Adaptive Model Order Reduction Techniques for the Broadband Finite Element Simulation of Electromagnetic Structures

Dissertation

zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Naturwissenschaftlich-Technischen Fakultät II
- Physik und Mechatronik -
der Universität des Saarlandes

von

Dipl.-Ing. Alwin Schultschik

Saarbrücken

2012

Tag des Kolloquiums:	08.02.2013
Dekan:	Univ.-Prof. Dr. rer. nat. Christian Wagner
Mitglieder des Prüfungsausschusses:	Univ.-Prof. Dr. Dr. Karsten Kruse Univ.-Prof. Dr. techn. Romanus Dyczij-Edlinger Univ.-Prof. Dr.-Ing. Michael Möller Dr. rer. nat. Herbert Wolf

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.

Saarbrücken, 27. Februar 2013

Danksagung

Als erstes möchte ich mich bei Herrn Prof. Dr. techn. Dyczij-Edlinger für die Möglichkeit diese Dissertation anzufertigen und die interessanten Aufgabenstellungen bedanken. Seine fachlichen Anregungen und seinen wissenschaftlichen Enthusiasmus schätze ich sehr.

Bei den weiteren Mitarbeitern des Lehrstuhls für Theoretische Elektrotechnik möchte ich mich für die fachlichen Anregungen und das gute Arbeitsklima am Institut bedanken. Ein besonderer Dank gilt Herrn Dr.-Ing. Ortwin Farle, der mich während meiner Studien- und Diplomarbeit betreut und auch während meiner Doktorarbeit beraten und motiviert hat. Herrn Robert Ginger danke ich für die professionelle Korrektur der englischen Sprache.

Ein weiterer Dank geht an meine Eltern. Sie haben mir mein Studium ermöglicht und mich entscheidend unterstützt. Als letztes möchte ich mich ganz besonders bei meiner Frau Alëna für die viele Geduld und die Motivierung, diese Arbeit fertig zu stellen, bedanken.

Summary

Model order reduction methods provide a powerful means for the broadband simulation of passive microwave devices. In particular projection-based moment matching methods are well-suited for the reduction of sparse finite element systems. However, for real-world problems, where high-dimensional systems of linear equations are assembled and a large number of excitations is considered in the right-hand side, the projection matrix may fill the main memory and render the process inefficient. In this thesis, techniques were developed which, as a result of reduced memory requirements, make model order reduction applicable to a large set of real-world problem simulations.

A new adaptive multi-point reduction method is introduced whose core is an incremental error measure. For the proposed single-point method, which is based on the well-conditioned asymptotic waveform evaluation, memory requirements are reduced by means of a block algorithm, whose moment matching properties are proven in this thesis. Memory swapping mechanisms for both approaches keep the main memory requirements for the projection matrix at a constant low level during the computations.

This thesis also includes an adaptive multi-point method for the broadband finite element simulation of waveguide problems and a broadband sensitivity analysis technique.

Kurzfassung

Verfahren der Modellordnungsreduktion stellen einen leistungsfähigen Ansatz für die breitbandige Simulation passiver Mikrowellenkomponenten dar. Insbesondere projektionsbasierte, momentenabgleichende Methoden eignen sich für die Reduktion der schwach besetzten Finite-Elemente Systeme. In praxisrelevanten Problemstellungen hingegen, bei denen hochdimensionale Gleichungssysteme assembliert werden und eine große Anzahl Anregungen in der rechten Seite berücksichtigt werden, kann die Projektionsmatrix den Arbeitsspeicher füllen und der Prozess ineffizient werden. In dieser Dissertation werden Algorithmen entwickelt, die aufgrund des reduzierten Speicherbedarfs Reduktionsverfahren auf eine große Auswahl praxisrelevanter Simulationen anwendbar machen.

Ein neues Mehrpunktverfahren wird eingeführt, dessen Kern ein inkrementelles Fehlermaß ist. Für das entwickelte Einpunktverfahren, welches auf der Well-Conditioned Asymptotic Waveform Evaluation basiert, wurde der Speicheraufwand mit Hilfe eines Blockalgorithmus reduziert, dessen momentenabgleichenden Eigenschaften in dieser Dissertation bewiesen werden. Datenauslagerungsmechanismen für beide Ansätze halten den Arbeitsspeicherbedarf für die Projektionsmatrix während der Berechnung konstant niedrig.

Diese Arbeit beinhaltet des Weiteren ein adaptives Mehrpunktverfahren für die breitbandige Finite-Elemente-Simulation von Wellenleiterproblemen und ein Verfahren zur breitbandigen Sensitivitätsanalyse.

Contents

Chapter 1

Introduction

1.1 Preliminary Words

Numerical simulation and computer-aided design processes enjoy an ever increasing area of application. This trend is driven by the more and more complex design of electronic devices. Hence, higher frequencies and coupling issues make the design process more challenging and the usage of electromagnetic field simulation demanding. Numerical field simulation techniques provide reliable solutions for electromagnetic problems, even for very complex structures. Circuit simulators are often employed in a subsequent process and may utilize extracted parameters from numerical field simulations.

Although the numerical computations in electromagnetics do exhibit high memory requirements as well as high computational costs, progress in hardware development allows field simulations to be run on a standard personal computer. However, for extended applications, e.g. parameter studies or optimization processes, the simulation methods often struggle with limited system resources. At the same time, the simulation of more and more complex electromagnetic devices is requested. The numerical simulation of complete structures such as printed circuit boards, integrated circuit packages or full machine wirings is the challenge of today's techniques. To fulfill these requirements, significant effort is devoted to both hardware development and improvement of numerical methods.

1.2 Numerical Simulation

The Finite Element Method

Among numerical simulation techniques for electromagnetic problems, the finite element (FE) method has become popular due to its flexibility in modeling geometry and material properties. FE simulation provides a numerical solution for electromagnetic boundary value problems, which may be derived from electromagnetic field theory. The focus of this thesis is on the simulation of passive microwave structures, where particularly driven problems [1], [2], [3] and waveguide problems [4], [5] are considered. The numerical simulation process using the FE method typically results in a large-scale sparse system of equations or an algebraic eigenvalue problem, respectively. The system of linear equations is assembled and solved for a fixed configuration. To simulate the frequency characteristics of a microwave component, e.g. a microwave filter, the solution process needs to be performed for a large number of evaluation points. This broadband simulation, however, may result in a time-consuming process. Parameterizing the frequency in the system of linear equation results in a problem well-suited for model order reduction (MORe) approaches.

Model Order Reduction

The main goal of MORe is to find a reduced order model (ROM), which approximates the large-scale system in a certain parameter range. The most common MORe approaches, which arise from the field of circuit simulation and control theory, may be categorized into singular value decomposition-based methods [6] and moment matching methods as the Asymptotic Waveform Evaluation (AWE) [7]. The focus of this thesis is on the moment matching approaches, which are more convenient for the considered large-scale FE systems. While early methods, as the AWE technique, suffered from a deficit in numerical stability, the later projection-based methods [8] resolved this numerical shortcoming. A numerically stable process for polynomial parameterized systems is provided by the Well-Conditioned Asymptotic Waveform Evaluation (WCAWE) [9]. Multi-point methods, which were originally developed to overcome the numerical instabilities of early single-point methods [10], are very general approaches and enjoy popularity due to their flexibility, see e.g. [11]. The generality of multi-point methods also supports MORe for polynomially parameterized system.

1.3 Problem Definition and Scientific Contribution

Problem Definition

The FE simulation of very complex structures, in which a system matrix of very large dimension is assembled, finds its limitation in the available system resources. This situation becomes worse if MORe techniques are employed for broadband simulations. While the FE system matrix as well as the right-hand side vectors are sparse, the computed projection matrix for the MORe process is a dense matrix, which may fill the main memory. Particularly for simulations in which a large number of moments need to be matched, the projection matrix memory requirements may become a limiting factor. Furthermore, devices with a large number of excitations result in systems with many right-hand sides, which also let the projection matrix dimension increase.

