homogenen Wellenleiter ist es in der Regel ausreichend, nur die ausbreitungsfähigen Wellenformen zu berücksichtigen. In manchen Fällen, zum Beispiel wenn sich Diskontinuitäten in der Nähe der Tore befinden und evaneszente Wellen in dieser Entfernung noch nicht ausreichend abgeklungen sind, werden mitunter zusätzliche Wellen mit $k_c > k_{max}$ mitgeführt. Alle weiteren Wellen mit $k_c \gg k_{max}$ können aufgrund der exponentiellen Dämpfung vernachlässigt werden. Die unendlichen Summen in (2.44) gehen damit in endliche Summen gemäß (2.47) über. Zur Herleitung des mathematischen Modells für die FE-Methode sind darüber hinaus nur die transversalen Komponenten der Wellenformen von Interesse, um die Randbedingung (2.19) vorgeben zu können. In diesem Zusammenhang lassen sich mit (2.41) die Verhältnisse im Wellenleiter in der Form

$$\boldsymbol{E}_t = \sum_{k=1}^{M} (a_k + b_k)\, \boldsymbol{e}_t^k, \tag{2.47a}$$

$$\boldsymbol{H}_t = \sum_{k=1}^{M} (a_k - b_k)\, \boldsymbol{h}_t^k \tag{2.47b}$$

darstellen.

2.6 Starke Formulierung des Randwertproblems für Mikrowellenstrukturen

Mit den bis hierher aufgezeigten Beziehungen lässt sich für das Modellproblem aus Abschnitt 2.1 das folgende Randwertproblem formulieren:

Problem 2.6.1. Sei $\Omega \subset \mathbb{R}^3$ ein beschränktes Gebiet und $\Gamma = \Gamma_{\text{WG}} \cup \Gamma_{\text{R}} \cup \Gamma_{\text{D}} \cup \Gamma_{\text{N}}$ der Rand von Ω, so dass gilt $\overline{\Omega} = \Omega \cup \Gamma$. Gesucht ist das Vektorfeld $\boldsymbol{E}$, das folgende

Bedingungen erfüllt:

$$\nabla \times \mu_r^{-1} \nabla \times \boldsymbol{E} + jk_0\eta_0\sigma\boldsymbol{E} - k_0^2\varepsilon_r\boldsymbol{E} = -jk_0\eta_0\boldsymbol{J}_e \qquad \text{in } \Omega, \tag{2.48a}$$

$$\hat{\boldsymbol{n}} \times (\boldsymbol{E} \times \hat{\boldsymbol{n}}) = 0 \qquad \text{auf } \Gamma_{\mathrm{D}}, \tag{2.48b}$$

$$\hat{\boldsymbol{n}} \times (\boldsymbol{H} \times \hat{\boldsymbol{n}}) = 0 \qquad \text{auf } \Gamma_{\mathrm{N}}, \tag{2.48c}$$

$$\hat{\boldsymbol{n}} \times (\boldsymbol{E} \times \hat{\boldsymbol{n}}) - \eta_0 \boldsymbol{H} \times \hat{\boldsymbol{n}} = 0 \qquad \text{auf } \Gamma_{\mathrm{R}}, \tag{2.48d}$$

$$(\mu_r^{-1} \nabla \times \boldsymbol{E}) \times \hat{\boldsymbol{n}} = -jk_0\eta_0 \sum_{k=1}^{M} (a_k - b_k) \boldsymbol{h}_t^k \times \hat{\boldsymbol{n}} \qquad \text{auf } \Gamma_{\mathrm{WG}}. \tag{2.48e}$$

Eine klassische Lösung für die *starke Formulierung* (2.48) existiert nur in einigen Spezialfällen. Daher wird in Abschnitt 3.1 eine Formulierung mit schwächeren Anforderungen an die gesuchte Lösung hergeleitet, welche auch die Grundlage für den numerischen Zugang darstellt.

2.7 Schaltung von Mikrowellenstrukturen auf Systemebene

Die Analyse komplexer Mikrowellenstrukturen ist häufig durch die verfügbare Rechenleistung und Speicherkapazität limitiert. Daher werden diese Strukturen häufig in Subsysteme untergliedert, die separat untersucht werden können.

Neben dem Feldverlauf ist vor allem das Übertragungsverhalten dieser Subsysteme von Interesse, welches durch verallgemeinerte Netzwerkmatrizen beschrieben werden kann. Hierzu wird in weiterer Folge die Analogie der Mikrowellenstrukturen zu *modalen Mehrtoren* herausgearbeitet und die Konstruktion von Netzwerkmatrizen gemäß der Klemmenmehrtortheorie aufgezeigt. Die Übertragungsfunktion der Gesamtstruktur wird schließlich durch Verschaltung der Einzelkomponenten auf der *Systemebene* gewonnen.

Mit Satz 2.5.2 ist gewährleistet, dass unterschiedliche Wellenformen in einem Wellenleiter voneinander entkoppelt, und somit als unabhängige Signale zu betrachten sind. Wie in Abbildung 2.3 gezeigt, kann somit jeder Wellenform $(\boldsymbol{E}_k, \boldsymbol{H}_k)$ ein eigenes Tor zugeordnet werden, welches im Modell des modalen Mehrtors genau einem Klemmenpaar (u_k, i_k) entspricht.

Ein Klemmenmehrtor (N-Tor) lässt sich beispielsweise mittels der *Impedanzmatrix* $\mathbf{Z}$ charakterisieren. Diese wird bestimmt, indem sukzessive jeweils an einem Tor ein Strom i eingeprägt wird, während die übrigen Tore im Leerlauf betrieben werden. Die sich somit einstellenden Tor-Spannungen liefern die zugehörigen Einträge der

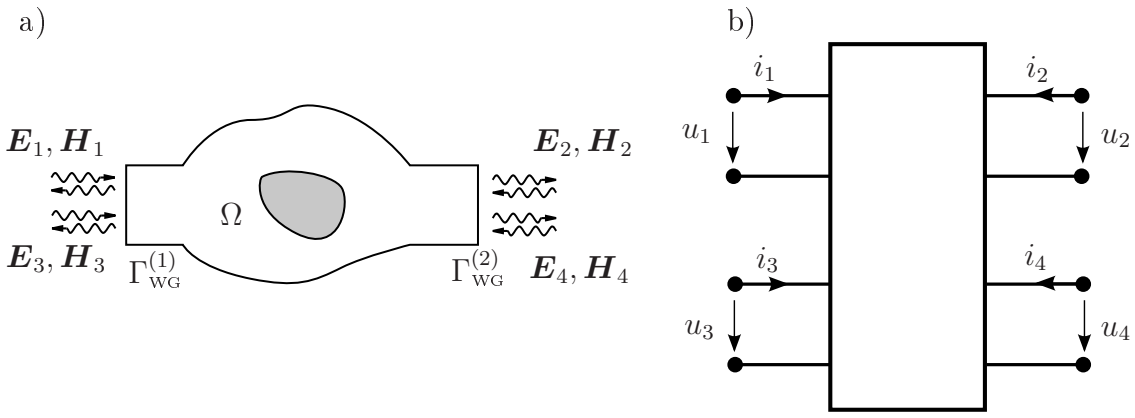


Abbildung 2.3: Äquivalente Darstellung von modalem Mehrtor und Klemmenmehrtor: a) Modales Zweitor; b) äquivalentes Klemmenmehrtor.

Impedanzmatrix [Poz05, S. 170 f],

$$z_{mn} = \left.\frac{u_m}{i_n}\right|_{i_k=0, k\neq n}, \qquad m = 1 \ldots N. \tag{2.49}$$

Der Zusammenhang von Strömen und Spannungen an den Toren des Netzwerks kann somit in der kompakten Form

$$\mathbf{u} = \mathbf{Z}\mathbf{i} \tag{2.50}$$

dargestellt werden. Hierin sind

$$\mathbf{u} = (u_1, \ldots, u_N)^T \qquad \text{und} \tag{2.51}$$

$$\mathbf{i} = (i_1, \ldots, i_N)^T \tag{2.52}$$

die Torspannungen bzw. -ströme in Vektorschreibweise.

Da im Bereich der Mikrowellentechnik vielfach Hohlleiter mit nur einer Elektrode Verwendung finden und somit Ströme und Spannungen nicht eindeutig festzulegen sind, wird das Konzept der *äquivalenten* Ströme und Spannungen eingeführt [Poz05, S. 162]. In Analogie zu den Leitungsgleichungen können diese mittels der komplexen Amplituden a und b aus (2.44) in der Form

$$u_k := a_k + b_k \qquad \text{bzw.} \tag{2.53}$$

$$i_k := a_k - b_k \tag{2.54}$$

definiert werden.

Bei der Betrachtung eines beschränkten Frequenzbereichs ist gemäß (2.47) die Zahl M der ausbreitungsfähigen Eigenformen endlich. Damit kann eine lineare Wirkungsfunktion gefunden werden, die den Zusammenhang zwischen den äquivalenten Strömen und Spannungen beschreibt. Diese Wirkungsfunktion wird in Analogie

zur Schaltungstechnik durch eine *verallgemeinerte* Netzwerkmatrix dargestellt. Beispielsweise ergibt sich für die verallgemeinerte Impedanzmatrix analog zu (2.50)

$$\mathbf{u} = \mathbf{Z}\mathbf{i}. \tag{2.55}$$

Entsprechend lässt sich auch die verallgemeinerte Admittanzmatrix $\mathbf{Y}$ aufstellen:

$$\mathbf{i} = \mathbf{Y}\mathbf{u}. \tag{2.56}$$

Die verallgemeinerte *Streumatrix* $\mathbf{S}$ verknüpft hingegen die Amplituden der reflektierten Wellen

$$\mathbf{b} = (b_1, \ldots, b_M)^T \tag{2.57}$$

mit jenen der eingeprägten Wellen

$$\mathbf{a} = (a_1, \ldots, a_M)^T \tag{2.58}$$

gemäß der Vorschrift

$$\mathbf{b} = \mathbf{S}\mathbf{a}. \tag{2.59}$$

Für die Netzwerkmatrizen von Mikrowellenmehrtoren gelten, wie für Klemmenmehrtore, die folgenden Umrechnungsvorschriften:

$$\mathbf{Z} = \mathbf{Y}^{-1} \tag{2.60}$$

und

$$\mathbf{S} = (\mathbf{Z} + \mathbf{I})^{-1}(\mathbf{Z} - \mathbf{I}). \tag{2.61}$$

Mit den hier aufgezeigten Beziehungen kann die Tor-Randbedingung (2.48e) des Randwertproblems 2.6.1 in der Form

$$\hat{\boldsymbol{n}} \times (\mu_r^{-1} \nabla \times \boldsymbol{E}) \times \hat{\boldsymbol{n}} = -jk_0\eta_0 \sum_{k=1}^{M} i_k \boldsymbol{h}_t^k \quad \text{auf } \Gamma_{\mathrm{WG}} \tag{2.62}$$

geschrieben werden. In Abschnitt 3.3 wird dargestellt, wie aus der Lösung $\boldsymbol{E}$ die äquivalenten Spannungen u_k auf Γ_{WG} folgen und somit die gesamte Impedanzmatrix $\mathbf{Z}$ der untersuchten Struktur zu bestimmen ist.

Kapitel 3

Finite-Elemente-Methode für Anregungsprobleme

Wie in Abschnitt 2.6 beschrieben, ist eine Lösung für das Randwertproblem 2.6.1 in der Regel nicht auf analytischem Wege zu finden. Stattdessen kann mit Hilfe numerischer Verfahren eine Approximation der gesuchten Feldgröße bestimmt werden.

Grundlage der in dieser Arbeit entwickelten Verfahren bildet die FE-Methode. Diese basiert auf einer Diskretisierung des Feldgebiets in primitive Grundkörper, auf denen lokale Ansatzfunktionen definiert sind. Mit den Ansatzfunktionen wird eine Approximation der gesuchten Feldgröße ermittelt.

Die Stärke der FE-Methode liegt insbesondere in der großen Flexibilität des Einsatzbereiches, da die Verwendung angepasster Grundkörper eine beliebig genaue Approximation der Geometrie erlaubt. Zusätzlich können die Ansatzfunktionen auf die vorliegende Problemstellung zugeschnitten werden. Insbesondere ist es möglich mittels geeigneter Fehlerschätzer festzustellen, in welchen Bereichen des Feldgebiets die Diskretisierung die geforderte Genauigkeit der approximierten Lösung nicht gewährleistet, so dass in jenen Bereichen die geometrische Diskretisierung verfeinert, oder der Grad der Approximationsfunktionen erhöht werden kann. Damit ist eine hohe Konvergenzrate des Verfahrens und das Erreichen einer hohen Genauigkeit der Approximation gewährleistet [GB86] [SB91] [Lös13]. Ein weiterer Vorzug der FE-Methode stellt die Möglichkeit dar, Materialsprünge im Feldgebiet durch die Diskretisierung sehr genau nachzubilden. Somit können die physikalischen Stetigkeitsbedingungen an Materialgrenzen exakt in die diskrete Domäne übertragen und auch orts- und richtungsabhängige Materialeigenschaften mit hoher Genauigkeit berücksichtigt werden.

In den folgenden Abschnitten werden ausgehend vom mathematischen Modell die Anforderungen an die Feldgrößen bestimmt und entsprechende Funktionenräume de-

finiert. Mit der Galerkin-Methode [Cia78, S. 37] erfolgt der Übergang zur schwachen Formulierung des Randwertproblems 2.6.1.

Entsprechend der Problemstellung 2.6.1 wird als gesuchte Funktion das elektrische Feld $\boldsymbol{E}$ gewählt. Alternativ kann auch eine Formulierung in den Potenzialen $\boldsymbol{A}$ und V aufgestellt werden, die über die Beziehungen

$$\boldsymbol{B} = \nabla \times \boldsymbol{A} \qquad \text{und} \tag{3.1}$$

$$\boldsymbol{E} = -\nabla V \tag{3.2}$$

definiert sind. Im Gegensatz zur $\boldsymbol{E}$-Feld-Formulierung wird mit der Potenzialformulierung eine bessere numerische Stabilität im Niederfrequenzbereich erzielt [DEB96]. Für die in der vorliegenden Arbeit gezeigten Anwendungen aus der Hochfrequenz- bzw. Mikrowellentechnik ist die Feldformulierung uneingeschränkt geeignet. Der Vorteil gegenüber einer stabileren Potenzialformulierungen liegt in der direkten physikalischen Interpretierbarkeit der Ergebnisse und der vergleichsweise einfachen Implementierung als Computer-Programm. Damit können an der $\boldsymbol{E}$-Feld-Formulierung alle wesentlichen Eigenschaften der Modellordnungsreduktion anschaulich aufgezeigt werden. Die Anwendung der Modellordnungsreduktionsverfahren auf Potenzialformulierungen ist ebenso möglich [Far07].

3.1 Funktionenräume und schwache Form

Zur Konstruktion der schwachen Form wird die Differenzialgleichung (2.48a) mit einer geeigneten *Testfunktion* $\boldsymbol{v}$ multipliziert und über das Gebiet Ω integriert:

$$\int_\Omega \boldsymbol{v} \cdot \left(\nabla \times \mu_r^{-1} \nabla \times \boldsymbol{E}\right) \,\mathrm{d}\Omega + jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot (\sigma \boldsymbol{E}) \,\mathrm{d}\Omega - k_0^2 \int_\Omega \boldsymbol{v} \cdot (\varepsilon_r \boldsymbol{E}) \,\mathrm{d}\Omega = -jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot (\boldsymbol{J}_e) \,\mathrm{d}\Omega. \tag{3.3}$$

Mittels der Identität

$$\nabla \cdot (\boldsymbol{u} \times \boldsymbol{v}) = \boldsymbol{v} \cdot (\nabla \times \boldsymbol{u}) - \boldsymbol{u} \cdot (\nabla \times \boldsymbol{v}) \tag{3.4}$$

und der Festlegung $\boldsymbol{u} = \mu_r^{-1} \nabla \times \boldsymbol{E}$ lässt sich dieser Ausdruck schreiben als

$$\int_\Omega (\nabla \times \boldsymbol{v}) \cdot \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \,\mathrm{d}\Omega + jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot (\sigma \boldsymbol{E}) \,\mathrm{d}\Omega - k_0^2 \int_\Omega \boldsymbol{v} \cdot (\varepsilon_r \boldsymbol{E}) \,\mathrm{d}\Omega = -jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot \boldsymbol{J}_e \,\mathrm{d}\Omega - \int_\Omega \nabla \cdot \left[\left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \times \boldsymbol{v}\right] \,\mathrm{d}\Omega. \tag{3.5}$$

Anwendung des Gaußschen Satzes auf der rechten Seite und Berücksichtigung des Induktionsgesetzes (2.10a) führt auf

$$\int_\Omega (\nabla \times \boldsymbol{v}) \cdot \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \,\mathrm{d}\Omega + jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot (\sigma \boldsymbol{E}) \,\mathrm{d}\Omega - k_0^2 \int_\Omega \boldsymbol{v} \cdot (\varepsilon_r \boldsymbol{E}) \,\mathrm{d}\Omega$$
$$= -jk_0\eta_0 \int_\Omega \boldsymbol{v} \cdot \boldsymbol{J}_e \,\mathrm{d}\Omega + jk_0\eta_0 \oint_\Gamma (\boldsymbol{H} \times \boldsymbol{v}) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma. \quad (3.6)$$

Damit gewährleistet ist, dass die auftretenden Integrale definierte Werte annehmen, müssen die Funktionen $\boldsymbol{v}, \boldsymbol{E}$ und deren Rotation quadratisch integrierbar sein. Hierzu wird der Raum

$$H(\mathrm{rot};\Omega) := \left\{ \boldsymbol{u} \in L^2(\Omega) \,\middle|\, \nabla \times \boldsymbol{u} \in L^2(\Omega) \right\} \quad (3.7)$$

eingeführt, wobei

$$L^2(\Omega) := \left\{ \boldsymbol{u} : \Omega \to \mathbb{C}^3 \,\middle|\, \int_\Omega |\boldsymbol{u} \cdot \boldsymbol{u}| \,\mathrm{d}\Omega < \infty \right\} \quad (3.8)$$

den Raum der quadratisch Lebesque-integrierbaren Funktionen darstellt. Mit dem Skalarprodukt

$$(\boldsymbol{u}, \boldsymbol{v})_{H(\mathrm{rot};\Omega)} := \int_\Omega \boldsymbol{u} \cdot \boldsymbol{v} \,\mathrm{d}\Omega + \int_\Omega (\nabla \times \boldsymbol{u}) \cdot (\nabla \times \boldsymbol{v}) \,\mathrm{d}\Omega \quad (3.9)$$

ist der Raum $H(\mathrm{rot};\Omega)$ vollständig und somit ein Hilbertraum [Bos98, S. 128]. Die eingeprägte Stromdichte muss der Bedingung

$$\boldsymbol{J}_e \in L^2(\Omega) \quad (3.10)$$

genügen.

Um das Randwertproblem 2.6.1 vollständig in seiner schwachen Form zu repräsentieren, sind auch die Randbedingungen im Ausdruck (3.6) zu berücksichtigen. Hierzu wird das Randintegral in die Anteile

$$\oint_\Gamma \cdots = \int_{\Gamma_\mathrm{D}} \cdots + \int_{\Gamma_\mathrm{N}} \cdots + \int_{\Gamma_\mathrm{R}} \cdots + \int_{\Gamma_\mathrm{WG}} \ldots \quad (3.11)$$

aufgespalten und die Terme getrennt betrachtet:

- Auf Γ_D verschwindet gemäß (2.48b) die Tangentialkomponente des elektrischen Feldes, so dass für die Lösung $\boldsymbol{E}$ die Einschränkung auf den Unterraum

$$H_\mathrm{D}(\mathrm{rot};\Omega) := \left\{ \boldsymbol{w} \in H(\mathrm{rot};\Omega) \middle| \hat{\boldsymbol{n}} \times (\boldsymbol{w} \times \hat{\boldsymbol{n}}) = 0 \text{ auf } \Gamma_\mathrm{D} \right\} \quad (3.12)$$

einzubringen ist. Entsprechend dürfen die Tangentialkomponenten der Testfunktionen $\boldsymbol{v}$ auf dem Dirichlet-Rand ebenfalls keinen Beitrag leisten, weshalb $H_{\mathrm{D}}(\mathrm{rot};\Omega)$ auch als Testraum geeignet ist. Die Integration mit $\boldsymbol{v} \in H_{\mathrm{D}}(\mathrm{rot};\Omega)$ über Γ_{D} liefert demnach keinen Beitrag, und es gilt

$$\int_{\Gamma_{\mathrm{D}}} (\boldsymbol{v} \times \hat{\boldsymbol{n}}) \cdot \boldsymbol{H} \,\mathrm{d}\Gamma = 0 \qquad \forall \boldsymbol{v} \in H_{\mathrm{D}}(\mathrm{rot};\Omega). \tag{3.13}$$

- Der Anteil des *homogenen* Neumann-Randes verschwindet wegen (2.48c) in natürlicher Weise, weshalb diese Randbedingung auch als *natürliche Randbedingung* bezeichnet wird.

- Auf dem absorbierenden Rand, der die Abstrahlung in den Freiraum beschreibt, kann gemäß (2.48d) die magnetische Erregung durch die elektrische Feldstärke ausgedrückt, und der Anteil

$$jk_0\eta_0 \int_{\Gamma_{\mathrm{R}}} (\boldsymbol{H} \times \boldsymbol{v}) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma = -jk_0 \int_{\Gamma_{\mathrm{R}}} (\boldsymbol{v} \times \hat{\boldsymbol{n}}) \cdot (\boldsymbol{E} \times \hat{\boldsymbol{n}}) \,\mathrm{d}\Gamma \tag{3.14}$$

  auf die linke Seite gebracht werden.

- Die Integration über die modalen Tore Γ_{WG} liefert mit (2.48e) und (2.54) den Anregungsterm

$$\int_{\Gamma_{\mathrm{WG}}} (\boldsymbol{H} \times \boldsymbol{v}) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma = \sum_{k=1}^{M} i_k \int_{\Gamma_{\mathrm{WG}}} \left(\boldsymbol{h}_t^k \times \boldsymbol{v}\right) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma. \tag{3.15}$$

Damit lautet die schwache Form des Randwertproblems 2.6.1 wie folgt:

Problem 3.1.1. Gesucht ist jene Funktion $\boldsymbol{E} \in H_{\mathrm{D}}(\mathrm{rot};\Omega)$, die die Beziehung

$$\begin{aligned}\int_{\Omega} (\nabla \times \boldsymbol{v}) \cdot \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \,\mathrm{d}\Omega + jk_0\eta_0 \int_{\Omega} \boldsymbol{v} \cdot (\sigma \boldsymbol{E}) \,\mathrm{d}\Omega \\ + jk_0 \int_{\Gamma_{\mathrm{R}}} (\boldsymbol{v} \times \hat{\boldsymbol{n}}) \cdot (\boldsymbol{E} \times \hat{\boldsymbol{n}}) \,\mathrm{d}\Gamma - k_0^2 \int_{\Omega} \boldsymbol{v} \cdot (\varepsilon_r \boldsymbol{E}) \,\mathrm{d}\Omega \\ = -jk_0\eta_0 \int_{\Omega} \boldsymbol{v} \cdot \boldsymbol{J}_e \,\mathrm{d}\Omega + jk_0\eta_0 \sum_{k=1}^{M} i_k \int_{\Gamma_{\mathrm{WG}}} \left(\boldsymbol{h}_t^k \times \boldsymbol{v}\right) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma,\end{aligned} \tag{3.16}$$

für alle $\boldsymbol{v} \in H_{\mathrm{D}}(\mathrm{rot};\Omega)$ erfüllt.

Die wesentliche Modifikation gegenüber der starken Form 2.6.1 liegt darin, dass nur noch Ableitungen erster Ordnung der beteiligten Funktionen auftreten. Darüber

hinaus muss die Lösung $\boldsymbol{E}$ lediglich das globale Kriterium (3.3) erfüllen, d. h. es wird gefordert, dass das mit $\boldsymbol{v}$ gewichtete und über Ω aufintegrierte Residuum

$$\boldsymbol{r} := \nabla \times \mu_r^{-1} \nabla \times \boldsymbol{E} + jk_0\eta_0\sigma\boldsymbol{E} - k_0^2\varepsilon_r\boldsymbol{E} - (-jk_0\eta_0\boldsymbol{J}_e) \tag{3.17}$$

verschwindet. Daher wird das Verfahren auch als *Methode der gewichteten Residuen* bezeichnet [ZT00, S. 42].

3.2 Finite-Elemente-Ansatz

3.2.1 Triangulierung des Feldgebiets

Die grundlegende Idee der FE-Methode besteht in der Zerlegung des Feldgebiets in primitive Grundkörper, der *Triangulierung* $T_h(\Omega)$. In der vorliegenden Arbeit entspricht die Triangulierung einer Zerlegung in eine endliche Anzahl von Simplices, welche für $\Omega \subset \mathbb{R}^2$ durch Dreiecke und für $\Omega \subset \mathbb{R}^3$ durch Tetraeder gegeben sind. Die Simplices S erfüllen folgende Eigenschaften [Mon03, S. 112]:

1. Jedes S ist eine offene Menge.
2. Sind $S, S' \in T_h(\Omega)$ zwei verschiedene Simplices, dann gilt: $S \cap S' = \emptyset$.
3. $\overline{\Omega} = \bigcup_{S \in T_h(\Omega)} \overline{S}$.

Hierbei bezeichnet $\overline{\Omega}$ den Abschluss von Ω und $\overline{S}$ entsprechend den Abschluss des Simplex S.

Ausgangspunkt der weiteren Betrachtungen in der vorliegenden Arbeit ist eine konsistente Triangulierung $T_h(\Omega)$ des Feldgebiets Ω. Eine Triangulierung wird als *konsistent* bezeichnet, wenn ausgehend von zwei Tetraedern $T, T' \in T_h(\Omega)$ mit $\overline{T} \cap \overline{T'} \neq \emptyset$ sich diese einzig auf eine der folgenden Arten berühren [Mon03, S. 112]:

- Beide Tetraeder berühren sich in genau einem Punkt und dieser Punkt ist ein Knoten beider Tetraeder.
- Beide Tetraeder berühren sich entlang einer gemeinsamen Kante, deren Knoten beiden Tetraedern eigen sind.
- Beide Tetraeder berühren sich an einem gemeinsamen Dreieck, dessen Knoten beiden Tetraedern eigen sind.

3.2.2 Finite-Elemente-Ansatzfunktionen

Zur numerischen Bestimmung einer Lösung $\boldsymbol{E} \in H_{\mathrm{D}}(\mathrm{rot};\Omega)$ für das Problem 3.1.1 wird ein Ansatz der Form

$$\tilde{\boldsymbol{E}} = \sum_{i=1}^{N} x_i \boldsymbol{w}_i \in W_h \tag{3.18}$$

gewählt, wobei der endlich-dimensionale Unterraum

$$W_h \subset H_{\mathrm{D}}(\mathrm{rot};\Omega) \tag{3.19}$$

von den linear unabhängigen Ansatzfunktionen $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N$ aufgespannt wird, also

$$W_h := \mathrm{span}\{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_N\} \quad \text{und} \quad \dim W_h = N. \tag{3.20}$$

Die Konstruktion des Unterraums W_h und damit auch die Anzahl der Freiheitsgrade N ist unmittelbar mit der Triangulierung $T_h(\Omega)$ des Feldgebiets verknüpft.

Zur Beschreibung des elektrischen Feldes $\boldsymbol{E}$ oder der magnetischen Erregung $\boldsymbol{H}$ in der FE-Methode hat sich die Wahl der *Kantenelemente* basierend auf den Whitney-Formen als der natürliche Ansatz herausgestellt [Whi57] [Ned80] [Bos88]. Die Bezeichnung als Kantenelemente ist auf die Eigenschaft zurückzuführen, dass die Freiheitsgrade x_i im Ansatz (3.18) dem Wert des Integrals entlang der zugeordneten Simplexkante entsprechen; zumindest gilt dies für die Ansatzfunktionen niedrigster Ordnung. Das wesentliche Merkmal der Kantenelemente liegt zudem darin, dass die Tangentialkomponenten der Funktionen $\boldsymbol{w}_i$ über Elementgrenzen hinweg stetig sind. In der Literatur wird daher der Begriff der $H(\mathrm{rot};\Omega)$-konformen Elemente verwendet. Die tangentiale Stetigkeit bildet exakt die physikalischen Eigenschaften der Feldgrößen $\boldsymbol{E}$ und $\boldsymbol{H}$ an Grenzflächen von Bereichen unterschiedlicher Materialien ab.

In [Ned80] wird dargestellt, wie endlich-dimensionale Räume mit oben beschriebenen Eigenschaften durch vektorwertige Polynome aufgespannt werden können. Vorschriften zur konkreten Konstruktion entsprechender Basen liefern [Lee90], [Web99] und [Ing06], wobei letztere im Rahmen der vorliegenden Arbeit verwendet wird.

Damit für die Formulierung 3.1.1 ein lineares Gleichungssystem aufgestellt werden kann, ist auch der Raum der Testfunktionen auf einen endlichdimensionalen Unterraum

$$V_h := \mathrm{span}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\} \subset H_{\mathrm{D}}(\mathrm{rot};\Omega) \tag{3.21}$$

einzuschränken.

Mit dem Ansatz (3.18) und der Vernachlässigung eingeprägter Ströme, $\boldsymbol{J}_e = 0$, ergibt sich somit die algebraisierte Darstellung des Randwertproblems zu

$$\left(\mathbf{S} + jk_0\mathbf{D} - k_0^2\mathbf{T}\right)\mathbf{x} = jk_0\eta_0\mathbf{b}, \tag{3.22}$$

mit den Systemmatrizen

$$\mathbf{S} = [s_{ij}] \in \mathbb{R}^{N\times N}, \tag{3.23}$$
$$\mathbf{D} = [d_{ij}] \in \mathbb{R}^{N\times N}, \tag{3.24}$$
$$\mathbf{T} = [t_{ij}] \in \mathbb{R}^{N\times N} \tag{3.25}$$

sowie

$$\mathbf{x} = [x_i] \in \mathbb{C}^N, \tag{3.26}$$
$$\mathbf{b} = [b_i] \in \mathbb{R}^N. \tag{3.27}$$

Die Matrixeinträge berechnen sich hierbei wie folgt:

$$s_{ij} = \int_\Omega (\nabla \times \boldsymbol{v}_i) \cdot \left(\mu_r^{-1} \nabla \times \boldsymbol{w}_j\right) \,\mathrm{d}\Omega, \tag{3.28a}$$
$$d_{ij} = \eta_0 \int_\Omega \boldsymbol{v}_i \cdot (\sigma \boldsymbol{w}_j) \,\mathrm{d}\Omega + \int_{\Gamma_\mathrm{R}} (\boldsymbol{v}_i \times \hat{\boldsymbol{n}}) \cdot (\boldsymbol{w}_j \times \hat{\boldsymbol{n}}) \,\mathrm{d}\Gamma, \tag{3.28b}$$
$$t_{ij} = \int_\Omega \boldsymbol{v}_i \cdot (\varepsilon_r \boldsymbol{w}_j) \,\mathrm{d}\Omega, \tag{3.28c}$$
$$b_i = \sum_{k=1}^{M} i_k \int_{\Gamma_\mathrm{WG}} \left(c_i^k \boldsymbol{w}_i \times \boldsymbol{v}_i\right) \cdot \hat{\boldsymbol{n}} \,\mathrm{d}\Gamma. \tag{3.28d}$$

In (3.28d) treten die noch nicht näher bestimmten *Wellenformkoeffizienten* c_i^k auf, deren Berechnung im folgenden Abschnitt erläutert wird.

3.3 Transfinite-Elemente-Methode

Wellenleitertore sind in der Modellierung so zu setzen, dass der Abstand zu Diskontinuitäten groß genug ist, um im Wellenleiter nicht ausbreitungsfähige modale Wellenformen ausreichend zu unterdrücken. Damit setzen sich an den Wellenleitertoren Γ_WG die Felder wegen (2.47) aus endlich vielen Wellenformen $(\boldsymbol{E}^m, \boldsymbol{H}^m, \gamma_m)$ zusammen. Entsprechend können die Freiheitsgrade, die den Toren zugeordnet sind, mit einem globalen Ansatzes so restringiert werden, dass nur modale Wellenformen durch die transversalen Feldkomponenten darstellbar sind. Durch gewichtete Überlagerung der modalen Wellenformen kann somit jede zulässige Feldverteilung am Wellenleitertor konstruiert werden.

Für beliebige Wellenleiterquerschnitte können durch numerische Lösung des Problems 2.5.1 die notwendigen Wellenformen für ein gewähltes Tor bestimmt werden. Das dieser Arbeit zugrundeliegende Computerprogramm basiert auf der in [FHDE04]

beschriebenen Vorgehensweise, die auch eine Behandlung wesentlich allgemeinerer Wellenleiterklassen, als die in Problem 2.5.1 definierten, erlaubt. An dieser Stelle wird lediglich darauf hingewiesen, dass die Berechnung der Wellenformen auf der FE-Triangulierung am Wellenleiterquerschnitt geschieht und somit die gefundenen Lösungen direkt in das Anregungsproblem eingebracht werden können. Zu diesem Zweck wird eine Partitionierung der Ansatzfunktionen in der Form

$$W_h = W_h^\Gamma \oplus W_h^\Omega \tag{3.29}$$

vorgenommen, wobei

$$W_h^\Gamma := \left\{ \boldsymbol{w} \in W_h \middle| \hat{\boldsymbol{n}} \times \boldsymbol{w} \neq 0 \text{ auf } \Gamma_{\text{WG}} \right\} \tag{3.30}$$

jene Funktionen umfasst, welche am Wellenleiter nicht verschwindende Tangentialkomponenten aufweisen, während

$$W_h^\Omega := \left\{ \boldsymbol{w} \in W_h \middle| \hat{\boldsymbol{n}} \times \boldsymbol{w} = 0 \text{ auf } \Gamma_{\text{WG}} \right\} \tag{3.31}$$

das Komplement zu dieser Menge darstellt, also all jene Funktionen, deren Tangentialkomponente auf Γ_{WG} keinen Beitrag leistet.

Die Zerlegung der gesuchten Lösung (3.18) lautet demnach

$$\tilde{\boldsymbol{E}} = \tilde{\boldsymbol{E}}^\Omega + \tilde{\boldsymbol{e}}_t \tag{3.32}$$

mit

$$\tilde{\boldsymbol{E}}^\Omega \in W_h^\Omega \quad \text{und} \quad \tilde{\boldsymbol{e}}_t \in W_h^\Gamma, \tag{3.33}$$

wobei hier die Nomenklatur dem Ansatz (2.29) folgt. Mit der Einschränkung, dass am Wellenleiterquerschnitt Γ_{WG} nur ein diskretes Wellenformenspektrum vorliegt, kann der zweite Term in (3.32) wegen (2.47a) in der Form

$$\tilde{\boldsymbol{e}}_t = \sum_{k=1}^{M} (a_k + b_k) \tilde{\boldsymbol{e}}_t^k \overset{(2.53)}{=} \sum_{k=1}^{M} u_k \tilde{\boldsymbol{e}}_t^k \tag{3.34}$$

geschrieben werden. Im numerischen Kontext stellt die approximierte Wellenform

$$\tilde{\boldsymbol{e}}_t^k := \sum_{j=1}^{N_{\text{WG}}} c_j^k \boldsymbol{w}_j \tag{3.35}$$

mit den modalen Koeffizienten $c_j^k \in \mathbb{C}$ eine globale Ansatzfunktion auf Γ_{WG} dar, eine sogenannte *TFE-Ansatzfunktion* (TFE: Transfinite-Elemente). Die Koeffizienten ergeben sich aus der Lösung des Eigenwertproblems 2.5.1 für Wellenformen, wobei eine Normierung derart vorgenommen wird, dass die Orthogonalitätsbeziehung (2.42) auch im diskreten erhalten bleibt, also

$$\int_{\Gamma_{\text{WG}}} \left(\tilde{\boldsymbol{e}}_t^m \times \tilde{\boldsymbol{h}}_t^n \right) \cdot \hat{\boldsymbol{n}} \, \mathrm{d}\Gamma = -\delta_{mn} \tag{3.36}$$

erfüllt ist. Das negative Vorzeichen resultiert aus der Tatsache, dass $\hat{\boldsymbol{n}}$ aus dem Feldgebiet heraus zeigt und somit der Ausbreitungsrichtung $\hat{\boldsymbol{z}}$ der Welle entgegengesetzt ist. Mit den hier aufgezeigten Beziehungen kann der Ansatz (3.32) in der Form

$$\tilde{\boldsymbol{E}} = \sum_{i=1}^{N_\Omega} x_i \boldsymbol{w}_i + \sum_{k=1}^{M} u_k \sum_{j=1}^{N_{\mathrm{WG}}} c_j^k \boldsymbol{w}_j = \sum_{i=1}^{N_\Omega} x_i \boldsymbol{w}_i + \sum_{k=1}^{M} u_k \tilde{\boldsymbol{e}}_t^k \tag{3.37}$$

mit

$$\boldsymbol{w}_i \in W_h^\Omega \quad \text{und} \quad \tilde{\boldsymbol{e}}_t^k \in W_h^{\mathrm{TF}} \tag{3.38}$$

geschrieben werden, wobei

$$W_h^{\mathrm{TF}} := \mathrm{span}\{\tilde{\boldsymbol{e}}_t^1, \ldots, \tilde{\boldsymbol{e}}_t^M\} \subset W_h^\Gamma \tag{3.39}$$

den durch die diskreten Wellenformen $\tilde{\boldsymbol{e}}_t^k$ aufgespannten Unterraum von W_h^Γ bezeichnet.