The Aim of This Thesis

The purpose of this thesis is to find MORe techniques suitable for the broadband FE simulation of complex real-world structures. Specifically, this means making projection-based MORe methods, which are actually very suitable for the FE simulation, accessible to high-dimensional systems with a large number of right-hand sides.

Scientific Contribution

An incremental error measure is introduced in Section 3.3, which offers a very efficient and reliable alternative to residual computations or error-bound evaluations [12]. This error measure is used in the bisection method-based adaptive multi-point algorithm developed in Section 3.5. Furthermore, the error measure is used as a termination criterion in an iterative WCAWE process, see Section 3.4. Numerical results indicate that the proposed adaptive multi-point algorithm requires fewer iterations to reach a given error limit than the single-point method, even if the WCAWE process is started at the optimum expansion frequency. Hence, the projection matrix in the multi-point MORe process is smaller and requires less memory.

To reduce the main memory requirements of the MORe process, out-of-core reduction algorithms are developed in Chapter 4. While in-core algorithms fill the system requirements in each iteration with the increasing projection matrix, the proposed algorithms keep the main memory requirements for the projection matrix constant on the level of the first iteration. Particularly for highly resonant structures, where the MORe process runs a large number of iterations, this technique is beneficial. In this framework of out-of-core techniques, this thesis gives a detailed proof that the extension of the WCAWE method to a block algorithm, provided in [13], does match moments.

An additional result of this thesis is a bisection-based adaptive multi-point MORe algorithm for waveguide problems in Chapter 5. Furthermore, in Chapter 6 a broadband sensitivity analysis theory is proposed and discussed in detail on computed numerical experiments.

Available Infrastructure and Implementation Work

For this thesis, a complete c++ FE code was available. Furthermore, an extensive mathematical library, including linkage to the Intel Math Kernel Library, was used.

During this thesis, a comprehensive framework for MORe was implemented, which includes single-point as well as multi-point methods. In particular, the adaptive methods in combination with the error measures as a termination criterion were implemented. Furthermore, a binary data-swapping process was coded for the out-of-core approaches. The code also includes an automatic parameterization for the FE system.

Chapter 2

Simulation of Electromagnetic Structures

The purpose of this second chapter is to provide an overview of the physical and mathematical tools for the FE simulation of passive microwave components. This chapter begins with a introduction to basic electromagnetics, which allows us to define a suitable boundary value problem for the considered field analysis. Applying the FE method to the electromagnetic boundary value problem results in a system of linear equations. The FE discretization allows us to assemble the system matrices parameterized in the frequency, which provides the basis for the later introduced MORe techniques.

2.1 Maxwell's Equations

Electric and magnetic macroscopic phenomena in classical electromagnetics are described by the Maxwell equations

$$\nabla \times \mathcal{E} = -\frac{\partial \mathcal{B}}{\partial t}, \tag{2.1}$$

$$\nabla \times \mathcal{H} = \mathcal{J} + \frac{\partial \mathcal{D}}{\partial t}, \tag{2.2}$$

$$\nabla \cdot \mathcal{D} = \varrho, \tag{2.3}$$

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

which are Faraday's law of induction, Ampère's circuit law, Gauss's law and Gauss's law for magnetism, respectively. The constitutive equations

$$\mathcal{D} = \varepsilon_0 \varepsilon_r \mathcal{E} = \varepsilon \mathcal{E}, \tag{2.5}$$

$$\mathcal{B} = \mu_0 \mu_r \mathcal{H} = \mu \mathcal{H}, \tag{2.6}$$

$$\mathcal{J} = \sigma \mathcal{E}, \tag{2.7}$$

which define material properties, complete the physical description for the electromagnetic fields. The material properties in this thesis are considered to be symmetric, linear and time invariant. Table 2.1 gives the definition of used physical quantities including their units.

Table 2.1: Symbols of electromagnetic quantities.

Symbol	Definition	Unit
$\mathcal{E}$	Electric field intensity	V/m
$\mathcal{D}$	Electric displacement field	C/m^2
$\mathcal{H}$	Magnetic field intensity	A/m
$\mathcal{B}$	Magnetic flux density	$T = Vs/m^2$
$\mathcal{J}$	Electric current density	A/m^2
ρ	Electric charge density	C/m^3
ε_0	Free space permittivity	$As/(Vm)$
ε_r	Relative permittivity	$-$
μ_0	Free space permeability	$Vs/(Am)$
μ_r	Relative permeability	$-$
σ	Electric conductivity	$A/(Vm)$
t	Time	s
ω	Angular frequency	rad/s

However, for the investigation of time harmonic settings, Maxwell's equations may be written in the frequency domain

$$\nabla \times \vec{E} = -j\omega \vec{B}, \tag{2.8}$$

$$\nabla \times \vec{H} = \vec{J} + j\omega \vec{D}, \tag{2.9}$$

$$\nabla \cdot \vec{D} = \rho, \tag{2.10}$$

$$\nabla \cdot \vec{B} = 0, \tag{2.11}$$

where the time dependent physical quantities become the phasors $\vec{E}$, $\vec{D}$, $\vec{H}$, $\vec{B}$, $\vec{J}$ and ρ. As the purpose of this thesis is to employ MORe techniques for the evaluation of fast frequency sweeps, the focus is on the frequency domain.

The introduced physical quantities fit into a mathematical framework of functional spaces. The space of square integrable scalar fields on the domain $\Omega \subset \mathbb{R}^3$ is defined as

$$\mathrm{L}^2(\Omega) := \{\mathbf{u}(x) |\, \|\mathbf{u}(x)\|_{\mathrm{L}^2} < \infty\}, \tag{2.12}$$

where the norm $\|\cdot\|_{\mathrm{L}^2}$ is induced by the inner product

$$(\mathbf{u}, \mathbf{v})_{\mathrm{L}^2} = \int_\Omega \mathbf{u}^H(x)\mathbf{v}(x) d\Omega, \ \mathbf{u}, \mathbf{v} \in \mathbb{C}^n. \tag{2.13}$$

Furthermore, the functional spaces

$$\mathrm{H}^1(\Omega) := \left\{u \in \mathrm{L}^2(\Omega) | \nabla u \in [\mathrm{L}^2(\Omega)]^3\right\}, \tag{2.14}$$

$$\mathbf{H}(\mathbf{curl};\Omega) := \left\{\mathbf{u} \in [\mathrm{L}^2(\Omega)]^3 | \nabla \times \mathbf{u} \in [\mathrm{L}^2(\Omega)]^3\right\}, \tag{2.15}$$

$$\mathbf{H}(\mathbf{div};\Omega) := \left\{\mathbf{u} \in [\mathrm{L}^2(\Omega)]^3 | \nabla \cdot \mathbf{u} \in \mathrm{L}^2(\Omega)\right\}, \tag{2.16}$$

are introduced, which together with the associated scalar products

$$(u, v)_1 = (\nabla u,\ \nabla v)_{\mathrm{L}^2} + (u, v)_{\mathrm{L}^2}, \tag{2.17}$$

$$(\mathbf{u}, \mathbf{v})_{curl} = (\nabla \times \mathbf{u},\ \nabla \times \mathbf{v})_{\mathrm{L}^2} + (\mathbf{u}, \mathbf{v})_{\mathrm{L}^2}, \tag{2.18}$$

$$(\mathbf{u}, \mathbf{v})_{div} = (\nabla \cdot \mathbf{u},\ \nabla \cdot \mathbf{v})_{\mathrm{L}^2} + (\mathbf{u}, \mathbf{v})_{\mathrm{L}^2}, \tag{2.19}$$

are Hilbert spaces. Since we have the inclusions

$$\nabla \mathrm{H}^1(\Omega) \subset \mathbf{H}(\mathbf{curl};\Omega), \tag{2.20}$$

$$\nabla \times \mathbf{H}(\mathbf{curl};\Omega) \subset \mathbf{H}(\mathbf{div};\Omega), \tag{2.21}$$

$$\nabla \times \mathbf{H}(\mathbf{div};\Omega) \subset \mathrm{L}^2(\Omega), \tag{2.22}$$

the introduced Hilbert spaces form the sequence

$$\mathrm{H}^1(\Omega) \xrightarrow{\nabla} \mathbf{H}(\mathbf{curl};\Omega) \xrightarrow{\nabla\times} \mathbf{H}(\mathbf{div};\Omega) \xrightarrow{\nabla\cdot} \mathrm{L}^2(\Omega). \tag{2.23}$$

The sequence is called exact, if

$$\nabla \times \mathbf{H}(\mathbf{curl};\Omega) = \ker(\mathrm{div}) := \{u \in \mathbf{H}(\mathbf{div};\Omega) | \nabla \cdot \mathbf{u} = 0\}, \tag{2.24}$$

$$\nabla \mathrm{H}^1(\Omega) = \ker(\mathrm{curl}) := \{u \in \mathbf{H}(\mathbf{curl};\Omega) | \nabla \times \mathbf{u} = 0\}, \tag{2.25}$$

as we have for the Euclidean space $\mathbb{E}^3$. However, the exactness property is lost for domains of non-trivial topology, which is explained in detail in textbooks as [14] and [15].

Maxwell's equations naturally fit in this framework of functional spaces. The electric field intensity $\vec{E}$ and the magnetic field intensity $\vec{H}$ belong to the space $\mathbf{H}(\mathbf{curl};\Omega)$

$$\vec{E}, \vec{H} \in \mathbf{H}(\mathbf{curl};\Omega). \tag{2.26}$$

The space $\mathbf{H}(\mathbf{div};\Omega)$ is associated with the magnetic flux density $\vec{B}$, the displacement field $\vec{D}$ and the current density $\vec{J}$

$$\vec{B}, \vec{D}, \vec{J} \in \mathbf{H}(\mathbf{div};\Omega). \tag{2.27}$$

Thus, the following diagram shows how Maxwell's equations, with the help of constitutive equations, fit into the framework of functional spaces:

$$\begin{tikzcd}
\varphi_e \arrow[r, "\nabla"] & \vec{E} \arrow[r, "\nabla\times"] \arrow[d, leftrightarrow, "\varepsilon"', "\sigma"] & \vec{B} \arrow[r, "\nabla\cdot"] \arrow[d, leftrightarrow, "\mu"'] & 0 \\
\rho & \vec{D}, \vec{J} \arrow[l, "\nabla\cdot"'] & \vec{H} \arrow[l, "\nabla\times"'] & \varphi_m \arrow[l, "\nabla"']
\end{tikzcd} . \tag{2.28}$$

In the diagram above, the electric scalar potential φ_e and the magnetic scalar potential φ_m are added into the framework.

2.2 Boundary Value Problem Definition

The boundary value problem is based on an electric field formulation. Therefore, the magnetic field intensity $\vec{H}$ in Maxwell's equations is eliminated, which results in the vector Helmholtz equation

$$\nabla \times \mu^{-1} \nabla \times \vec{E} - \omega^2 \varepsilon \vec{E} = -j\omega \vec{J}. \tag{2.29}$$

Furthermore, assuming the source free domain Ω, with the non-overlapping boundaries $\Gamma = \Gamma_E \cup \Gamma_H \cup \Gamma_\Theta \cup \Gamma_Z$, the boundary value problem yields

$$\begin{aligned}
\nabla \times \mu^{-1} \nabla \times \vec{E} - \omega^2 \varepsilon \vec{E} &= 0 && \text{in } \Omega, && (2.30\text{a})\\
\hat{e}_n \times \vec{E} &= 0 && \text{on } \Gamma_E, && (2.30\text{b})\\
\hat{e}_n \times \vec{H} &= 0 && \text{on } \Gamma_H, && (2.30\text{c})\\
\vec{H} \times \hat{e}_n &= \vec{H}_{T\Theta} \times \hat{e}_n && \text{on } \Gamma_\Theta, && (2.30\text{d})\\
\vec{H} \times \hat{e}_n &= \frac{1}{Z_s} \hat{e}_n \times (\vec{E} \times \hat{e}_n) && \text{on } \Gamma_Z, && (2.30\text{e})
\end{aligned}$$

where $\hat{e}_n$ denotes the outward directed normal unit vector on the boundary. Γ_E and Γ_H represent the electric walls and magnetic walls, respectively. Γ_Z is the impedance boundary with the impedance parameter Z_s and $\vec{H}_{T\Theta}$ stands for the impressed tangential magnetic field strength on the wave port region Γ_Θ.

The weak form of the boundary value problem above is obtained by weighting (2.30a) with $\vec{w}_i$ and integrating over the domain Ω

$$\int_\Omega \vec{w}_i \cdot \nabla \times \mu^{-1} \nabla \times \vec{E} d\Omega - \omega^2 \int_\Omega \vec{w}_i \cdot \varepsilon \vec{E} d\Omega = 0, \tag{2.31}$$

where

$$\vec{w}_i \in \mathbf{H}(\mathbf{curl}; \Omega, \Gamma_E) := \left\{ \vec{u} \in [\mathrm{L}^2(\Omega)]^3 | \vec{u} \times \nabla \in [\mathrm{L}^2(\Omega)]^3 \text{ and } \hat{e}_n \times \vec{u} = 0 \text{ on } \Gamma_E \right\}. \tag{2.32}$$

Employing Gauss' theorem and the vector identity

$$\nabla \cdot (\vec{v} \times \vec{u}) = \vec{u} \cdot (\nabla \times \vec{v}) - \vec{v} \cdot (\nabla \times \vec{u}), \tag{2.33}$$

we can write

$$\int_\Omega \nabla \times \vec{w}_i \cdot \mu^{-1} \nabla \times \vec{E} d\Omega - \omega^2 \int_\Omega \vec{w}_i \cdot \varepsilon \vec{E} d\Omega - \oint_\Gamma (\vec{w}_i \times \mu^{-1} \nabla \times \vec{E}) \cdot \hat{e}_n d\Gamma = 0. \tag{2.34}$$

Plugging Faraday's law of induction into the boundary integral results in

$$\int_\Omega \nabla \times \vec{w}_i \cdot \mu^{-1} \nabla \times \vec{E} d\Omega - \omega^2 \int_\Omega \vec{w}_i \cdot \varepsilon \vec{E} d\Omega + j\omega \oint_\Gamma (\vec{w}_i \times \vec{H}) \cdot \hat{e}_n d\Gamma = 0. \tag{2.35}$$

We rewrite the boundary integral as

$$j\omega \oint_\Gamma \ldots d\Gamma = j\omega \int_{\Gamma_E} \ldots d\Gamma + j\omega \int_{\Gamma_H} \ldots d\Gamma + j\omega \int_{\Gamma_\Theta} \ldots d\Gamma + j\omega \int_{\Gamma_Z} \ldots d\Gamma, \tag{2.36}$$

and look on each boundary separated. On Γ_E, we have

$$j\omega \int_{\Gamma_E} (\vec{w}_i \times \vec{H}) \cdot \hat{e}_n d\Gamma = j\omega \int_{\Gamma_E} \vec{H} \cdot (\hat{e}_n \times \vec{w}_i) d\Gamma = 0, \tag{2.37a}$$

as $\vec{w}_i \in \mathbf{H}(\mathbf{curl}; \Omega, \Gamma_E)$. On the boundary Γ_H we have by definition

$$j\omega \int_{\Gamma_H} (\vec{w}_i \times \vec{H}) \cdot \hat{e}_n d\Gamma = -j\omega \int_{\Gamma_H} \vec{w}_i \cdot (\hat{e}_n \times \vec{H}) d\Gamma = 0. \tag{2.37b}$$