Einsetzen von (3.37) in (3.16) führt schließlich auf die diskretisierte schwache Formulierung des Randwertproblems:

Problem 3.3.1. Gesucht ist jene Funktion $\boldsymbol{E} \in W_h^\Omega \oplus W_h^{\mathrm{TF}}$, die

$$\begin{aligned} &\int_\Omega (\nabla \times \boldsymbol{w}_i) \cdot \left(\mu_r^{-1} \nabla \times \boldsymbol{E}\right) \mathrm{d}\Omega + jk_0\eta_0 \int_\Omega \boldsymbol{w}_i \cdot (\sigma \boldsymbol{E}) \, \mathrm{d}\Omega \\ &\qquad + jk_0 \int_{\Gamma_{\mathrm{R}}} (\boldsymbol{w}_i \times \hat{\boldsymbol{n}}) \cdot (\boldsymbol{E} \times \hat{\boldsymbol{n}}) \, \mathrm{d}\Gamma - k_0^2 \int_\Omega \boldsymbol{w}_i \cdot (\varepsilon_r \boldsymbol{E}) \, \mathrm{d}\Omega \\ &\qquad = -jk_0\eta_0 \int_\Omega \boldsymbol{w}_i \cdot \boldsymbol{J}_e \, \mathrm{d}\Omega + jk_0\eta_0 \sum_{k=1}^{M} i_k \int_{\Gamma_{\mathrm{WG}}} \left(\tilde{\boldsymbol{h}}_t^k \times \boldsymbol{w}_i\right) \cdot \hat{\boldsymbol{n}} \, \mathrm{d}\Gamma, \end{aligned} \tag{3.40}$$

für alle $\boldsymbol{w}_i \in W_h^\Omega \oplus W_h^{\mathrm{TF}}$ erfüllt.

Im Sinne der Galerkin-Methode wird der Testraum identisch zum Ansatzraum $W_h^\Omega \oplus W_h^{\mathrm{TF}}$ gewählt. Das Randintegral über Γ_{WG} liefert wegen (3.36) für Testfunktionen $\tilde{\boldsymbol{e}}_t^n \in W_h^\Gamma$ mit

$$\sum_{k=1}^{M} i_k \int_{\Gamma_{\mathrm{WG}}} \left(\tilde{\boldsymbol{h}}_t^k \times \tilde{\boldsymbol{e}}_t^n\right) \cdot \hat{\boldsymbol{n}} \, \mathrm{d}\Gamma = \sum_{k=1}^{M} i_k \delta_{nk} = i_n \tag{3.41}$$

gerade die äquivalenten Ströme und somit die komplexen Amplituden der modalen Wellenformen an der Tor-Randbedingung.

3.4 Die FE-Formulierung als LTI-System

Im Gleichungssystem (3.22), in dem die Restriktionen am Wellenleiterquerschnitt noch nicht explizit eingebracht sind, lassen sich die Systemmatrizen $\mathbf{S}, \mathbf{D}, \mathbf{T}$ derart umsortieren, dass eine Partitionierung in der Form

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{II} & \mathbf{A}_{IT} \\ \mathbf{A}_{TI} & \mathbf{A}_{TT} \end{pmatrix}, \mathbf{A} \in \{\mathbf{S}, \mathbf{D}, \mathbf{T}\}, \tag{3.42}$$

vorliegt, wobei $\mathbf{A}_{II}$ den Funktionen aus W_h^{Ω}, und $\mathbf{A}_{TT}$ denjenigen aus W_h^{Γ} zugeordnet ist. $\mathbf{A}_{IT}$ und $\mathbf{A}_{TI} = \mathbf{A}_{IT}^T$ stellen die *Koppelmatrizen* dar, die die Kopplung zwischen dem Inneren der Struktur und den Tor-Randbedingungen gewährleisten. Gemäß der Restriktion (3.37) werden die Freiheitsgrade an den Tor-Randbedingungen im Sinne globaler Ansätze verstanden. Hierzu werden die modalen Vektoren

$$\mathbf{c}^k = (c_1^k, c_2^k, \ldots, c_{N_{\mathrm{WG}}}^k)^T, k = 1, 2, \ldots, M, \tag{3.43}$$

und mit diesen die Restriktionsmatrix

$$\mathbf{C} := \begin{bmatrix} \mathbf{c}^1 & \mathbf{c}^2 & \cdots & \mathbf{c}^M \end{bmatrix} \tag{3.44}$$

eingeführt. Die TFE-Systemmatrizen ergeben sich dann durch Anwendung von $\mathbf{C}$ in der Form

$$\mathbf{A}_{\mathrm{TF}} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C}^T \end{pmatrix} \begin{pmatrix} \mathbf{A}_{II} & \mathbf{A}_{IT} \\ \mathbf{A}_{TI} & \mathbf{A}_{TT} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} \end{pmatrix} = \begin{pmatrix} \mathbf{A}_{II} & \mathbf{A}_{IC} \\ \mathbf{A}_{CI} & \mathbf{A}_{CC} \end{pmatrix}. \tag{3.45}$$

Hierin sind die Einträge mit Index C den modalen Ansätzen aus W_h^{TF} zugeordnet. Durch die Restriktion der Matrizen verringert sich die Anzahl der Unbekannten dahingehend, dass jede Wellenform nur noch mit *einem* Koeffizienten, der äquivalenten Spannung u_i, verknüpft ist. Für den Lösungsvektor aus (3.22) führt die Partitionierung auf die Darstellung

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_I \\ \mathbf{x}_T \end{pmatrix} \rightarrow \mathbf{x}_{\mathrm{TF}} = \begin{pmatrix} \mathbf{x}_I \\ \mathbf{u} \end{pmatrix}, \tag{3.46}$$

mit

$$\mathbf{u} := (u_1, u_2, \ldots, u_M)^T. \tag{3.47}$$

Vollkommen analog folgt für den Anregungsvektor $\boldsymbol{b}$ die Darstellung

$$\mathbf{b} = \begin{pmatrix} \mathbf{0} \\ \mathbf{b}_T \end{pmatrix} \rightarrow \mathbf{b}_{\mathrm{TF}} = \begin{pmatrix} \mathbf{0} \\ \mathbf{i} \end{pmatrix}, \tag{3.48}$$

mit

$$\mathbf{i} := (i_1, i_2, \ldots, i_M)^T. \tag{3.49}$$

Das vollständige TFE-Gleichungssystem lautet schließlich

$$\left(\mathbf{S}_{\text{TF}} + jk_0\mathbf{D}_{\text{TF}} - k_0^2\mathbf{T}_{\text{TF}}\right)\mathbf{x}_{\text{TF}} = jk_0\eta_0\mathbf{b}_{\text{TF}}, \tag{3.50}$$

wobei sich die Matrixeinträge wie in (3.28) berechnen, jedoch mit $\boldsymbol{v}_i, \boldsymbol{w}_j \in W_h^\Omega \oplus W_h^{\text{TF}}$. Die Dimension des TFE-Systems ergibt sich zu

$$N_{\text{TF}} := \dim W_h^\Omega \oplus W_h^{\text{TF}} = N_\Omega + M. \tag{3.51}$$

Im Anregungsvektor $\mathbf{b}_{\text{TF}}$ entspricht jeder Eintrag i_k der Anregung der zugeordneten Wellenform am entsprechenden Tor. Durch sukzessive Anregung der einzelnen Wellenformen mit dem Einheitsstrom $i_k = 1$ kann die jeweilige Systemantwort und somit die gesamte Übertragungsmatrix des Systems bestimmt werden. In kompakter Blockschreibweise ist das TFE-System gegeben durch

$$\Sigma_{\text{TF}}(jk_0) = \left\{ \begin{array}{rl} \left(\mathbf{S}_{\text{TF}} + jk_0\mathbf{D}_{\text{TF}} + (jk_0)^2\mathbf{T}_{\text{TF}}\right)\mathbf{X}_{\text{TF}} & = jk_0\eta_0\mathbf{B}_{\text{TF}}, \\ \mathbf{Z} & = \mathbf{B}_{\text{TF}}^T\mathbf{X}_{\text{TF}}, \end{array} \right. \tag{3.52}$$

mit

$$\mathbf{X}_{\text{TF}} := \begin{pmatrix} \mathbf{X}_I \\ \mathbf{U} \end{pmatrix} \in \mathbb{C}^{N_{\text{TF}} \times M}, \quad \mathbf{U} = \text{diag}\left(u_1, u_2, \ldots, u_M\right), \tag{3.53}$$

und

$$\mathbf{B}_{\text{TF}} := \begin{pmatrix} \mathbf{0}_I \\ \mathbf{I} \end{pmatrix} \in \mathbb{R}^{N_{\text{TF}} \times M}, \quad \mathbf{I} = \begin{pmatrix} 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}. \tag{3.54}$$

Die Ausgangsmatrix $\mathbf{Z} \in \mathbb{C}^{M \times M}$ stellt die äquivalenten Spannungen und Ströme in Relation und wird daher als *verallgemeinerte Impedanzmatrix* bezeichnet.

Gemäß der in Kapitel 4 näher erläuterten Eigenschaften ist (3.52) ein linear-zeitinvariantes System zweiter Ordnung.

Aufgrund des hier angewandten Galerkin-Ansatzes erfüllt die Lösung $\tilde{\boldsymbol{E}} \in W_h^\Omega \oplus W_h^{\text{TF}}$ die anschauliche Eigenschaft, dass das Residuum (3.17) im Sinne des L_2-Skalarprodukts orthogonal auf dem Ansatzraum $W_h^\Omega \oplus W_h^{\text{TF}}$ steht, also

$$\boldsymbol{r} \perp W_h^\Omega \oplus W_h^{\text{TF}}. \tag{3.55}$$

Aus der Definition des Gleichungssystems (3.22) lassen sich folgende weitere Eigenschaften der Systemmatrizen ablesen:

- Die *Steifigkeitsmatrix* $\mathbf{S}_{\text{TF}}$ ist symmetrisch positiv-semidefinit. Der Nullraum wird im Wesentlichen durch die Gradienten $\nabla\phi \in W_h^{\text{TF}}$ gebildet, da diese im Kern des Rotationsoperators liegen.

- Die *Dämpfungsmatrix* $\mathbf{D}_{\text{TF}}$ ist symmetrisch positiv-semidefinit.
- Die *Massenmatrix* $\mathbf{T}_{\text{TF}}$ ist stets symmetrisch positiv definit.

Das resultierende algebraische Problem bewahrt somit die wichtigen Systemeigenschaften des Ausgangsmodells, wie Passivität [VCK98, S. 348] und Reziprozität. Letztere folgt direkt aus der Symmetrie der Ausgangsmatrix $\mathbf{Z}$ [Poz05, S. 170].

3.4.1 TE- und TM-Wellen im FE-System

Das Einbringen der TFE-Restriktionen in das FE-System ist direkt mit der Normierung (3.36) verbunden, die nur im Zusammenhang mit TEM-Wellen von der Frequenz unabhängig ist. Wenn im betrachteten Problem TE- und TM-Wellen auftreten, sind zwar nach wie vor die transversalen Feldverläufe, und damit die modalen Vektoren in (3.43) frequenzunabhängig,

$$\mathbf{c}^k = \text{const}(k_0), \tag{3.56}$$

jedoch nicht die Ausbreitungskoeffizient,

$$\gamma \not\sim k_0. \tag{3.57}$$

Dies wird unmittelbar aus der Dispersionsgleichung (2.40) ersichtlich. Jedoch ist bei Vorhandensein von dispersiven Wellenformen die Normierung bzw. die Restriktion $\mathbf{C}$ nicht in jedem Frequenzpunkt neu zu bestimmen. Stattdessen kann mittels einer geeigneten Skalierung der Anregung $\mathbf{B}$ die explizite Frequenzabhängigkeit eingebracht werden.

Zunächst wird hierzu die Feldwellenimpedanz Z eingeführt, die die modalen Felder in der Form

$$\boldsymbol{e}_t^k(k_0) = Z_W^k(k_0)\boldsymbol{h}_t^k(k_0) \times \hat{\boldsymbol{z}},\ W \in \{\text{TE}, \text{TM}\} \tag{3.58}$$

verknüpft. Aus den Maxwell-Gleichungen lassen sich die konkreten Darstellungen

$$Z_{\text{TE}} = \frac{j\omega\mu}{\gamma} \qquad \text{für TE-Wellen,} \tag{3.59}$$

$$Z_{\text{TM}} = \frac{\gamma}{j\omega\epsilon} \qquad \text{für TM-Wellen} \tag{3.60}$$

ableiten. Um die explizite Frequenzabhängigkeit der TFE-Restriktion zu bestimmen, wird in einem nächsten Schritt eine Normierung für ein festes $k_0 = \hat{k}_0$ in der Form

$$\begin{aligned} -1 &= \int_{\Gamma_{\text{WG}}} \left(\boldsymbol{e}_t^k(\hat{k}_0) \times \boldsymbol{h}_t^k(\hat{k}_0)\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}\Gamma \\ &= \int_{\Gamma_{\text{WG}}} \left(\boldsymbol{e}_t^k(\hat{k}_0) \times \left[\frac{\hat{\boldsymbol{z}} \times \boldsymbol{e}_t^k(\hat{k}_0)}{Z_W^k(\hat{k}_0)}\right]\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}\Gamma \end{aligned} \tag{3.61}$$

vorgenommen. Für $k_0 \neq \hat{k}_0$ gilt dann wegen (3.56)

$$\begin{aligned}\int\limits_{\Gamma_{\mathrm{WG}}} \left(\boldsymbol{e}_t^k(k_0) \times \boldsymbol{h}_t^k(k_0)\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}\Gamma &\overset{(3.56)}{=} \int\limits_{\Gamma_{\mathrm{WG}}} \left(\boldsymbol{e}_t^k(\hat{k}_0) \times \boldsymbol{h}_t^k(k_0)\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}\Gamma \\ &= \int\limits_{\Gamma_{\mathrm{WG}}} \left(\boldsymbol{e}_t^k(\hat{k}_0) \times \left[\frac{\hat{\boldsymbol{z}} \times \boldsymbol{e}_t^k(k_0)}{Z_W^k(k_0)}\right]\right) \cdot \hat{\boldsymbol{n}}\, \mathrm{d}\Gamma. \end{aligned} \tag{3.62}$$

Ein Vergleich von (3.61) und (3.4.1) führt schließlich auf die gesuchte Skalierung

$$\xi_k(\hat{k}_0, k_0) := \frac{Z_W^k(\hat{k}_0)}{Z_W^k(k_0)} = \begin{cases} \frac{\sqrt{k_{c,k}^2 - k_0^2}}{\sqrt{k_{c,k}^2 - \hat{k}_0^2}} & \text{für TE-Wellen} \\ \frac{\sqrt{k_{c,k}^2 - \hat{k}_0^2}}{\sqrt{k_{c,k}^2 - k_0^2}} & \text{für TM-Wellen,} \end{cases} \tag{3.63}$$

wobei $k_{c,k}$ die Grenzwellenzahl zur entsprechenden Wellenform ist.

Die Zustandsgleichung in (3.52) ist somit bei Anwesenheit von TE- bzw. TM-Wellen um die Skalierung der Anregung in der Form

$$\left(\mathbf{S}_{\mathrm{TF}} + jk_0\mathbf{D}_{\mathrm{TF}} + (jk_0)^2\mathbf{T}_{\mathrm{TF}}\right)\mathbf{X}_{\mathrm{TF}} = j\hat{k}_0\eta_0\mathbf{B}_{\mathrm{TF}}\boldsymbol{\Xi}(k_0) \tag{3.64}$$

zu erweitern, wobei

$$\boldsymbol{\Xi}(\hat{k}_0, k_0) := \begin{pmatrix} \xi_1(\hat{k}_0, k_0) & & \\ & \ddots & \\ & & \xi_M(\hat{k}_0, k_0) \end{pmatrix} \in \mathbb{R}^{M \times M} \tag{3.65}$$

die frequenzabhängige Skalierung der auftretenden Wellenformen vorgibt.

Kapitel 4

Schnelle Frequenzgangberechnung linearer Systeme

Für eine Vielzahl technischer Anwendungen führt die mathematische Modellierung physikalischer Problemstellungen auf ein System von Differenzialgleichungen, wobei hier der wichtige Spezialfall der *LTI*-Systeme (engl. *Linear Time-invariant*) von besonderer Bedeutung ist. Solche Systeme können beispielsweise aus der Beschreibung elektrischer Netzwerke oder der Diskretisierung partieller Differenzialgleichungen resultieren. Aufgrund der praktischen Relevanz von LTI-Systemen existieren effiziente Methoden zur Charakterisierung derselben. LTI-Systeme erlauben beispielsweise die Beschreibung im Frequenzbereich, womit eine Algebraisierung der Zeitableitung einhergeht. Für hochdimensionale Systeme ist die Konstruktion einer expliziten Übertragungsfunktion in der Regel nicht praktikabel. Vielmehr wird das Übertragungsverhalten in diskreten Stützstellen in einem festgelegten Parameterintervall, zum Beispiel dem Frequenzbereich

$$\mathcal{I}^f = [f_{\min}, f_{\max}] \subset \mathbb{R}^+ \tag{4.1}$$

gesucht. Um schmalbandige Effekte, wie scharfe Resonanzen (engl. Spikes), aufzulösen oder die Rücktransformation in den Zeitbereich zu ermöglichen, muss der diskrete Frequenzbereich

$$\mathcal{I}_h^f = \left\{f_1, f_2, \ldots, f_{N_f}\right\} \quad \text{mit } f_1 = f_{\min} < f_2 < \cdots < f_{N_f} = f_{\max} \tag{4.2}$$

ausreichend hoch aufgelöst sein.

Aufgrund der kompakteren Darstellung wird für die folgenden Betrachtungen auch der entsprechende Wellenzahlbereich

$$\mathcal{I}_h^{k_0} = \left\{k_{0,1}, k_{0,2}, \ldots, k_{0,N_f}\right\} \text{ mit } k_{0,i} = 2\pi f_i \sqrt{\varepsilon_0 \mu_0},\ i = 1, \ldots, N_f, \tag{4.3}$$

bzw. das Intervall des komplexen Frequenzparameters s

$$\mathcal{I}_h^s = \left\{s_1, s_2, \ldots, s_{N_f}\right\} \text{ mit } s_i = jk_{0,i},\ i = 1, \ldots, N_f, \tag{4.4}$$

herangezogen.

Ein SISO-LTI-System (SISO, engl. *Single Input - Single Output*) kann mit den vorangegangenen Erläuterungen allgemein beschrieben werden durch

$$\Sigma(s) = \begin{cases} \sum_{i=0}^{L} s^i \mathbf{A}_i \mathbf{x}(s) &= \sum_{i=0}^{L} s^i \mathbf{b}_i u(s), \\ z(s) &= \sum_{i=0}^{L} s^i \mathbf{c}_i^T \mathbf{x}(s) + du(s) \end{cases} \tag{4.5}$$

mit Systemmatrizen $\mathbf{A}_i \in \mathbb{C}^{N \times N}$, *Anregungsvektoren* $\mathbf{b}_i$, *Ausgangsvektoren* $\mathbf{c}_i \in \mathbb{C}^N$, der *Eingangsgröße* $u(s)$, der *Ausgangsgröße* $z(s)$ und dem Lösungs- bzw. *Zustandsvektor* $\mathbf{x} \in \mathbb{C}^N$. Die *Systemordnung* ist definiert durch $L \in \mathbb{N}$, also der höchsten Potenz in s, die in den Termen von (4.5) vorkommt. Ist die maximale Potenz in einem der Terme $I < L$, sind die Komponenten mit Index $i > I$ zu Null zu setzen, also $\mathbf{g}_i = 0$ für $i > I, \mathbf{g}_i \in \{\mathbf{A}_i, \mathbf{b}_i, \mathbf{c}_i\}$. Für die weiteren Betrachtungen wird o. B. d. A.

$$u(s) \equiv 1 \tag{4.6}$$

angenommen und die Regularität der Matrix $\mathbf{A}_0$ vorausgesetzt. Außerdem wird zur Vereinfachung auch der Durchgangsfaktor

$$d \equiv 0 \tag{4.7}$$

gesetzt, da dieser durch die Operationen im MOR-Kontext ohnehin unverändert bleibt, und somit keinen Beitrag in den Fehlerbetrachtungen liefert. Die Einführung der Abkürzungen

$$\mathbf{A}(s) := \sum_{i=0}^{L} s^i \mathbf{A}_i, \tag{4.8}$$

$$\mathbf{b}(s) := \sum_{i=0}^{L} s^i \mathbf{b}_i \qquad \text{und} \tag{4.9}$$

$$\mathbf{c}(s) := \sum_{i=0}^{L} s^i \mathbf{c}_i \tag{4.10}$$

erlaubt die kompakte Darstellung des Systems (4.5) in der Form

$$\Sigma(s) = \begin{cases} \mathbf{A}(s)\mathbf{x}(s) &= \mathbf{b}(s), \\ z(s) &= \mathbf{c}^T(s)\mathbf{x}(s). \end{cases} \tag{4.11}$$

Die *Übertragungsfunktion* des LTI-Systems (4.5) ist als das Verhältnis von Ausgangs- zu Eingangsgröße definiert und lautet wegen (4.6)

$$H(s) := z(s) = \mathbf{c}(s)^T \mathbf{A}^{-1}(s)\mathbf{b}(s). \tag{4.12}$$

Zur Charakterisierung eines LTI-Systems im Parameterintervall $\mathcal{I}^s$ ist demnach die Übertragungsfunktion für alle Stützstellen $s_i \in \mathcal{I}_h^s$ auszuwerten. Entsprechend muss das Gleichungssystem

$$\mathbf{A}(s)\mathbf{x}(s) = \mathbf{b}(s) \tag{4.13}$$

N_f mal gelöst werden, wobei die Systemdimension N leicht Größenordnungen von mehreren Millionen aufweisen kann. Bei mehreren Hundert oder gar Tausend Stützstellen s_i wird die Simulation daher sehr zeitaufwändig, insbesondere wenn im Rahmen eines Optimierungsprozesses für unterschiedliche Systemvarianten die Übertragungsfunktion immer wieder neu berechnet werden muss.

Mit Hilfe der nachfolgend vorgestellten *Modellordnungsreduktionsverfahren* wird eine sehr gute Approximation der Übertragungsfunktion konstruiert, die im Vergleich zu vollen FE-Lösungen um ein Vielfaches schneller ausgewertet werden kann.

In weiterer Folge werden ausschließlich Projektionsverfahren behandelt, die für hochdimensionale, schwach besetzte Systeme besonders geeignet sind.

Aus Gründen der Anschaulichkeit wird vorerst der SISO-Fall mit skalaren Eingangs- und Ausgangsgrößen $u(s)$ und $z(s)$ betrachtet. Die vorgeschlagenen Verfahren lassen sich mittels Block-Algorithmen auf MIMO-Systeme (engl. *Multiple Input - Multiple Output*) erweitern, die entsprechend auf Ausgangs*matrizen* $\mathbf{Z} \in \mathbb{C}^{M\times M}$ führen. Hierbei ist M die Anzahl der betrachteten System-Tore.

4.1 Projektionsbasierte Modellordnungsreduktion

Ein auf Projektion basierendes ordnungsreduziertes Modell von (4.5) unter Berücksichtigung von (4.6) ist gegeben durch

$$\bar{\Sigma}(s) = \left\{ \begin{array}{rl} \sum\limits_{i=0}^{L} s^i \bar{\mathbf{A}}_i \bar{\mathbf{x}}(s) & = \sum\limits_{i=0}^{L} s^i \bar{\mathbf{b}}_i, \\ \bar{z}(s) & = \sum\limits_{i=0}^{L} s^i \bar{\mathbf{c}}_i^T \bar{\mathbf{x}}(s), \end{array} \right. \tag{4.14}$$

mit

$$\bar{\mathbf{A}}_i := \mathbf{P}^T \mathbf{A}_i \mathbf{Q}, \tag{4.15}$$

$$\bar{\mathbf{b}}_i := \mathbf{P}^T \mathbf{b}_i, \tag{4.16}$$

$$\bar{\mathbf{c}}_i := \mathbf{Q}^T \mathbf{c}_i \tag{4.17}$$

für $i = 1, \ldots, L$ und $\mathbf{P}, \mathbf{Q} \in \mathbb{C}^{N\times K}$, $K \ll N$. Entsprechend gilt auch hier die kompakte Schreibweise

$$\bar{\Sigma}(s) = \left\{ \begin{array}{rl} \bar{\mathbf{A}}(s)\bar{\mathbf{x}}(s) & = \bar{\mathbf{b}}(s), \\ \bar{z}(s) & = \bar{\mathbf{c}}^T(s)\bar{\mathbf{x}}(s). \end{array} \right. \tag{4.18}$$

Die Übertragungsfunktion in ordnungsreduzierter Form lautet schließlich

$$\bar{H}(s) := \bar{\mathbf{c}}(s)^T \bar{\mathbf{A}}^{-1}(s)\bar{\mathbf{b}}(s). \tag{4.19}$$

Aufgrund der im Vergleich zu N wesentlich niedrigeren Dimension K des Systems (4.18) ist der numerische Aufwand für die Auswertung vieler Stützstellen s_i um Größenordnungen geringer.

Der Ansatz der Projektionsmethoden ist wie folgt motiviert: Für das System (4.5) wird eine Näherungslösung $\tilde{\mathbf{x}} \approx \mathbf{x}$ mit

$$\tilde{\mathbf{x}} = \mathbf{Q}\bar{\mathbf{x}} \tag{4.20}$$

derart gesucht, dass das Residuum

$$\mathbf{r} := \sum_{i=0}^{L} s^i \mathbf{A}_i \tilde{\mathbf{x}}(s) - \sum_{i=0}^{L} s^i \mathbf{b}_i u(s) \tag{4.21}$$

im Sinne der Galerkin-Bedingung orthogonal auf dem Unterraum $\operatorname{bild}\mathbf{P}$ steht, also

$$\mathbf{P}^T\mathbf{r} = 0. \tag{4.22}$$

Aufgrund der Beziehungen (4.20) und (4.22) werden die Bezeichnungen *Ansatzraum* für $\operatorname{bild}\mathbf{Q}$ und *Testraum* für $\operatorname{bild}\mathbf{P}$ eingeführt.

Die konkrete Wahl der Unterräume $\operatorname{bild}\mathbf{P}, \operatorname{bild}\mathbf{Q} \subset \mathbb{C}^N$ hat einen entscheidenden Einfluss auf die Qualität der Näherungslösung $\tilde{\mathbf{x}}$. Bei der Konstruktion sind daher drei Aspekte zu beachten:

1. Der Ansatzraum $\operatorname{bild}\mathbf{Q}$ muss die wesentlichen Komponenten des gesuchten Lösungsvektors $\mathbf{x}(s)$ beinhalten.
2. Das Verfahren muss numerisch robust und effizient zu analysieren sein.
3. Das resultierende ROM soll die physikalischen Eigenschaften des Ausgangssystems (4.5) erhalten.

Für Punkt 1 werden hier zwei unterschiedliche Ansätze vorgestellt. Beim *Mehrpunktverfahren* wird $\operatorname{bild}\mathbf{Q}$ durch Lösungsvektoren $\boldsymbol{x}(s_i)$ an definierten Stützstellen $s_i \in \mathcal{I}_h^s$ aufgespannt, während beim *Einpunktverfahren* geeignete Eigenvektornäherungen, nämlich Krylov-Vektoren, die lineare Hülle der gesuchten Näherungslösung festlegen.

Damit Punkt 2 erfüllt ist, werden für $\operatorname{bild}\mathbf{Q}$ und $\operatorname{bild}\mathbf{P}$ orthonormale Basen bestimmt, wobei die Basisvektoren explizit paarweise auf Orthogonalität getestet werden.

Zuletzt ist im Hinblick auf Punkt 3 die Erhaltung der Systemmatrixeigenschaften, wie positive (semi-)Definitheit und Symmetrie zu gewährleisten. Auch die polynomielle Struktur des ROM bleibt unverändert zum Ausgangssystem (4.5), womit die korrekte Frequenzabhängigkeit der Systemkomponenten sichergestellt ist.

4.1.1 Mehrpunktverfahren

Die Idee beim Mehrpunktverfahren besteht darin, an unterschiedlichen Stützstellen $\hat{s}_1, \ldots, \hat{s}_K$ im Parameterbereich $\mathcal{I}_h^s$ die Zustandsvektoren

$$\mathbf{x}_k := \mathbf{x}(\hat{s}_k), k = 1, \ldots, K, \tag{4.23}$$

des Systems (4.5) zu bestimmen. Diese spannen den Suchraum $\operatorname{bild}\mathbf{Q}$ auf [Gri97]. Aus Gründen der numerischen Stabilität wird z. B. mit dem modifizierten Gram-Schmidt-Verfahren [TB97, S. 58] eine orthonormale Basis $\{\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K\}$ für den durch $\operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$ aufgespannten Raum konstruiert. Damit kann $\mathbf{Q}$ als eine aus den Spalten $\mathbf{q}_i$ generierte Matrix

$$\mathbf{Q} := \begin{bmatrix} \mathbf{q}_1 & \mathbf{q}_2 & \cdots & \mathbf{q}_K \end{bmatrix} \tag{4.24}$$

dargestellt werden, wobei

$$\operatorname{bild}\mathbf{Q} = \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\} \tag{4.25}$$

und

$$\mathbf{q}_i^H \mathbf{q}_j = \delta_{ij} \tag{4.26}$$

gilt. Das Mehrpunktverfahren erweist sich insbesondere bei breitbandigen Anwendungen als sehr robust. Diese Eigenschaft lässt sich dadurch erklären, dass die Lösungsvektoren zu unterschiedlichen Parameterwerten $\hat{s}$ ergänzende Informationen über das System liefern. Das Systemverhalten kann somit im gesamten Parameterbereich sehr gut wiedergegeben werden. Der Nachteil dieses Verfahrens liegt in der Notwendigkeit, das hochdimensionale Ausgangssystem (4.5) an K Stützstellen lösen zu müssen, um eine ausreichend gute Approximation des ROM gewährleisten zu können. Bei resonanten Strukturen muss die Dimension des ordnungsreduzierten Modells gegebenenfalls in der Größenordnung von $K > 50$ liegen, was zu einem großen Rechenaufwand führen kann. In vielen Anwendungen ist jedoch für $K \ll N_f$ ein sehr gutes Ergebnis erzielbar, so dass mit dem Mehrpunktverfahren ein effizientes Werkzeug zur schnellen und äußerst robusten Berechnung der Übertragungsfunktion gegeben ist.

Beim Mehrpunktverfahren können zudem iterative Gleichungslöser zur Bestimmung der Lösungen $\mathbf{x}_k$ eingesetzt werden. Damit ist insbesondere die Simulation sehr hochdimensionaler Systeme effizient durchzuführen. Wie in Abschnitt 4.2.1 beschrieben, ist dies ein wesentlicher Vorteil gegenüber der Einpunktverfahren.

Eine besondere Bedeutung kommt beim Mehrpunktverfahren der Wahl der *Entwicklungspunkte* $\hat{s}_1, \ldots, \hat{s}_K$ zu, da diese erheblichen Einfluss auf die Qualität des ordnungsreduzierten Modells haben. Unterschiedliche Strategien zur Bestimmung der Stützstellen sowie Kriterien zur Bewertung ordnungsreduzierter Modelle werden in Abschnitt 5 vorgestellt und untersucht.

4.1.2 Einpunktverfahren

Im Gegensatz zu den Mehrpunktverfahren wird mit den *Einpunktverfahren* eine Näherungslösung allein aus der Systeminformation im Parameterpunkt s_0 gewonnen. Aufgrund der einfachen Darstellung wird für die folgenden Herleitungen ohne Beschränkung der Allgemeinheit $s_0 = 0$ gesetzt. Um problemangepasst den Entwicklungspunkt in $s \neq s_0$ zu legen, kann ein Parameter-Shift $s \to \sigma := s - s_0$ vollzogen werden. Details zur Shift-Invert-Vorkonditionierung werden in Abschnitt 4.2.5 erläutert.

Die grundlegende Idee der Einpunkt-Modellordnungsreduktionsverfahren besteht in der Konstruktion einer Übertragungsfunktion, deren *Momente* in einer hinreichend kleinen Umgebung von $s_0 = 0$ mit denen der Taylor-Entwicklung übereinstimmen. Aus der Darstellung der Übertragungsfunktion mittels der Taylor-Reihe

$$H(s) = \sum_{n=0}^{\infty} \frac{1}{n!} \frac{\mathrm{d}^n H(s)}{\mathrm{d}s^n}\Big|_{s=0} s^n = \sum_{n=0}^{\infty} m_n s^n \tag{4.27}$$

können die Momente m_n als Taylor-Koeffizienten direkt abgelesen werden. Demnach wird eine ROM-Übertragungsfunktion $\tilde{H}(s)$ derart gesucht, dass die Bedingung

$$\frac{\mathrm{d}^n H(s)}{\mathrm{d}s^n}\Big|_{s=0} = \frac{\mathrm{d}^n \tilde{H}(s)}{\mathrm{d}s^n}\Big|_{s=0} \text{ für } n = 0, 1, \ldots, R-1 \tag{4.28}$$

erfüllt ist. Ordnungsreduktionsverfahren mit dieser Eigenschaft werden daher auch als *Momentabgleichende Verfahren* (engl. *Moment Matching Methods*) bezeichnet.

Die Berechnung einer solchen MOR-Übertragungsfunktion lässt sich über eine Rekursionsbeziehung bilden. Hierzu wird in einem ersten Schritt der Zustandsvektor $\mathbf{x}(s)$ ebenfalls als Taylor-Entwicklung

$$\mathbf{x}(s) = \sum_{n=0}^{\infty} \mathbf{x}_n s^n \tag{4.29}$$

um $s_0 = 0$ mit vektoriellen Taylor-Koeffizienten $\mathbf{x}_n$ dargestellt. Einsetzen in (4.5) und Vergleich der Terme selber Potenz in s liefert die Rekursionsvorschrift

$$\mathbf{x}_n = \begin{cases} \mathbf{0} & \text{für } n < 0, \\ \mathbf{A}_0^{-1}(\mathbf{b}_n - \sum_{i=1}^{L} \mathbf{A}_i \mathbf{x}_{n-i}) & \text{für } n \geq 0 \end{cases} \tag{4.30}$$

für die Taylor-Koeffizienten. Wegen

$$H(s) = z(s) = \mathbf{c}^T(s)\mathbf{x}(s) \tag{4.31}$$

können auch die Momente m_k über einen Koeffizientenvergleich ermittelt werden. Einsetzen von (4.29) in (4.31) führt auf

$$H(s) = \sum_{i=0}^{L} \mathbf{c}_i^T s^i \sum_{n=0}^{\infty} \mathbf{x}_n s^n, \tag{4.32}$$

so dass mit der Festlegung

$$\mathbf{c}_i = \mathbf{0} \text{ für } i > L \tag{4.33}$$

durch

$$m_n = \sum_{i=0}^{n} \mathbf{c}_i^T \mathbf{x}_{n-i}, \text{ für } n = 1, 2, \ldots \tag{4.34}$$

eine Vorschrift zur Bestimmung der Momente gegeben ist. Durch (4.30) und (4.34) liegen demnach sämtliche Schritte vor, die zur Bestimmung der Momente benötigt werden. An dieser Stelle gilt es festzuhalten, dass in (4.30) nur die *Wirkung* von $\mathbf{A}_0^{-1}$ auszuwerten ist. Es genügt $\mathbf{A}_0$ zu faktorisieren und die Wirkung von $\mathbf{A}_0^{-1}$ durch Vorwärts-Rückwärtseinsetzen zu berechnen. Gegenüber einer expliziten Berechnung der Inversen $\mathbf{A}_0^{-1}$ reduzieren sich Rechenaufwand und Speicheranforderungen erheblich. Weitere Operationen bestehen im Wesentlichen in der Bildung von Matrix-Vektor-Produkten, die im Falle dünnbesetzter Matrizen effizient zu berechnen sind.

Aufgrund des häufig sehr kleinen Konvergenzradius der Taylor-Reihe, wird die Übertragungsfunktion nicht in Form eines Taylor-Polynoms approximiert. Stattdessen wird die gebrochen-rationale Struktur der Übertragungsfunktion (4.12) des Systems (4.5) gemäß

$$H(s) = \mathbf{c}^T(s)\frac{\operatorname{adj}\mathbf{A}(s)}{\det\mathbf{A}(s)}\mathbf{b}(s) = \frac{\sum_{i=0}^{P} a_i(s-s_0)^i}{1+\sum_{j=1}^{Q} b_j(s-s_0)^j} \tag{4.35}$$

auch für die ROM-Übertragungsfunktion angenommen. Dies wird über den Padé-Ansatz

$$\tilde{H}_{pq}(s) = \frac{\sum_{i=0}^{p} \tilde{a}_i(s-s_0)^i}{1+\sum_{j=1}^{q} \tilde{b}_j(s-s_0)^j} \tag{4.36}$$

gewährleistet, wobei das Verhältnis von Zähler- zu Nennergrad in (4.36) der Übertragungsfunktion (4.35) anzupassen sind, also

$$\frac{p}{q} \approx \frac{P}{Q}. \tag{4.37}$$

Für den Padé-Ansatz wird weiterhin die Aufrechterhaltung der momentabgleichenden Eigenschaft (4.28) gefordert.

4.2 Einpunktverfahren für linear parametrierte Systeme

Eine wichtige Klasse in der Modellierung technischer Anwendungen bilden die LTI-Systeme erster Ordnung. Diese werden durch eine Systemmatrix

$$\mathbf{A}(s) = \mathbf{A}_0 + s\mathbf{A}_1 \tag{4.38}$$

beschrieben, die linear vom Parameter s abhängt. Insbesondere ist es stets möglich, Systeme höherer Ordnung in linear parametrierte Systeme zu überführen [Saa92b, S. 299 f], so dass Werkzeuge, die zur Behandlung von Systemen erster Ordnung entwickelt wurden, oftmals auch auf Systeme höherer Ordnung anwendbar sind.

Bei der Definition des Systems erster Ordnung werden zunächst die Anregung $\mathbf{b} = \text{const}(s)$ und der Ausgangsvektor $\mathbf{c} = \text{const}(s)$ als konstante, vom Parameter s unabhängige Größen angenommen. In Abschnitt 4.4 wird gezeigt, dass dies keine notwendige Einschränkung darstellt. Die momentabgleichende Eigenschaft der Modellordnungsreduktion kann auch für Systeme aufrecht erhalten werden, bei denen $\mathbf{b}$ beziehungsweise $\mathbf{c}$ eine explizite Abhängigkeit von s der Form

$$\mathbf{b}(s) = \alpha(s)\mathbf{b}_0 \tag{4.39}$$

mit $\mathbf{b}_0 = \text{const}(s)$ aufweisen.

Mit den hier getroffenen Vereinbarungen kann ein LTI-System erster Ordnung somit geschrieben werden als

$$\Sigma_1(s) = \left\{ \begin{array}{rl} (\mathbf{A}_0 + s\mathbf{A}_1)\mathbf{x}(s) & = \mathbf{b} \\ z(s) & = \mathbf{c}^T\mathbf{x}(s), \end{array} \right. \tag{4.40}$$

mit konstanten Matrizen $\mathbf{A}_0, \mathbf{A}_1 \in \mathbb{C}^{N\times N}$. Für die weiteren Betrachtungen wird vorausgesetzt, dass

$$\operatorname{rang} \mathbf{A}_0 = N \tag{4.41}$$

gilt und damit $\mathbf{A}_0$ invertierbar ist. Aus der Definition der Übertragungsfunktion (4.35) ist zu sehen, dass sich diese für (4.40) als eine gebrochen-rationale Funktion mit Zählergrad

$$P := \operatorname{grad}(\operatorname{adj} \mathbf{A}(s)) = N - 1 \tag{4.42}$$

und Nennergrad

$$Q := \operatorname{grad}(\det \mathbf{A}(s)) = N \tag{4.43}$$

ergibt. Die in weiterer Folge vorgestellten Verfahren liefern als Näherung der Übertragungsfunktion daher einen Padé-Ansatz, der die Forderung (4.37) berücksichtigt. Als weitere Nebenbedingung sind die Koeffizienten $\tilde{a}_i, \tilde{b}_j$ so zu finden, dass die Momente $m_0, m_1, \ldots, m_R$ der Padé-Näherung (4.36) mit jenen von (4.27) übereinstimmen. Die wesentlichen Unterschiede liegen in den Anforderungen an das untersuchte System und in den numerischen Eigenschaften, wie beispielsweise Robustheit oder Rechenaufwand.

4.2.1 Asymptotic Waveform Evaluation

Ein früher Ansatz, der vor allem in der Berechnung großer linearer Netzwerke aus der Schaltungstechnik Anwendung fand, stellt das AWE-Verfahren (engl. *Asymptotic Waveform Evaluation*) [PR90] dar. Die Modellierung eines Schaltungsnetzwerks führt auf ein System der Form (4.40), so dass dieses als Beispiel zur Verfahrensbeschreibung herangezogen wird. Die Rekursion (4.34) zur expliziten Berechnung der Momente vereinfacht sich für Systeme erster Ordnung zu

$$m_n = \mathbf{c}^T \mathbf{A}^n \mathbf{g} \tag{4.44}$$

mit den Abkürzungen

$$\mathbf{A} = -\mathbf{A}_0^{-1}\mathbf{A}_1 \tag{4.45}$$

und

$$\mathbf{g} = \mathbf{A}_0^{-1}\mathbf{b}. \tag{4.46}$$

Durch Koeffizientenvergleich kann damit das lineare Gleichungssystem

$$\begin{pmatrix} m_0 & m_1 & \dots & m_{n-1} \\ m_1 & m_2 & \dots & m_n \\ \vdots & \vdots & \ddots & \vdots \\ m_{n-1} & m_n & \dots & m_{2n-2} \end{pmatrix} \begin{pmatrix} \tilde{b}_n \\ \tilde{b}_{n-1} \\ \vdots \\ \tilde{b}_1 \end{pmatrix} = \begin{pmatrix} m_n \\ m_{n+1} \\ \vdots \\ m_{2n-1} \end{pmatrix} \tag{4.47}$$

zur Bestimmung der Padé-Koeffizienten $\tilde{b}_j$ aufgestellt werden. Die Padé-Koeffizienten $\tilde{a}_i$ des Zählerpolynoms resultieren aus der Rekursion

$$\tilde{a}_0 = m_0, \tag{4.48}$$

$$\tilde{a}_1 = m_1 + \tilde{b}_1 m_0 \tag{4.49}$$

$$\vdots \tag{4.50}$$

$$\tilde{a}_{n-1} = m_{n-1} \sum_{k=1}^{n-1} \tilde{b}_k m_{n-k-1}. \tag{4.51}$$

Wie aus (4.44) und (4.45) zu sehen ist, muss nicht explizit $\mathbf{A}_0^{-1}$, sondern nur die Wirkung der Inversen von $\mathbf{A}_0$ auf Vektoren berechnet werden. D. h. nach einmaliger Faktorisierung ist die Wirkung von $\mathbf{A}_0^{-1}$ in Sinne einer Vorwärts-Rückwärts-Einsetzung einzubringen. Die maßgeblichen Operationen bestehen daher in der Berechnung von Matrix-Vektor-Produkten, wobei die beteiligten Matrizen typischerweise dünnbesetzt sind. Der numerische Aufwand im Vergleich zum Mehrpunktverfahren, bei dem in jeder Stützstelle $\hat{s}_k$ ein Gleichungssystem $\mathbf{A}(\hat{s}_k)\mathbf{x}(\hat{s}_k) = \mathbf{b}(\hat{s}_k)$ zu lösen ist, fällt daher deutlich geringer aus. Demgegenüber steht allerdings die Einschränkung,

dass ausreichend Speicherkapazität für die Faktorisierung verfügbar sein muss. Ein iterativer Ansatz zur Berechnung von (4.45) stellt hierbei keine erfolgversprechende Alternative dar, da zur Bestimmung jedes Moments das Gleichungssystem gelöst werden müsste und damit der Vorteil gegenüber dem Mehrpunktverfahren verloren wäre. Die Notwendigkeit der einmaligen Faktorisierung von $\mathbf{A}_0$ ist bestimmend für alle hier vorgestellten Einpunktverfahren.

Auch wenn das AWE-Verfahren sich in vielen Anwendungen bewährt hat, stößt es insbesondere bei breitbandigen Charakterisierungen schnell an seine Grenzen und für Ordnungen im Bereich von $n = 10 \ldots 15$ bricht das Verfahren erfahrungsgemäß zusammen [ZC02, S. 327]. Der Grund hierfür liegt in der Tatsache, dass die Momente m_n durch Ausdrücke der Form $\mathbf{A}^n\mathbf{g}$ gebildet werden, welche für $n \to \infty$ nach der Von-Mises-Iteration gegen den Eigenvektor zum betragsgrößten Eigenwert konvergieren. Damit werden Momente höherer Ordnung immer ähnlicher, bzw. sind im Rahmen der Maschinengenauigkeit kaum mehr voneinander zu unterscheiden. Das Koeffizientengleichungssystem (4.47) des AWE-Ansatzes wird durch diese numerischen Auslöschungseffekte zunehmend schlecht konditioniert. In praktischen Anwendungen hat dies zur Folge, dass die gefundene Padé-Näherung keine ausreichende Übereinstimmung mit der tatsächlichen Übertragungsfunktion über den geforderten Parameterbereich aufweist. Mehrpunktverfahren unterliegen dieser Limitierung nicht.

Ein Kompromiss für eine breitbandigere Gültigkeit der AWE-Übertragungsfunktion ist, ähnlich wie bei den in Abschnitt 4.1.1 geschilderten Mehrpunktverfahren, durch Hinzunahme weiterer Entwicklungspunkte $\hat{s}_k$ erreichbar. In den zusätzlichen Entwicklungspunkten werden ebenfalls Momente und entsprechende Übertragungsfunktionen mittels (4.44) und (4.47) bestimmt, die in einer hinreichend kleinen Umgebung von $\hat{s}_k$ eine sehr gute Approximation darstellen. Anschließend werden die so bestimmten Teilübertragungsfunktionen zu einer Gesamtübertragungsfunktion verknüpft. Ein weit verbreitetes Verfahren dieser Art stellt das *Complex Frequency Hopping (CFH)* dar [KSN+96]. Bei diesem Ansatz muss jedoch für jeden Entwicklungspunkt erneut die Systemmatrix faktorisiert werden, was einem erheblichen Mehraufwand entspricht und den Vorteil gegenüber den Mehrpunktverfahren relativiert. Eine weitergehende Diskussion solch gemischter Ansätze ist beispielsweise in [Gri97] zu finden.

4.2.2 Krylov-Unterraum-Verfahren

Einen wesentlich robusteren Ansatz im Vergleich zum AWE-Verfahren stellt das Verfahren Padé-Via-Lanczos (PVL) [FF95] oder der Arnoldi-Algorithmus [GL96, S. 499] dar. Für diese Verfahren wird explizite eine Basis des zugrundeliegenden Krylov-Unterraums konstruiert, und die momentabgleichende Eigenschaft erfolgt durch eine Projektion auf diesen Unterraum (vgl. Abschnitt 4.1).

Der n-dimensionalen Krylov-Unterraum zu einem Paar $(\mathbf{A}, \mathbf{g})$, $\mathbf{A} \in \mathbb{C}^{N\times N}$, $\mathbf{g} \in \mathbb{C}^N$ ist definiert als

$$\mathcal{K}_n(\mathbf{A}, \mathbf{g}) := \operatorname{span}\{\mathbf{g}, \mathbf{A}\mathbf{g}, \mathbf{A}^2\mathbf{g}, \ldots, \mathbf{A}^{n-1}\mathbf{g}\}, \quad n \leq N. \tag{4.52}$$

Die ROM-Übertragungsfunktionen, die mittels Projektion generiert werden, verfügen im Entwicklungspunkt über dieselben Eigenschaften wie die AWE-Näherung, jedoch ist aufgrund der numerisch robusteren Formulierung eine deutlich höhere Ordnung des Padé-Ansatzes, und damit eine größere Anzahl abgeglichener Momente erreichbar. Ein bestimmender Faktor in der numerischen Stabilität resultiert aus der Forderung der Unitaritätseigenschaft

$$\mathbf{Q}^H\mathbf{Q} = \mathbf{I}, \tag{4.53a}$$

$$\mathbf{P}^H\mathbf{P} = \mathbf{I}. \tag{4.53b}$$

Es ist dabei $\mathbf{I} = \operatorname{diag}(1, 1, \ldots, 1)$ die $n \times n$-Einheitsmatrix. Der Momentenabgleich wird erreicht, indem eine Lösung $\tilde{\mathbf{x}}$ gemäß (4.20) mit

$$\operatorname{bild}\mathbf{Q} \supset \operatorname{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_q\} \tag{4.54}$$

gefunden wird. Hierin entsprechen die Vektoren $\mathbf{x}_i$, $i = 0, 1, \ldots, q$, den Taylor-Koeffizienten in (4.30). Für Systeme erster Ordnung sind diese wegen (4.44) gegeben durch

$$\mathbf{x}_i = \mathbf{A}^i\mathbf{g} = -(\mathbf{A}_0^{-1}\mathbf{A}_1)^i\mathbf{A}_0^{-1}\mathbf{b}. \tag{4.55}$$

Abkürzend wird die Definition

$$\mathcal{K}_q(\Sigma_1) := \mathcal{K}_q(\mathbf{A}_0^{-1}\mathbf{A}_1, \mathbf{A}_0^{-1}\mathbf{b}) \tag{4.56}$$

des *Ansatz-Krylovraums* für Systeme erster Ordnung eingeführt. Damit kann die für Krylov-Unterraum-Verfahren zentrale Aussage formuliert werden:

Für das System Σ_1 führt ein Projektionsverfahren basierend auf der *Ansatzmatrix* $\mathbf{Q}$ mit

$$\operatorname{bild}\mathbf{Q} \supseteq \mathcal{K}_q(\Sigma_1) \tag{4.57}$$

unter der Bedingung

$$\det(\mathbf{P}^T\mathbf{A}_0\mathbf{Q}) \neq 0 \tag{4.58}$$

auf ein ordnungsreduziertes System $\bar{\Sigma}_1$, das auf Taylor-Koeffizienten

$$\tilde{\mathbf{x}}_k := \mathbf{Q}\bar{\mathbf{x}}_k \tag{4.59}$$

führt, die bis zum Grad $q-1$ mit jenen des Originalsystems übereinstimmen, also

$$\mathbf{x}_k = \tilde{\mathbf{x}}_k \quad \text{für} \quad k = 0, 1, \ldots, q-1. \tag{4.60}$$

Hierbei sind die ROM-Taylor-Koeffizienten $\bar{\mathbf{x}}_k$ definiert über die Entwicklung

$$\bar{\mathbf{x}}(s) = \sum_{k=0}^{\infty} \bar{\mathbf{x}}_k s^k. \tag{4.61}$$

Die ROM-Momente berechnen sich entsprechend zu

$$\bar{m}_k = \bar{\mathbf{c}}^T \bar{\mathbf{x}}_k \tag{4.62}$$

und es gilt

$$\bar{m}_k = m_k \quad \text{für} \quad k = 0, 1, \ldots, q-1. \tag{4.63}$$

Zusätzliche Freiheitsgrade stehen mit dem *Testraum* bild $\mathbf{P}$ zur Verfügung. Eine geeignete Wahl des aufgespannten Unterraums erlaubt die Anzahl abgeglichener Momente weiter zu erhöhen [Far07, S. 114]. Die Konstruktion des Testraums erfolgt analog zur vorangehend beschriebenen Vorgehensweise, jedoch wird in diesem Fall das transponierte System

$$\Sigma_1^T(s) = \left\{ \begin{array}{rl} (\mathbf{A}_0^T + s\mathbf{A}_1^T)\mathbf{w}(s) & = \mathbf{c}, \\ v(s) & = \mathbf{b}^T\mathbf{w}(s), \end{array} \right. \tag{4.64}$$

betrachtet. Wegen

$$v(s) = v^T(s) = \mathbf{w}^T(s)\mathbf{b} = \mathbf{c}^T(\mathbf{A}_0 + s\mathbf{A}_1)^{-1}\mathbf{b} = z(s) \tag{4.65}$$

ist die Übertragungsfunktion des transponierten Systems identisch zu derjenigen des Originalsystems (4.40). Die Rekursion für die Taylor-Koeffizienten des Systems (4.64) lautet in Analogie zu (4.55)

$$\mathbf{w}_i = (\mathbf{A}_0^{-T}\mathbf{A}_1^T)^i \mathbf{A}_0^{-T}\mathbf{c}, \tag{4.66}$$

womit der geeignete Ansatzraum für $\mathbf{w}$ durch

$$\mathcal{K}_p(\Sigma_1^T) := \operatorname{span}\{\mathbf{w}_0, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{p-1}\} \tag{4.67}$$

gegeben ist. Wegen (4.65) gilt daher für das reduzierte Modell (4.14) mit

$$\text{bild}\,\mathbf{P} \supseteq \mathcal{K}_p(\Sigma_1^T) \tag{4.68}$$

unter der Bedingung (4.58), dass die ROM-Momente $\bar{m}_0$, $\bar{m}_1$, …, $\bar{m}_{p-1}$ mit jenen des Originalmodells übereinstimmen.

Mit den am transponierten System aufgezeigten Eigenschaften kann die Aussage (4.63) wie folgt erweitert werden:

Die Verwendung von Projektionsmatrizen $\mathbf{P}$ und $\mathbf{Q}$ mit den Eigenschaften (4.57) bzw. (4.68) führt unter der Bedingung (4.58) auf ein ordnungsreduziertes Modell, mit der Eigenschaft

$$\bar{m}_k = m_k \quad \text{für} \quad k = 0, 1, \ldots, p+q-1. \tag{4.69}$$

Die ROM-Übertragungsfunktion stimmt also in den ersten $q+p$ Momenten mit jenen der Originalübertragungsfunktion überein.

Der Beweis für die momentabgleichende Eigenschaft der Krylov-Unterraumverfahren wird beispielsweise in [FF95] erbracht. In [SLL03a] wird dieser Beweis auf Krylov-Verfahren für Matrix-Polynome höherer Ordnung erweitert.

Neben der Beibehaltung der momentabgleichenden Eigenschaften überwinden die Krylov-Verfahren die in Abschnitt 4.2.1 aufgezeigten Schwächen der AWE-Methode und ermöglichen die effiziente Anwendbarkeit der Modellordnungsreduktion auf eine Vielzahl praktischer Problemstellungen.

Eine Einschränkung des PVL-Verfahrens besteht in der Drei-Schritt-Rekursion, die dem Lanczos-Algorithmus bei der Bildung der Projektionsbasis zugrunde liegt. Aufgrund numerischer Effekte kann damit ein Orthogonalitätsverlust der Basisvektoren einhergehen und ein vorzeitiger Einbruch des Verfahrens eintreten. Zwar existieren modifizierte Algorithmen [GL96, S. 482 ff], die diesen Nachteil eliminieren, für die vorliegende Arbeit wird jedoch die Arnoldi-Iteration gemäß Algorithmus 1 gewählt, da diese einen guten Kompromiss aus Robustheit und Rechenaufwand darstellt. Im Gegensatz zur Lanczos-Iteration ist zudem keine Symmetrie der Systemmatrizen gefordert, und die Rekursion zur Berechnung der Projektionsbasis bezieht *alle* zuvor ermittelten Basisvektoren mit ein. Damit ist beim Arnoldi-Algorithmus die Bedingung (4.53) stets gewährleistet, und er ist unverändert auch auf nicht-symmetrische Systeme anwendbar.

Algorithmus 1 Arnoldi-Algorithmus zur Berechnung von $\mathcal{K}_n(\mathbf{A}, \mathbf{g})$ gemäß (4.52)

1: $\mathbf{q}_1 = \mathbf{g} / \|\mathbf{g}\|_2$
2: **for** $k = 1, 2, \ldots, n-1$ **do**
3: $\quad \tilde{\mathbf{q}}_{k+1} = \mathbf{A}\mathbf{q}_k$;
4: $\quad$ **for** $j = 1, 2, \ldots, k$ **do**
5: $\quad\quad h_{jk} = \mathbf{q}_j^H \mathbf{A}\mathbf{q}_k$;
6: $\quad\quad \tilde{\mathbf{q}}_{k+1} = \tilde{\mathbf{q}}_{k+1} - h_{jk}\mathbf{q}_j$;
7: $\quad$ **end for**
8: $\quad h_{k+1,k} = \|\tilde{\mathbf{q}}_{k+1}\|_2$;
9: $\quad \mathbf{q}_{k+1} = \tilde{\mathbf{q}}_{k+1} / h_{k+1,k}$
10: **end for**

Die mit dem Arnoldi-Algorithmus berechneten Koeffizienten h_{jk} stellen die Einträge einer oberen *Hessenberg-Matrix* dar,

$$\mathbf{H}_n := \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1n} \\ h_{21} & h_{22} & \ldots & \vdots \\ 0 & \ddots & \ddots & \vdots \\ 0 & \ldots & h_{n,n-1} & h_{nn} \end{pmatrix}. \tag{4.70}$$

Diese entspricht der Projektion von $\mathbf{A}$ auf den Krylov-Unterraum $\mathcal{K}_n(\mathbf{A}, \mathbf{g})$. Der Arnoldi-Algorithmus bildet die Grundlage effizienter iterativer Gleichungs- und Ei-

genwertlöser, wie beispielsweise der GMRES-Verfahren (Generalized Minimal Residual Method) [Saa96, S. 157]. Die Lösungsverfahren haben gemeinsam, dass Näherungen der gesuchten Größen anhand der niedrigdimensionalen Hessenberg-Matrix bestimmt werden. Bei geeigneten Matrizen $\mathbf{A}$, d. h. bei Matrizen die am Rand des Spektrums keine Eigenwert-Cluster aufweisen, stellen die betragsgrößten Eigenwerte von $\mathbf{H}_n$ bereits für niedrige Dimensionen $n \ll N$ sehr gute Approximationen der Eigenwerte von $\mathbf{A}$ dar [Saa92b, S. 204]. Aufgrund der niedrigen Dimension von $\mathbf{H}_n$ können die Eigenwerte zum Problem

$$\mathbf{H}_n \bar{\mathbf{v}} = \lambda \bar{\mathbf{v}} \tag{4.71}$$

mit klassischen Ansätzen, wie dem QR-Verfahren effizient bestimmt werden. Die so approximierten Eigenwerte von $\mathbf{A}$ heißen *Ritz-Werte* und die zugehörigen Eigenvektornäherungen

$$\tilde{\mathbf{v}} = \mathbf{Q} \bar{\mathbf{v}} \tag{4.72}$$

entsprechend *Ritz-Vektoren* [Saa92b, S. 175].

4.2.3 Residuenberechnung im Arnoldi-Verfahren

Eine wichtige Größe zur Bewertung von Näherungslösungen, beziehungsweise der Qualität ordnungsreduzierter Modelle, stellt das Residuum dar. Für die Zustandsgleichung des Systems (4.40) ist das Residuum zu einer Näherungslösung $\tilde{\mathbf{x}} = \mathbf{Q}\bar{\mathbf{x}}$ durch die Beziehung

$$\mathbf{r} := (\mathbf{A}_0 + s\mathbf{A}_1)\tilde{\mathbf{x}} - \mathbf{b} \tag{4.73}$$

gegeben. Aus der Arnoldi-Iteration

$$\mathbf{q}_1 = \mathbf{A}_0^{-1}\mathbf{b}, \tag{4.74}$$

$$\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{Q}_n = \mathbf{Q}_n\mathbf{H}_n + h_{n+1,n}\hat{\mathbf{q}}_{n+1}\hat{\mathbf{e}}_n^T \tag{4.75}$$

folgt mit $\hat{\mathbf{e}}_n^T = (0, 0, \ldots, 0, 1) \in \mathbb{R}^{1\times n}$,

$$\mathbf{A}_0^{-1}\mathbf{b} = \mathbf{Q}_n\hat{\mathbf{e}}_1 h_{11} \tag{4.76}$$

und der Näherungslösung

$$\tilde{\mathbf{x}} = \mathbf{Q}_n\bar{\mathbf{x}} \tag{4.77}$$

das Residuum in der Form

$$\begin{aligned} \mathbf{r} &= (\mathbf{A}_0 + s\mathbf{A}_1)\mathbf{Q}_n\bar{\mathbf{x}} - \mathbf{b} && (4.78)\\ &= (\mathbf{A}_0 + s\mathbf{A}_1)\mathbf{Q}_n\bar{\mathbf{x}} - \mathbf{A}_0\mathbf{Q}_n\hat{\mathbf{e}}_1 h_{11} \\ &= \mathbf{A}_0\mathbf{Q}_n\left[(\mathbf{I} + s\mathbf{H}_n)\bar{\mathbf{x}} - \hat{\mathbf{e}}_1 h_{11}\right] + s h_{n+1,n}\mathbf{A}_0\hat{\mathbf{q}}_{n+1}\hat{\mathbf{e}}_n^T\bar{\mathbf{x}}. && (4.79) \end{aligned}$$

Mit der Bedingung

$$\mathbf{Q}_n^T \mathbf{r} = \mathbf{0} \tag{4.80}$$

kann der Ausdruck geschrieben werden als

$$\mathbf{Q}_n^T \mathbf{A}_0 \left[\mathbf{Q}_n (\mathbf{I} + s\mathbf{H}_n) + s h_{n+1,n} \hat{\mathbf{q}}_{n+1} \hat{\mathbf{e}}_n^T \right] \bar{\mathbf{x}} = \mathbf{Q}_n^T \mathbf{b}, \tag{4.81}$$

$$\Leftrightarrow \quad \left[\mathbf{I} + s\mathbf{H}_n + s h_{n+1,n} (\mathbf{Q}_n^T \mathbf{A}_0 \mathbf{Q}_n)^{-1} \mathbf{Q}_n^T \mathbf{A}_0 \hat{\mathbf{q}}_{n+1} \hat{\mathbf{e}}_n^T \right] \bar{\mathbf{x}} = \hat{\mathbf{e}}_1 h_{11}. \tag{4.82}$$

Durch Einsetzen in (4.79) gilt für das Residuum schließlich

$$\mathbf{r} = s h_{n+1,n} \mathbf{A}_0 \left[\mathbf{I} - \mathbf{Q}_n (\mathbf{Q}_n^T \mathbf{A}_0 \mathbf{Q}_n)^{-1} \mathbf{Q}_n^T \mathbf{A}_0 \right] \hat{\mathbf{q}}_{n+1} (\hat{\mathbf{e}}_n^T \bar{\mathbf{x}}). \tag{4.83}$$

Damit sind in der Darstellung des Residuums nur noch Terme der Dimension $n \ll N$ oder Produkte, die ohnehin im Arnoldi-Verfahren berechnet werden, enthalten. Die Residuenauswertung für eine große Zahl an Parameterstützstellen s_i ist daher mit geringem Aufwand zu bewerkstelligen. Durch eine Eigenzerlegung der Systemmatrix lässt sich der Rechenaufwand nochmals verringern, wie im folgenden Abschnitt gezeigt wird.

4.2.4 Schnelle Auswertung von Ausgangsgröße und Residuum

Die Bewertung der Qualität der Näherungslösung aus einem ordnungsreduzierten Modell ist umso aussagekräftiger, je feiner der untersuchte Parameterbereich abgetastet ist. Um den Aufwand im Vergleich zum gesamten MOR-Prozess möglichst gering zu halten, werden effiziente Methoden zur Auswertung der Residuen (4.83) und der Ausgangsgröße $z(s)$ vorgeschlagen. Grundlage der Vorgehensweise ist die Berechnung der Eigenpaare $(\bar{\lambda}_i, \bar{\mathbf{v}}_i), i = 1, \ldots, n$ zum verallgemeinerten Eigenwertproblem in der ROM-Domäne:

$$\bar{\mathbf{A}}_0 \bar{\mathbf{V}} = \bar{\mathbf{A}}_1 \bar{\mathbf{V}} \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \qquad \text{mit } \bar{\mathbf{V}} = [\bar{\mathbf{v}}_1 \ldots \bar{\mathbf{v}}_n], \tag{4.84}$$

$$\bar{\mathbf{V}}^T \bar{\mathbf{A}}_1 \bar{\mathbf{V}} = \mathbf{I}. \tag{4.85}$$

Da $\bar{\mathbf{A}}_0$ und $\bar{\mathbf{A}}_1$ symmetrisch sind und außerdem $\bar{\mathbf{A}}_1$ positiv definit ist, ist die Diagonalisierbarkeit des Paars $(\bar{\mathbf{A}}_0, \bar{\mathbf{A}}_1)$ gegeben [Saa92b, S. 295]. Der Basiswechsel

$$\bar{\mathbf{x}} = \bar{\mathbf{V}} \bar{\mathbf{w}} \tag{4.86}$$

führt auf die als Funktion von s schnell auswertbaren Beziehungen

$$\bar{\mathbf{w}} = \operatorname{diag}\left(\frac{1}{\bar{\lambda}_i + s} \right) \bar{\mathbf{V}}^T \bar{\mathbf{b}}, \tag{4.87}$$

$$z = s (\bar{\mathbf{V}}^T \bar{\mathbf{b}})^T \operatorname{diag}\left(\frac{1}{\bar{\lambda}_i + s} \right) \bar{\mathbf{V}}^T \bar{\mathbf{b}}, \tag{4.88}$$

$$\mathbf{r} = s h_{n+1,n} \mathbf{A}_0 \left[\mathbf{I} - \mathbf{Q}_n \bar{\mathbf{V}} \operatorname{diag}\left(\frac{1}{\bar{\lambda}_i} \right) \bar{\mathbf{V}}^T \mathbf{Q}_n^T \mathbf{A}_0 \right] \hat{\mathbf{q}}_{n+1} (\hat{\mathbf{e}}_n^T \bar{\mathbf{V}} \bar{\mathbf{w}}). \tag{4.89}$$

Insbesondere die Auswertung des Residuums kann sehr effizient gestaltet werden, da lediglich der konstante Anteil

$$\mathbf{r}_0 := h_{n+1,n}\mathbf{A}_0\left[\mathbf{I} - \mathbf{Q}_n\bar{\mathbf{V}}\,\mathrm{diag}\left(\frac{1}{\tilde{\lambda}_i}\right)\bar{\mathbf{V}}^T\mathbf{Q}_n^T\mathbf{A}_0\right]\hat{\mathbf{q}}_{n+1} \tag{4.90}$$

mit einem s-abhängigen Faktor in der Form

$$\mathbf{r} = s(\hat{\mathbf{e}}_n^T\bar{\mathbf{V}}\bar{\mathbf{w}})\mathbf{r}_0 \tag{4.91}$$

skaliert werden muss.

4.2.5 Shift- und Invert-Vorkonditionierung

Die bisher aufgezeigten Eigenschaften der Modellordnungsreduktionsverfahren basieren auf einem Momentenabgleich im Ursprung $s = 0$. Dabei wird die Approximationseigenschaft der Krylov-Unterraum-Verfahren unter anderem mit der Tatsache erklärt, dass die Krylov-Vektoren den Raum jener Eigenvektoren aufspannen, die zu isolierten Eigenwerten am Rand des Spektrums gehören [BDD+00, Kapitel 7]. Im konkreten Fall des Raums $\mathcal{K}_n(\Sigma_1) = \mathcal{K}_n(\mathbf{A}_0^{-1}\mathbf{A}_1, \mathbf{A}_0^{-1}\mathbf{b})$ heißt das, es werden Approximationen zum verallgemeinerten Eigenwertproblem

$$(\mathbf{A}_0 - \lambda\mathbf{A}_1)\mathbf{v} = 0 \tag{4.92}$$

$$\Leftrightarrow \qquad \mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{v} = \frac{1}{\lambda}\mathbf{v} \tag{4.93}$$

bestimmt, wobei die gefundenen Ritzwerte $\tilde{\lambda}$ im Wesentlichen in der Abfolge

$$|\tilde{\lambda}_1| \leq |\tilde{\lambda}_2| \leq |\tilde{\lambda}_3| \leq \ldots \tag{4.94}$$

konvergieren [BDD+00, S. 369]. Mit der Invertierung von $\mathbf{A}_0$, also der Invert-Vorkonditionierung, werden die betragskleinsten Eigenwerte an den Rand des Spektrums überführt. Mit dem Arnoldi-Algorithmus konvergieren daher zunächst die Eigenwerte in der Umgebung von $s = 0$. Ist für ein gegebenes System das Verhalten in einer Umgebung

$$\mathcal{I}_s := [s_{\min}, s_{\max}] = [s_0 - \Delta s, s_0 + \Delta s] \tag{4.95}$$

um den *Entwicklungspunkt* $s_0 \neq 0$ von Interesse, kann durch einen Parameter-Shift

$$\hat{s} : s \longmapsto s - s_0 \tag{4.96}$$

das Spektrum des Systems derart verschoben werden, dass zuerst die Eigenwerte in der Umgebung von s_0 konvergieren und auch der Momentenabgleich im Entwicklungspunkt s_0 stattfindet. Der Shift (4.96) führt auf die Darstellung der Systemgleichung in der Form

$$\mathbf{A}_0 + (s - s_0 + s_0)\mathbf{A}_1\mathbf{x} = \mathbf{b} \tag{4.97}$$

$$\Leftrightarrow \quad (\mathbf{A}_0 + s_0\mathbf{A}_1) + \hat{s}\mathbf{A}_1\mathbf{x} = \mathbf{b}. \tag{4.98}$$

Mit der Abkürzung

$$\hat{\mathbf{A}}_0 := \mathbf{A}_0 + s_0\mathbf{A}_1 \tag{4.99}$$

wird das Eigenwertproblem (4.92) in die Form

$$\hat{\mathbf{A}}_0 - \hat{\lambda}\mathbf{A}_1\mathbf{v} = 0 \tag{4.100}$$

überführt, wobei der hierin auftretende Eigenwert über die Beziehung

$$\hat{\lambda} = \lambda + s_0. \tag{4.101}$$

mit dem Eigenwert λ in (4.92) zusammenhängt. Das LTI-System

$$\hat{\Sigma}(\hat{s}) = \left\{ \begin{array}{rl} (\hat{\mathbf{A}}_0 + \hat{s}\mathbf{A}_1)\mathbf{x} & = \mathbf{b} \\ z & = \mathbf{c}^T\mathbf{x} \end{array} \right. \tag{4.102}$$

ist nach dem Parameter-Shift wieder mit derselben Parametrierung wie in (4.40) gegeben und es gelten dieselben Aussagen bezüglich der Eigenwert-Konvergenz, in diesem Fall jedoch für die Umgebung von $\hat{s} = 0$. Der mittels *Shift- und Invertvorkonditionierung* generierte Krylov-Unterraum zum System (4.102) lautet entsprechend

$$\mathcal{K}_n(\hat{\Sigma}_1) = \mathrm{span}\{\hat{\mathbf{g}}, \hat{\mathbf{A}}\hat{\mathbf{g}}, \hat{\mathbf{A}}^2\hat{\mathbf{g}}, \ldots, \hat{\mathbf{A}}^{n-1}\hat{\mathbf{g}}\} \tag{4.103}$$

mit

$$\hat{\mathbf{A}} := -\hat{\mathbf{A}}_0^{-1}\mathbf{A}_1 \quad \text{und} \quad \hat{\mathbf{g}} := \hat{\mathbf{A}}_0^{-1}\mathbf{b}. \tag{4.104}$$

Die Shift-Invert-Vorkonditionierung ist genauso auch auf Systeme höherer Ordnung anwendbar, also für die Verfahren aus Abschnitt 4.3 zugänglich. Hierzu werden nach Anwendung des Parameter-Shifts (4.96) Terme selber Ordnung in $\hat{s}$ umsortiert und ein äquivalentes System um den Entwicklungspunkt s_0 aufgestellt. Im MOR-Prozess findet dann, wie hier für linear parametrierte Systeme gezeigt, der Momentenabgleich im Punkt s_0 statt.

4.3 Systeme höherer Ordnung

In Abschnitt 4.1.2 wird die Anwendung des Arnoldi- und des Lanczos-Algorithmus auf Systeme mit linearer Parametrierung, das heißt $L = 1$, gezeigt. Allgemein polynomiell parametrierte Systeme (4.5) mit $L > 1$ bedürfen eines erweiterten Ansatzes. Eine Möglichkeit stellt beispielsweise die in [ZC02] aufgezeigte Linearisierung des Systems dar. Allerdings wird hierbei die Systemdimension mit dem Faktor L vervielfacht und die symmetrische Struktur muss aufgegeben werden. Strukturerhaltende Projektionsverfahren für Systeme höherer Ordnung stellen für $L = 2$ das *Second Order Arnoldi-* (SOAR) [BS05b] und für beliebiges L das *Well Conditioned AWE* - Verfahren (WCAWE) [SLL03a] dar. Diese basieren beide auf der Konstruktion einer orthogonalen Basis für den Krylov-Unterraum höherer Ordnung. Für Matrizen $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_L \in \mathbb{C}^{N\times N}$ und Vektoren $\mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_L \in \mathbb{C}^N$ ist dieser definiert als

$$\mathcal{K}_q(\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_L; \mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_L) := \mathrm{span}\{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_{q-1}\} \tag{4.105}$$

mit

$$\mathbf{x}_0 = \mathbf{b}_0 \tag{4.106a}$$
$$\mathbf{x}_1 = \mathbf{b}_1 - \mathbf{A}_1\mathbf{x}_0 \tag{4.106b}$$
$$\mathbf{x}_2 = \mathbf{b}_2 - \mathbf{A}_1\mathbf{x}_1 - \mathbf{A}_2\mathbf{x}_0 \tag{4.106c}$$
$$\vdots$$
$$\mathbf{x}_\beta = \mathbf{b}_\beta - \sum_{\alpha=1}^{L} \mathbf{A}_\alpha \mathbf{x}_{\beta-\alpha} \tag{4.106d}$$

und

$$\mathbf{b}_\beta = \mathbf{0} \quad \text{für } \beta > L. \tag{4.107}$$

Damit lässt sich die Aussage (4.60) des Momentenabgleichs auch auf Systeme höherer Ordnung entsprechend der Definition (4.5) übertragen: Erfüllt ein ordnungsreduziertes Modell gemäß (4.14) die Bedingung

$$\det(\mathbf{P}^T\mathbf{A}_0\mathbf{Q}) \neq 0 \tag{4.108}$$

und ist die Ansatzmatrix $\mathbf{Q}$ derart gewählt, dass

$$\operatorname{span}\mathbf{Q} \quad \supseteq \quad \mathcal{K}_q(\Sigma) \quad := \quad \mathbf{A}_0^{-1}\mathcal{K}_q(\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_L; \mathbf{b}_0, \mathbf{b}_1, \ldots, \mathbf{b}_L) \tag{4.109}$$

gilt, dann lässt sich der Momentenabgleich in der Form

$$\bar{m}_k = m_k \quad \text{für} \quad k = 0, 1, \ldots, q-1 \tag{4.110}$$

auch für Systeme höherer Ordnung aufrecht erhalten.