The integral on Γ_Θ results in

$$\begin{aligned} j\omega \int_{\Gamma_\Theta} (\vec{w}_i \times \vec{H}) \cdot \hat{e}_n d\Gamma =& j\omega \int_{\Gamma_\Theta} \vec{w}_i \cdot (\vec{H} \times \hat{e}_n) d\Gamma \\ =& j\omega \int_{\Gamma_\Theta} \vec{w}_i \cdot (\vec{H}_{T\Theta} \times \hat{e}_n) d\Gamma, \end{aligned} \tag{2.37c}$$

and will be moved to the right-hand side. On the boundary Γ_Z the integral yields

$$\begin{aligned} j\omega \int_{\Gamma_Z} (\vec{w}_i \times \vec{H}) \cdot \hat{e}_n d\Gamma =& j\omega \int_{\Gamma_Z} \vec{w}_i \cdot (\vec{H} \times \hat{e}_n) d\Gamma \\ =& j\omega \int_{\Gamma_Z} \vec{w}_i \cdot \frac{1}{Z_s} \hat{e}_n \times (\vec{E} \times \hat{e}_n) d\Gamma \\ =& j\omega \int_{\Gamma_Z} (\hat{e}_n \times \vec{w}_i) \cdot (\frac{1}{Z_s} \hat{e}_n \times \vec{E}) d\Gamma. \end{aligned} \tag{2.37d}$$

Thus, the weak form of the boundary value problem reads

$$\begin{aligned} &\int_\Omega \nabla \times \vec{w}_i \cdot \mu^{-1} \nabla \times \vec{E} d\Omega + j\omega \int_{\Gamma_Z} (\hat{e}_n \times \vec{w}_i) \cdot (\frac{1}{Z_s} \hat{e}_n \times \vec{E}) d\Gamma \\ &- \omega^2 \int_\Omega \vec{w}_i \cdot \varepsilon \vec{E} d\Omega = j\omega \int_{\Gamma_\Theta} \vec{w}_i \cdot (\hat{e}_n \times \vec{H}_{T\Theta}) d\Gamma, \ \forall \vec{w}_i \in \mathbf{H}(\mathbf{curl}; \Omega, \Gamma_E). \end{aligned} \tag{2.38}$$

2.2.1 Impedance Boundary Condition

In some electromagnetic configurations, solids with highly conductive material properties, lumped network elements or unbounded domains need to be considered. To model these physical effects, the impedance boundary condition (2.30e) may be used.

First we turn to the modeling of solids with highly conductive material properties, as for instance metals exhibit. Due to the skin effect, alternating currents have the tendency to flow mostly near the surface of conductors. The skin depth

$$\delta = \sqrt{\frac{2}{\omega\mu\sigma}}, \tag{2.39}$$

defines the depth under the surface where the field components have dropped by a factor e. Hence, in case of large geometrical dimensions of the conductive solid, compared to the skin depth, the solid may be removed and modeled with impedance boundary conditions on its surfaces. The impedance parameter is then set to

$$Z_s = (1+j)\sqrt{\frac{\omega\mu}{2\sigma}}, \tag{2.40}$$

as evaluated in [16].

In the free space, the electromagnetic field components of a plane wave satisfy

$$\hat{e} \times \vec{E} = \sqrt{\frac{\mu_0}{\varepsilon_0}} \vec{H}, \tag{2.41}$$

where $\hat{e}$ denotes the propagation direction. This property is used to model infinite computational domains with absorbing boundary conditions [17]. Thus, the parameter Z_s in the impedance boundary condition is set to

$$Z_s = \sqrt{\frac{\mu_0}{\varepsilon_0}}. \tag{2.42}$$

Unlike the impedance boundaries above, sheets associated with lumped elements are allowed to be defined in the interior of the computational domain. The tangential components of the magnetic field intensity in (2.30e) are therefore rather $\vec{H} \times \hat{e}_n = (\vec{H}_2 - \vec{H}_1) \times \hat{e}_n$, the difference of the fields on both sides of the sheet. Ampère's circuit law allows us to write

$$\oint_{\partial \Gamma_Z} (\vec{H}_2 - \vec{H}_1) \times \hat{e}_n d\vec{s} = Kb, \tag{2.43}$$

where K and b are the surface current density and the width of the sheet, respectively. Furthermore, we have an electric voltage along the sheet of length l. With these results, we can derive the impedance parameter in (2.30e) from Ohm's law as

$$Z_s = Z\frac{b}{l}, \tag{2.44}$$

where Z is the impedance of the lumped element.

2.2.2 Excitation on the Wave Port Region

Waveguide Modes

For the field analysis on the boundary Γ_Θ, the surface is subdivided into

$$\Gamma_\Theta = \bigcup_{\vartheta=1}^{N_\vartheta} \Gamma_\vartheta, \tag{2.45}$$

where each Γ_ϑ is defined to be the transverse plane of an axially uniform waveguide, which is connected to the device. For each waveguide a local coordinate system is defined, where the z-axis is set as the uniform axis of the waveguide and the transverse plane Γ_ϑ is at $z = 0$. Hence, the normal vector $\hat{e}_n$ on Γ_ϑ in the global coordinates is equivalent to the local unit vector $\hat{e}_z$ of each waveguide.

Each supported modal waveform $(\vec{e}_\xi, \vec{h}_\xi)$ in a waveguide may be separated into transversal and axial field components

$$\vec{e}_\xi = \vec{e}_{T\xi} + \hat{e}_{z\xi}, \tag{2.46}$$

$$\vec{h}_\xi = \vec{h}_{T\xi} + \hat{h}_{z\xi}, \tag{2.47}$$

where the subscripts T and z denote the transversal and normal components, respectively. The transversal modal field patterns $\vec{e}_{T\xi}$ and $\vec{h}_{T\xi}$ are normalized beyond their waveform cut-off frequency such that

$$\int_{\Gamma_\vartheta} (\vec{e}_{T\xi} \times \vec{h}_{T\xi}) \cdot \hat{e}_z d\Gamma = 1. \tag{2.48}$$

For the superposition of forward and backward traveling transverse field components $\vec{E}_{T\xi}$ and $\vec{H}_{T\xi}$ in the waveguide, we write

$$\vec{E}_{T\xi}(z) = u_\xi^f \vec{e}_{T\xi}\, \mathrm{e}^{-\gamma_\xi z} + u_\xi^b \vec{e}_{T\xi}\, \mathrm{e}^{+\gamma_\xi z}, \tag{2.49}$$

$$\vec{H}_{T\xi}(z) = u_\xi^f \vec{h}_{T\xi}\, \mathrm{e}^{-\gamma_\xi z} - u_\xi^b \vec{h}_{T\xi}\, \mathrm{e}^{+\gamma_\xi z}, \tag{2.50}$$

where γ_ξ, u_ξ^f and u_ξ^b stand for the propagation constant, and the complex amplitude of the forward and backward traveling waves, respectively. Furthermore, the transversal fields of two different waveforms $(\vec{e}_\xi, \vec{h}_\xi)$ and $(\vec{e}_\zeta, \vec{h}_\zeta)$ satisfy the orthogonality condition [18]

$$\int_{\Gamma_\vartheta} (\vec{e}_{T\xi} \times \vec{h}_{T\zeta}) \cdot \hat{e}_z d\Gamma = 0. \tag{2.51}$$

Hence, defining the equivalent modal voltage V_ξ and equivalent modal current I_ξ

$$V_\xi = u_\xi^b + u_\xi^f, \tag{2.52}$$

$$I_\xi = u_\xi^b - u_\xi^f, \tag{2.53}$$

the electromagnetic field $(\vec{E}_T, \vec{H}_T)$ on Γ_ϑ, i.e. the transverse field at $z = 0$ in the local waveguide coordinates, may be written as the expansion

$$\vec{E}_T = \sum_{\xi=1}^{N_\xi} \vec{E}_{T\xi} = \sum_{\xi=1}^{N_\xi} V_\xi \vec{e}_{T\xi}, \text{ with } N_\xi \to \infty, \tag{2.54}$$

$$\vec{H}_T = \sum_{\xi=1}^{N_\xi} \vec{H}_{T\xi} = \sum_{\xi=1}^{N_\xi} I_\xi \vec{h}_{T\xi}, \text{ with } N_\xi \to \infty. \tag{2.55}$$

Network Parameter Evaluation

The introduction of complex modal amplitudes of traveling waves and equivalent voltages V_ζ and currents I_ξ, allows us to determine network parameters for the modeled microwave device [19]. The generalized impedance $Z_{\zeta\xi}$ is defined to be the quotient of V_ζ and I_ξ

$$Z_{\zeta\xi} = \left.\frac{V_\zeta}{I_\xi}\right|_{I_\nu=0,\,\forall\,\nu\neq\xi}, \tag{2.56}$$

where only one single non-zero current is set as excitation. Thus, we have

$$\underbrace{\begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ Z_{N1} & \dots & \dots & Z_{NN} \end{bmatrix}}_{\mathbf{Z}} \begin{bmatrix} I_1 \\ I_2 \\ \vdots \\ I_N \end{bmatrix} = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_N \end{bmatrix}, \tag{2.57}$$

where **Z** denotes the generalized impedance matrix. Therefore, for the boundary value problem only the transverse magnetic field $\vec{h}_{T\xi}$ of one single modal waveform is employed as excitation on Γ_Θ

$$\vec{H}_{T\Theta} = \vec{h}_{T\xi}, \tag{2.58}$$

which means to plug $\vec{h}_{T\xi}$ into (2.38). Hence, the boundary value problem solution, i.e. the electric field, is denoted as $\vec{E}(\vec{h}_{T\xi})$ and in particular, the electric field on the boundary Γ_Θ is denoted as $\vec{E}_T(\vec{h}_{T\xi})$. Plugging $\vec{E}_T(\vec{h}_{T\xi})$ into the orthogonality relation (2.51) and using the expansion (2.54), we have

$$\begin{aligned}
\int_{\Gamma_\Theta} (\vec{E}_T(\vec{h}_{T\xi}) \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma &= \int_{\Gamma_\vartheta} (\vec{E}_T(\vec{h}_{T\xi}) \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma && \text{(2.59a)}\\
&= \int_{\Gamma_\vartheta} \left(\left(\sum_{\nu=1}^{N_\xi} \vec{E}_{T\nu}(\vec{h}_{T\xi}) \right) \times \vec{h}_{T\zeta} \right) \cdot \hat{e}_n d\Gamma && \text{(2.59b)}\\
&= \sum_{\nu=1}^{N_\xi} \left(\int_{\Gamma_\vartheta} \left(\vec{E}_{T\nu}(\vec{h}_{T\xi}) \times \vec{h}_{T\zeta} \right) \cdot \hat{e}_n d\Gamma \right) && \text{(2.59c)}\\
&= \int_{\Gamma_\vartheta} (\vec{E}_{T\zeta}(\vec{h}_{T\xi}) \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma. && \text{(2.59d)}
\end{aligned}$$

Rewriting the transverse field with its amplitude, i.e. $\vec{E}_{T\zeta}(\vec{h}_{T\xi}) = V_\zeta(\vec{h}_{T\xi})\vec{e}_{T\zeta}$, we have due to linearity

$$V_{T\zeta}(\vec{h}_{T\xi}) = V_{T\zeta}(\frac{1}{I_\xi}\vec{H}_{T\xi}) = \frac{1}{I_\xi}V_{T\zeta}(\vec{H}_{T\xi}). \tag{2.60}$$

This linearity property together with (2.48), allows us to evaluate the generalized impedance

$$\begin{aligned}
\int_{\Gamma_\vartheta} (\vec{E}_{T\zeta}(\vec{h}_{T\xi}) \times \vec{h}_{T\zeta}) \cdot \vec{e}_n d\Gamma &= \int_{\Gamma_\vartheta} (V_\zeta(\vec{h}_{T\xi})\vec{e}_{T\zeta} \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma && \text{(2.61a)}\\
&= V_\zeta(\vec{h}_{T\xi}) \int_{\Gamma_\vartheta} (\vec{e}_{T\zeta} \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma && \text{(2.61b)}\\
&= \frac{V_\zeta(\vec{H}_{T\xi})}{I_\xi} \int_{\Gamma_\vartheta} (\vec{e}_{T\zeta} \times \vec{h}_{T\zeta}) \cdot \hat{e}_n d\Gamma && \text{(2.61c)}\\
&= \frac{V_\zeta(\vec{H}_{T\xi})}{I_\xi} && \text{(2.61d)}\\
&= Z_{\zeta\xi}, && \text{(2.61e)}
\end{aligned}$$

where the definition (2.56) is used.

An alternative way to describe microwave devices, which is probably even more common, are generalized scattering parameters, defined as

$$S_{\zeta\xi} = \left. \frac{u_\zeta^f}{u_\xi^b} \right|_{u_\nu^b=0,\,\forall\,\nu\neq\xi}. \tag{2.62}$$

Thus, we have

$$\underbrace{\begin{bmatrix} S_{11} & S_{12} & \dots & S_{1N} \\ S_{21} & S_{22} & & \vdots \\ \vdots & & \ddots & \vdots \\ S_{N1} & \dots & \dots & S_{NN} \end{bmatrix}}_{\mathbf{S}} \underbrace{\begin{bmatrix} u_1^b \\ u_2^b \\ \vdots \\ u_N^b \end{bmatrix}}_{\mathbf{u}^b} = \underbrace{\begin{bmatrix} u_1^f \\ u_2^f \\ \vdots \\ u_N^f \end{bmatrix}}_{\mathbf{u}^f}, \tag{2.63}$$

where **S** denotes the generalized scattering matrix. Once the generalized impedance matrix is available, the generalized scattering matrix is obtained by

$$\mathbf{u}^f + \mathbf{u}^b = \mathbf{Z}(\mathbf{u}^b - \mathbf{u}^f), \tag{2.64a}$$

$$(\mathbf{Z} - \mathbf{I})\mathbf{u}^b = (\mathbf{Z} + \mathbf{I})\mathbf{u}^f, \tag{2.64b}$$

$$\mathbf{u}^f = (\mathbf{Z} + \mathbf{I})^{-1}(\mathbf{Z} - \mathbf{I})\mathbf{u}^b, \tag{2.64c}$$

$$\mathbf{S} = (\mathbf{Z} + \mathbf{I})^{-1}(\mathbf{Z} - \mathbf{I}). \tag{2.64d}$$

2.3 Finite Element Simulation

The purpose of the FE method is to approximate the solution of the boundary value problem (2.38), i.e. the electric field $\vec{E}$, in a finite-dimensional subspace of $\mathbf{H}(\mathbf{curl};\Omega,\Gamma_E)$. For the discretization of the domain Ω, a triangulation $\mathcal{T}_h(\Omega)$ into elements of polyhedral geometry is employed, where Ω is assumed to be a bounded domain. The electric field is thus discretized in the global FE space

$$\mathcal{W}(\mathcal{T}_h(\Omega)) \subset \mathbf{H}(\mathbf{curl};\Omega,\Gamma_E). \tag{2.65}$$

The FE shape functions in the FE space are denoted as

$$\vec{w}_j \in \mathcal{W}(\mathcal{T}_h(\Omega)), \tag{2.66}$$

where for all numerical calculations in this thesis the hierarchical basis functions of [20] are used and a triangulation into elements of tetrahedral geometry is employed. In the global FE space, basis functions which belong to the wave port are separated for further work

$$\mathcal{W}(\mathcal{T}_h(\Omega)) = \mathcal{W}_T \oplus \mathcal{W}_\Theta, \tag{2.67}$$

where

$$\mathcal{W}_T := \{\vec{w} \in \mathcal{W}(\mathcal{T}_h(\Omega)) \mid \hat{e}_n \times \vec{w} = 0 \text{ on } \Gamma_\Theta\}, \tag{2.68a}$$
$$\mathcal{W}_\Theta := \{\vec{w} \in \mathcal{W}(\mathcal{T}_h(\Omega)) \mid \hat{e}_n \times \vec{w} \neq 0 \text{ on } \Gamma_\Theta\}. \tag{2.68b}$$