Entsprechend kann die Anzahl abgeglichener Momente vergrößert werden, wenn zusätzlich

$$\mathcal{K}_p(\Sigma^T) \subseteq \operatorname{span}\mathbf{P} \tag{4.111}$$

gilt. Beweise zu den getroffenen Aussagen finden sich in [Far07].

Darüber hinaus sind Einpunktverfahren auch sehr effizient bei der Ordnungsreduktion multivariater Systeme einsetzbar [Far07].

4.4 Systeme in der Simulation elektrodynamischer Felder

In Kapitel 3 wird gezeigt, wie sich die Maxwell-Gleichungen in das Finite-Elemente LTI-System (3.52) überführen lassen. Grundlage der folgenden Betrachtungen wie auch der Computerimplementierung bildet ausschließlich das TFE-System; aus

Gründen der Übersichtlichkeit wird jedoch auf die Angabe des Index „TF“ bei den Systemkomponenten verzichtet.

Damit das LTI-System den Werkzeugen der Modellordnungsreduktion zugänglich gemacht werden kann, sind vorab spezielle Eigenschaften des Systems zu erläutern. Die hier gewählte FE-Formulierung (3.28) ist auf die Analyse von Wellenphänomenen in Mikrowellenstrukturen, also auf Hochfrequenzanwendungen zugeschnitten. Für diese Formulierung ist aus (3.28a) zu erkennen, dass aufgrund des Rotationsoperators alle Gradienten $\operatorname{grad}\phi \in W_h$ im Nullraum von $\mathbf{S}$ liegen und damit $\mathbf{S}$ singulär wird. Das System hat für den Grenzfall $jk_0 \to 0$ keine eindeutige Lösung. Unter Beachtung dieser Einschränkung wird für die Charakterisierung des FE-Systems der Wellenzahlbereich

$$\mathcal{I}_h^{k_0} = \left\{k_{0,1}, k_{0,2}, \ldots, k_{0,N_f}\right\} \text{ mit } k_{0,i} = 2\pi f_i \sqrt{\varepsilon_0 \mu_0},\ i = 1, \ldots, N_f \tag{4.112}$$

festgelegt.

Die in der Systemdefinition (4.5) geforderte Regularität der Matrix $\mathbf{A}_0$ kann für das FE-System gemäß der Betrachtungen in Abschnitt 4.2.5 durch einen Parameter-Shift

$$\hat{s} : jk_0 \mapsto j(k_0 - \hat{k}_0) \tag{4.113}$$

bewerkstelligt werden. Dadurch wird zum einen die Invertierung der parameterunabhängigen Matrix

$$\hat{\mathbf{S}} := \mathbf{S} + j\hat{k}_0\mathbf{D} + (j\hat{k}_0)^2\mathbf{T} \tag{4.114}$$

ermöglicht und zum anderen, bei entsprechender Wahl von $\hat{k}_0 \in \mathcal{I}_h^{k_0}$, eine verbesserte Konvergenz in der Modellordnungsreduktion erzielt. Nach Anwendung des Parameter-Shifts (4.113) lässt sich das FE-System in der Form

$$\Sigma_{FE}(\hat{s}) = \left\{ \begin{array}{rl} (\hat{\mathbf{S}} + \hat{s}\hat{\mathbf{D}} + \hat{s}^2\mathbf{T})\mathbf{X} & = (j\hat{k}_0 + \hat{s})\eta_0\mathbf{B} \\ \mathbf{Z} & = \mathbf{B}^T\mathbf{X} \end{array} \right. \tag{4.115}$$

schreiben, wobei

$$\hat{\mathbf{D}} := \mathbf{D} + 2j\hat{k}_0\mathbf{T} \tag{4.116}$$

gilt. Mit (4.115) liegt somit ein System *zweiter* Ordnung vor, das alle Eigenschaften der Definition (4.5) erfüllt und daher den projektionsbasierten Modellordnungsreduktionsverfahren zugänglich ist. Damit sind auch die momentabgleichenden Eigenschaften der entsprechenden MOR-Verfahren gewährleistet.

Wie in Abschnitt 3.4.1 erläutert, ist die polynomielle Parametrierung der rechten Seite nicht mehr aufrecht zu erhalten, wenn TE- oder TM-Moden im System zu berücksichtigen sind. Diese wirken gemäß (3.64) mit der Skalierung $\boldsymbol{\Xi}(\hat{k}_0, k_0)$ auf die rechte Seite. Eine genaue Betrachtung der Definition von Krylov-Unterräumen mit höherer Ordnung in $\mathbf{B}$ gemäß (4.106) zeigt jedoch, dass diese Räume invariant

gegenüber Skalierungen der rechten Seite sind. Die lineare Hülle des Krylov-Unterraums zum System $\Sigma_{FE}(\hat{s})$ wird von der konkreten Skalierung der Anregung $\mathbf{B}$ nicht beeinflusst. Es gilt

$$\begin{aligned}\mathcal{K}_n(\Sigma_{FE}) &= \mathcal{K}_n(\hat{\mathbf{S}}^{-1}\hat{\mathbf{D}}, \hat{\mathbf{S}}^{-1}\mathbf{T}; jk_0\eta_0\hat{\mathbf{S}}^{-1}\mathbf{B}) \\ &= \mathcal{K}_n(\hat{\mathbf{S}}^{-1}\hat{\mathbf{D}}, \hat{\mathbf{S}}^{-1}\mathbf{T}; (j\hat{k}_0 + \hat{s})\eta_0\hat{\mathbf{S}}^{-1}\mathbf{B}) \\ &= \mathcal{K}_n(\hat{\mathbf{S}}^{-1}\hat{\mathbf{D}}, \hat{\mathbf{S}}^{-1}\mathbf{T}; (j\hat{k}_0 + \hat{s})\eta_0\hat{\mathbf{S}}^{-1}\mathbf{B}\boldsymbol{\Xi}(\hat{k}_0, k_0)) \\ &= \mathcal{K}_n(\hat{\mathbf{S}}^{-1}\hat{\mathbf{D}}, \hat{\mathbf{S}}^{-1}\mathbf{T}; \hat{\mathbf{S}}^{-1}\mathbf{B}).\end{aligned} \tag{4.117}$$

Da die Skalierung $\boldsymbol{\Xi}(\hat{k}_0, k_0)$ identisch auf die Übertragungsfunktionen des Original- und des ROM-Systems wirkt, bleibt die momentabgleichende Eigenschaft unverändert erhalten.

Für den wichtigen Spezialfall des verlustfreien elektrodynamischen Systems, bei dem Leitungs- und Abstrahlungsverluste vernachlässigt werden können, gilt $\mathbf{D} = 0$. Dies erlaubt die Einführung des Parameters

$$\kappa := \hat{k}_0^2 - k_0^2, \tag{4.118}$$

und damit die Formulierung eines elektrodynamischen Systems mit linear parametrierter Systemmatrix

$$(\mathbf{S} - \hat{k}_0^2\mathbf{T}) + (\hat{k}_0^2 - k_0^2)\mathbf{T} = \hat{\mathbf{S}} + \kappa\mathbf{T}. \tag{4.119}$$

In Kapitel 6 wird die Anwendung von MOR-Verfahren für diese wichtige Systemklasse beschrieben und in diesem Zusammenhang ein neuer Fehlerschätzers vorgestellt.

4.5 Definition und Eigenschaften von ROM-Fehlern

Um eine Aussage hinsichtlich der Approximationseigenschaften eines ROMs treffen zu können, sind geeignete Bewertungskriterien zu definieren. In der vorliegenden Arbeit werden ausschließlich jene Fehler betrachtet, die beim Übergang vom originalen System (4.5) zum ordnungsreduzierten Modell (4.14) entstehen. Andere Modellierungsfehler, wie beispielsweise Diskretisierungsfehler der FE-Methode, sind nicht Bestandteil der nachfolgenden Untersuchungen.

Als naheliegende Kriterien werden zunächst der *Fehler im Lösungsvektor*

$$\mathbf{e}_{\mathbf{x}}(s) := \mathbf{Q}\bar{\mathbf{x}}(s) - \mathbf{x}(s) \tag{4.120}$$

und der *Fehler in der Übertragungsfunktion*

$$e_H(s) := \bar{H}(s) - H(s) \tag{4.121}$$

eingeführt. Für ordnungsreduzierte Systeme mit den Eigenschaften (4.57) und (4.58) gilt dann

$$\|\mathbf{e}_\mathbf{x}(s)\| \in \mathcal{O}(s^q), \tag{4.122}$$

$$|e_H(s)| \in \mathcal{O}(s^q). \tag{4.123}$$

In der Umgebung des Entwicklungspunkts $s = 0$ verhält sich der Fehler also proportional zu s^q.

Die Eigenschaft (4.122) lässt sich durch Einsetzen von (4.60) in die Taylor-Entwicklung (4.29) überprüfen. Die Gültigkeit von (4.123) folgt dann wegen

$$\left\|\mathbf{c}^T(s)\right\| = \left\|\sum_{i=0}^{L} s^i \mathbf{c}_i^T(s)\right\| \in \mathcal{O}(1) \tag{4.124}$$

zu

$$|e_H(s)| = \left|\mathbf{c}^T(s)\mathbf{e}_\mathbf{x}(s)\right| \leq \left\|\mathbf{c}^T(s)\right\| \|\mathbf{e}_\mathbf{x}(s)\| \in \mathcal{O}(s^{0+q}) = \mathcal{O}(s^q). \tag{4.125}$$

Ebenso wie die Anzahl der abgeglichenen Momente, verdoppelt sich auch die Fehlerordnung in der Übertragungsfunktion, wenn zusätzlich der Testraum die Bedingung (4.68) erfüllt.

Um dies zu zeigen, wird zunächst das ordnungsreduzierte Modell zum transponierten System (4.64) aufgestellt,

$$\bar{\Sigma}^T(s) = \left\{ \begin{array}{rl} \bar{\mathbf{A}}^T(s)\bar{\mathbf{w}}(s) & = \bar{\mathbf{c}}(s), \\ \bar{v}(s) & = \bar{\mathbf{b}}^T(s)\bar{\mathbf{w}}(s) \end{array} \right. \tag{4.126}$$

und der *Fehlervektor des transponierten Systems*

$$\mathbf{e}_\mathbf{w} := \mathbf{P}\bar{\mathbf{w}} - \mathbf{w} \tag{4.127}$$

eingeführt. Für den gesuchten Ausgangsfehler e_H folgt damit

$$\begin{aligned} e_H &\overset{(4.121)}{=} \bar{H} - H = \mathbf{c}^T(\mathbf{Q}\bar{\mathbf{x}} - \mathbf{x}) \overset{(4.120)}{=} \mathbf{c}^T\mathbf{e}_\mathbf{x} \overset{(4.126)}{=} \mathbf{w}^T\mathbf{A}\mathbf{e}_\mathbf{x} \\ &\overset{(4.127)}{=} (\mathbf{P}\bar{\mathbf{w}} - \mathbf{e}_\mathbf{w})^T\mathbf{A}\mathbf{e}_\mathbf{x} = \bar{\mathbf{w}}^T\mathbf{P}^T\mathbf{A}\mathbf{e}_\mathbf{x} - \mathbf{e}_\mathbf{w}^T\mathbf{A}\mathbf{e}_\mathbf{x}. \end{aligned} \tag{4.128}$$

Aufgrund der Galerkin-Bedingung (4.22) gilt

$$\mathbf{P}^T\mathbf{A}\mathbf{e}_\mathbf{x} = \mathbf{P}^T\mathbf{r} = \mathbf{0} \tag{4.129}$$

und somit

$$e_H = -\mathbf{e}_\mathbf{w}^T\mathbf{A}\mathbf{e}_\mathbf{x}. \tag{4.130}$$

Da $\mathbf{A}$ ein Polynom in s ist und die Fehlervektoren den Beziehungen $\|\mathbf{e}_\mathbf{x}(s)\| \in \mathcal{O}(s^q)$ bzw. $\|\mathbf{e}_\mathbf{w}(s)\| \in \mathcal{O}(s^p)$ genügen, ist die Behauptung

$$e_H \in \mathcal{O}(s^{q+p}) \tag{4.131}$$

erfüllt.

Kapitel 5

Adaptive Strategien in der Mehrpunkt-Modellordnungsreduktion

Modellordnungsreduktionsverfahren haben sich in vielen Anwendungsfeldern als ein sehr effizientes Werkzeug zur schnellen Frequenzgangberechnung linearer Systeme bewährt. Wie in Abschnitt 4.1.1 erläutert, zeichnen sich Mehrpunktverfahren durch eine hohe Flexibilität aus und erlauben die effiziente Anwendbarkeit von iterativen Lösern. Neben FE-Anregungsproblemen können Mehrpunktverfahren auch zur Charakterisierung inhomogener Wellenleiter eingesetzt werden. Die Wellenleiterformulierung führt auf ein verallgemeinertes Eigenwertproblem, dessen Lösungen die Ausbreitungskoeffizienten und zugehörige modale Wellenformen sind [SFDE08].

Ein zentraler Aspekt im Rahmen der Modellordnungsreduktion besteht in der Formulierung eines geeigneten Abbruchkriteriums, welches die folgenden Anforderungen berücksichtigen muss:

- Das Kriterium darf nicht zum Beenden der MOR-Iteration führen, bevor das ordnungsreduzierte Modell die gewünschte Genauigkeit erreicht hat.
- Das Kriterium soll die Generierung redundanter Informationen unterbinden, also die Anzahl zeitintensiver MOR-Iterationen klein halten, um eine effiziente Auswertung zu gewährleisten
- Die Auswertung des Kriteriums soll die zusätzlichen Rechenzeiten und Speicheranforderungen möglichst gering halten.

In den folgenden Abschnitten werden in Anlehnung an [KFK+11] und [SFKDE11] unterschiedliche Strategien zur adaptiven Konstruktion des ROMs untersucht. Hier-

bei kommen residuenbasierte sowie inkrementelle Fehlerindikatoren zur Anwendung. In der vorliegenden Arbeit wird ein direkter Vergleich der genannten Strategien dargestellt, der insbesondere auf die spezifischen Eigenschaften elektrodynamischer Problemstellungen eingeht.

Der naive Ansatz, Entwicklungspunkte s_i gleichmäßig über den Frequenzbereich zu verteilen, führt in der Regel nicht auf ein optimales Ergebnis in der Mehrpunkt-Iteration. Insbesondere werden neue Punkte an Stellen gesetzt, an denen die Systemlösung weitestgehend redundante Information liefert. Gleichzeitig wird mit diesem Vorgehen in Frequenzbereichen mit hoher Systemdynamik das Systemverhalten nicht hinreichend aufgelöst. Dieser Ansatz führt demnach auf ordnungsreduzierte Modelle unnötig hoher Dimension oder unzureichender Genauigkeit.

Mit der *Greedy-Methode* [RRM09] wird eine Strategie vorgeschlagen, die diese Einschränkung überwindet. Hierbei werden neue Stützstellen in jene Punkte gelegt, in denen ein geeignetes Fehlerkriterium ein Maximum annimmt.

Alternativ zur Greedy-Methode wird das Bisektions-Verfahren untersucht. Dieses beruht auf der sukzessiven Teilung jener Intervalle, die den maximalen Wert für ein Fehlerkriterium beinhalten [SFDE09]. Im Gegensatz zur Greedy-Methode müssen hierbei die Fehlerkriterien nicht punktweise betrachtet werden. Stattdessen ist es auch möglich ein integrales Fehlermaß für Teilintervalle heranzuziehen. Die Vorgehensweise des Bisektions-Verfahrens wird dadurch begründet, dass der Interpolationsfehler für glatte Felder dazu tendiert, um den Mittelpunkt zwischen zwei Stützstellen sein Maximum anzunehmen. Daraus resultiert die Annahme, dass durch die Wahl einer Stützstelle in diesem Mittelpunkt ein Maximum an neuer Information zur ROM-Basis hinzugefügt werden kann.

Die numerischen Experimente in Abschnitt 5.4 belegen, dass für beide Ansätze, die Greedy- wie auch die Bisektions-Methode, exponentielle Konvergenz und somit eine vergleichbare Effizienz erreicht wird.

Unabhängig von der Punktwahlstrategie muss in jedem Adaptionsschritt ein Fehlerkriterium in einer Vielzahl von Auswertungspunkten berechnet werden, weshalb der schnellen Berechnung des Kriteriums eine zentrale Bedeutung zukommt. Ein zu hoher Aufwand würde die Vorteile der Modellordnungsreduktion zunichte machen.

5.1 Mehrpunkt-Modellordnungsreduktion in der Simulation elektrodynamischer Felder

Wie bereits bei den Einpunktverfahren festgestellt, kann die mathematische Behandlung der FE-Systeme ohne Beschränkung der Allgemeinheit anhand des SISO-

Falls hinreichend erläutert werden. Im Falle mehrerer Ein- und Ausgangsgrößen sind entsprechende Blockalgorithmen anzuwenden, die in trivialer Weise aus den SISO-Gleichungen abzuleiten sind [SFDE09].

5.1.1 Anregungsprobleme

Basierend auf den in Abschnitt 4.1.1 gezeigten Zusammenhängen wird nachfolgend das Mehrpunkt-Verfahren auf elektrodynamische Problemstellungen angewendet.

Dazu wird als Ausgangspunkt das FE-Gleichungssystem (3.52) herangezogen, das mit der Abkürzung

$$s := jk_0 \tag{5.1}$$

und unter Weglassen der „TF“-Indizes in der Form

$$\mathbf{A}(s) := \left(\mathbf{S} + s\mathbf{D} + s^2\mathbf{T}\right)\mathbf{x} = s\eta_0\mathbf{b} \tag{5.2}$$

gegeben ist. Wie in Abschnitt 4.1.1 beschrieben, werden Lösungen

$$\mathbf{x}_k := \mathbf{x}(s_k) \tag{5.3}$$

an Stützstellen $s_k \in \mathcal{I}_h^s$ bestimmt, die die lineare Hülle des Ansatzraumes

$$\operatorname{bild}\mathbf{Q} := \operatorname{span}\{\mathbf{x}_1, \ldots, \mathbf{x}_K\} \tag{5.4}$$

mit

$$\mathbf{Q} = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_K], \quad \mathbf{q}_i^H\mathbf{q}_j = \delta_{ij}, \tag{5.5}$$

aufspannen.

Im Sinne einer Galerkin-Methode wird der Testraum mit dem Ansatzraum gleichgesetzt,

$$\mathbf{P} := \mathbf{Q}. \tag{5.6}$$

Entsprechend der Definition in Abschnitt 4.1 ergibt sich so das ordnungsreduzierte Modell der FE-Gleichung durch Projektion in der Form

$$\bar{\Sigma}(s) = \left\{ \begin{array}{rl} \bar{\mathbf{A}}(s)\bar{\mathbf{x}}(s) & = \eta_0 s\bar{\mathbf{b}}, \\ \bar{z}(s) & = \bar{\mathbf{b}}^T\bar{\mathbf{x}}(s) \end{array} \right. \tag{5.7}$$

mit der Systemmatrix

$$\bar{\mathbf{A}}(s) := \mathbf{Q}^T\left(\mathbf{S} + s\mathbf{D} + s^2\mathbf{T}\right)\mathbf{Q} \tag{5.8}$$

und der Anregung

$$\bar{\mathbf{b}} := \mathbf{Q}^T \mathbf{b}. \tag{5.9}$$

Die Näherungslösung $\tilde{\mathbf{x}}(s) \approx \mathbf{x}(s)$ zum FE-System (3.52) ist dann durch

$$\tilde{\mathbf{x}}(s) = \mathbf{Q}\bar{\mathbf{x}}(s) \quad \text{mit} \quad \tilde{\mathbf{x}}(s) \in \mathbb{C}^N \quad \text{und} \quad \bar{\mathbf{x}}(s) \in \mathbb{C}^K \tag{5.10}$$

gegeben. So lange die Dimension K des ROM wesentlich kleiner als die des Originalsystems ist, $K \ll N$, kann das ROM um ein Vielfaches schneller ausgewertet, und somit eine hohe Auflösung im relevanten Frequenzbereich erreicht werden.

5.1.2 Radial inhomogene Wellenleiter

In den bisherigen Ausführungen wurden lediglich homogene Wellenleiter betrachtet, bei denen die Feldverläufe von der Frequenz unabhängig sind und der Ausbreitungskoeffizient über die Dispersionsgleichung (2.40) einfach bestimmt werden kann. Entsprechend Abschnitt 4.4 erlaubt diese explizite Frequenzabhängigkeit die Anwendung der Modellordnungsreduktion ohne weitere Einschränkungen.

Bei inhomogenen Wellenleitern treten hingegen allgemein dispersive Wellenformen auf, deren transversale Feldverläufe frequenzabhängig sind. Der Ausbreitungskoeffizient kann für diesen Wellenleiter in der Regel nicht mehr durch einen geschlossenen Ausdruck angegeben werden, dieser ist daher numerisch zu bestimmen. Die Methode zur Behandlung allgemein dispersiver Wellenformen im Rahmenwerk der Modellordnungsreduktion wurde in [SFDE08] vorgestellt. An dieser Stelle wird lediglich die grundlegende Idee der Vorgehensweise angedeutet, da der Fokus nachfolgender Untersuchungen auf den anzuwendenden Abbruchkriterien, jedoch nicht auf dem Verfahren selbst liegt.

Das Ziel der Modellordnungsreduktion im Zusammenhang mit inhomogenen Wellenleitern besteht darin, den Feldverlauf und der zugehörige Ausbreitungskoeffizient einer Wellenform über einen breiten Frequenzbereich zu berechnen. Grundlage bildet hierbei das verallgemeinerte Eigenwertproblem, wie es gemäß [LLL03] und [FHDE04] aus der FE-Diskretisierung resultiert:

$$\sum_{r=0}^{2} f^r \mathbf{S}_r \mathbf{x}(f) = \gamma^2(f) \mathbf{T} \mathbf{x}(f). \tag{5.11}$$

Hierin ist $\gamma \in \mathbb{C}$ der Ausbreitungskoeffizient und $\mathbf{x} \in \mathbb{C}^{N_{\mathrm{WG}}}$ der zugehörige Eigenvektor. Analog zum Anregungsproblem in Abschnitt 3.4 werden die Komponenten $\mathbf{S}_{0,1,2}\mathbb{R}^{N_{\mathrm{WG}} \times N_{\mathrm{WG}}}$ als Steifigkeitsmatrizen bezeichnet, während $\mathbf{T} \in \mathbb{R}^{N_{\mathrm{WG}} \times N_{\mathrm{WG}}}$ die Massenmatrix darstellt. Die vorliegende Struktur erlaubt somit die Anwendung des Mehrpunktverfahrens in gleicher Weise wie für den Anregungsfall. Allerdings werden hier modale Lösungen $\mathbf{x}(f_j), j = 1, 2, \ldots, M$, also Eigenvektoren

in den Entwicklungspunkten $f_1, f_2, \ldots, f_M \in \mathcal{I}_f^h$ gesucht, um eine Projektionsbasis $\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$ mit der Eigenschaft

$$\operatorname{span}\{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\} \subseteq \operatorname{span}\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_M\} \tag{5.12}$$

zu konstruieren.

Für die numerische Berechnung muss gemäß der Ausführungen in [SFDE08] gewährleistet sein, dass die Basisvektoren $\mathbf{w}_j$ $\mathbf{T}$-orthogonal zum a priori bekannten Unterraum

$$\mathcal{N} := \operatorname{span}\{\mathbf{n}_1(f), \mathbf{n}_2(f), \ldots, \mathbf{n}_{N_n}\} \tag{5.13}$$

von *Nullfeldern* $\mathbf{n}_i(f)$ sind. Die Nullfelder ergeben sich als nicht-triviale Lösungen aus dem Gleichungssystem

$$\sum_{r=0}^{2} f^r \mathbf{S}_r \mathbf{n}(f) = 0 \tag{5.14}$$

und müssen zur Vermeidung unphysikalischer Lösungen im MOR-Prozess aus den Projektionsräumen ausgeschlossen werden. Es wird daher gefordert, dass die Basisvektoren $\mathbf{w}_j(f)$ der Bedingung

$$\mathbf{n}_i^T(f)\mathbf{T}\mathbf{w}_j(f) = \mathbf{0} \tag{5.15}$$

für alle Frequenzen f genügen. Wie in [SFDE08] gezeigt, kann dieser Forderung mittels einer frequenzabhängigen Transformation der Form

$$\mathbf{q}_j(f) = (\mathbf{Q}_0 + f\mathbf{Q}_1)\mathbf{w}_j \quad \text{mit geeigneten Matrizen } \mathbf{Q}_0, \mathbf{Q}_1 \in \mathbb{R}^{N_{\mathrm{WG}} \times N_{\mathrm{WG}}} \tag{5.16}$$

genüge getan werden. Mit den so ermittelten Basisvektoren werden, wie im Anregungsfall, die Projektionsmatrizen

$$\mathbf{Q} = [\mathbf{q}_1 \mathbf{q}_2 \cdots \mathbf{q}_K] \text{ und } \mathbf{P} = \mathbf{Q} \tag{5.17}$$

zusammengesetzt und das ordnungsreduzierte Modell

$$\sum_{r=0}^{4} f^r \bar{\mathbf{S}}_r \bar{\mathbf{x}}(f) = \gamma^2 \sum_{r=0}^{2} f^r \bar{\mathbf{T}}_r \bar{\mathbf{x}}(f) \tag{5.18}$$

mit $\bar{\mathbf{S}}_{0\ldots4}, \bar{\mathbf{T}}_{0\ldots2} \in \mathbb{C}^{M \times M}$ gebildet. Hierin berechnet sich die linke Seite über die Beziehung

$$\sum_{r=0}^{4} f^r \bar{\mathbf{S}}_r = \mathbf{Q}^H (\mathbf{Q}_0 + f\mathbf{Q}_1)^H \Big(\sum\nolimits_{r=0}^{2} f^r \mathbf{S}_r \Big) (\mathbf{Q}_0 + f\mathbf{Q}_1)\mathbf{Q} \tag{5.19a}$$

und für die rechte Seite gilt

$$\sum_{r=0}^{2} f^r \bar{\mathbf{T}}_r = \mathbf{Q}^H (\mathbf{Q}_0 + f\mathbf{Q}_1)^H \mathbf{T} (\mathbf{Q}_0 + f\mathbf{Q}_1)\mathbf{Q}. \tag{5.19b}$$

Die ROM-Näherungslösung $\tilde{\mathbf{x}} \approx \mathbf{x}$ ist über die Beziehung

$$\tilde{\mathbf{x}}(f) = (\mathbf{Q}_0 + f\mathbf{Q}_1)\mathbf{Q}\bar{\mathbf{x}}(f) \tag{5.20}$$

gegeben.

5.2 Fehlerindikatoren

Die Modellordnungsreduktion ermöglicht eine breitbandige Charakterisierung elektrodynamischer Systeme bei gleichzeitig hoher Auflösung im betrachteten Frequenzbereich $\mathcal{I}_f$. Für typische Anwendungen wird eine Auflösung von $N_f = 1000 \ldots 5000$ Frequenzpunkten angestrebt, um auch schmalbandige Effekte darstellen zu können. Damit wird deutlich, dass neben der schnellen Auswertbarkeit des ROMs auch die Berechnung des Fehlerkriteriums effizient gestaltet werden muss. Nachfolgend werde zwei Ansätze vorgeschlagen, die diese Voraussetzung erfüllen.

5.2.1 Residuen-basierte Indikatoren

Die Residuen $\mathbf{r}$ zu den FE-Gleichungen (5.2) bzw. (5.11) bezüglich der ROM-Approximationen $\bar{\mathbf{x}}$ genügen den folgenden Beziehungen:

- **Anregungsfall**: Für (5.2) gilt mit (5.10) für das Residuum

$$\mathbf{r}(s) = \mathbf{A}(s)\tilde{\mathbf{x}}(s) - \eta_0 s\mathbf{b}. \tag{5.21a}$$

- **Wellenleiterprobleme:** Für (5.11) gilt mit (5.20)

$$\mathbf{r}(f) = \Big(\sum\nolimits_r f^r \mathbf{S}_r - \gamma^2 \mathbf{T}\Big)(\mathbf{Q}_0 + f\mathbf{Q}_1)\tilde{\mathbf{x}}(f). \tag{5.21b}$$

Als *lokaler* Indikator wird das relative Residuum

$$r(s) = \frac{\|\mathbf{r}(s)\|_2}{\rho_0}, \tag{5.22}$$

herangezogen; analog $r(f)$ für das Wellenleiterproblem. Hierin wird das Referenzresiduum ρ_0 von einer initialen Schätzung der FE-Lösung $\mathbf{x}_0$ abgeleitet. Im Anregungsfall berechnet sich dieses über

$$\rho_0 = \max_{s_i \in \mathcal{I}_s^h} \|\mathbf{A}(s_i)\mathbf{x}_0 - \eta_0 s_i \mathbf{b}\|_2 \tag{5.23}$$

und für Wellenleiterprobleme ergibt sich entsprechend

$$\rho_0 = \max_{f_i \in \mathcal{I}_f^h} \left\| \left(\sum\nolimits_r f_i^r \mathbf{S}_r - \gamma_0^2(f_i)\mathbf{T} \right) \mathbf{x}_0 \right\|_2 . \tag{5.24}$$

Zur Berechung des Residuums im Wellenleiterfall muss zusätzlich eine Approximation des Eigenwerts γ^2 bestimmt werden. Diese erfolgt über den Rayleigh-Koeffizienten in der Form

$$\gamma^2(f_i) \approx \gamma_0^2(f_i) = \frac{\mathbf{x}_0^T \mathbf{S}(f_i)\mathbf{x}_0}{\mathbf{x}_0^T \mathbf{T}\mathbf{x}_0}. \tag{5.25}$$

Mit den eingeführten Größen lässt sich der *globale* Indikator $\mathcal{R}$ über dem gesamten Parameterbereich als skalierte L_2-Norm auf der Menge der Evaluierungspunkte $\mathcal{I}_h^s$ bzw. $\mathcal{I}_h^f$

$$\mathcal{R} = \sqrt{\frac{1}{N_f} \sum_{i=1}^{N_f} r_i^2} \tag{5.26}$$

formulieren, wobei hier die Abkürzung $r_i = r(f_i)$ bzw. $r_i = r(s_i)$ verwendet wird.

Häufig wird für iterative Lösungsverfahren als Abbruchkriterium das Residuum herangezogen, was sich für praktische Anwendungen als sehr effizient und zuverlässig erweist. Daher ist auch im Rahmen der Modellordnungsreduktion diese Vorgehensweise naheliegend, auch wenn ein kleines Residuum nicht notwendig auf einen kleinen Fehler schließen lässt.

Wie zuvor schon erläutert, ist eine wesentliche Anforderung durch die schnelle Auswertbarkeit der Indikatoren gegeben. In der Residuengleichung (5.21) erfolgt die Berechnung anhand hochdimensionaler FE-Matrizen und -Vektoren. Obwohl die Matrizen dünnbesetzt sind und somit Produkte effizient berechnet werden können, ist der Aufwand zur Indikatorberechnung nicht mehr zu vernachlässigen. Dies liegt insbesondere an der großen Anzahl N_f an Auswertepunkten. Das Auswerten des Kriteriums in jedem ROM-Iterationsschritt würde somit den gesamten ROM-prozess dominieren.

In weiterer Folge wird gezeigt, wie die Auswertung des Residuums und damit auch des globalen Fehlerkriteriums sehr effizient gestaltet werden kann. Mit Hilfe der Darstellung der Näherungslösung (5.10) kann die quadrierte Norm des Residuums

im Anregungsfall (5.21a) geschrieben werden als

$$\begin{aligned}\|\mathbf{r}\|_2^2 = \bar{\mathbf{x}}^H \left(\mathbf{Q}^H \mathbf{A}^H(s)\mathbf{A}(s)\mathbf{Q}\right) \bar{\mathbf{x}} \\ + 2\,\mathrm{Re}\left[\eta_0 s \left(\mathbf{b}^H \mathbf{A}(s)\mathbf{Q}\right) \bar{\mathbf{x}}\right] + \eta_0^2 \left|s\right|^2 \|\mathbf{b}\|_2^2 .\end{aligned} \tag{5.27a}$$

Entsprechend führt das Wellenleiterproblem mit (5.20) und (5.21b) auf

$$\begin{aligned}\|\mathbf{r}\|_2^2 = \sum\nolimits_s \sum\nolimits_r f^{r+s} \bar{\mathbf{x}}^H \left(\mathbf{Q}^H \mathbf{S}_s^H \mathbf{S}_r \mathbf{Q}\right) \bar{\mathbf{x}} \\ - 2\,\mathrm{Re}\left[\gamma^2 \sum\nolimits_s f^s \bar{\mathbf{x}}^H \left(\mathbf{Q}^H \mathbf{S}_s^H \mathbf{T}\mathbf{Q}\right) \bar{\mathbf{x}}\right] \\ + |\gamma|^4 \bar{\mathbf{x}}^H \left(\mathbf{Q}^H \mathbf{T}^H \mathbf{T}\mathbf{Q}\right) \bar{\mathbf{x}}.\end{aligned} \tag{5.27b}$$

Damit weisen alle auftretenden Matrizen und Vektoren in (5.27) lediglich die Dimension K auf, so dass die notwendigen Berechnungen mit den übrigen ROM-Operationen vergleichbar und damit sehr schnell auszuwerten sind. Außerdem fallen die Produkte $\mathbf{A}(s)\mathbf{Q}, \mathbf{S}_r\mathbf{Q}$ und $\mathbf{T}\mathbf{Q}$ ohnehin während der Generierung der ROM-Matrizen $\bar{\mathbf{A}}(s), \bar{\mathbf{S}}_r$ und $\bar{\mathbf{T}}_r$ in (5.8) bzw. (5.19) an und müssen nicht zusätzlich berechnet werden.

Die Kosten der Berechnung residuenbasierter Fehlerindikatoren fallen somit sehr gering aus, was die Eignung des globalen Kriteriums $\mathcal{R}$ als Indikator in einem adaptiven Prozess unterstreicht.

5.2.2 Inkrementelle Indikatoren

Inkrementelle Indikatoren, wie in [SFDE09] beschrieben, basieren auf der Differenz der Ausgangsgrößen zweier aufeinanderfolgender ROM-Iterationen $k-1$ und k. Im spezifischen Fall von Anregungsproblemen in der elektromagnetischen Feldsimulation stellen Netzwerkparameter, insbesondere S-Parameter S_{mn}, geeignete Größen dar, um ein ROM zu bewerten. Der Zusammenhang zwischen der Ausgangsgröße $\mathbf{Z}$ in (3.52) und S-Parametern wird in Abschnitt 6.5 näher erläutert. Für Wellenleitern bildet der Ausbreitungskoeffizient γ jene Größe, die zur Bewertung des ROM herangezogen wird.