Thus, the electric field intensity $\vec{E}$ is discretized in the expansion

$$\vec{E} = \vec{E}_I + \vec{E}_\Theta, \tag{2.69}$$

with

$$\vec{E}_I = \sum_{j=1}^{N_I} u_j \vec{w}_j, \qquad \vec{w}_j \in \mathcal{W}_T, \tag{2.70a}$$
$$\vec{E}_\Theta = \sum_{j=1}^{N_\Theta} u_j \vec{w}_j, \qquad \vec{w}_j \in \mathcal{W}_\Theta, \tag{2.70b}$$

where u_j are the coefficients of the basis functions. N_I and N_Θ denote the number of degrees of freedom associated with $\mathcal{W}_T$ and $\mathcal{W}_\Theta$, respectively. Plugging the electric field expansion (2.69) into the weak form of the boundary value problem (2.38), the FE discretization results in the system of linear equations

$$(\mathbf{A}_0 + j\eta_0 k_0 \mathbf{A}_1 - k_0^2 \mathbf{A}_2)\mathbf{x} = -j\eta_0 k_0 \mathbf{b}, \tag{2.71}$$

where $k_0 = \omega\sqrt{\varepsilon_0\mu_0}$ is the wavenumber in the free space and $\eta_0 = \sqrt{\frac{\mu_0}{\varepsilon_0}}$. The structures of the matrices in the system of linear equations appear as

$$\mathbf{A}_0 = \begin{bmatrix} \mathbf{A}^\nu_{TT} & \mathbf{A}^\nu_{T\Theta} \\ \mathbf{A}^\nu_{\Theta T} & \mathbf{A}^\nu_{\Theta\Theta} \end{bmatrix} \in \mathbb{C}^{N_f \times N_f}, \qquad \mathbf{A}_0^T = \mathbf{A}_0, \tag{2.72a}$$
$$\mathbf{A}_1 = \begin{bmatrix} \mathbf{A}^z & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{C}^{N_f \times N_f}, \qquad \mathbf{A}_1^T = \mathbf{A}_1, \tag{2.72b}$$
$$\mathbf{A}_2 = \begin{bmatrix} \mathbf{A}^\varepsilon_{TT} & \mathbf{A}^\varepsilon_{T\Theta} \\ \mathbf{A}^\varepsilon_{\Theta T} & \mathbf{A}^\varepsilon_{\Theta\Theta} \end{bmatrix} \in \mathbb{C}^{N_f \times N_f}, \qquad \mathbf{A}_2^T = \mathbf{A}_2, \tag{2.72c}$$

where $N_f = N_I + N_\Theta$, and the matrix entries are defined with the bilinear forms

$$[A^{\nu}_{TT}]_{ij} = \int_\Omega \nabla\times\vec{w}_i\cdot\mu_r^{-1}\nabla\times\vec{w}_j d\Omega, \qquad \forall\ \vec{w}_i,\vec{w}_j\in\mathcal{W}_T, \tag{2.73a}$$

$$[A^{\nu}_{T\Theta}]_{ij} = \int_\Omega \nabla\times\vec{w}_i\cdot\mu_r^{-1}\nabla\times\vec{w}_j d\Omega, \qquad \forall\ \vec{w}_i\in\mathcal{W}_T,\ \vec{w}_j\in\mathcal{W}_\Theta, \tag{2.73b}$$

$$[A^{\nu}_{\Theta\Theta}]_{ij} = \int_{\Gamma_\Theta} \nabla\times\vec{w}_i\cdot\mu_r^{-1}\nabla\times\vec{w}_j d\Gamma, \qquad \forall\ \vec{w}_i,\vec{w}_j\in\mathcal{W}_\Theta, \tag{2.73c}$$

$$[A^{z}]_{ij} = \int_{\Gamma_Z} \frac{1}{Z_s}(\hat{e}_n\times\vec{w}_i)\cdot(\hat{e}_n\times\vec{w}_j)d\Gamma, \qquad \forall\ \vec{w}_i,\vec{w}_j\in\mathcal{W}_T, \tag{2.73d}$$

$$[A^{\varepsilon}_{TT}]_{ij} = \int_\Omega \vec{w}_i\cdot\varepsilon_r\vec{w}_j d\Omega, \qquad \forall\ \vec{w}_i,\vec{w}_j\in\mathcal{W}_T, \tag{2.73e}$$

$$[A^{\varepsilon}_{T\Theta}]_{ij} = \int_\Omega \vec{w}_i\cdot\varepsilon_r\vec{w}_j d\Omega, \qquad \forall\ \vec{w}_i\in\mathcal{W}_T,\ \vec{w}_j\in\mathcal{W}_\Theta, \tag{2.73f}$$

$$[A^{\varepsilon}_{\Theta\Theta}]_{ij} = \int_{\Gamma_\Theta} \vec{w}_i\cdot\varepsilon_r\vec{w}_j d\Gamma, \qquad \forall\ \vec{w}_i,\vec{w}_j\in\mathcal{W}_\Theta. \tag{2.73g}$$

For the right-hand side vector $\mathbf{b}$ we have

$$\mathbf{b} = \begin{bmatrix} \mathbf{b}_T \\ \mathbf{b}_\Theta \end{bmatrix} \in \mathbb{C}^{N_f}, \tag{2.74}$$

with

$$[b_T]_i = 0, \qquad \forall\ \vec{w}_i\in\mathcal{W}_T, \tag{2.75a}$$

$$[b_\Theta]_i = \int_{\Gamma_\Theta} (\vec{w}_i\times\vec{H}_{T\Theta})\cdot\hat{e}_n d\Gamma, \qquad \forall\ \vec{w}_i\in\mathcal{W}_\Theta. \tag{2.75b}$$

Setting the coefficients $u_j = m_{\xi j}$, for $j = 1,\dots,N_\Theta$, such that the electric field expansion on the wave port area in (2.70b) form the transverse electric field of the modal waveform $(\vec{e}_\xi, \vec{h}_\xi)$, we have

$$\vec{e}_{T\xi} = \sum_{j=1}^{N_\Theta} m_{\xi j}\vec{w}_j, \tag{2.76}$$

and define the coefficient vector

$$\mathbf{m}_\xi^T = [m_{\xi 1}, m_{\xi 2}, \dots,\ m_{\xi N_\Theta}]. \tag{2.77}$$

Furthermore, we define the matrix

$$\mathbf{M} = [\mathbf{m}_1,\ \mathbf{m}_2,\dots,\ \mathbf{m}_{N_\xi}] \in \mathbb{C}^{N_\Theta\times N_\xi}, \tag{2.78}$$

where each column is associated with a modal waveform. In an approach similar to the transfinite element method [2], the degrees of freedom on Γ_Θ in the system (2.71) are reduced, which results in

$$(\mathbf{A}_{M0} + jk_0\mathbf{A}_{M1} - k_0^2\mathbf{A}_{M2})\mathbf{x}_M = -jk_0\eta_0\mathbf{b}_M, \tag{2.79}$$

where

$$\mathbf{A}_{M0} = \begin{bmatrix} \mathbf{A}^{\nu}_{TT} & \mathbf{A}^{\nu}_{T\Theta}\mathbf{M} \\ \mathbf{M}^T\mathbf{A}^{\nu}_{\Theta T} & \mathbf{M}^T\mathbf{A}^{\nu}_{\Theta\Theta}\mathbf{M} \end{bmatrix} \in \mathbb{C}^{N_r \times N_r}, \tag{2.80a}$$

$$\mathbf{A}_{M1} = \begin{bmatrix} \mathbf{A}^{\nu}_{ZZ} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{C}^{N_r \times N_r}, \tag{2.80b}$$

$$\mathbf{A}_{M2} = \begin{bmatrix} \mathbf{A}^{\varepsilon}_{TT} & \mathbf{A}^{\varepsilon}_{T\Theta}\mathbf{M} \\ \mathbf{M}^T\mathbf{A}^{\varepsilon}_{\Theta T} & \mathbf{M}^T\mathbf{A}^{\varepsilon}_{\Theta\Theta}\mathbf{M} \end{bmatrix} \in \mathbb{C}^{N_r \times N_r}, \tag{2.80c}$$

$$\mathbf{b}_M = \begin{bmatrix} \mathbf{b}_T \\ \mathbf{M}^T\mathbf{b}_\Theta \end{bmatrix} \in \mathbb{C}^{N_r}, \mathbf{b}_T = \mathbf{0}, \tag{2.80d}$$

and $N_r = N_I + N_\xi$.