Für die weiteren Ausführungen wird der lokale Indikator $e(f)$ eingeführt:

$$e(f) := \begin{cases} |S_{mn}^k(f) - S_{mn}^{k-1}(f)| & \text{im Anregungsfall,} \\ |\gamma^k(f) - \gamma^{k-1}(f)| & \text{für Wellenleiter.} \end{cases} \tag{5.28}$$

Aus Gründen der besseren Lesbarkeit wird in weiterer Folge nur noch die Frequenz f als System-Parameter gewählt. Da der Ausbreitungskoeffizient in $\mathcal{I}_h^f$ beschränkt ist und für S-Parameter einer passiven Struktur stets $|S_{mn}| \leq 1$ gilt, ergibt sich der

entsprechende globale Indikator $\mathcal{E}$ für den gesamten Frequenzbereich als Maximum von $e(f)$ auf der Menge $\mathcal{I}_f^h$ zu

$$\mathcal{E} = \max_{f_i \in \mathcal{I}_f^h} e(f_i). \tag{5.29}$$

Gleichung (5.28) zeigt, dass inkrementelle Indikatoren mit sehr geringem Rechenaufwand zu bestimmen sind. Darüber hinaus sind diese, im Gegensatz zu den residuenbasierten Indikatoren, direkt mit der Systemlösung verknüpft. Demgegenüber steht die Tatsache, dass der tatsächliche Fehler insbesondere bei Verfahren mit langsamer Konvergenz *unter*schätzt wird. Dies führt zu einem vorzeitigen Abbruch der MOR-Iteration und somit zu einem ROM, das nicht die gewünschte Approximationsgenauigkeit aufweist. Wie sich jedoch bei den numerischen Beispielen in Abschnitt 5.4 zeigt, kann das Risiko des verfrühten Beendens der Iteration im MOR-Kontext eingeschränkt werden. Der globale Indikator $\mathcal{E}$ berücksichtigt die Änderung in vielen Frequenzpunkten gleichzeitig. Damit der Indikator versagt, müsste demnach die Lösung über dem gesamten Parameterbereich, also in allen N_f Evaluierungspunkten, stagnieren. Zudem zeigen die untersuchten Ordnungsreduktionsverfahren eine sehr schnelle Konvergenz, so dass betragsmäßig kleine Differenzen in aufeinanderfolgenden Iterationsschritten in der Regel erst dann auftreten, wenn das ROM schon in sehr guter Näherung mit dem Originalmodell übereinstimmt.

5.3 Adaptive Strategien zur Bestimmung von Entwicklungspunkten

Wie in Abschnitt 4.1.2 gezeigt, wird bei Einpunktverfahren lediglich die Lösung in einem einzelnen Entwicklungspunkt benötigt. In den meisten Anwendungen basiert die Festlegung der Entwicklungsfrequenz auf Erfahrungswerten, so dass zum Beispiel die Mittenfrequenz gewählt wird. Ein ungünstig positionierter Entwicklungspunkt kann dabei das Konvergenzverhalten des Verfahrens negativ beeinflussen, so dass das resultierende ROM bezüglich der Modellordnung nicht optimal ist.

Im Gegensatz dazu ist bei Mehrpunktverfahren die Möglichkeit der *adaptiven Punktwahl* gegeben. So kann im MOR-Prozess das Systemverhalten aus bereits bekannten Lösungsanteilen bewertet und zur Wahl weiterer Entwicklungspunkte herangezogen werden. Die folgenden Abschnitte behandeln zwei konkurrierende Strategien zur Positionierung der Entwicklungspunkte, wobei die Bewertung der Zwischenergebnisse anhand der Fehlerindikatoren aus Abschnitt 5.2 erfolgt.

5.3.1 Bisektionsmethode

Im ersten Schritt der Bisektionsmethode erfolgt eine Aufteilung des Frequenzbereichs in zwei gleichgroße Teilintervalle [SFDE09]. Für die jeweiligen Abschnitte werden die globalen Fehlerindikatoren aus Abschnitt 5.2 ausgewertet und miteinander verglichen. In jenem Teilintervall, das den größten Wert für das Fehlermaß aufzeigt, wird ein neuer Entwicklungspunkt in der Intervallmitte platziert. Das entsprechende Teilintervall wird also erneut in zwei gleichgroße Abschnitte halbiert.

Anhand des Anregungsproblems kann diese Strategie wie folgt motiviert werden: Sei $[f_1, f_2]$ ein Intervall der Breite $W := f_2 - f_1$ und $\mathbf{x}(f)$ ausreichend glatt, so dass die zweite Ableitung $\mathbf{x}''(f)$ beschränkt ist,

$$\|\mathbf{x}''(f)\|_\infty \leq L < \infty \qquad \text{für } f \in [f_1, f_2]. \tag{5.30}$$

Unter der Voraussetzung, dass der durch $\mathbf{Q}$ aufgespannte Spaltenraum wesentliche Komponenten in Richtung von $\mathbf{x}(f)$ aufweist, was für ein ROM ausreichender Dimension angenommen werden kann, folgt aus der diskreten Version von [BS94, Theorem 5.7.6] unter schwachen Annahmen für $\mathbf{A}(f)$ [BS94, Abschnitt 5.7], dass das ROM dahingehend quasi-optimal ist, dass eine normabhängige positive Konstante $C < \infty$ derart existiert, dass

$$\|\tilde{\mathbf{x}}(f) - \mathbf{x}(f)\| \leq C \min_{\mathbf{y} \in \operatorname{bild} \mathbf{Q}} \|\mathbf{y}(f) - \mathbf{x}(f)\| \tag{5.31}$$

gilt.

Da die FE-Lösungen $\mathbf{x}(f_1)$ und $\mathbf{x}(f_2)$ in den Entwicklungspunkten f_1 und f_2 in $\operatorname{bild} \mathbf{Q}$ liegen, gilt für die lineare Interpolation $\mathbf{p}_I(f)$,

$$\mathbf{p}_I(f) = \left[(f - f_1)\mathbf{x}(f_2) - (f - f_2)\mathbf{x}(f_1)\right]/W. \tag{5.32}$$

Gemäß [SB92, Theorem 2.1.4.1] ist der entsprechende Interpolationsfehler $\mathbf{p}_I(f) - \mathbf{x}(f)$ beschränkt durch

$$\|\mathbf{p}_I(f) - \mathbf{x}(f)\| \leq C_I L W^2 \qquad \text{für } f \in [f_0, f_1], \tag{5.33}$$

mit einer normabhängigen positiven Konstante $C_I < \infty$. Substitution von $\mathbf{y}$ durch $\mathbf{p}_I$ in (5.31) liefert für den ROM-Fehler die Schranke

$$\|\bar{\mathbf{x}}(f) - \mathbf{x}(f)\| \leq C C_I L W^2 \qquad \text{für } f \in [f_0, f_1]. \tag{5.34}$$

Aus Plausibilitätsgründen kann die Bisektionsmethode als sehr effiziente Strategie angenommen werden, da hierbei die maximale Breite des neuen Subintervalls auf $W/2$ minimiert wird. Durch (5.34) wird damit die Konvergenz garantiert. Gleichzeitig wird zudem die minimale Distanz zwischen dem neuen Entwicklungspunkt und dem nächsten Nachbarn maximiert, was Auslöschungseffekte vermeidet.

5.3.2 Greedy-Methode

Im Vergleich zur Bisektionsmethode verfolgt die Greedy-Methode [RRM09] eine aggressivere Strategie: In jedem Iterationsschritt wird ein neuer Entwicklungspunkt $\check{f}$ an jener Stelle platziert wird, an der der *lokale* Fehlerindikator sein Maximum annimmt,

$$\check{f} = \arg\max_{f_i \in \mathcal{I}_f^h} e(f_i). \tag{5.35}$$

Die Attraktivität dieses Ansatzes liegt in dem Versuch begründet, im Gegensatz zur Bisektionsmethode mehr lokale Information in den adaptiven Zyklus einfließen zu lassen. Die mathematischen Grundlagen und Eigenschaften sind in [BMP+12] [BCD+11] dargestellt.

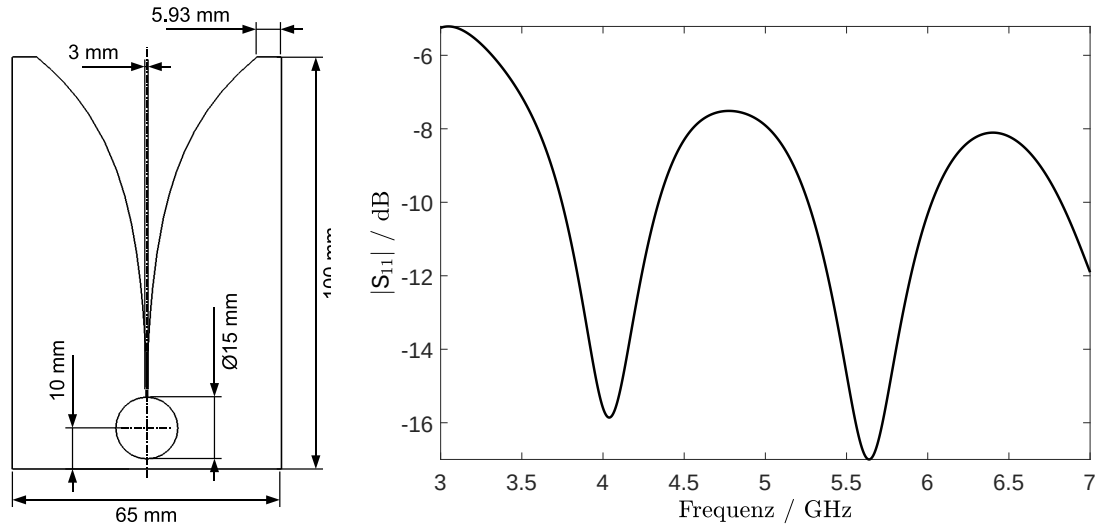
Ein potenzielles Problem besteht jedoch in der möglichen numerischen Auslöschung, da Entwicklungspunkte beliebig nah beieinander liegen können. In diesem Fall wird der Großteil der neuen Information einer FE-Lösung durch Rundungsfehler eliminiert. Für Interpolationsansätze kann ein solches Verhalten zum Versagen des ganzen Verfahrens führen. Zwar ist im Rahmenwerk der Modellordnungsreduktion durch Projektion aufgrund der Orthogonalisierung in (5.5) diese Gefahr nicht gegeben, dennoch wächst durch die Hinzunahme redundanter Ansatzvektoren die ROM-Dimension mehr als nötig und die ROM-Generierung nimmt zusätzlich Rechenzeit und Speicherkapazität in Anspruch.

5.4 Numerische Untersuchungen

In den folgenden Beispielen ist die Anzahl der Evaluierungspunkte stets auf $N_f = 2001$ gesetzt. Zur Beurteilung der untersuchten Verfahren wird jeweils auch der Bezug zum tatsächlichen Fehler entsprechend der Definition (4.121), also der Vergleich zur klassischen FE-Lösung, in den Evaluierungspunkten aufgezeigt. Der dargestellte Fehler beinhaltet somit nicht den Diskretisierungsfehler der aus der FE-Methode selbst resultiert. Für alle Berechnungen erfolgt der Abbruch erst, wenn der Fehler im Bereich des numerischen Rauschpegels liegt. Damit wird gewährleistet, dass die unterschiedlichen Varianten der betrachteten adaptiven MOR-Verfahren bestmöglich zutage treten. In praktischen Anwendungen wird ein adäquates Abbruchkriterium festzulegen sein, das eine bestmögliche Effizienz der Verfahren gewährleistet.

Der Fokus liegt bei den numerischen Untersuchungen auf

- den Konvergenzraten,
- dem Rundungsverhalten der schnellen Residuenberechnung gemäß (5.27),



(a) Geometrie. (b) $|S_{11}|$ über der Frequenz.

Abbildung 5.1: Vivaldi-Antenne: Geometrische Struktur und Amplitudenantwort über der Frequenz.

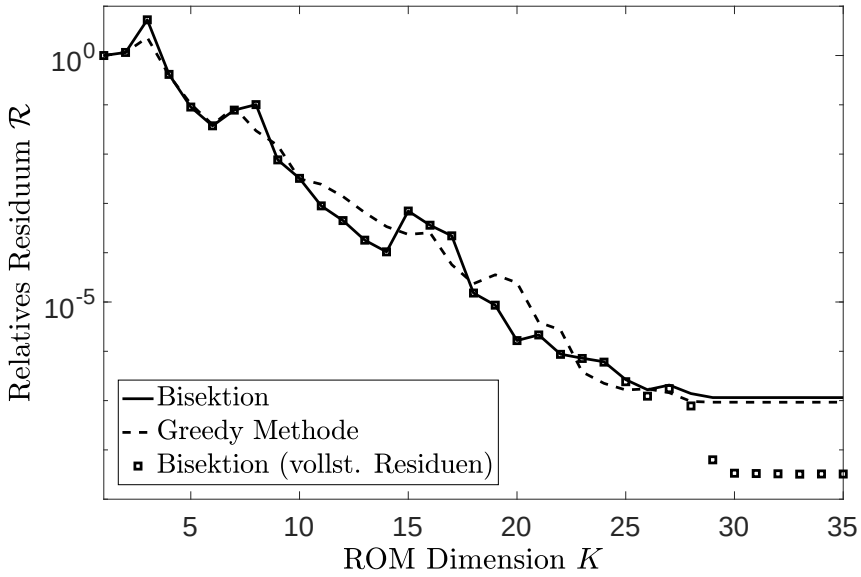
- der Effizienz der residuellen wie auch der inkrementellen Fehlerschätzer und
- den Unterschieden zwischen den adaptiven Strategien, also der Greedy- und der Bisektionsmethode.

5.4.1 Vivaldi-Antenne

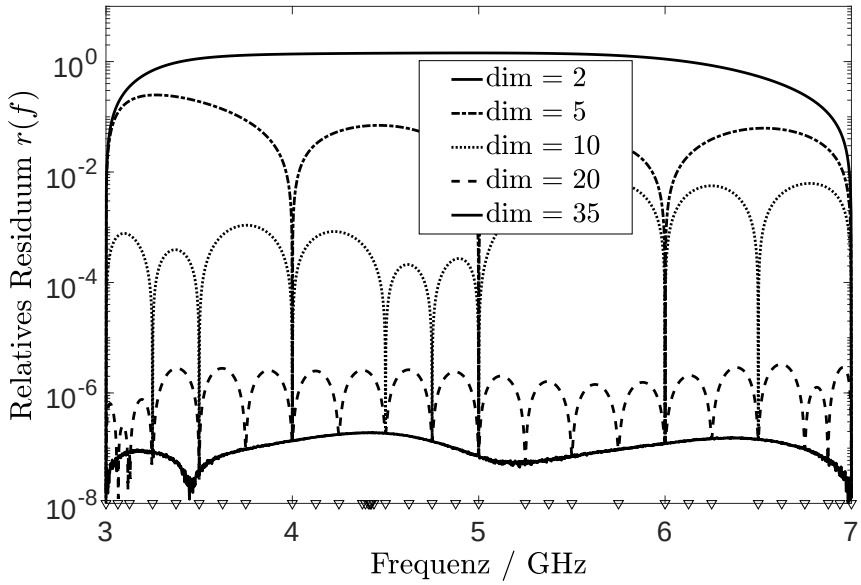
Das Beispiel beschreibt eine einfache Breitbandantenne gemäß Abbildung 5.1(a), wobei die Metallisierung als idealer elektrischer Leiter modelliert ist. Die Frequenzantwort von $|S_{11}|$ in Abbildung 5.1(b) zeigt ein nicht-resonantes Verhalten der abstrahlenden Struktur. In Abbildung 5.2(a) und Abbildung 5.3(a) wird das Verhalten des residuenbasierten und des inkrementellen Fehlerindikators bezüglich der ROM-Dimension K gezeigt. In beiden Fällen kann kein Vorteil der Greedy- gegenüber der Bisektionsmethode festgestellt werden. Zusätzlich ist aus dem Vergleich mit dem tatsächlichen Fehler in S_{11} gemäß Abbildung 5.4(a) zu erkennen, dass die residuen- und die inkrementell gesteuerten adaptiven Strategien ähnlich gute Ergebnisse erzielen: Die Linien konstanter durchschnittlicher Steigung in semi-logarithmischen Darstellung weisen darauf hin, dass eine *exponentielle* Konvergenz erzielt wird, mit

$$\left\| S_{11}^{\mathrm{ROM}}(M) - S_{11}^{\mathrm{FE}} \right\|_{\infty} \propto 10^{-0.55M}. \tag{5.36}$$

In Abbildung 5.2(a) ist zu erkennen, dass die schnelle Residuenberechnung (5.27a) ab der ROM-Dimension $K = 28$ im numerischen Rauschen resultiert und ein relatives Residuum von $r \approx 10^{-7}$ anzeigt. Demgegenüber liefert die naive, wenn auch unpraktikable Berechnung des Residuums noch aussagekräftige Werte bis zur ROM-

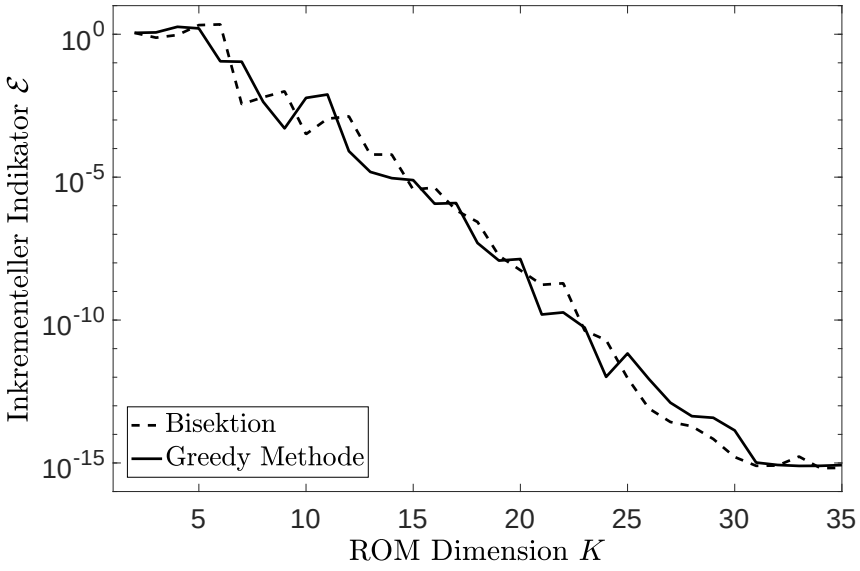


(a) Globaler Indikator vs. ROM-Dimension.

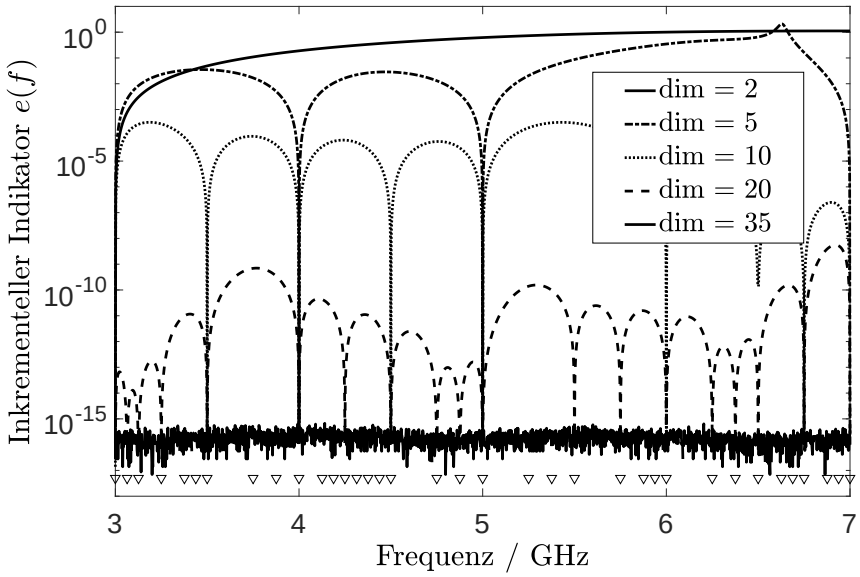


(b) Bisektion: Lokaler Indikator vs. Frequenz.

Abbildung 5.2: Vivaldi-Antenne: adaptive ROM-Generierung gesteuert von residuenbasiertem Fehlerindikator in der euklidischen Norm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

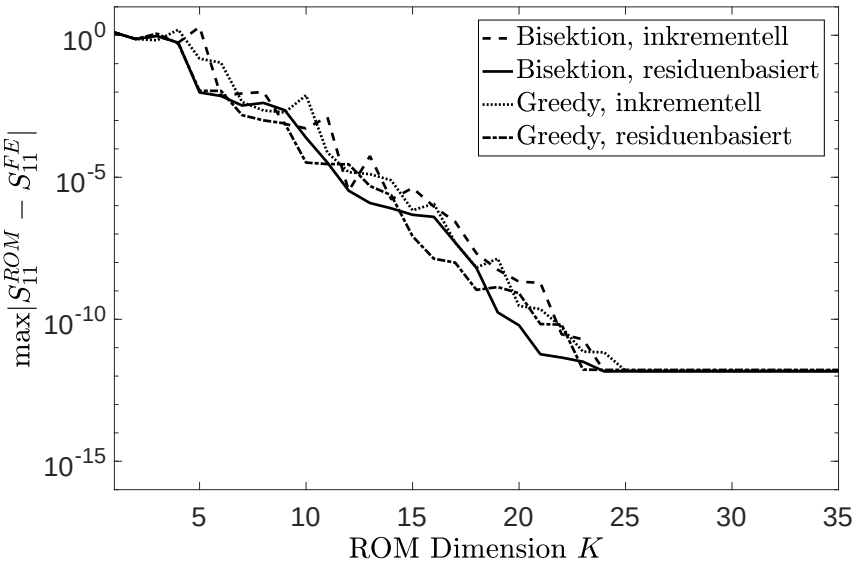


(a) Globaler Indikator vs. ROM-Dimension.

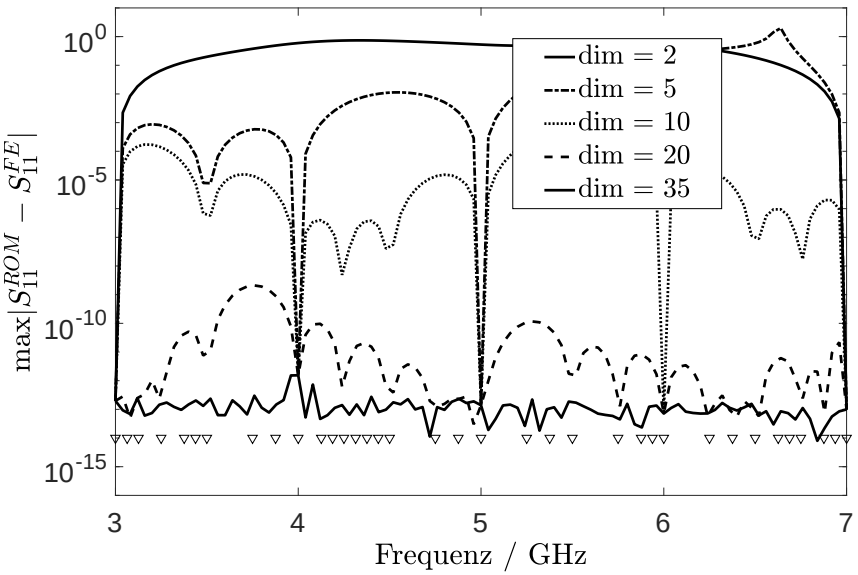


(b) Bisektion: lokaler Indikator vs. Frequenz.

Abbildung 5.3: Vivaldi-Antenne: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.



(a) Globaler Fehler vs. ROM-Dimension.



(b) Bisektion: lokaler Fehler vs. Frequenz.

Abbildung 5.4: Vivaldi-Antenne: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

Dimension $K = 30$ und bis zu einem Niveau von etwa $r \approx 3 \cdot 10^{-9}$. Jedoch zeigt Abbildung 5.4(a), dass bereits für $K = 24$ das Fehlerniveau bei 10^{-12} liegt. Damit ist der Fehler des Modellordnungsprozesses unterhalb typischer Modellierungsfehler der FE-Methode und damit für praktische Anwendungen ausreichend klein.

Die Abbildungen 5.2(b), 5.3(b) und 5.4(b) zeigen die lokalen Indikatoren und den tatsächlichen Fehler in S_{11} als Funktion der Frequenz für verschiedene Stadien des bisektion-basierten adaptiven Prozesses. Die Dreiecke auf der Abszissenachse zeigen dabei jeweils die Entwicklungspunkte des finalen reduzierten Modells an.

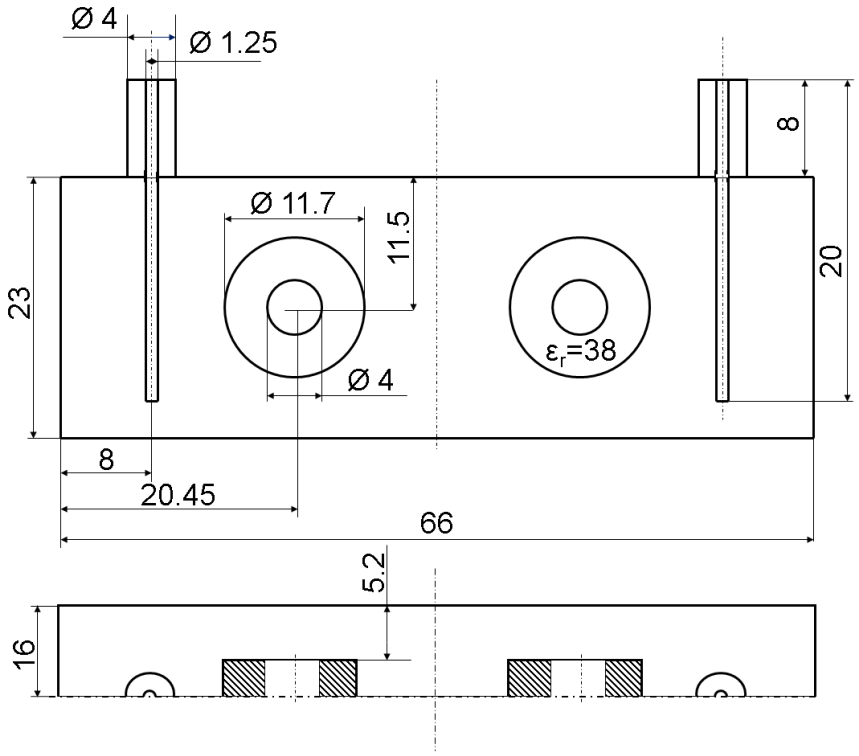
5.4.2 Bandpassfilter

Ein weiteres Untersuchungsbeispiel für den Anregungsfall stellt das Bandpassfilter mit dielektrischen Resonatoren gemäß Abbildung 5.5(a) dar [BL97]. Die Frequenzantwort der hochresonanten Struktur ist in Abbildung 5.5(b) gezeigt. Die Abbildungen 5.6(a), 5.7(a) und 5.8(a) illustrieren das Verhalten der Fehlerindikatoren und des tatsächlichen Fehlers gegenüber der ROM-Dimension. Auch hier zeigt sich, dass sich der residuenbasierte Indikator sehr ähnlich zum inkrementellen Indikator verhält. In beiden Fällen liefert die Greedy-Methode etwas bessere Ergebnisse. Abbildung 5.6(a) zeigt, dass wie schon im vorangegangenen Beispiel die schnelle Residuenauswertung früher den Pegel des numerischen Grundrauschen erreicht als die naive Berechnung. Allerdings gilt auch hier wieder, dass entsprechend Abbildung 5.8(a) dieser Effekt erst auftritt, wenn der tatsächliche Fehler bereits unterhalb eines für die Praxis relevanten Niveaus liegt.

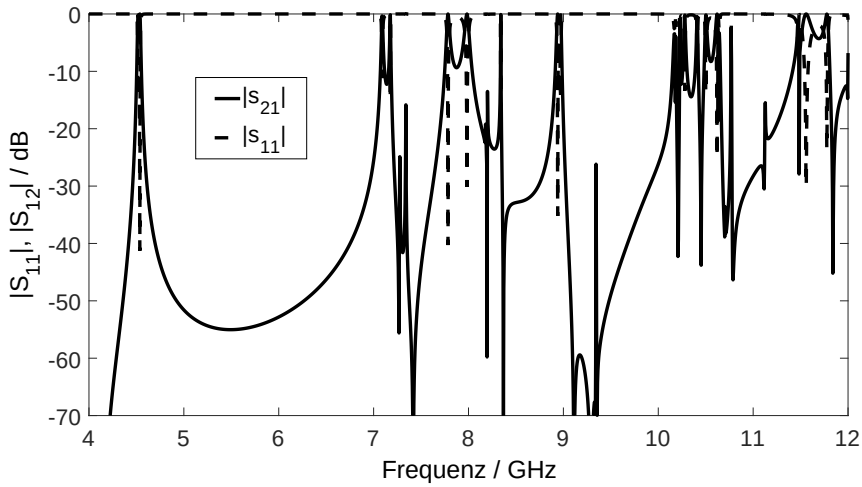
Der einzig nennenswerte Unterschied zum Untersuchungsbeispiel aus Abschnitt 5.4.1 liegt im Gesamtverlauf der Konvergenzkurven. Im Fall des Bandpassfilters tritt eine längere Periode der Stagnation ein, auf welche die nahezu sofortige Konvergenz folgt. Begründen lässt sich dies wie folgt: Die scharfen Resonanzen in Abbildung 5.5(b) zeigen, dass die Übertragungsfunktion im betrachteten Frequenzbereich eine große Anzahl an Polstellen in enger Nachbarschaft aufweist. So lange die MOR-Iteration nicht in der Lage ist, alle diese Polstellen mit einer hohen Genauigkeit aufzulösen, liegt in der Nähe dieser Polstellen eine große Dynamik in den Frequenzantworten unterschiedlicher Iterationsstufen vor. Die Fehlerindikatoren liefern hier zuverlässig große Werte, so dass ein vorzeitiger Abbruch des Verfahrens verhindert wird.

Insbesondere beim Vergleich des tatsächlichen Fehlers in Abbildung 5.8(b) mit der Frequenzantwort gemäß Abbildung 5.5(b) wir dieser Effekt deutlich: Sogar bei der ROM-Dimension von $N_f = 50$ weist der Fehler signifikante Spitzen in vielen der Resonanzfrequenzen auf.

Die Abbildungen 5.6(b) und 5.7(b) demonstrieren, dass beide Ansätze, der residuenbasierte, wie auch der inkrementelle Indikator, dieses Verhalten sehr gut wi-

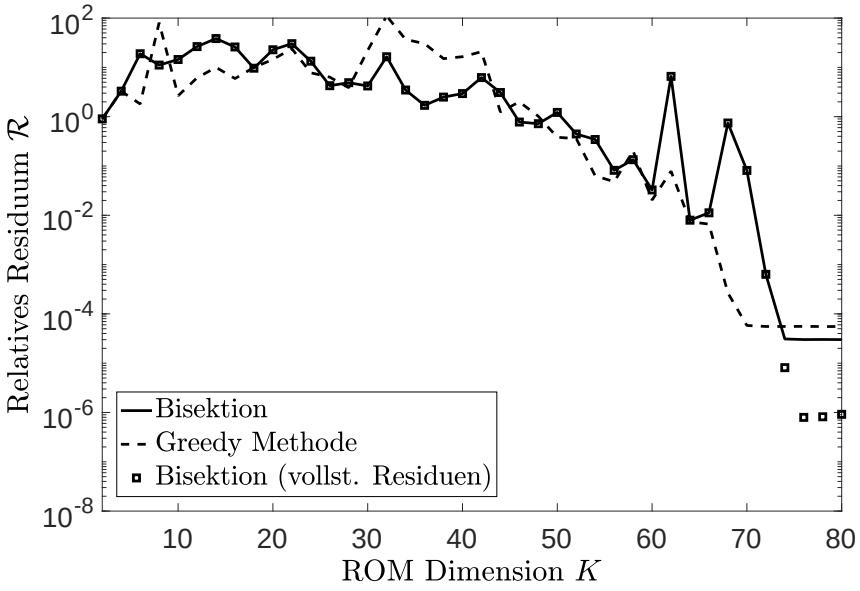


(a) Dimensionen in mm.

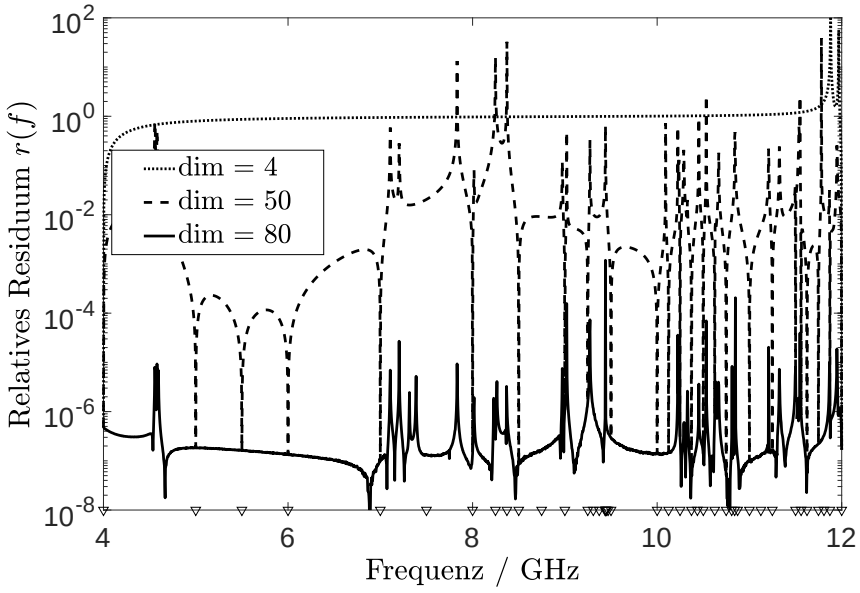


(b) Filterantwort.

Abbildung 5.5: Bandpassfilter: Struktur und Amplitudenantwort über der Frequenz.

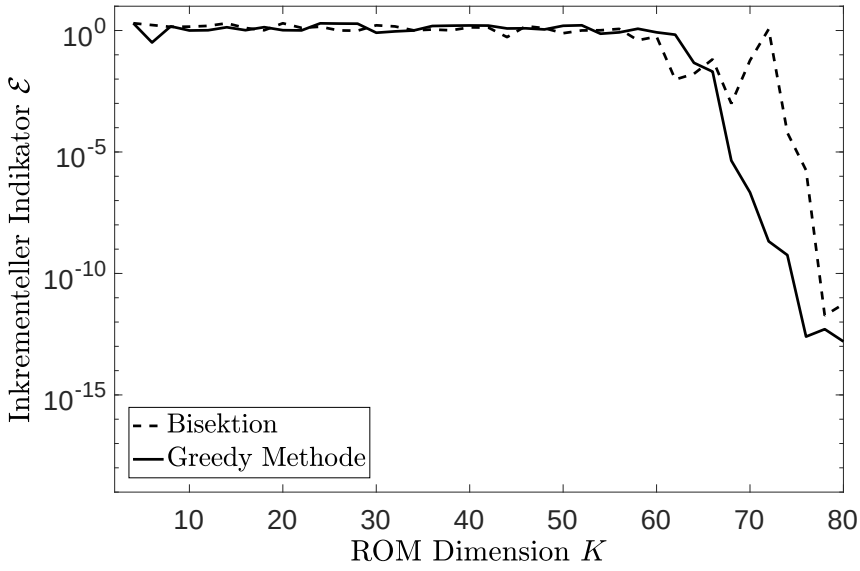


(a) Globaler Indikator vs. ROM-Dimension.

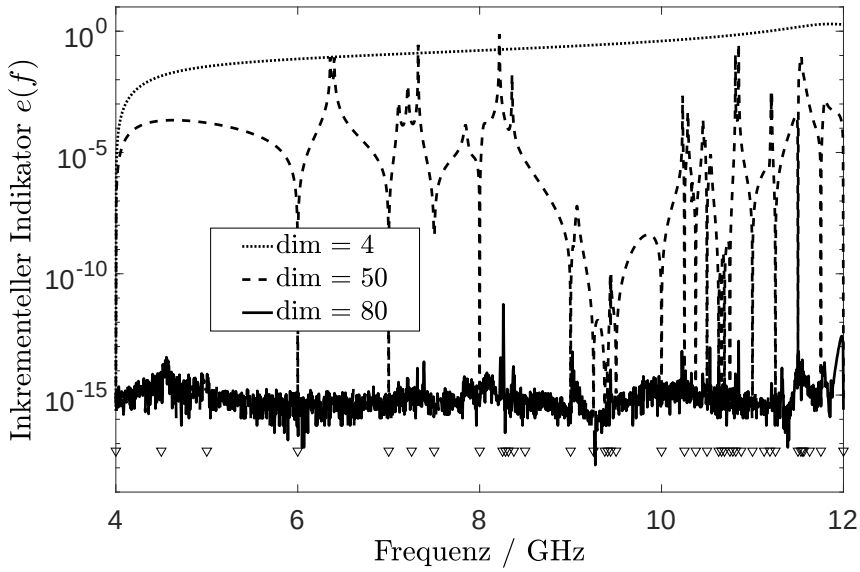


(b) Bisektion: lokaler Indikator vs Frequenz.

Abbildung 5.6: Bandpassfilter: adaptive ROM-Generierung gesteuert von residualbasiertem Fehlerindikator in der euklidischen Norm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

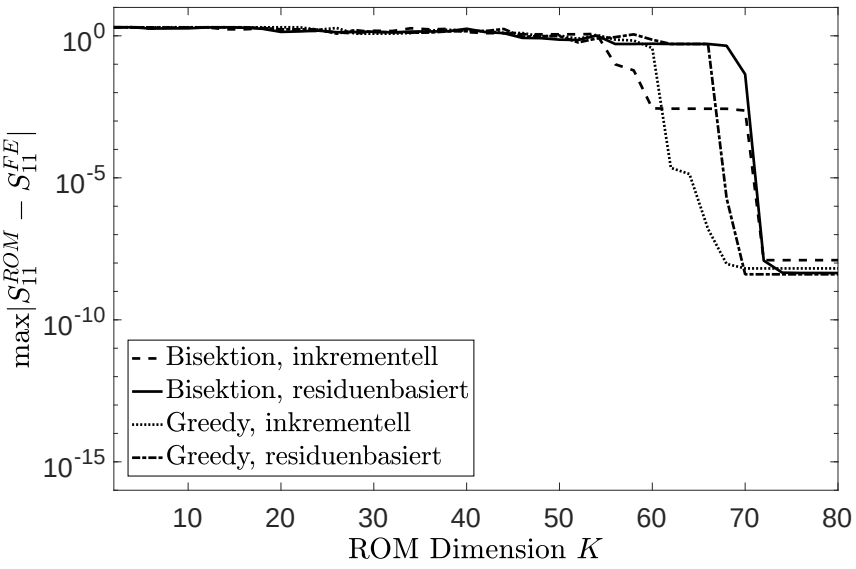


(a) Globaler Indikator vs. ROM-Dimension.

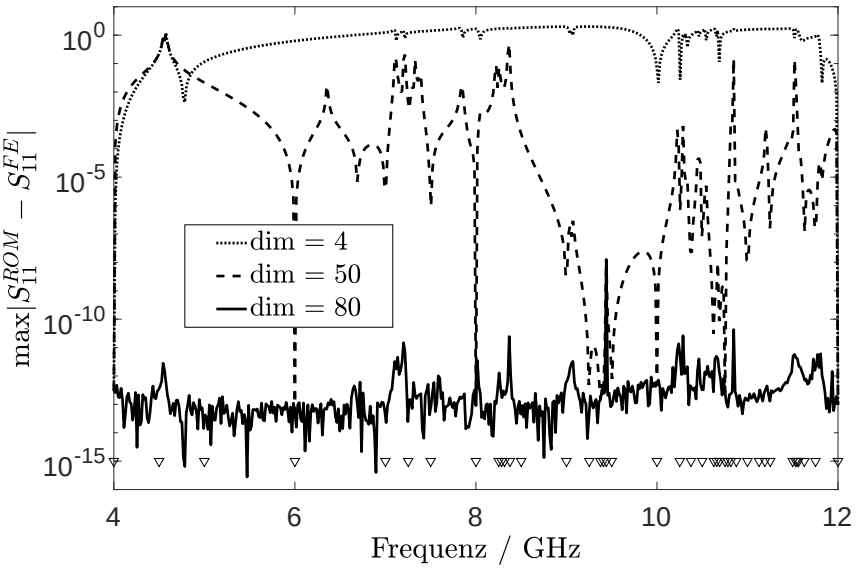


(b) Bisektion: lokaler Indikator vs Frequenz.

Abbildung 5.7: Bandpassfilter: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.



(a) Globaler Fehler vs. ROM-Dimension.



(b) Bisektion: lokaler Fehler vs Frequenz.

Abbildung 5.8: Bandpassfilter: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

derspiegeln. Im weiteren Verlauf des adaptiven Prozesses sammelt das ROM stets Informationen zu zusätzlichen Lösungskomponenten, so dass sukzessive die Resonanzen aufgelöst werden können. Sobald sämtliche Polstellen im ROM darstellbar sind, erfolgt die Konvergenz unmittelbar.

Tatsächlich verkleinert sich der Fehler mit einem Faktor von 10^8 innerhalb nur zweier Iterationen. Dies zeigt, dass die Konvergenz des Verfahrens offensichtlich von der Anzahl und Verteilung der Polstellen im betrachteten Frequenzbereich abhängt.

5.4.3 Wellenleiter mit dielektrischem Einsatz

Als numerisches Beispiel für das Eigenwertproblem (5.11) wird der radial inhomogene Wellenleiter in Abbildung 5.9(a) untersucht, vergleiche hierzu [SA85]. Die Dispersionskurve in Abbildung 5.9(b) zeigt den Verlauf der dominanten Wellenformen.

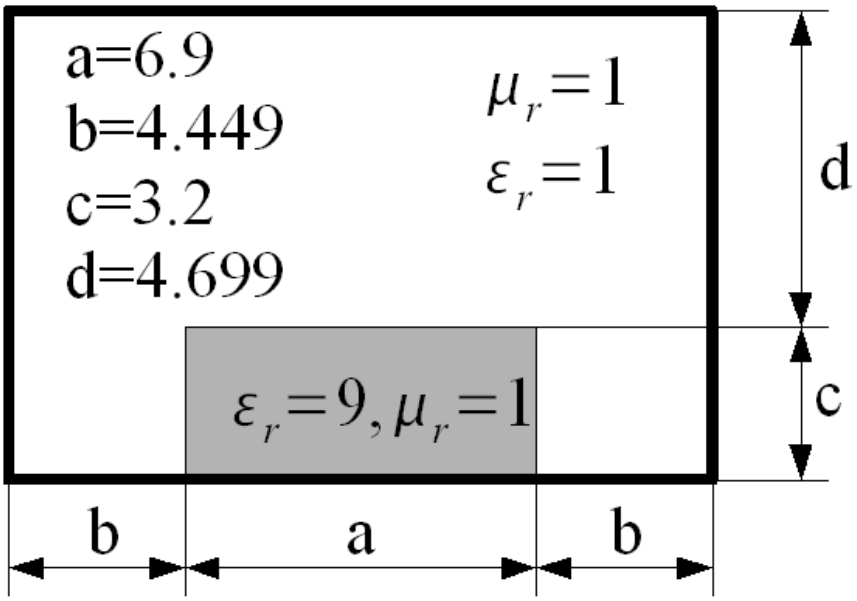
Das Verhalten des globalen Fehlerindikators und des tatsächlichen Fehlers in Abhängigkeit der ROM-Dimension ist in Abbildung 5.10 bzw. Abbildung 5.12 dargestellt. Auch im Wellenleiterfall ist festzustellen, dass der residuenbasierte wie auch der inkrementelle Indikator vergleichbar gute Ergebnisse liefern. Zudem können keine signifikanten Unterschiede zwischen der Greedy- und der Bisektionsmethode festgestellt werden.

Wie bereits im Anregungsfall aufgezeigt, kann aus Abbildung 5.10(a) auch für das Wellenleiterproblem das Verhalten der unterschiedlichen Residuenberechnungen reproduziert werden. Mit der schnellen Residuenberechnung liegt der Pegel des Grundrauschens etwas höher als bei der naiven Residuenberechnung. Allerdings erfolgt auch hier die Stagnation der schnellen Residuenauswertung erst, wenn der tatsächliche Fehler bereits unterhalb typischer FE-Modellierungsfehler liegt. Der tatsächliche Fehler in Abhängigkeit von der ROM-Dimension ist in Abbildung 5.12(a) dargestellt.

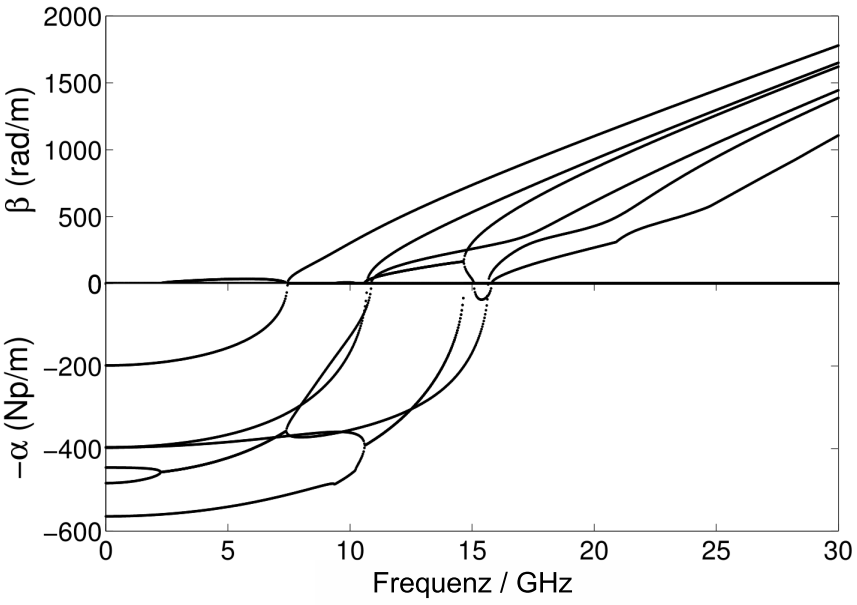
Darüber hinaus bestätigt Abbildung 5.12(a) erneut, dass alle Varianten der adaptiven Strategien exponentielle Konvergenz erzielen, mit

$$\left\|\gamma^{\mathrm{ROM}}(M) - \gamma^{\mathrm{FE}}\right\|_\infty \propto 10^{-0.15M}. \tag{5.37}$$

Dies entspricht auch der Erwartung, da der Verlauf von $\gamma^2(f)$, ausgenommen für die Bifurkationspunkte, glatt ist.

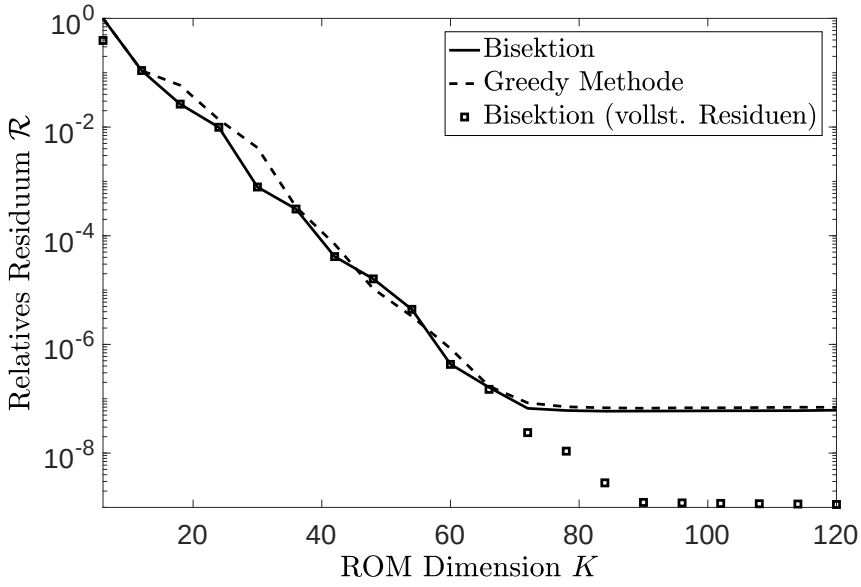


(a) Dimensionen in mm.

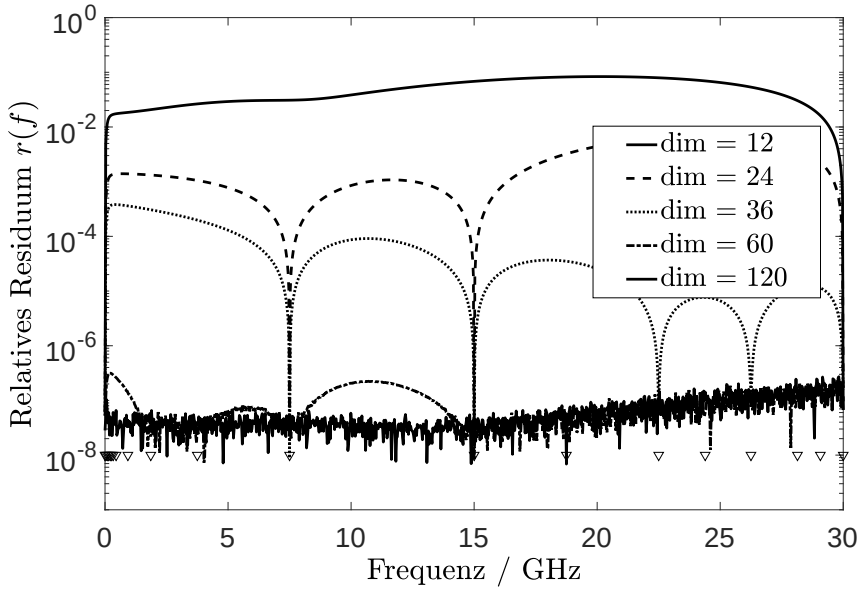


(b) Dispersionskurven.

Abbildung 5.9: Wellenleiter mit dielektrischem Einsatz: Geometrie der Struktur, Maße in mm und Dispersionskurven der dominanten Wellenformen.

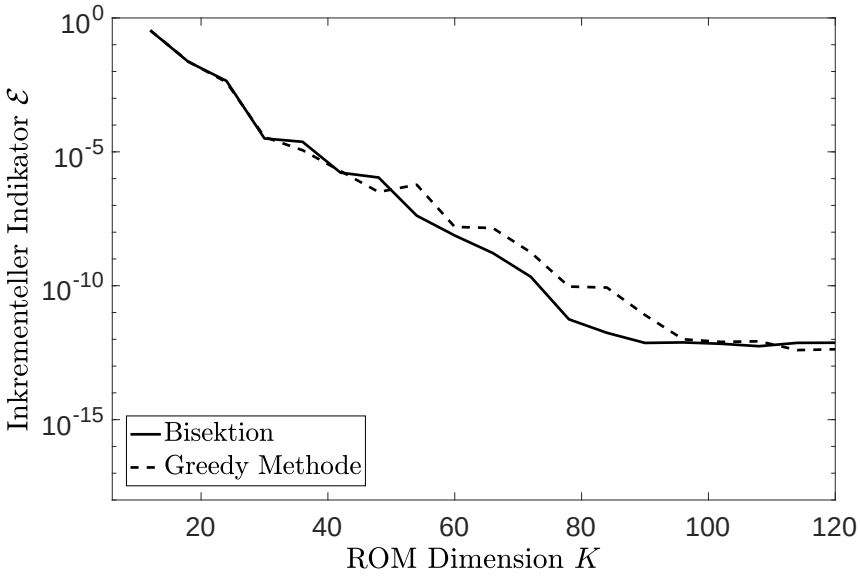


(a) Globaler Indikator vs. ROM-Dimension.

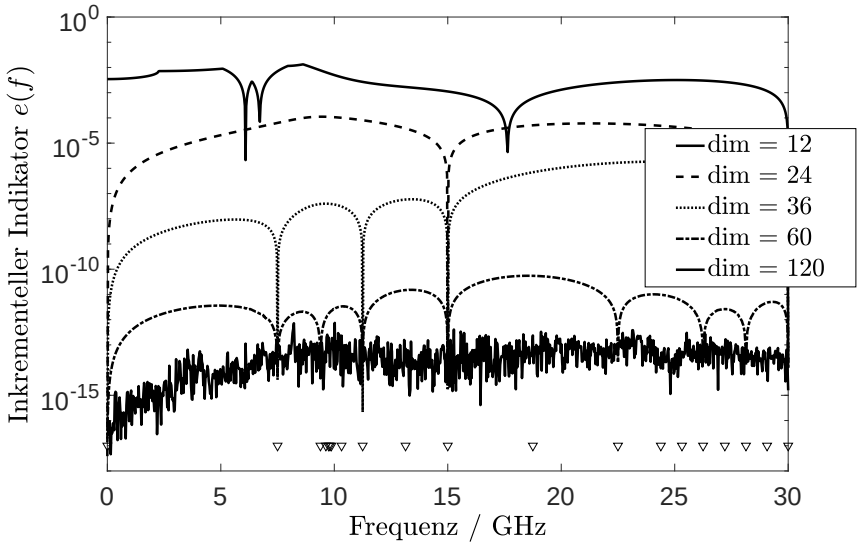


(b) Bisektion: lokaler Indikator vs Frequenz.

Abbildung 5.10: Wellenleiter mit dielektrischem Einsatz: adaptive ROM-Generierung gesteuert von residuenbasiertem Fehlerindikator in der euklidischen Norm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

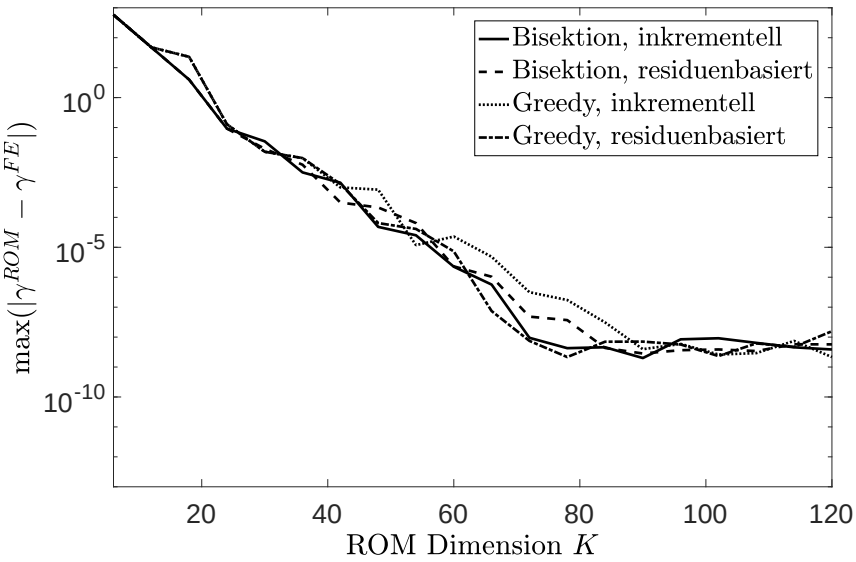


(a) Globaler Indikator vs. ROM-Dimension.

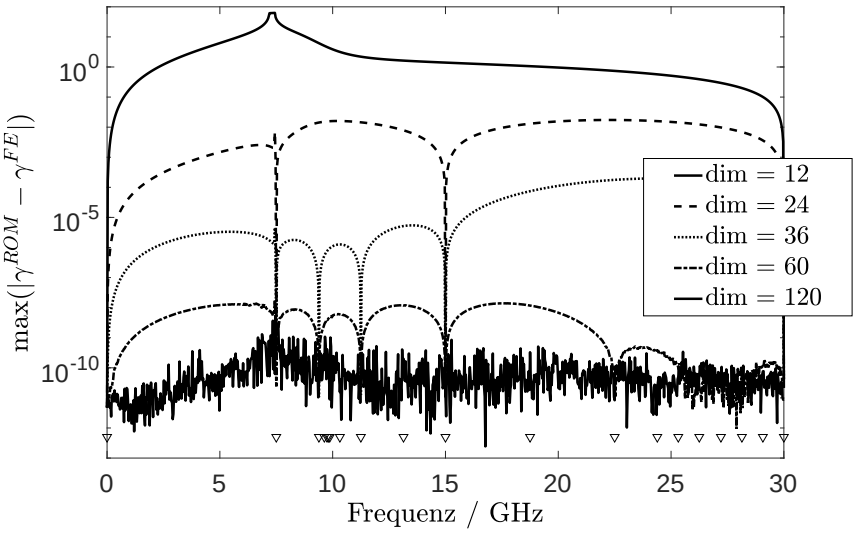


(b) Bisektion: lokaler Indikator vs Frequenz.

Abbildung 5.11: Wellenleiter mit dielektrischem Einsatz: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.



(a) Globaler Fehler vs. ROM-Dimension.



(b) Lokaler Fehler vs. Frequenz.

Abbildung 5.12: Wellenleiter mit dielektrischem Einsatz: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. Dreiecke auf der Abszissenachse markieren die Entwicklungspunkte.

5.5 Fazit

Die Untersuchung der in diesem Kapitel beschriebenen adaptiven Prozesse für die Mehrpunkt-Modellordnungsreduktion im Anregungs- wie auch im Wellenleiterfall lassen die nachfolgend dargestellten Schlussfolgerungen zu.

5.5.1 Konvergenzraten

Für den Fall, dass die Frequenzantwort hinreichend glatt ist, zeigen die numerischen Beispiele in Abschnitt 5.4 exponentielle Konvergenz. Lediglich bei hoch-resonanten Strukturen ist für den Anregungsfall eine präasymptotische Zone in den Konvergenzkurven festzustellen, auf die eine nahezu plötzliche Konvergenz folgt. Eine qualitative Einschätzung zu diesem Verhalten wird in Abschnitt 5.4.2 beschrieben.

5.5.2 Vergleich von Bisektion und Greedy-Methode

Die beiden Ansätze weisen nur geringe Unterschiede in der Effizienz auf. Aus Gründen der numerischen Stabilität ist gemäß den Erläuterungen in Abschnitt 5.3 die Bisektionsmethode zu favorisieren.

5.5.3 Vergleich residuenbasierter und inkrementeller Fehlerindikatoren

Beide Methoden liefern vergleichbare Ergebnisse. Allerdings sind für die inkrementellen Indikatoren zwei Vorteile festzustellen:

- Die inkrementellen Indikatoren basieren direkt auf den gesuchten Ausgangsgrößen.
- Die Berechnung der inkrementellen Indikatoren ist trivial und hat minimale Anforderungen an Speicher und Rechenzeit.

Es sei an dieser Stelle jedoch darauf hingewiesen, dass mit den in Abschnitt 5.2.1 entwickelten Gleichungen (5.27) auch die Berechnung der Residuen sehr effizient und verlässlich gestaltet werden kann.

5.5.4 Abbruchkriterien

Ein Vergleich der Darstellungen in Abschnitt 5.4 zeigt schließlich, dass die inkrementellen Indikatoren den Verlauf des tatsächlichen Fehlers gut wiedergeben. Dennoch sollte in praktischen Anwendungen ein eher konservatives Abbruchkriterium gewählt werden; aufgrund der hohen Konvergenzraten ist der rechnerische Mehraufwand als sehr gering einzuschätzen. Die Erfahrung aus einer Vielzahl an Untersuchungen zeigt, dass für den Anregungsfall ein Wert von $\mathcal{E}_{\min}=10^{-4}$ als Abbruchkriterium in der Regel ausreichend ist.

Aus den gewonnenen Daten ergibt sich für den Autor die Schlussfolgerung, dass für praktische Anwendungen die Bisektionsmethode in Verbindung mit dem inkrementellen Fehlerindikator zu empfehlen ist.

Kapitel 6

A-posteriori-Fehlerschätzer in der Einpunkt-Modellordnungsreduktion

Wie in den vorangegangenen Kapiteln gezeigt, stellt das Rahmenwerk der Modellordnungsreduktion sehr effiziente Werkzeuge zur Verfügung, um eine schnelle und breitbandige Charakterisierung von LTI-Systemen durchführen zu können. Zudem wird in Kapitel 5 die Praktikabilität der adaptiven Methoden im Bereich der elektrodynamischen Feldsimulation gezeigt. Die hierbei eingesetzten globalen Fehlerindikatoren können völlig analog auch als Abbruchkriterium im Zusammenhang mit Einpunktverfahren verwendet werden: Residuen sind mittels (4.89) sehr effizient zu bestimmen und der inkrementelle Fehlerindikator ergibt sich, wie beim Mehrpunktverfahren, auf triviale Weise. Durch die Eigenzerlegung (4.88) des ordnungsreduzierten Modells kann auch bei den Einpunktverfahren die Auswertung zusätzlich beschleunigt werden. Weitere Fehlerindikatoren für Einpunktverfahren werden beispielsweise in [BSSY99] und [RRM09] vorgestellt.

Die zuvor genannten Indikatoren bzw. Abbruchkriterien sind für praktische Anwendungen von großem Nutzen und auch sehr zuverlässig, allerdings basieren sie auf heuristischen Annahmen und können somit keine garantierte Fehlerschranke für die ordnungsreduzierten Modelle liefern. Sowohl residuenbasierte wie auch inkrementelle Indikatoren können zum vorzeitigen Abbruch der MOR-Iteration führen, obwohl im Vergleich zum Originalmodell ein nicht vernachlässigbarer Fehler vorliegt.

Beweisbare Fehlerschranken, die solche Effekte ausschließen, werden in [CHMR10] und [PS10] vorgeschlagen. Die Fehlerschätzer basieren hierbei auf der unteren Schranke für die Inf-Sup-Konstante, welche mittels der *Successive Constraint Method* (SCM) [HRSP07], [CHMR10] effizient bestimmt wird. Die SCM basiert auf der Lösung von Eigenwertproblemen der Dimension des Originalsystems. Daher kann der Rechenaufwand insbesondere bei resonanten Strukturen größer sein, als zur ROM-Generierung an sich. Zudem liegen für Immitanzformulierungen die Pole verlustloser Strukturen

auf der Frequenzachse und die Inf-Sup-Konstante kann beliebig klein werden. Dies limitiert die Praktikabilität dieser Fehlerschätzmethoden zusätzlich.

In weiterer Folge wird ein neuer Fehlerschätzer im Kontext der krylov-unterraumbasierten Modellordnungsreduktion vorgestellt. Es wird gezeigt, dass der Aufwand zur Berechnung der Fehlerschranke wesentlich effizienter durchzuführen ist als bei den zuvor genannten Fehlerschätzverfahren. Insbesondere die Online-Kosten (vgl. Abschnitt 6.3) können durch Ausnutzen der Systemeigenschaften gering gehalten werden. Damit wird die Laufzeit des MOR-Iterationsprozesses nur wenig beeinflusst, und die Effizienz der Modellordnungsreduktion bleibt erhalten. Der vorgestellte Fehlerschätzer ist zwar auf die Anwendung verlustloser Systeme eingeschränkt, allerdings stellt diese Systemklasse einen sehr wichtigen Fall in der Modellierung elektrodynamischer Systeme dar. Im Kontext von Optimierungsaufgaben können beispielsweise Verluste in guten Leitern und Polarisationsverluste vernachlässigt werden. Die nachfolgenden Ausführungen basieren auf den Methoden, wie sie vom Autor in [KFDE10] beziehungweise in einer überarbeiteten Version in [KFDE13] und [KFS+14] veröffentlicht wurden. Ein komplementärer Ansatz im Zeitbereich, der auf einer ähnlichen Berechnung basiert, wird in [Saa92a] gezeigt.

Wie bereits in Abschnitt 4.1.2 erläutert, sind Krylov-Unterraum-Methoden sehr effizient, da die Systemmatrix nur in einem einzigen Entwicklungspunkt zu faktorisieren ist. Alle weiteren Berechnungsschritte erfolgen dann mittels Vor- und Rückeinsetzen, d. h. es wird lediglich die *Wirkung* der inversen Systemmatrix - jedoch nicht die Inverse selbst benötigt. Für den Fehlerschätzprozess wird dann die zentrale Eigenschaft der Krylov-Unterraum-Methoden ausgenutzt: die schnelle Auflösung der Eigenvektoren in der Umgebung des Entwicklungspunktes (vgl. Abschnitt 4.2.5). Aus der spektralen Zerlegung des Residuenvektors in konvergierte und nicht-konvergierte Eigenvektoren kann dann eine obere Schranke für den tatsächlichen Fehler gefunden werden.

Der zusätzliche Aufwand zur Berechnung der Fehlerschranke fällt hierbei verhältnismäßig klein aus: Es ist lediglich die zusätzliche Faktorisierung einer symmetrisch-positiv-definiten Matrix der Dimension des Originalsystems durchzuführen. Alle weiteren Größen können innerhalb der ROM-Domäne bestimmt werden.

Im Zusammenhang mit elektrodynamischen Feldproblemen ist die vorgeschlagene Vorgehensweise in direkter Form nur auf die Immitanzformulierung anwendbar, d. h. die Fehlerschranke bezieht sich auf eine unbeschränkte Größe, die Singularitäten aufweisen kann. Um diese Einschränkung zu umgehen, wird der Fehlerschätzer dahingehend erweitert, dass auch das asymptotische Verhalten des Fehlers in den Streumatrizen aufgezeigt werden kann.

6.1 MOR für verlustlose elektrodynamische Systeme

In vielen praktischen Anwendungen können elektrische Leiter als ideal leitfähig angenommen werden, was deren Modellierung als Dirichlet-Randbedingung ermöglicht (vgl. Abschnitt 2.4.1). Außerdem ist in vielen Fällen zusätzlich das Vernachlässigen dielektrischer Verluste erlaubt, womit $\varepsilon \in \mathbb{R}$ gilt. Weist die Struktur zudem keine Abstrahlung in den Freiraum auf, also $\Gamma_{\mathrm{R}} = \emptyset$, dann liegt ein *verlustloses System* vor und (3.52) vereinfacht sich im SISO-Fall zu

$$\Sigma_{\mathrm{TF}}(jk_0) = \left\{ \begin{array}{rl} (\mathbf{S}_{\mathrm{TF}} + (jk_0)^2\mathbf{T}_{\mathrm{TF}})\,\mathbf{x}_{\mathrm{TF}} & = jk_0\eta_0\mathbf{b}_{\mathrm{TF}}, \\ z & = \mathbf{b}_{\mathrm{TF}}^T\mathbf{x}_{\mathrm{TF}}, \end{array} \right. \tag{6.1}$$

mit

$$\mathbf{S}_{\mathrm{TF}}, \mathbf{T}_{\mathrm{TF}} \in \mathbb{R}^{N\times N}. \tag{6.2}$$

Gemäß (4.117) hat die Skalierung der rechten Seite $\mathbf{b}_{\mathrm{TF}}$ keinen Einfluss auf den Momentenabgleich oder die Konstruktion der Krylov-Vektoren. Auf der linken Seite kann somit der Shift-Parameter

$$\kappa := k_0^2 - \hat{k}_0^2 \tag{6.3}$$

eingeführt werden, woraus sich das verlustlose FE-System um den Entwicklungspunkt $\hat{k}_0$ in der Form

$$\Sigma_1(\kappa) = \left\{ \begin{array}{rl} (\mathbf{A} - \kappa\mathbf{T})\mathbf{x} & = \mathbf{b} \\ z & = jk_0\eta_0\mathbf{b}^T\mathbf{x} \end{array} \right. \tag{6.4}$$

ergibt. Hierin ist

$$\mathbf{A} := \mathbf{S} - \hat{k}_0^2\mathbf{T} \tag{6.5}$$

die Systemmatrix im Entwicklungspunkt $\hat{k}_0$ und die Wellenzahl

$$k_0 = k_0(\kappa) = \sqrt{\kappa + \hat{k}_0} \tag{6.6}$$

ist als eine Funktion von κ zu sehen. Das verlustlose FE-System ist somit ein LTI-System *erster* Ordnung ($L = 1$). Wie in den vorangegangenen Kapiteln wird auch hier aus Gründen der Übersichtlichkeit der Index „$_{\mathrm{TF}}$“ weggelassen und zudem der Faktor $jk_0\eta_0$ in die Ausgangsgröße überführt.

Da (6.1) dieselbe Struktur wie das System (4.40) aufweist, kann zur Konstruktion der Projektionsmatrizen der Arnoldi-Algorithmus (Algorithmus 1) herangezogen werden. Um im Rahmen der Modellordnungsreduktion den Momentenabgleich zu gewährleisten, ist dabei eine Basis für den Krylov-Unterraum

$$\mathcal{K}_K(\Sigma_1) := \mathcal{K}_K(\mathbf{A}^{-1}\mathbf{T}, \mathbf{A}^{-1}\mathbf{b}) \tag{6.7}$$

zu bestimmen. Aufgrund der Symmetrie der rellen Systemmatrizen $\mathbf{A}, \mathbf{T} \in \mathbb{R}^{N\times N}$ und dem, bis auf Skalierung, identischen Anregungs- und Ausgangsvektor $\mathbf{b} \in \mathbb{R}^N$ gilt

$$\mathcal{K}_K(\Sigma_1) = \mathcal{K}_K(\Sigma_1^T), \tag{6.8}$$

so dass mit Projektionsmatrizen

$$\mathbf{P} = \mathbf{Q} \qquad \text{und} \tag{6.9}$$

$$\operatorname{bild} \mathbf{Q} = \mathcal{K}_K(\Sigma_1) \tag{6.10}$$

die ersten $2K-1$ Momente des ROMs mit dem Originalsystem übereinstimmen. Die Komponenten des ordnungsreduzierten Systems

$$\bar{\Sigma}_1(\kappa) = \left\{ \begin{array}{rl} (\bar{\mathbf{A}} - \kappa\bar{\mathbf{T}})\bar{\mathbf{x}} & = \bar{\mathbf{b}} \\ \bar{z} & = jk_0\eta_0\bar{\mathbf{b}}^T\bar{\mathbf{x}} \end{array} \right. \tag{6.11}$$

lauten entsprechend

$$\bar{\mathbf{A}} := \mathbf{Q}^T\mathbf{A}\mathbf{Q}, \tag{6.12}$$

$$\bar{\mathbf{T}} := \mathbf{Q}^T\mathbf{T}\mathbf{Q} \qquad \text{und} \tag{6.13}$$

$$\bar{\mathbf{b}} := \mathbf{Q}^T\mathbf{b}, \tag{6.14}$$

mit $\bar{\mathbf{A}}, \bar{\mathbf{T}} \in \mathbb{R}^{K\times K}$ und $\bar{\mathbf{b}} \in \mathbb{R}^K$.

6.2 Fehlerschätzer für die Impedanzmatrix

Der vorgeschlagene Fehlerschätzer basiert auf der Annahme, dass im betrachteten Frequenzbereich $\mathcal{I}_f$, bzw. im äquivalenten Parameterbereich $\mathcal{I}_\kappa$, die Eigenvektoren und Eigenwerte eines konvergierten ordnungsreduzierten Modells mit denen des Originalmodells übereinstimmen. Diese Annahme lässt sich mittels zweier Argumente rechtfertigen:

1. Das Arnoldi-Verfahren konvergiert im Zusammenhang mit der Shift-Invert-Vorkonditionierung gegen die Eigenwerte in der Umgebung des Shift-Parameters $k_0 = \hat{k}_0$ bzw. $\kappa = 0$, vergleiche hierzu Abschnitt 4.2.5
2. Nicht-aufgelöste Eigenwerte liefern einen wesentlichen Beitrag zum Fehler in der Übertragungsfunktion, da die Eigenwerte auf der Frequenzachse liegen. Damit wird ein vorzeitiges Abbrechen verhindert.