For the network parameter evaluation only one single mode is excited on Γ_Θ, which means to set $\vec{H}_{T\Theta} = \vec{h}_{T\xi}$ in (2.38). Hence, in (2.75b) we set

$$[b_\xi]_i = \int_{\Gamma_\Theta} (\vec{w}_i \times \vec{h}_{T\xi}) \cdot \hat{e}_n d\Gamma, \qquad \forall\ \vec{w}_i \in \mathcal{W}_\Theta. \tag{2.81}$$

and denote the right-hand side excitation with

$$\mathbf{b}_{M\xi} = \begin{bmatrix} \mathbf{b}_T \\ \mathbf{M}^T\mathbf{b}_\xi \end{bmatrix} \in \mathbb{C}^{N_r}, \mathbf{b}_T = \mathbf{0}. \tag{2.82}$$

The coefficient vector $\mathbf{x}_M(\vec{h}_{T\xi})$ denotes the solution of an excitation with the modal waveform $(\vec{e}_\xi, \vec{h}_\xi)$. The generalized impedance parameter evaluation is done by the output functional

$$Z_{\zeta\xi} := Z_\zeta(\mathbf{x}_M(\vec{h}_{T\xi})) = \mathbf{b}^T_{M\zeta}\mathbf{x}_M(\vec{h}_{T\xi}). \tag{2.83}$$

This can be verified by writing

$$\begin{aligned}
\mathbf{b}^T_{M\zeta}\mathbf{x}_M(\vec{h}_{T\xi}) &= \mathbf{b}^T_\zeta \mathbf{M}\mathbf{x}_M(\vec{h}_{T\xi}) && (2.84\text{a}) \\
&= \sum_{\nu=1}^{N_\xi} \mathbf{b}^T_\zeta \mathbf{m}_\nu V_\nu(\vec{h}_{T\xi}) && (2.84\text{b}) \\
&= \sum_{\nu=1}^{N_\xi} V_\nu(\vec{h}_{T\xi}) \left(\sum_{i=1}^{N_\Theta} m_{\nu i} \int_{\Gamma_\Theta} \vec{w}_i \times \vec{h}_{T\zeta} \cdot \hat{e}_n d\Gamma \right) && (2.84\text{c}) \\
&= \sum_{\nu=1}^{N_\xi} V_\nu(\vec{h}_{T\xi}) \int_{\Gamma_\Theta} \left(\sum_{i=1}^{N_\Theta} m_{\nu i}\vec{w}_i \right) \times \vec{h}_{T\zeta} \cdot \hat{e}_n d\Gamma && (2.84\text{d}) \\
&= \sum_{\nu=1}^{N_\xi} V_\nu(\vec{h}_{T\xi}) \int_{\Gamma_\Theta} \vec{e}_{T\nu} \times \vec{h}_{T\zeta} \cdot \hat{e}_n d\Gamma && (2.84\text{e}) \\
&= \sum_{\nu=1}^{N_\xi} \left(\int_{\Gamma_\Theta} \vec{E}_{T\nu}(\vec{h}_{T_\xi}) \times \vec{h}_{T\zeta} \cdot \hat{e}_n d\Gamma \right), && (2.84\text{f})
\end{aligned}$$

where the last line can be plugged into (2.59).

2.4 System Parameterization

For the MORe process in this thesis the frequency is chosen as parameter, actually the wavenumber k_0, which allows us to evaluate fast frequency sweeps. The parameterization of the linear system of equations is dependent on the considered structure. To discuss this, the FE system together with its excitations and output functionals is denoted as

$$(\sum_{i=0}^{a} s^i \mathbf{A}_i)\mathbf{X}(s) = s^t \mathbf{B}\mathbf{u}(s), \qquad \mathbf{A}_i \in \mathbb{C}^{N_r \times N_r}, \tag{2.85a}$$

$$\mathbf{Y}(s) = \mathbf{B}^T \mathbf{X}(s), \qquad \mathbf{B} \in \mathbb{C}^{N_r \times N_\xi}, \tag{2.85b}$$

where we use s as parameter, a for the polynomial degree and t is the parameter exponent of the right-hand side. A device which is modeled without any impedance boundary Γ_Z has only a linear parameter dependency, $a = 1$, where the parameter is the substituted square of the wavenumber, $s = k_0^2$. An absorbing boundary condition causes a linear wavenumber dependency, thus we have $s = k_0$ and $a = 2$. However, in case of modeled surface losses, we need to use the square root of the wavenumber as parameter, i.e. $s = \sqrt{k_0}$, and the polynomial degree rises to $a = 4$.

In the FE simulation of passive microwave structures, impedance as well as scattering formulations are commonly used. While the introduced approach is based on an impedance formulation, the above cited original transfinite element formulation [2] was introduced in a scattering formulation. This means, that the amplitudes u_ξ^b are used in the excitation and u_ξ^f fill the solution vector, which directly makes the scattering parameters available. However, these approaches result in different parameterizations, which is explicitly discussed in [21], also in context with MORe techniques.

Another issue is the modal excitation on the boundary Γ_Θ. While the exciting field on a fixed frequency may be evaluated, analytically or numerically, the frequency dependencies on the boundaries need to be considered for a system parameterization. As long as excitations with non-varying modal field patterns are employed, i.e. TE-, TM- or TEM-Modes, scaling approaches as [22] are suitable. If however the modal field patterns are frequency dependent, the exciting fields need to be found on each evaluation point.

Chapter 3

Adaptive Model Order Reduction Methods

3.1 Preliminary Words on Model Order Reduction

The FE method is a well-established technique for solving driven time-harmonic field problems. Since FE matrices are sparse, the resulting systems of linear equations, though large in size, can be solved very efficiently by direct or iterative methods. However, FE matrices depend on the working frequency, and hence computing the system response over a wide frequency range turns out to be very expensive, because it involves FE solutions at a large number of frequency points.

To overcome this difficulty, which arises also in large-scale circuit simulations, methods of MORe have been developed [7]. Amongst the techniques available, projection-based MORe methods as [8], [23], [24] are particularly attractive, because they are well-suited for large-scale systems and constitute Petrov-Galerkin methods, as the FE method itself. The underlying idea is to restrict the FE solution to a carefully chosen subspace and apply a (Petrov)-Galerkin method to reduce the original problem to a ROM of low dimension. Single-point methods, such as [25], [9], construct the projection matrix or matrices from a Krylov expansion about one frequency point. For smaller problems, single-point methods are very attractive, because the FE matrix needs to be factorized only once: all Krylov vectors required are generated by forward-back substitutions. However, the quality of the ROM depends strongly on the expansion frequency, the optimal location of which is a priori unknown. Moreover, for large-scale problems, matrix factorizations are prohibitively expensive. Then, (semi)-iterative solvers [26] must be employed, and the cost of computing one Krylov vector becomes comparable to that of a full FE run. In this situation, multi-point methods [27], [28], which utilize FE solutions at multiple frequencies to build the projection matrix, become very attractive, because they offer flexibility in choosing the expansion frequencies and great numerical robustness.

In this chapter, the mathematical background knowledge for MORe is presented. Furthermore, adaptive MORe algorithms, for both single-point and multi-point techniques are introduced. Numerical experiments, employing some error measures, compare the adaptive multi-point technique to single-point methods and show the following: The dimension of a ROM to reach a given error limit, is significantly smaller, if the adaptive multi-point algorithm is employed, even if the expansion frequency for the single-point method is chosen at its optimum. Hence, the adaptive multi-point method is superior to the best possible single-point method with regard to memory requirements for generating the ROM.