Für die Herleitung des Fehlerschätzers wird der Fehlervektor $\mathbf{e}$, also die Differenz der ROM-Lösung zur Lösung des originalen FE-Systems gemäß

$$\mathbf{e} := \mathbf{Q}\bar{\mathbf{x}} - \mathbf{x} \tag{6.15}$$

betrachtet. Darüber hinaus ist vor allem die Beurteilung des Fehlers in der Ausgangsgröße gemäß Definition (4.121) gesucht, im Fall der Formulierung (6.4) also der Impedanzfehler

$$e_z := \bar{z} - z. \tag{6.16}$$

Die nachfolgend aufgezeigte Vorgehensweise basiert im Wesentlichen auf der Idee, dass die gesuchte Lösung $\mathbf{x}$ in zwei Anteile aufgespalten werden kann: in einen Anteil $\mathbf{x}_\parallel$, der durch das ROM im Rahmen der numerischen Genauigkeit exakt bekannt ist, und einen Anteil $\mathbf{x}_\perp$, dessen Fehler gegen eine obere Schranke abgeschätzt werden kann. Für den Fehlervektor gilt damit

$$\begin{aligned}\mathbf{e} &= \mathbf{Q}(\bar{\mathbf{x}}_\parallel + \bar{\mathbf{x}}_\perp) - (\mathbf{x}_\parallel + \mathbf{x}_\perp)\\ &= \underbrace{\mathbf{Q}\bar{\mathbf{x}}_\parallel - \mathbf{x}_\parallel}_{=\mathbf{0}} + \mathbf{Q}\bar{\mathbf{x}}_\perp - \mathbf{x}_\perp\\ &= \mathbf{Q}\bar{\mathbf{x}}_\perp - \mathbf{x}_\perp.\end{aligned} \tag{6.17}$$

Entsprechend lässt sich auch der Impedanzfehler zerlegen:

$$\begin{aligned}e_z &= jk_0\eta_0(\bar{\mathbf{b}}^T\bar{\mathbf{x}} - \mathbf{b}^T\mathbf{x})\\ &= jk_0\eta_0(\bar{\mathbf{b}}^T(\bar{\mathbf{x}}_\parallel + \bar{\mathbf{x}}_\perp) - \mathbf{b}^T(\mathbf{x}_\parallel + \mathbf{x}_\perp))\\ &= jk_0\eta_0(\underbrace{\bar{\mathbf{b}}^T\bar{\mathbf{x}}_\parallel - \mathbf{b}^T\mathbf{x}_\parallel}_{=0}) + jk_0\eta_0(\bar{\mathbf{b}}^T\bar{\mathbf{x}}_\perp - \mathbf{b}^T\mathbf{x}_\perp)\\ &= jk_0\eta_0(\bar{\mathbf{b}}^T\bar{\mathbf{x}}_\perp - \mathbf{b}^T\mathbf{x}_\perp).\end{aligned} \tag{6.18}$$

Die Aufteilung der Lösung in der gezeigten Form wird über eine Eigenzerlegung der Systemmatrizen erreicht. Hierzu wird zunächst das verallgemeinerte Eigenwertproblem

$$\mathbf{A}\mathbf{v} = \kappa\mathbf{T}\mathbf{v} \tag{6.19}$$

zum FE-System (6.1) herangezogen. Mit Hilfe der Eigenvektormatrix

$$\mathbf{V} = \begin{bmatrix}\mathbf{v}_1 & \dots & \mathbf{v}_N\end{bmatrix} \tag{6.20}$$

lassen sich die Systemmatrizen $\mathbf{A}$ und $\mathbf{T}$ gemäß der Vorschrift

$$\mathbf{V}^T\mathbf{A}\mathbf{V} = \operatorname{diag}(\kappa_1, \kappa_2, \dots, \kappa_N), \tag{6.21}$$

$$\mathbf{V}^T\mathbf{T}\mathbf{V} = \mathbf{I} \tag{6.22}$$

diagonalisieren. Dies führt auf die Darstellung der inversen Systemmatrix in der Form

$$(\mathbf{A} - \kappa\mathbf{T})^{-1} = \mathbf{V}\operatorname{diag}\left(\frac{1}{\kappa_i - \kappa}\right)\mathbf{V}^T. \tag{6.23}$$

Im nächsten Schritt wird mit dem Residuum

$$\mathbf{r}(\kappa) := (\mathbf{A} - \kappa\mathbf{T})\mathbf{Q}\bar{\mathbf{x}} - \mathbf{b} \tag{6.24}$$

die Fehlergleichung

$$(\mathbf{A} - \kappa\mathbf{T})\mathbf{e} = \mathbf{r} \tag{6.25}$$

aufgestellt, wobei aus Gründen der Übersichtlichkeit das Argument κ weggelassen wird. Mit (6.23) und der Galerkin-Bedingung $\mathbf{Q}^T\mathbf{r} = \mathbf{0}$ gilt somit für den Impedanzfehler

$$\begin{aligned} e_z &= jk_0\eta_0\mathbf{b}^T(\mathbf{A} - \kappa\mathbf{T})^{-1}\mathbf{r} \\ &= jk_0\eta_0(\mathbf{Q}\bar{\mathbf{x}} - \mathbf{e})^T(\mathbf{A} - \kappa\mathbf{T})(\mathbf{A} - \kappa\mathbf{T})^{-1}\mathbf{r} \\ &= jk_0\eta_0\left[\bar{\mathbf{x}}^T\mathbf{Q}^T(\mathbf{A} - \kappa\mathbf{T})(\mathbf{A} - \kappa\mathbf{T})^{-1}\mathbf{r} - \mathbf{e}^T(\mathbf{A} - \kappa\mathbf{T})(\mathbf{A} - \kappa\mathbf{T})^{-1}\mathbf{r}\right] \\ &= jk_0\eta_0\big[\bar{\mathbf{x}}^T \underbrace{\mathbf{Q}^T\mathbf{r}}_{=0} - \mathbf{r}^T(\mathbf{A} - \kappa\mathbf{T})^{-1}\mathbf{r}\big] \\ &= -jk_0\eta_0\mathbf{r}^T\mathbf{V}\,\mathrm{diag}\left(\frac{1}{\kappa_i - \kappa}\right)\mathbf{V}^T\mathbf{r}. \end{aligned} \tag{6.26}$$

Anhand dieser Darstellung kann die Aufspaltung des Fehlers in Anlehnung an (6.18) erfolgen. Seien hierzu $\{\mathbf{v}_1, \ldots, \mathbf{v}_B\}$ Eigenvektoren und $\kappa_1, \ldots, \kappa_B$ die zugehörigen Eigenwerte zum Problem (6.19) mit der Eigenschaft $\kappa_i \in \mathcal{I}_\kappa, i = 1, \ldots, B$. Dann kann unter der Voraussetzung, dass Linearkombinationen

$$\tilde{\mathbf{v}}_i = \mathbf{Q}\bar{\mathbf{v}}_i \tag{6.27}$$

mit

$$\tilde{\mathbf{v}}_i \approx \mathbf{v}_i \quad \text{für } i = 1, \ldots, B \tag{6.28}$$

existieren, die Beziehung

$$\mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_B) \subset \mathrm{bild}\,\mathbf{Q} \tag{6.29}$$

angenommen werden. Dies bedeutet daher auch, dass das Residuum $\mathbf{r}$ orthogonal auf $\mathrm{span}(\mathbf{v}_1, \ldots, \mathbf{v}_B)$ steht, und es gilt:

$$\begin{aligned} \mathbf{V}^T\mathbf{r} &= \left[\mathbf{v}_1 \cdots \mathbf{v}_B\ \mathbf{v}_{B+1} \cdots \mathbf{v}_N\right]^T\mathbf{r} \\ &= \left[\mathbf{V}_\parallel\ \mathbf{V}_\perp\right]^T\mathbf{r} \\ &= \mathbf{V}_\perp^T\mathbf{r}. \end{aligned} \tag{6.30}$$

An dieser Stelle wird die Eigenschaft der Krylov-Unterraumverfahren ausgenutzt, dass – wie in Abschnitt 4.2.5 beschrieben – im ROM zunächst die betragskleinsten Eigenwerte konvergieren. Durch Anwendung der Shift-Invert-Vorkonditionierung heißt das im konkreten Fall des Systems (6.1), dass mit wachsender ROM-Dimension K die Eigenwerte in der Umgebung von $\kappa = 0$ mit zunehmender Genauigkeit bestimmt werden können. Sei $\{(\mathbf{v}_i, \kappa_i) | i = 1, \ldots, B\}$ die Menge aller konvergierten Eigenpaare für ein gegebenes ROM mit der Eigenschaft

$$0 < |\kappa_1| \leq |\kappa_2| \leq \cdots \leq |\kappa_B| \leq |\kappa_{B+1}| \leq \cdots \leq |\kappa_N|\,. \tag{6.31}$$

Dann gilt wegen (6.30) für den Impedanzfehler (6.26) die Abschätzung

$$\begin{aligned} |e_z| &\leq k_0\eta_0 \mathbf{r}^T \mathbf{V}_\perp \operatorname{diag}\left(\frac{1}{|\kappa_{B+1}-\kappa|},\ldots,\frac{1}{|\kappa_N-\kappa|}\right)\mathbf{V}_\perp^T\mathbf{r} \\ &\leq \frac{k_0\eta_0}{\min\limits_{\kappa_i\notin\mathcal{U}_B}|\kappa_i-\kappa|}\mathbf{r}^T\mathbf{V}_\perp\mathbf{V}_\perp^T\mathbf{r}, \end{aligned} \tag{6.32}$$

wobei

$$\mathcal{U}_B := \{\kappa\in\mathcal{I}_\kappa \big|\, |\kappa| \leq |\kappa_B|\} \tag{6.33}$$

die Umgebung von $\kappa = 0$ beschreibt, in der sämtliche Eigenwerte konvergiert sind. Die Konvergenz der Eigenwerte wird in Abschnitt 6.2.1 näher erläutert. Bei der Berechnung einer Schranke für $|e_z|$ wird aus Gründen der numerischen Effizienz die Aufspaltung $\mathbf{V} = \begin{bmatrix}\mathbf{V}_\parallel & \mathbf{V}_\perp\end{bmatrix}$ nicht explizit durchgeführt. Stattdessen wird die Beziehung $\mathbf{T}^{-1} = \mathbf{V}\mathbf{V}^T$ ausgenutzt, die aus der $\mathbf{T}$-Normierung (6.22) resultiert. Wegen

$$\mathbf{r}^T\mathbf{V}_\perp\mathbf{V}_\perp^T\mathbf{r} = \mathbf{r}^T\mathbf{V}\mathbf{V}^T\mathbf{r} = \mathbf{r}^T\mathbf{T}^{-1}\mathbf{r} \tag{6.34}$$

lässt sich (6.32) somit schreiben als

$$|e_z(\kappa)| \leq \frac{k_0\eta_0}{\min\limits_{\kappa_i\notin\mathcal{U}_B}|\kappa_i-\kappa|}\mathbf{r}^T(\kappa)\mathbf{T}^{-1}\mathbf{r}(\kappa). \tag{6.35}$$

Mit (6.35) ist schließlich eine obere Fehlerschranke für die Ausgangsgröße $\bar{z}(\kappa)$ für alle $\kappa\in\mathcal{U}_B$ gegeben.

6.2.1 Bewertung der ROM-Eigenwerte

In der Zerlegung (6.30) wird die Annahme getroffen, dass die Eigenpaare $(\mathbf{v}_i,\kappa_i), i = 1,\ldots,B$ bekannt sind. Bei der numerischen Auswertung im MOR-Verfahren muss demnach ein Kriterium herangezogen werden, das die Konvergenz der Eigenwerte und -vektoren prüft. Ein solches Kriterium kann aus dem folgende Satz abgeleitet werden:

Satz 6.2.1. *Seien $\mathbf{x}\in\mathbb{C}^N\neq\mathbf{0}$ ein beliebiger Vektor und $\tilde{\kappa}\in\mathbb{C}$ eine beliebige Zahl, $\mathbf{A}\in\mathbb{C}^{N\times N}$ eine hermitesche Matrix und $\mathbf{T}\in\mathbb{C}^{N\times N}$ eine hermitesche positiv definite Matrix. Dann existiert für das verallgemeinerte Eigenwertproblem*

$$\mathbf{A}\mathbf{x} = \kappa\mathbf{T}\mathbf{x} \tag{6.36}$$

stets ein Eigenwert κ mit

$$|\kappa-\tilde{\kappa}| \leq \frac{\|\mathbf{A}\mathbf{x}-\tilde{\kappa}\mathbf{T}\mathbf{x}\|_{\mathbf{T}^{-1}}}{\|\mathbf{x}\|_{\mathbf{T}}}. \tag{6.37}$$

Beweis. Der Fall $\tilde{\kappa} = \kappa$ ist offensichtlich. Andernfalls folgt mit der Spektralzerlegung

$$\mathbf{V}^H \mathbf{A} \mathbf{V} = \operatorname{diag}(\kappa_i) =: \mathbf{\Lambda}, \tag{6.38}$$

$$\mathbf{V}^H \mathbf{T} \mathbf{V} = \mathbf{I} \tag{6.39}$$

die Beziehung

$$\begin{aligned} \|(\mathbf{A} - \tilde{\kappa}\mathbf{T})\mathbf{x}\|_{\mathbf{T}^{-1}}^2 &= \left\|\mathbf{V}^{-H}\mathbf{V}^H(\mathbf{A} - \tilde{\kappa}\mathbf{T})\mathbf{V}\mathbf{V}^{-1}\mathbf{x}\right\|_{\mathbf{T}^{-1}}^2 \\ &= \mathbf{x}^H\mathbf{V}^{-H} \operatorname{diag}(\kappa_i - \tilde{\kappa})^H \left(\mathbf{V}^H\mathbf{T}\mathbf{V}\right)^{-1} \operatorname{diag}(\kappa_i - \tilde{\kappa})\, \mathbf{V}^{-1}\mathbf{x} \\ &= \mathbf{x}^H\mathbf{V}^{-H} \operatorname{diag}(\kappa_i - \tilde{\kappa})^H \operatorname{diag}(\kappa_i - \tilde{\kappa})\, \mathbf{V}^{-1}\mathbf{x} \\ &\geq \min_i |\kappa_i - \tilde{\kappa}|^2\, \mathbf{x}^T\mathbf{V}^{-H}\mathbf{V}^{-1}\mathbf{x}, \end{aligned} \tag{6.40}$$

und mit $\mathbf{V}^{-H}\mathbf{V}^{-1} = \mathbf{T}^{-1}$ die Behauptung (6.37). □

Sei $(\tilde{\kappa}_i, \bar{\mathbf{v}}_i)$ das Eigenpaar zum ROM-System (6.11), welches aufgrund der niedrigen Systemdimension K beispielsweise mit dem QZ-Verfahren effizient bestimmt werden kann [GL96, S. 374ff]. Dann gilt für den Eigenwertfehler im Originalsystem

$$\epsilon_{\kappa,i} := |\tilde{\kappa}_i - \kappa_i| \tag{6.41}$$

die Abschätzung

$$\epsilon_{\kappa,i} \leq \frac{\|(\mathbf{A} - \tilde{\kappa}_i\mathbf{T})\mathbf{Q}\bar{\mathbf{v}}_i\|_{\mathbf{T}^{-1}}}{\|\mathbf{Q}\bar{\mathbf{v}}_i\|_{\mathbf{T}}} = \left[\boldsymbol{\rho}_i^T\mathbf{T}^{-1}\boldsymbol{\rho}_i\right]^{1/2}. \tag{6.42}$$

Hierin ist

$$\boldsymbol{\rho}_i = (\mathbf{A} - \tilde{\kappa}_i\mathbf{T})\mathbf{Q}\bar{\mathbf{v}}_i. \tag{6.43}$$

das Residuum zum Eigenwertproblem (6.19). In den numerischen Beispielen werden Eigenwerte als konvergiert betrachtet, wenn der relative Fehler durch

$$|\bar{\kappa}_i - \kappa_i|/\bar{\kappa}_i \leq 10^{-3} \tag{6.44}$$

abgeschätzt werden kann. Wie im Falle des Impedanzfehlerschätzers (6.35) muss auch hier das Residuum nicht über die Komponenten der Originaldomäne bestimmt werden. Der folgende Abschnitt zeigt, wie die effiziente Auswertung unter Verwendung der ROM-Komponenten erfolgt.

6.3 Numerischer Aufwand bei der Fehlerschätzung

Hinsichtlich des numerischen Aufwands kann der Prozess im Rahmen der Modellordnungsreduktion in zwei grundlegende Kostenfaktoren unterteilt werden:

1. *Offline*-Kosten umfassen jene Berechnungen, die zur Generierung des ROMs notwendig sind. Die Offline-Kosten sind unabhängig von der Anzahl N_f der Evaluationspunkte und beziehen sich auf Operationen in der Domäne des hochdimensionalen Originalsystems.

2. *Online*-Kosten umfassen die Berechnungen innerhalb der ROM-Domäne, das heißt insbesondere die Auswertung des ROMs an den N_f Evaluationspunkten.

Diese Aufteilung ist insbesondere im Hinblick auf hochdimensionale Systeme von Bedeutung, da beispielsweise für Optimierungsprobleme die Offline-Schritte auf Rechen-Clustern ausgeführt werden können während die Auswertung des Online-Anteils auf Arbeitsplatzrechnern erfolgt. In weiterer Folge wird gezeigt, dass der numerische Aufwand zur Auswertung des Fehlerschätzers (6.35) ebenfalls in Offline- und Online-Kosten aufteilbar ist. Außerdem wird der zusätzliche Aufwand im Vergleich zum reinen MOR-Prozess diskutiert.

Die Offline-Kosten im MOR-Prozess werden insbesondere von der einmaligen Faktorisierung der symmetrischen Systemmatrix $\mathbf{A}$ dominiert. Wie aus (6.35) abzulesen ist, wird für den Fehlerschätzer zusätzlich die Wirkung von $\mathbf{T}^{-1}$ benötigt. Da $\mathbf{T}$ symmetrisch positiv definit ist, kann die Faktorisierung mittels der Cholesky-Zerlegung erfolgen. Die Kosten für die Faktorisierung von $\mathbf{T}$ sind daher etwa halb so hoch wie für die Faktorisierung von $\mathbf{A}$ [TB97, Lecture 23], und deshalb, gemessen am Gesamtaufwand für den MOR-Prozess, akzeptabel.

Bei den Online-Kosten des Fehlerschätzers ist insbesondere die Auswertung des frequenzabhängigen Terms $\mathbf{r}(\kappa)^T\mathbf{T}^{-1}\mathbf{r}(\kappa) = \|\mathbf{r}(\kappa)\|^2_{\mathbf{T}^{-1}}$ zu untersuchen. Gemäß der Vorschrift (4.83) kann das Residuum $\mathbf{r}$ direkt aus Komponenten der ROM-Domäne bestimmt werden, die ohnehin im Arnoldi-Algorithmus anfallen. Für das System (6.4) gilt somit

$$\begin{aligned}\mathbf{r}(\kappa) &= \kappa h_{n+1,n}\mathbf{A}\left[\mathbf{I} - \mathbf{Q}_n\bar{\mathbf{A}}^{-1}\mathbf{Q}_n^T\mathbf{A}\right]\hat{\mathbf{q}}_{n+1}(\hat{\mathbf{e}}_n^T\bar{\mathbf{x}}(\kappa)) \\ &= \kappa\mathbf{r}_0(\hat{\mathbf{e}}_n^T\bar{\mathbf{x}}(\kappa)) \\ &= \kappa\mathbf{r}_0\bar{x}_n(\kappa)\end{aligned} \tag{6.45}$$

mit einem frequenzunabhängigen Term

$$\mathbf{r}_0 = h_{n+1,n}\mathbf{A}\left[\mathbf{I} - \mathbf{Q}_n\bar{\mathbf{A}}^{-1}\mathbf{Q}_n^T\mathbf{A}\right]\hat{\mathbf{q}}_{n+1}. \tag{6.46}$$

Die Auswertung des Residuums für alle Stützstellen $\kappa \in \mathcal{I}_\kappa^h$ besteht demnach lediglich aus eine Skalierung von $\mathbf{r}_0$ mit dem Parameter κ und der n-ten Komponente $\bar{x}_n(\kappa)$ des ROM-Lösungsvektors.

Die Abschätzung des Eigenwertfehlers (6.42) erfolgt analog: Das Residuum (6.43) lässt sich schreiben als

$$\boldsymbol{\rho}_i = \tilde{\kappa}_i\mathbf{r}_0\bar{v}_{i,n}, \tag{6.47}$$

wobei $\mathbf{r}_0$ identisch zu (6.46) ist und $\bar{v}_{i,n}$ die n-te Komponente des ROM-Eigenvektors $\bar{\mathbf{v}}_i$ zum ROM-Eigenwert $\tilde{\kappa}_i$ beschreibt.

Damit gilt für die Fehlerschätzer folgende Berechnungsvorschrift:

- Berechne einmalig in jeder ROM-Iteration

$$\|\mathbf{r}_0\|_{\mathbf{T}^{-1}} = (\mathbf{r}_0^T \mathbf{T}^{-1} \mathbf{r}_0)^{1/2}. \tag{6.48}$$

- Bewerte die Eigenwerte $\tilde{\kappa}_i \in \mathcal{U}_B$ mit

$$\epsilon_{\kappa,i} \leq |\tilde{\kappa}_i \bar{v}_{i,n}| \, \|\mathbf{r}_0\|_{\mathbf{T}^{-1}} \, . \tag{6.49}$$

- Bewerte den Impedanzfehler für alle Stützstellen $\kappa \in \mathcal{I}_\kappa^h$ mit

$$|e_z(\kappa)| \leq \frac{k_0 \eta_0}{\min\limits_{\kappa_i \notin \mathcal{U}_B} |\kappa_i - \kappa|} \kappa^2 \bar{x}_n^2(\kappa) \, \|\mathbf{r}_0\|_{\mathbf{T}^{-1}}^2 \, . \tag{6.50}$$

Aus dieser Berechnungsvorschrift ist zu erkennen, dass einmalig die quadratische Form gemäß (6.48) zu bestimmen ist. Die zusätzlichen Online-Kosten der Residuenauswertung reduzieren sich daher auf Multiplikationen von Skalaren.

Im hier vorgeschlagenen Ablauf ist noch eine weitere Beschleunigung der Berechnungen zu erreichen. Wie in Abschnitt 6.2.1 beschrieben, wird für das ROM eine vollständige Eigenzerlegung mit dem QZ-Verfahren in der Form

$$\bar{\mathbf{V}}^T \bar{\mathbf{A}} \bar{\mathbf{V}} = \operatorname{diag}(\tilde{\kappa}_i) \, , \tag{6.51}$$

$$\bar{\mathbf{V}}^T \bar{\mathbf{T}} \bar{\mathbf{V}} = \mathbf{I} \tag{6.52}$$

bestimmt. Daher muss das ROM-Gleichungssystem nicht explizit durch Invertierung der Systemmatrix $(\bar{\mathbf{A}} - \kappa \bar{\mathbf{T}})$ für alle Stützstellen gelöst werden. Stattdessen erlaubt der Basiswechsel

$$\bar{\mathbf{x}} = \bar{\mathbf{V}} \bar{\mathbf{w}} \tag{6.53}$$

die schnelle Auswertung der Beziehungen

$$\bar{\mathbf{w}} = \operatorname{diag}\left(\frac{1}{\tilde{\kappa}_i - \kappa}\right) \bar{\mathbf{V}}^T \bar{\mathbf{b}}, \tag{6.54}$$

$$z = j k_0 \eta_0 (\bar{\mathbf{V}}^T \bar{\mathbf{b}})^T \bar{\mathbf{w}}, \tag{6.55}$$

$$\mathbf{r}_0 = h_{n+1,n} \mathbf{A} \big[\mathbf{I} - \mathbf{Q}_n \bar{\mathbf{V}} \operatorname{diag}\left(\frac{1}{\tilde{\kappa}_i}\right) \bar{\mathbf{V}}^T \mathbf{Q}_n^T \mathbf{A} \big] \hat{\mathbf{q}}_{n+1}. \tag{6.56}$$

Zusammenfassend lässt sich feststellen, dass der zusätzliche Aufwand für den gesamten Fehlerschätzungsprozess, sowohl bei den Offline- wie auch den Online-Kosten, jeweils deutlich unter denen des reinen MOR-Prozesses liegen.

6.4 Fehlerschätzer für MIMO-Systeme

Die Fehlerschranke (6.32) ist ohne weiteres auch auf MIMO-Systeme anwendbar: Der Vektor $\mathbf{b}_{\mathrm{TF}} \in \mathbb{R}^N$ in (6.1) wird, wie in (3.54), zu einer Matrix $\mathbf{B} \in \mathbb{R}^{N\times M}$, worin M die Anzahl modaler Tore definiert. Der Residuenvektor $\mathbf{r}$ wird entsprechend zu einer Residuenmatrix $\mathbf{R} \in \mathbb{R}^{N\times M}$. Damit nimmt (6.26) die Form

$$\mathbf{E}_Z = -jk_0\eta_0\mathbf{R}^T\mathbf{V}\mathbf{D}\mathbf{V}^T\mathbf{R} \tag{6.57}$$

an. Hierin ist

$$\mathbf{E}_Z = \bar{\mathbf{Z}} - \mathbf{Z} \tag{6.58}$$

die Impedanzfehlermatrix und

$$\mathbf{D} = \mathrm{diag}\,(d_1, \ldots, d_M) \tag{6.59}$$

mit

$$d_i = \begin{cases} \left(\min\limits_{\kappa_i\notin\mathcal{U}_B} |\kappa_i - \kappa|\right)^{-1}, & \text{wenn } \kappa_i \notin \mathcal{U}_B, \\ 0, & \text{wenn } \kappa_i \in \mathcal{U}_B \end{cases} \tag{6.60}$$

eine Diagonalmatrix mit Abschätzungen für die Eigenwerte außerhalb der konvergenten Umgebung $\mathcal{U}_B$. Der Ausdruck (6.57) führt somit zur Fehlerschranke

$$\begin{aligned} |e_{z,mn}| &= k_0\eta_0\|\mathbf{V}^T\mathbf{R}\mathbf{e}_m\|_2\|\mathbf{D}\|_2\|\mathbf{V}^T\mathbf{R}\mathbf{e}_n\|_2 \\ &\leq \frac{k_0\eta_0}{\min\limits_{\kappa_i\notin\mathcal{U}_B}|\kappa_i - \kappa|}\|\mathbf{R}\mathbf{e}_m\|_{\mathbf{T}^{-1}}\|\mathbf{R}\mathbf{e}_n\|_{\mathbf{T}^{-1}}. \end{aligned} \tag{6.61}$$

Auch im MIMO-Fall kann $\mathbf{R}$ geschrieben werden als ein konstanter Faktor $\mathbf{R}_0$, dessen Spaltenvektoren mit dem Parameter κ sowie Komponenten der ROM-Lösungsvektoren skaliert werden. Die Berechnung des Terms $\|\cdot\|_{\mathbf{T}^{-1}}$ benötigt also auch hier nur $\mathcal{O}(K)$ Operationen, wobei K die Dimension des ordnungsreduzierten Modells beschreibt.

6.5 Asymptotischer Fehlerschätzer für Streuparameter

In der Hochfrequenz- bzw. Mikrowellentechnik werden zur Systemcharakterisierung meist Streumatrizen herangezogen. Für Problemstellungen in der Elektrodynamik kann die FE-Formulierung auch so gewählt werden, dass als Resultat unmittelbar

die Streumatrix folgt [CL88]. Im Gegensatz zur Immitanzformulierung (3.64) muss dann nicht der Umweg über die Impedanz- bzw. Admittanzmatrix gegangen werden. Allerdings führt die Streuformulierung auf frequenzabhängige Systemmatrizen, wenn an den Tor-Randbedingungen Γ_{WG} Nicht-TEM-Wellenformen zu berücksichtigen sind. Auch bei verlustlosen Systemen resultiert die Formulierung in einer echtquadratischen Frequenzabhängigkeit. Damit müssen für diese Formulierung alternative MOR-Ansätze gewählt werden, beispielsweise die in Abschnitt 4.3 erwähnten SOAR- oder WCAWE-Verfahren. Eine detaillierte Diskussion der Streuparameterformulierung ist außerdem in [FLDE10] gegeben.

Die unmittelbare Anwendung des in Abschnitt 6.2 vorgestellten Fehlerschätzers ist für Systeme, wie sie aus der Streuformulierung resultieren, nicht möglich. Um mittels des hier präsentierten Fehlerschätzers dennoch Aussagen bezüglich der Genauigkeit von S-Parametern treffen zu können, wird ein Verfahren zur asymptotischen Fehlerschätzung der Streumatrix vorgeschlagen.

Die Streumatrizen des Originalmodells und des ordnungsreduzierten Modells, $\mathbf{S}$ und $\bar{\mathbf{S}}$, ergeben sich aus der Impedanzmatrix gemäß der Vorschrift

$$\mathbf{S} = (\mathbf{Z}+\mathbf{I})^{-1}(\mathbf{Z}-\mathbf{I}) = \mathbf{I} - 2(\mathbf{Z}+\mathbf{I})^{-1}, \tag{6.62a}$$
$$\bar{\mathbf{S}} = (\bar{\mathbf{Z}}+\mathbf{I})^{-1}(\bar{\mathbf{Z}}-\mathbf{I}) = \mathbf{I} - 2(\bar{\mathbf{Z}}+\mathbf{I})^{-1}. \tag{6.62b}$$

Der zugehörige Fehler $\mathbf{E}_S = \bar{\mathbf{S}} - \mathbf{S}$ ist entsprechend durch

$$\begin{aligned}\mathbf{E}_S &= -2(\bar{\mathbf{Z}}+\mathbf{I})^{-1} + 2(\mathbf{Z}+\mathbf{I})^{-1} \\ &= -2\big[(\bar{\mathbf{Z}}+\mathbf{I})^{-1} - (\bar{\mathbf{Z}}-\mathbf{E}_Z+\mathbf{I})^{-1}\big]\end{aligned} \tag{6.63}$$

gegeben. Unter Einbeziehung von $\bar{\mathbf{Z}}+\mathbf{I}-\mathbf{E}_Z = \big[\mathbf{I}-\mathbf{E}_Z(\bar{\mathbf{Z}}+\mathbf{I})^{-1}\big](\bar{\mathbf{Z}}+\mathbf{I})$ und der Entwicklung von $\big[\mathbf{I}-\mathbf{E}_Z(\bar{\mathbf{Z}}+\mathbf{I})^{-1}\big]^{-1}$ in einer Neumann-Reihe folgt daraus

$$(\bar{\mathbf{Z}}-\mathbf{E}_Z+\mathbf{I})^{-1} = (\bar{\mathbf{Z}}+\mathbf{I})^{-1}\big[\mathbf{I}+\mathbf{E}_Z(\bar{\mathbf{Z}}+\mathbf{I})^{-1}+\mathcal{O}(\epsilon_z^2)\big] \tag{6.64}$$

mit

$$\epsilon_z = \max_{m,n} |z_{mn} - \bar{z}_{mn}|\,. \tag{6.65}$$

Somit ist durch

$$\mathbf{E}_S \to 2(\bar{\mathbf{Z}}+\mathbf{I})^{-1}\mathbf{E}_Z(\bar{\mathbf{Z}}+\mathbf{I})^{-1} \tag{6.66}$$

eine asymptotische Approximation für (6.63) gegeben. Substituieren von $\mathbf{E}_Z$ durch (6.57) führt auf

$$\mathbf{E}_S \to -2jk_0\eta_0\big[\mathbf{V}^T\mathbf{R}(\bar{\mathbf{Z}}+\mathbf{I})^{-1}\big]^T\mathbf{D}\big[\mathbf{V}^T\mathbf{R}(\bar{\mathbf{Z}}+\mathbf{I})^{-1}\big]. \tag{6.67}$$

Mit der Abkürzung

$$\mathbf{p}_i = \mathbf{R}(\bar{\mathbf{Z}}+\mathbf{I})\mathbf{e}_i \tag{6.68}$$

und der Beziehung

$$\|\mathbf{V}^T\mathbf{p}_i\|_2^2 \overset{\mathbf{V}\in\mathbb{R}^{N\times N}}{=} \mathbf{p}_i^H\mathbf{V}\mathbf{V}^T\mathbf{p}_i = \mathbf{p}_i^H\mathbf{T}^{-1}\mathbf{p}_i = \|\mathbf{p}_i\|_{\mathbf{T}^{-1}}^2 \tag{6.69}$$

lautet der asymptotische Fehlerschätzer für die Streumatrix schließlich

$$|\bar{s}_{mn} - s_{mn}| \leq 2k_0\eta_0\|\mathbf{p}_m\|_{\mathbf{T}^{-1}}\|\mathbf{p}_n\|_{\mathbf{T}^{-1}}\|\mathbf{D}\|_2. \tag{6.70}$$

Die Auswertung von $\|\mathbf{p}_m\|_{\mathbf{T}^{-1}}$ und $\|\mathbf{p}_n\|_{\mathbf{T}^{-1}}$ benötigt nur $\mathcal{O}(n)$ Operationen für einen festen Parameter κ, weshalb auch der asymptotische Fehlerschätzer für S-Parameter in das MOR-Rahmenwerk integriert werden kann, ohne den Rechenaufwand in erheblichem Maße zu vergrößern.

6.6 Numerische Beispiele

Die in der vorliegenden Arbeit präsentierten Verfahren sind als MATLAB-Programm realisiert. Über eine Schnittstelle zur institutseigenen FE-Software können die Komponenten des FE-Gleichungssystems, also die Systemmatrizen und Anregungsvektoren, an den MOR-Prozess übergeben werden. Aufgrund der Dimension der FE-Gleichungssysteme wird zudem die Anbindung an einen externen Gleichungssystemlöser (Intel MKL PARDISO) notwendig. Die Schnittstellen sind in Abbildung 6.1 grafisch dargestellt. In den folgenden Abschnitten werden anhand numerischer Beispiele die

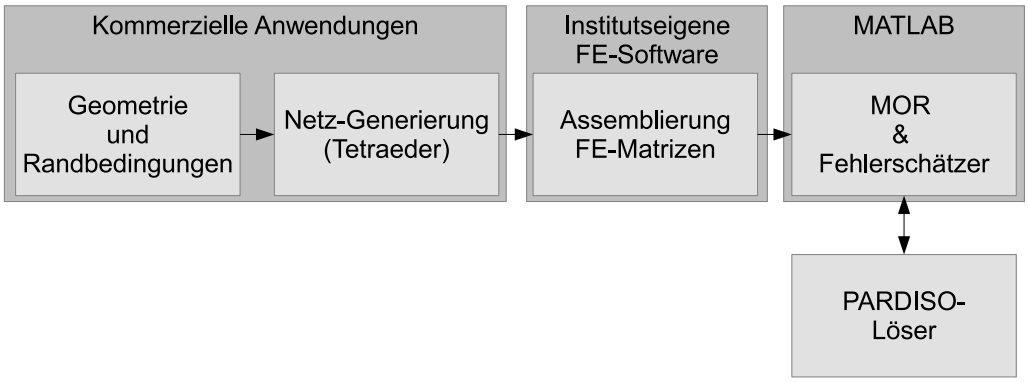


Abbildung 6.1: Programmschnittstellen im MOR-Fehlerschätzprozess.

Verlässlichkeit des Fehlerschätzers und die Effizienz der Algorithmen untersucht und bewertet.

6.6.1 Bandpassfilter

In Abbildung 6.2 ist die Struktur eines Bandpassfilters dargestellt. Die FE-Diskretisierung führt auf ein System mit $N = 213472$ Freiheitsgraden. Betrachtet werden

$N_f = 201$ äquidistante Evaluierungspunkte im Frequenzbereich $\mathcal{I}_f = [0.58, 0.63]$ GHz mit dem Entwicklungspunkt bei $\hat{f} = 0.6$ GHz. Um den gesamten Fehler über das

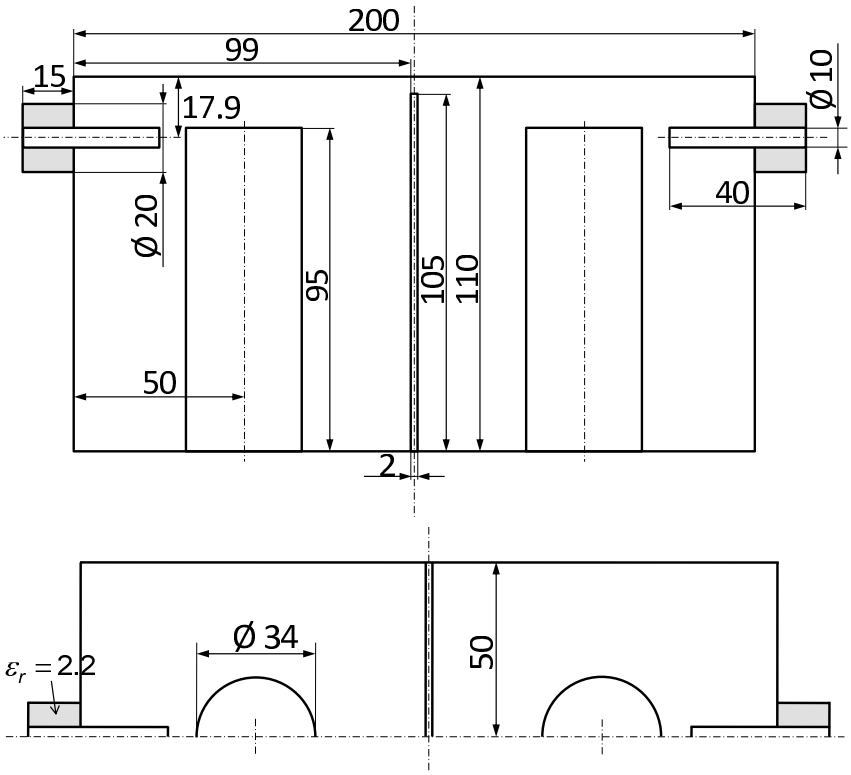


Abbildung 6.2: Bandpass Filter. Alle Dimensionen sind in mm.

Frequenzband zu berücksichtigen, wird das Fehlermaß

$$E_\infty(\mathbf{N}; \mathcal{B}_f) = \max_{m,n,i} |\bar{\nu}_{mn}(f_i) - \nu_{mn}(f_i)|\,, \tag{6.71}$$

als Bewertungskriterium herangezogen. Hierin steht $\mathbf{N}$ als Platzhalter für eine frequenzabhängige Netzwerkmatrix, also $\mathbf{N} \in \{\mathbf{Z}, \mathbf{S}, \mathbf{Y}\}$. Die geschätzten Fehler sind mittels (6.61) und (6.70) berechnet und werden gegen die tatsächlichen Fehler

$$\mathbf{E}_N = \begin{cases} \bar{\mathbf{Z}} - \mathbf{Z} & \text{für Impedanzmatrizen,} \\ \bar{\mathbf{S}} - \mathbf{S} & \text{für Streumatrizen} \end{cases} \tag{6.72}$$

verglichen. Darüber hinaus sind den so ermittelten Fehlern die Ergebnisse basierend auf dem Fehlerschätzer von [CHMR10] und [PS10] gegenübergestellt. Letzterer basiert auf der Ermittlung einer unteren Schranke für die Inf-Sup-Konstante, welche durch den kleinsten Singulärwert der Systemmatrix abgeschätzt wird. Für die Schnelle Online-Auswertung der Fehlerschranke wird die SCM Methode [CHMR10], [HRSP07] herangezogen. Abbildung 6.3 zeigt einen Vergleich zwischen dem tatsächlichen Fehler, der State-of-the-Art-Schätzung und der vorgeschlagenen Fehlerschranke. Die vertikalen Linien in Abbildung 6.3(e) und Abbildung 6.3(f) zeigen jene ROM-

Dimension an, bei welcher sämtliche Eigenwerte im Frequenzbereich $\mathcal{I}_f$ im Sinne der Schranke (6.44) konvergiert, und damit die Fehlerschätzungen (6.32) und (6.70) über den *gesamten* Bereich $\mathcal{I}_f$ gültig sind. So lange der tatsächliche Fehler über dem numerischen Grundrauschen liegt, stellt der vorgeschlagene Fehlerschätzer entsprechend der theoretischen Herleitung eine obere Schranke dar. Der State-of-the-Art-Schätzer ist im Vergleich um drei bis fünf Größenordnungen ungenauer. Das numerische Grundrauschen wird von beiden Fehlerschätzern nicht realistisch erfasst. Letztere Tatsache stellt jedoch für den praktischen Nutzen keine Einschränkung dar, da ein *Unterschätzen* des Fehlers erst eintritt, wenn das ROM bereits vollständig konvergiert ist.

Ein Vergleich der Rechenzeiten ist in Tabelle 6.1 dargestellt. Beide Fehlerschätzer erlauben eine effiziente Auswertung der 201 Frequenzpunkte im Online-Schritt, allerdings führt die SCM-basierte Methode im Offline-Teil zu einem beträchtlichen Mehraufwand: die Rechenzeit von 18943 s wird benötigt, um Lösungen für 418 Eigenwertprobleme der Dimension N zu berechnen. Der in dieser Arbeit vorgeschlagene Ansatz benötigt hingegen im Offline-Teil nur 3 s, die im Wesentlichen der Faktorisierung der Massenmatrix $\mathbf{T}$ zuzuschreiben sind.

Tabelle 6.1: Berechnungsdaten[1]

		Bandpassfilter	Dielektrischer Resonator
Parameter	N	213 472	426 878
	K	12	108
	N_f	201	401
Offline Rechenzeit	Neuer Schätzer	3 s	10.5 s
	State-of-the-Art	18 943 s	n.a. (vgl. 6.6.2)
Online Rechenzeit	Neuer Schätzer	1.36 s	31.8 s
	State-of-the-Art	1.57 s	n.a. (vgl. 6.6.2)

[1] MATLAB 2010 Programm auf Intel Core i7-3370KCPU, 3.5 GHz Prozessor. Intel MKL PARDISO zum Lösen dünnbesetzter Gleichungssysteme.

6.6.2 Filter mit dielektrischen Resonatoren

Als weiteres Beispiel dient ein Filter mit dielektrischen Resonatoren gemäß Abbildung 6.4. Um die Zuverlässigkeit des vorgeschlagenen Fehlerschätzers auch für breitbandige Anwendungen zu zeigen, wird das Filter im Frequenzband $\mathcal{I}_f = [4, 12]$ GHz mit $N_f = 401$ äquidistanten Evaluierungspunkten untersucht. Aus den Abbildungen 6.5(a) und 6.5(b) ist zu sehen, dass im untersuchten Frequenzband eine Vielzahl scharfer Resonanzen vorliegt. Der Vergleich zwischen tatsächlichem und geschätzten Fehler in den Abbildungen 6.5(c) und 6.5(d) bestätigt, dass der vorgeschlagene

Fehlerschätzer eine obere Schranke darstellt, solange der tatsächliche Fehler Werte oberhalb des numerischen Grundrauschens aufweist. Die vorgeschlagene Methode zeigt zuverlässig die Konvergenz des ordnungsreduzierten Modells an. Ein Vergleich zur State-of-the-Art-Methode kann hier nicht angegeben werden, da die SCM bei der Berechnung einer gültigen unteren Schranke für die Inf-Sup-Konstante versagt. Dies lässt sich damit begründen, dass Schätzwerte für den kleinsten Singulärwert sehr sensitiv von der Qualität der Lösung des linearen Programms abhängen, welches Bestandteil dieses Verfahrens ist.

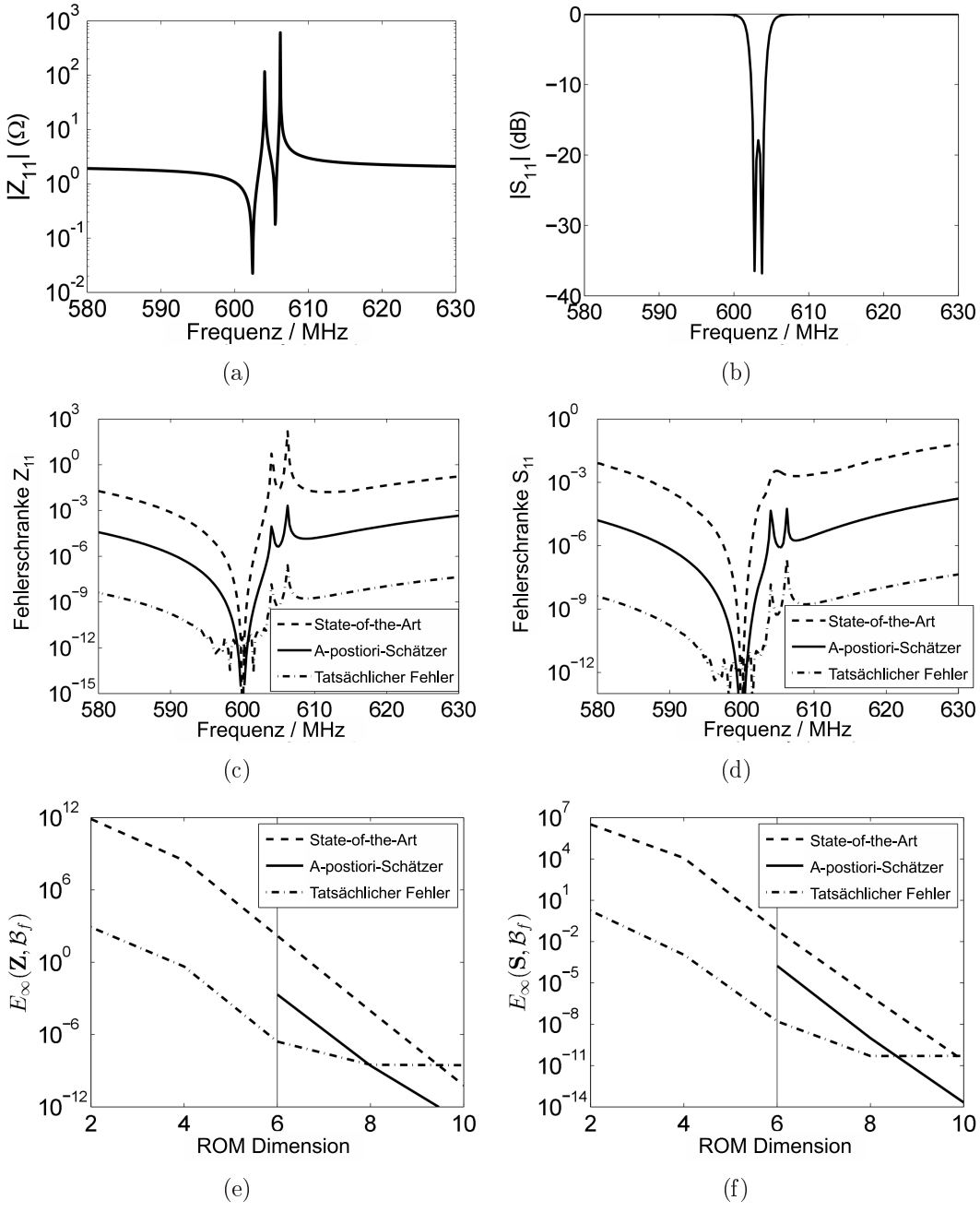


Abbildung 6.3: Bandpassfilter. (a) Betrag von Z_{11}, (b) Betrag von S_{11}, (c) Fehlerschätzung für Z_{11} at $n = 6$, (d) Fehlerschätzung für S_{11} at $n = 6$, (e) Fehlermaß $E_\infty(\mathbf{Z}; \mathcal{B}_f)$, (f) Fehlermaß $E_\infty(\mathbf{S}; \mathcal{B}_f)$.

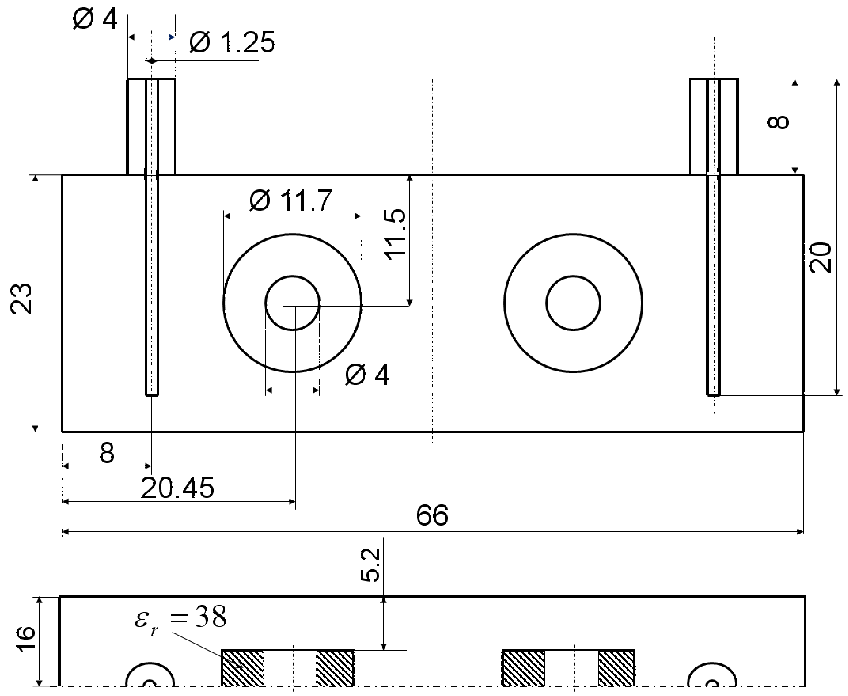


Abbildung 6.4: Filter mit dielektrischen Resonatoren. Dimensionen in mm.

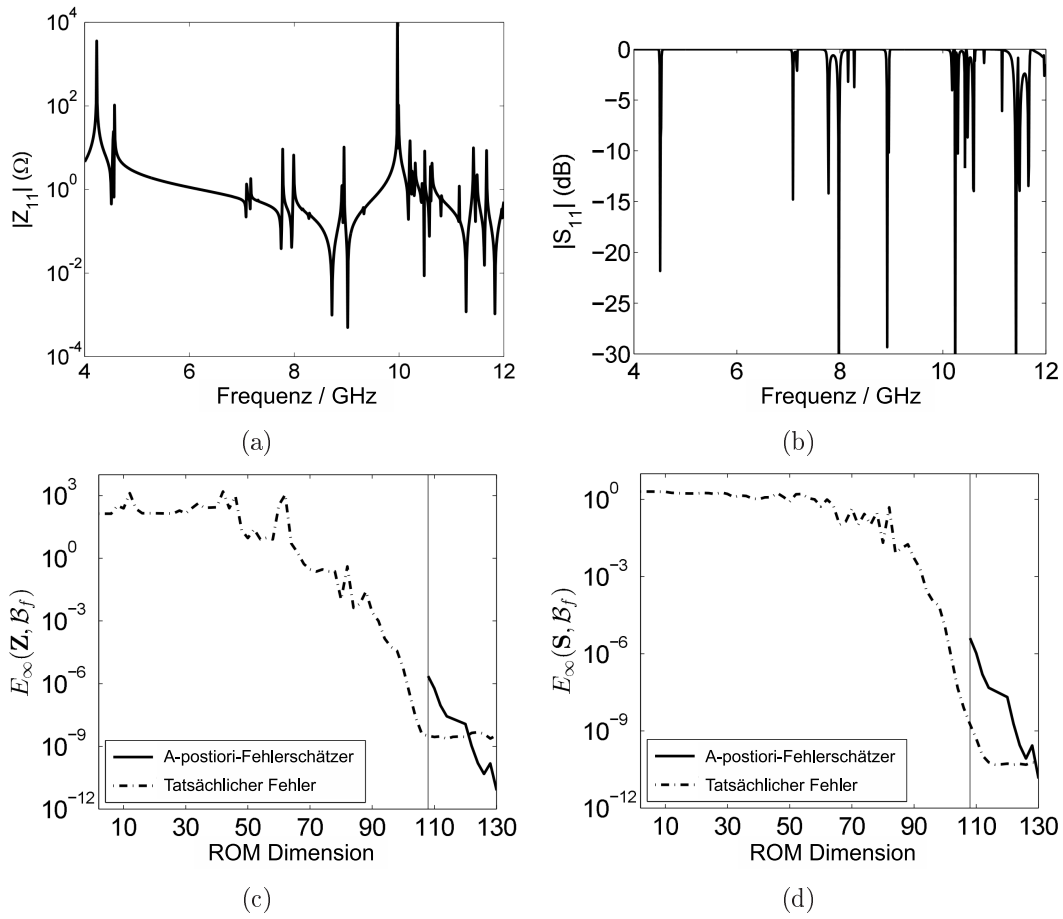


Abbildung 6.5: Filter mit dielektrischen Resonatoren. (a) Betrag von Z_{11}, (b) Betrag von S_{11}, (c) Fehlermaß $E_\infty(\mathbf{Z}; \mathcal{B}_f)$, (d) Fehlermaß $E_\infty(\mathbf{S}; \mathcal{B}_f)$.

Kapitel 7

Zussammenfassung

In der vorliegenden Arbeit werden Verfahren vorgestellt, die die automatisierte Anwendung von projektionsbasierten Modellordnungsreduktionsverfahren ermöglichen. Untersuchungen der Strategien und Verfahren anhand numerischer Beispiele belegen die Funktionalität und Effizienz.

In Kapitel 4 werden die Grundlagen für die effiziente Berechnung von Residuen und die schnelle Auswertung der ordnungsreduzierten Modelle entwickelt. Außerdem werden die speziellen Eigenschaften der Systeme aus der elektrodynamischen FE-Simulation aufgezeigt, insbesondere die Beibehaltung des Momentenabgleichs der MOR-Verfahren auch bei vorliegen von TE- und TM-Wellen.

Bei der Untersuchung in Kapitel 5 zeigen die unterschiedlichen Strategien zur Stützstellenwahl in Mehrpunkt-Anwendungen ähnliches Konvergenzverhalten. Der Autor schließt aus den Untersuchungen, dass die Bisektionsmethode gegenüber der Greedy-Methode zu bevorzugen ist, da eine geringere Tendenz zur Häufung von Stützstellen in einem schmalen Parameterbereich vorliegt. Bei den Fehlerindikatoren zeigt der inkrementelle Ansatz im Vergleich zum residuenbasierten Verfahren eine höhere numerische Effizienz bei gleicher Zuverlässigkeit.

Für den wichtigen Spezialfall verlustloser Systeme in der Elektrodynamik wird in Kapitel 6 als weiteres Abbruchkriterium ein a-posteriori-Fehlerschätzer für Einpunkt-Verfahren vorgestellt. Durch konsequente Anwendung schneller Auswerteverfahren und durch Wiederverwendung von Zwischenergebnissen, die im MOR-Prozess anfallen, kann der zusätzliche Rechen- und Speicheraufwand gering gehalten werden. Damit die notwendige lineare Parametrierung des untersuchten Systems aufrecht erhalten werden kann, ist der Fehlerschätzer in der Impedanzformulierung anzuwenden. Die so bestimmten Fehler können mit beschränkten Größen in Bezug gesetzt werden, indem zusätzlich ein asymptotischer Fehlerschätzer in den S-Parametern eingesetzt wird. Die Abschätzung des Fehlers in den S-Parametern ist insbesondere

für praktische Anwendungen von großer Bedeutung. Numerische Beispiele belegen die Zuverlässigkeit und Effizienz des vorgestellten Verfahrens.

Eine Erweiterung der Fehlerschranke auf verlustbehaftete Systeme der Elektrodynamik beziehungsweise auf allgemein polynomiell parametrierte Systeme ist Gegenstand aktueller Forschung. Die bei der Erstellung der Arbeit bekannten Ansätze sind nur mit erheblichem numerischen Aufwand umzusetzen und daher für praktische Anwendungen unattraktiv.

Abbildungsverzeichnis

2.1 Modales Mehrtor mit Randbedingungen. 6

2.2 Axial homogener, zylindrischer Wellenleiter. 12

2.3 Äquivalente Darstellung von modalem Mehrtor und Klemmenmehrtor 18

5.1 Vivaldi-Antenne: Geometrische Struktur und Amplitudenantwort über der Frequenz. 68

5.2 Vivaldi-Antenne: adaptive ROM-Generierung gesteuert von residuenbasiertem Fehlerindikator in der euklidischen Norm. 69

5.3 Vivaldi-Antenne: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. 70

5.4 Vivaldi-Antenne: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. 71

5.5 Bandpassfilter: Struktur und Amplitudenantwort über der Frequenz. 73

5.6 Bandpassfilter: adaptive ROM-Generierung gesteuert von residualbasiertem Fehlerindikator in der euklidischen Norm. 74

5.7 Bandpassfilter: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. 75

5.8 Bandpassfilter: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. . 76

5.9 Wellenleiter mit dielektrischem Einsatz. 78

5.10 Wellenleiter mit dielektrischem Einsatz: adaptive ROM-Generierung gesteuert von residuenbasiertem Fehlerindikator in der euklidischen Norm. 79

5.11 Wellenleiter mit dielektrischem Einsatz: adaptive ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. . 80

5.12 Wellenleiter mit dielektrischem Einsatz: tatsächlicher Fehler bei adaptiver ROM-Generierung gesteuert von inkrementellem Fehlerindikator in der Maximumnorm. 81

6.1 Programmschnittstellen im MOR-Fehlerschätzprozess. 97

6.2 Bandpass Filter. Alle Dimensionen sind in mm. 98

6.3 Bandpassfilter: Übertragungsverhalten und Ergebnisse der Fehlerschätzung 101

6.4 Filter mit dielektrischen Resonatoren. Dimensionen in mm. 102

6.5 Filter mit dielektrischen Resonatoren: Übertragungsverhalten und Ergebnisse der Fehlerschätzung 103

Literaturverzeichnis

[BCD+11] Binev, P. ; Cohen, A. ; Dahmen, W. ; DeVore, R. ; Petrova, G. ; Wojtaszczyk, P.: Convergence Rates for Greedy Algorithms in Reduced Basis Methods. In: *SIAM Journal on Mathematical Analysis* 43 (2011), Nr. 3, 1457-1472. http://dx.doi.org/10.1137/100795772. – DOI 10.1137/100795772

[BDD+00] Bai, Z. ; Demmel, J. ; Dongarra, J. ; Ruhe, A. ; Vorst, H. van d.: *Templates for the Solution of Algebraic Eigenvalue Problems.* SIAM, Philadelphia, 2000

[Beb00] Bebendorf, M.: Approximation of boundary element matrices. In: *Numer. Math.* 86 (2000), S. 565–589

[Ben04] Benner, P.: Solving large-scale control problems. In: *IEEE Control Systems Magazine* 14 (2004), Februar, Nr. 1, S. 44–59

[Ber94] Berenger, Jean-Pierre: A perfectly matched layer for the absorption of electromagnetic waves. In: *Journal of Computational Physics* 114 (1994), Nr. 2, 185 - 200. http://dx.doi.org/http://dx.doi.org/10.1006/jcph.1994.1159. – DOI http://dx.doi.org/10.1006/jcph.1994.1159. – ISSN 0021–9991

[Bet77] Bettess, P.: Infinite elements. In: *International Journal for Numerical Methods in Engineering* 11 (1977), Nr. 1, S. 53–64

[BL97] Brauer, J. R. ; Lizalek, G. C.: Microwave filter analysis using a new 3-D finite-element modal frequency method. In: *IEEE Transactions on Microwave Theory and Techniques* 45 (1997), Mai, Nr. 5, S. 810–818

[BMP+12] Buffa, Annalisa ; Maday, Yvon ; Patera, Anthony T. ; Prud homme, Christophe ; Turinici, Gabriel: A priori convergence of the Greedy algorithm for the parametrized reduced basis method. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46 (2012), 5, 595–603. http://dx.doi.org/10.1051/m2an/2011056. – DOI 10.1051/m2an/2011056. – ISSN 1290–3841

[Bos88] Bossavit, A.: Whitney forms: a class of finite elements for three-dimensional computations in electromagnetism. In: *IEE Proceedings A - Physical Science, Measurement and Instrumentation, Management and Education - Reviews* 135 (1988), Nov, Nr. 8, S. 493–500. http://dx.doi.org/10.1049/ip-a-1.1988.0077. – DOI 10.1049/ip–a–1.1988.0077. – ISSN 0143–702X

[Bos98] Bossavit, A.: *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*. Academic Press, 1998

[BS94] Brenner, S.C. ; Scott, L.R.: *The Mathematical Theory of Finite Element Methods*. Springer, 1994

[BS05a] Bai, Z. ; Su, Y.: Dimension reduction of large-scale second-order dynamical systems via a second-order Arnoldi method. In: *SIAM Journal on Scientific Computing* 26 (2005), Nr. 5, S. 1692–1709

[BS05b] Bai, Z. ; Su, Y.: SOAR: A second Order Arnoldi Method for the Solution of the quadratic Eigenvalue Problem. In: *SIAM J. Matrix Anal. Appl.* 26 (2005), Nr. 3, S. 640–659

[BSSY99] Bai, Z. ; Slone, R. D. ; Smith, W. T. ; Ye, Q.: Error Bound for Reduced System Model by Padé Approximation via the Lanczos Process. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 18 (1999), S. 133–141

[BT80] Bayliss, A. ; Turkel, E.: Radiation boundary conditions for wave-like equations. In: *Communications on Pure and Applied Mathematics* 33 (1980), November, S. 707–725

[CHMR10] Chen, Y. ; Hesthaven, J. S. ; Maday, Y. ; Rodriguez, J.: Certified Reduced Basis Methods and Output Bounds for the Harmonic Maxwells Equations. In: *SIAM Journal on Scientific Computing* (2010)

[Cia78] Ciarlet, P.: *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam, 1978

[CL88] Cendes, Z. J. ; Lee, J.-F.: The Transfinite Element Method for Modeling MMIC Devices. In: *IEEE Trans. on Microwave Theory tech* 12 (1988), S. 1639–1649

[Col91] Collin, R. E.: *Field Theory of Guided Waves*. 2. Auflage. Piscataway, NY : IEEE Press, 1991

[DEB96] Dyczij-Edlinger, R. ; Biro, O.: A joint vector and scalar potential formulation for driven high frequency problems using hybrid edge and nodal finite elements. In: *Microwave Theory and Techniques, IEEE Transactions on* 44 (1996), Nr. 1, S. 15–23. http://dx.doi.org/10.1109/22.481380. – DOI 10.1109/22.481380. – ISSN 0018–9480

[EM77] Engquist, Bjorn ; Majda, Andrew: Absorbing Boundary Conditions for the Numerical Simulation of Waves. In: *Mathematics of Computation* 31 (1977), Nr. 139, 629-651. http://www.jstor.org/stable/2005997. – ISSN 00255718, 10886842

[Far07] Farle, O.: *Ordnungsreduktionsverfahren für die Finite-Elemente-Simulation parameterabhängiger passiver Mikrowellenstrukturen*, Universität des Saarlandes, Diss., 2007

[FF95] Feldmann, P. ; Freund, R. W.: Efficient linear circuit analysis by Padé approximation via the Lanczos process. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 14 (1995), S. 639–649

[FHDE04] Farle, O. ; Hill, V. ; Dyczij-Edlinger, R.: Finite Element waveguide solvers revisited. In: *IEEE Transactions on Magnetics* 40 (2004), S. 1468–1471

[FLDE10] Farle, Ortwin ; Lösch, Markus ; Dyczij-Edlinger, Romanus: Efficient Fast Frequency Sweep Without Nonphysical Resonances. In: *Electromagnetics* 30 (2010), Nr. 1-2, 51-68. http://dx.doi.org/10.1080/02726340903485307. – DOI 10.1080/02726340903485307

[GB86] Gui, W. ; Babuška, I.: The h, p and h-p versions of the finite element method in 1 dimension. Part I: The error analysis of the p-version, Part II: The error analysis of the h- and h-p versions, Part III: The adaptive h-p version. In: *Numerische Mathematik* 49 (1986), S. 577–683

[GL96] Golub, G. H. ; Loan, C. van: *Matrix Computations*. 3rd. The John Hopkins University Press, Baltimore, Maryland, 1996

[Glo84] Glover, K.: Optimal Hankel-norm approximations of linear multivariable systems and their L-ininity-error bounds. In: *International Journal of Control* 39 (1984), Juni, Nr. 6, S. 1115–1193

[Gri97] Grimme, E. J.: *Krylov Projections Methods for Model Reduction*, University of Illinois at Urbana-Champaign, Diss., 1997

[Hac99] Hackbusch, W.: A Sparse Matrix Arithmetic Based on H(curl)-conforming-Matrices. Part I: Introduction to H-Matrices. In: *Computing* 62 (1999), Apr, Nr. 2, 89–108. http://dx.doi.org/10.1007/s006070050015. – DOI 10.1007/s006070050015. – ISSN 1436–5057

[HEL94] Hoppe, D. J. ; Epp, L. W. ; Lee, Jin-Fa: A hybrid symmetric FEM/MOM formulation applied to scattering by inhomogeneous bodies of revolution. In: *IEEE Transactions on Antennas and Propagation* 42 (1994), Jun, Nr. 6, S. 798–805. http://dx.doi.org/10.1109/8.301698. – DOI 10.1109/8.301698. – ISSN 0018–926X

[HFDE03] Hill, V. ; Farle, O. ; Dyczij-Edlinger, R.: A stabilized multilevel vector finite-element solver for time-harmonic electromagnetic waves. In: *IEEE Transactions on Magnetics* 39 (2003), Mai, S. 1203–1206

[HFDE04] Hill, V. ; Farle, O. ; Dyczij-Edlinger, R.: Finite element basis functions for nested meshes of nonuniform refinement level. In: *IEEE Transactions on Magnetics* 40 (2004), März, S. 981–984

[HRSP07] Huynh, D.B.P. ; Rozza, G. ; Sen, S. ; Patera, A.T.: A successive constraint linear optimization method for lower bounds of parametric coercivity and inf-sup stability constants. In: *C. R. Acad. Sci. Paris* Ser. I 345 (2007), S. 1–6

[Ing06] Ingelström, P.: A new set of H(curl)-conforming, hierarchical basis functions for tetrahedral meshes. In: *IEEE Transactions on Microwave Theory and Techniques* 54 (2006), Nr. 1, S. 106–114

[KFDE10] Konkel, Y. ; Farle, O. ; Dyczij-Edlinger, R.: An Error Estimator for Krylov-Based Fast Frequency Sweeps. In: *10th International Workshop on Finite Elements for Microwave Engineering, Paper S2-P4*, 2010

[KFDE13] Konkel, Yves ; Farle, Ortwin ; Dyczij-Edlinger, Romanus: A Posteriori Error Bounds for Krylovbased Fast Frequency Sweeps of Finite Element Systems. In: *COMPUMAG 2013, Paper PC2-4*. Budapest, Hungary : International Compumag Society, June 30 - July 4 2013

[KFK+11] Konkel, Y. ; Farle, O. ; Köhler, A. ; Schultschik, A. ; Dyczij-Edlinger, R.: Adaptive strategies for fast frequency sweeps. In: *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering* 30 (2011), Nr. 6, S. 1855–1869

[KFS+14] Konkel, Y. ; Farle, O. ; Sommer, A. ; Burgard, S. ; Dyczij-Edlinger, R.: A Posteriori Error Bounds for Krylov-Based Fast Frequency Sweeps of Finite-Element Systems. In: *IEEE Transactions on Magnetics* 50 (2014), Feb, Nr. 2, S. 441–444. `http://dx.doi.org/10.1109/TMAG.2013.2285442`. – DOI 10.1109/TMAG.2013.2285442. – ISSN 0018–9464

[KSN+96] Kolbehdari, M.A. ; Srinivasan, M. ; Nakhla, M.S. ; Zhang, Qi-Jun ; Achar, R.: Simultaneous time and frequency domain solutions of EM problems using finite element and CFH techniques. In: *Microwave Theory and Techniques, IEEE Transactions on* 44 (1996), Nr. 9, S. 1526–1534. `http://dx.doi.org/10.1109/22.536600`. – DOI 10.1109/22.536600. – ISSN 0018–9480

[Lee90] Lee, J. F.: Analysis of passive microwave devices by using three-dimensional tangential vector finite elements. In: *International Journal of Numerical Modeling: Electronic Networks, Devices and Fields* 3 (1990), S. 235–246

[LLL03] Lee, S.-C. ; Lee, J.-F. ; Lee, R.: Hierarchical vector finite elements for analyzing waveguiding structures. In: *IEEE Transactions on Microwave Theory and Techniques* 51 (2003), August, S. 1897–1905

[Lös13] Lösch, M.: *Selbst-adaptive Finite-Elemente-Verfahren zur Simulation passiver Mikrowellenstrukturen*, Universität des Saarlandes, Diss., 2013

[LW02] Li, J.-R. ; White, J.: Low rank solution of Lyapunov equations. In: *SIAM Journal on Matrix Analysis and Applications* 24 (2002), Nr. 1, S. 260–280

[Met] Metis: *[online]*. `http://glaros.dtc.umn.edu/gkhome/views/metis`

[Mon03] Monk, P.: *Finite element methods for Maxwell's equations*. Oxford Science Publications, 2003 (Numerical mathematics and scientific computation)

[Moo81] Moore, B. C.: Principal component analysis in linear system: controllability, observability and model reduction. In: *IEEE Transactions on Automatic Control* 26 (1981), Februar, Nr. 1, S. 17–32

[Ned80] Nedelec, J. C.: Mixed finite elements in R3. In: *Numer. Math.* 35 (1980), S. 315–341

[Par] Pardiso: *[online]*. `www.computational.unibas.ch/cs/scicomp`

[Pet88] Peterson, Andrew F.: Absorbing boundary conditions for the vector wave equation. In: *Microwave and Optical Technology Letters* 1 (1988), Nr. 2, 62–64. `http://dx.doi.org/10.1002/mop.4650010206`. – DOI 10.1002/mop.4650010206. – ISSN 1098–2760

[Poz05] Pozar, D. M.: *Microwave Engineering*. 2nd Edition. John Wiley & Sons, Inc., 2005

[PR90] Pillage, L. T. ; Rohrer, R. A.: Asymptotic waveform evaluation for timing analysis. In: *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* 33 (1990), Apr, Nr. 9, S. 352–366

[PS10] Pomplun, J. ; Schmidt, F.: Accelerated A Posteriori Error Estimation for the Reduced Basis Method with Application to 3D Electromagnetic Scattering Problems. In: *SIAM Journal on Scientific Computing* 32 (2010), Nr. 2, 498-520. `http://dx.doi.org/10.1137/090760271`. – DOI 10.1137/090760271

[RH73] Reed, W. H. ; Hill, T. R.: Triangular mesh methods for the neutron transport equation / Los Alamos Scientific Laboratory. 1973 (LA-UR-73-479). – Forschungsbericht

[Rok85] Rokhlin, V.: Rapid solution of integral equations of classical potential theory. In: *Journal of Computational Physics* 60 (1985), S. 187–207

[RRM09] Rubia, V. de l. ; Razafison, U. ; Maday, Y.: Reliable Fast Frequency Sweep for Microwave Devices via the Reduced-Basis Method. In: *IEEE Transactions on Microwave Theory and Techniques* 57 (2009), Dec., Nr. 12, S. 2923–2937

[SA85] Strube, J. ; Arndt, F.: Rigorous hybrid-mode analysis of the transition from rectangular waveguide to shielded dielectric image guide. In: *IEEE Transactions on Microwave Theory and Techniques* 33 (1985), Mai, S. 391–401

[Saa92a] Saad, Y.: Analysis of some Krylov subspace approximations to the matrix exponential operator. In: *SIAM J. Numer. Anal.* 29 (1992), S. 209–228

[Saa92b] Saad, Y.: *Numerical Methods for large Eigenvalue Problems.* First Edition. Manchester University Press, 1992

[Saa96] Saad, Y.: *Iterative Methods for Sparse Linear Systems.* First Edition. SIAM, 1996

[SB91] Szabó, B. A. ; Babuška, I.: *Finite Element Analysis.* John Wiley & Sons, 1991

[SB92] Stoer, J. ; Bulirsch, R.: *Introduction to Numerical Analysis.* 2. Springer, 1992

[SCNZ94] Sanaie, R. ; Chiprout, E. ; Nakhla, M. S. ; Zhang, Q.-J.: A fast method for frequency and time domain simulation of high-speed VLSI interconnects. In: *IEEE Transactions on Microwave Theory and Techniques* 42 (1994), Dezember, S. 2562–2571

[SF73] Strang, W. G. ; Fix, G. J.: *Analysis of the Finite Element Method.* Wellesley Cambridge Press, 1973

[SFDE08] Schultschik, A. ; Farle, O. ; Dyczij-Edlinger, R.: A Model Order Reduction Method for the Finite-Element Simulation of Inhomogeneous Waveguides. In: *IEEE Transactions on Magnetics* 44 (2008), Nr. 6, S. 1394–1397

[SFDE09] Schultschik, A. ; Farle, O. ; Dyczij-Edlinger, R.: An adaptive multi-point fast frequency sweep for large-scale finite element models. In: *IEEE Transactions on Magnetics* 45 (2009), März, Nr. 3, S. 1108–1111

[SFKDE11] Schultschik, A. ; Farle, O. ; Konkel, Y. ; Dyczij-Edlinger, R.: Self-adaptive fast frequency sweep for the finite element analysis of waveguide modes. In: *Radio Science* 46 (2011), Nr. 5, n/a–n/a. http://dx.doi.org/10.1029/2010RS004638. – DOI 10.1029/2010RS004638. – ISSN 1944–799X. – RS0E09

[Sim79] Simonyi, Kàroly: *Theoretische Elektrotechnik*. Deutscher Verlag der Wissenschaften, 1979

[SKLL95] Sacks, Z. S. ; Kingsland, D. M. ; Lee, R. ; Lee, J.-F.: A perfectly matched anisotropic absorber for use as an absorbing boundary condition. In: *IEEE Trans. Antennas Propag.* 43 (1995), Dezember, Nr. 12, S. 1460–1463

[SL06] Salimbahrami, B. ; Lohmann, B.: Order reduction of large scale second order systems using Krylov subspace methods. In: *Linear Algebra and its Application* 415 (2006), Juni, Nr. 2-3, S. 385–405

[SLL02] Slone, R. D. ; Lee, Jin-Fa ; Lee, R.: Automating multipoint Galerkin AWE for a FEM fast frequency sweep. In: *IEEE Transactions on Magnetics* 38 (2002), March, Nr. 2, S. 637–640. http://dx.doi.org/10.1109/20.996166. – DOI 10.1109/20.996166. – ISSN 0018–9464

[SLL03a] Slone, R. D. ; Lee, R. ; Lee, J.-F.: Broadband Model Order Reduction of Polynomial Matrix Equation using Single-Point Well-Conditioned Asymptotic Waveform Evaluation: Derivation and Theory. In: *International Journal for Numerical Methods in Engineering* 58 (2003), December, S. 2325 – 2342

[SLL03b] Slone, R. D. ; Lee, R. ; Lee, J.-F.: Well-conditioned asymptotic waveform evaluation for finite elements. 51 (2003), September, Nr. 9, S. 2442–2447

[TB97] Trefethen, L. N. ; Bau, D.: *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics, Philadelphia, 1997

[TH05] Taflove, A. ; Hagness, S. C.: *Computational Electrodynamics: The Finite-Difference Time-Domain Method*. 3. Auflage. Artech House Publishers, 2005

[VCK98] Volakis, J. L. ; Chatterjee, A. ; Kempel, L. C.: *Finite Element Method for Electromagnetics*. IEEE Press, 1998

[Web99] Webb, J. P.: Hierarchal Vector Basis Functions of Arbitrary Order for Triangular and Tetrahedral Finite Elements. In: *IEEE Transactions on Antennas and Propagation* 47 (1999), S. 1244–1253

[Wei96] Weiland, T.: Time domain electromagnetic field computation with finite difference methods. In: *International Journal of Numerical Modelling* 9 (1996), S. 295–319

[Whi57] Whitney, Hassler: *Geometric integration theory.* Princeton, N. J. : Princeton University Press, 1957. – xv+387 S.

[WK89] Webb, J. P. ; Kanellopoulos, V. N.: Absorbing boundary conditions for the finite element solution of the vector wave equation. In: *Microwave and Optical Technology Letters* (1989)

[Yee66] Yee, K.: Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. In: *IEEE Trans. Antennas Propag.* 14 (1966), Mai, Nr. 3, S. 302–307

[ZC02] Zhu, Y. ; Cangellaris, A. C.: Finite element-based model order reduction of electromagnetic devices. In: *International Journal of Numerical Modelling Electronic Networks, Devices and Fields* 15 (2002), S. 73–92

[ZT00] Zienkiewicz, O.C. ; Taylor, R.L: *Finite Element Method: Volume 1 - The Basis.* 5th. Oxford : Butterworth-Heinemann, 2000

[ZVL06] Zhao, Kezhong ; Vouvakis, M. N. ; Lee, Jin-Fa: Solving electromagnetic problems using a novel symmetric FEM-BEM approach. In: *IEEE Transactions on Magnetics* 42 (2006), April, Nr. 4, S. 583–586. http://dx.doi.org/10.1109/TMAG.2006.872489. – DOI 10.1109/TMAG.2006.872489. – ISSN 0018–9464