Another argument for employing multi-point methods is the lower dimension of the resulting ROM. There are applications that require very large numbers of ROM evaluations, such as repetitive calls to time-domain recovery algorithms in circuit simulators, utilizing ROMs as library elements, or stochastic optimization methods employing multi-variate ROMs [29] for computing cost functions. For such purposes, it is very desirable to minimize the size of the ROM because, in contrast to the original FE matrices, ROM matrices are full, and matrix factorization costs grow proportionally to the third power of the ROM dimension.

3.2 Introduction to Moment Matching

Single-point methods employ one expansion point, which means that the solution of the large-scale system is only needed at one single frequency point for the MORe process. These methods are based on the moment matching principle, which is closely related to Krylov subspace methods. In this section, explicit moment matching [7] as well as the projection-based implicit moment matching [8], [23] principle is introduced. Furthermore, MORe techniques for first-order systems will be shown and extended in the end of the section to higher-order systems.

3.2.1 Explicit Moment Matching

For the introduction of the moment matching principle, the single-input single-output system of first-order

$$(\mathbf{A}_0 + s\mathbf{A}_1)\mathbf{x}(s) = \mathbf{b}u(s), \qquad \mathbf{A}_0, \mathbf{A}_1 \in \mathbb{C}^{N\times N}, \tag{3.1a}$$

$$y(s) = \mathbf{c}^T\mathbf{x}(s), \qquad \mathbf{b}, \mathbf{c}, \mathbf{x} \in \mathbb{C}^N, \tag{3.1b}$$

is considered, where s is the employed parameter. The transfer function of this system takes the form

$$H(s) = \mathbf{c}^T(\mathbf{A}_0 + s\mathbf{A}_1)^{-1}\mathbf{b}. \tag{3.2}$$

By employing a Taylor expansion, at expansion point s_0, we have

$$H(s) = \sum_{i=0}^{\infty} \frac{1}{i!}\frac{\mathrm{d}^i H(s)}{\mathrm{d}s^i}\Big|_{s=s_0}(s-s_0)^i, \tag{3.3}$$

which allows us to define the ith moment as

$$\mu_i = \frac{1}{i!}\frac{\mathrm{d}^i H(s)}{\mathrm{d}s^i}\Big|_{s=s_0}. \tag{3.4}$$

In the following, the expansion point $s_0 = 0$ is employed for simplicity. However, for all following techniques an expansion point $s_0 \neq 0$ can be chosen, which causes only a simple substitution.

Assuming the matrix $\mathbf{A}_0$ in the system (3.1) non-singular, we can write

$$H(s) = \mathbf{c}^T\left(\mathbf{A}_0(\mathbf{I} + s\mathbf{A}_0^{-1}\mathbf{A}_1)\right)^{-1}\mathbf{b} = \mathbf{c}^T(\mathbf{I} + s\mathbf{A}_0^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_0^{-1}\mathbf{b}. \tag{3.5}$$

A basic approach to make the moments μ_i explicitly available is to employ a Neumann series expansion

$$(\mathbf{I} - (-s\mathbf{A}_0^{-1}\mathbf{A}_1))^{-1} = \sum_{n=0}^{\infty}(-s\mathbf{A}_0^{-1}\mathbf{A}_1)^n. \tag{3.6}$$

Hence, we write the transfer function as

$$H(s) = \sum_{i=0}^{\infty} \mu_i s^i, \tag{3.7}$$

where the moments μ_i yield

$$\mu_i = \mathbf{c}^T(-\mathbf{A}_0^{-1}\mathbf{A}_1)^i\mathbf{A}_0^{-1}\mathbf{b}. \tag{3.8}$$

Thus, the transfer function $H(s)$ can be approximated by the transfer function $H_k(s)$, which matches in the first k moments

$$H(s) \approx H_k(s) = \sum_{i=0}^{k-1} \mu_i s^i. \tag{3.9}$$

One of the main goals of this first explicit approach is to find a rational function $H_k(s)$, which matches the first k moments of the transfer function $H(s)$. Therefore, a Padé approximation is employed to match the first $k = 2n$ moments of the function

$$H_{2n}(s) = \frac{a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s + a_0}{b_n s^n + b_{n-1}s^{n-1} + \cdots + b_1 s + 1}, \tag{3.10}$$

with the moments μ_i of the transfer function $H(s)$. explicitly available. This is nothing more than writing

$$\sum_{i=0}^{2n-1} \mu_i s^i = \frac{a_{n-1}s^{n-1} + a_{n-2}s^{n-2} + \cdots + a_1 s + a_0}{b_n s^n + b_{n-1}s^{n-1} + \cdots + b_1 s + 1}, \tag{3.11}$$

and choose the coefficients in (3.10) such that they match the first $2n$ moments. Comparing the coefficients for the resulting polynomials in s, the coefficients b_i can be evaluated through solving the system of linear equations

$$\begin{bmatrix} \mu_0 & \mu_1 & \cdots & \mu_{n-1} \\ \mu_1 & \mu_2 & \cdots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_{n-1} & \mu_n & \cdots & \mu_{2n-2} \end{bmatrix} \begin{bmatrix} b_n \\ b_{n-1} \\ \vdots \\ b_1 \end{bmatrix} = \begin{bmatrix} \mu_n \\ \mu_{n+1} \\ \vdots \\ \mu_{2n-1} \end{bmatrix}, \tag{3.12}$$

and the coefficients a_i follow from the recursion

$$\begin{aligned} a_0 &= \mu_0 \\ a_1 &= \mu_1 + b_1\mu_0 \\ &\vdots \\ a_{n-1} &= \mu_{n-1} + \sum_{i=1}^{n-1} b_i \mu_{n-i-1}. \end{aligned} \tag{3.13}$$

This explicit moment matching technique was introduced in [7] as Asymptotic Waveform Evaluation (AWE) and employed for circuit system timing analysis. Unfortunately, the practical application of the AWE is limited due to a numerical deficit. The recursive computation of $(\mathbf{A}_0^{-1}\mathbf{A}_1)^i\mathbf{A}_0^{-1}\mathbf{b}$ converges to that eigenvector of the matrix $\mathbf{A}_0^{-1}\mathbf{A}_1$, which corresponds to the eigenvalue with largest absolute value. The limited precision of floating point numbers in numerical computation leads to a loss of information during this recursive computation, which causes the wrong convergence. Hence, only a small number of moments are properly matched and therefore, the reduced transfer function is only accurate in a small bandwidth around the expansion point. Extending this single-point approach to a multi-point AWE technique was employed in [30] to overcome this difficulty. A final remark is that the AWE technique can also be employed for polynomial parameterized system matrices and right-hand sides [31].

3.2.2 Implicit Moment Matching

Krylov Subspace Methods

The explicit moment representation in (3.8) may be written as

$$\mu_i = \mathbf{c}^T \mathbf{v}_i, \tag{3.14}$$

where $\mathbf{v}_i$ are the Krylov vectors defined as

$$\mathbf{v}_i = \mathbf{P}^i \mathbf{u}, \tag{3.15}$$

with $\mathbf{P} = -\mathbf{A}_0^{-1}\mathbf{A}_1$ and $\mathbf{u} = \mathbf{A}_0^{-1}\mathbf{b}$. This property motivates us to employ Krylov subspace methods to generate numerically stable moments for the MORe process.

One of the first approaches is the Padé via Lanczos process [8], where a Lanczos algorithm-based method was introduced to improve moment matching. As the name of the algorithm already claims, the algorithm generates a Padé approximation by employing the Lanczos algorithm, which computes the moments numerically stable. The resulting transfer function matches the first $2q$ moments, as the AWE process introduced in the section above.

Another Krylov subspace-based MORe technique was introduced in [23], where instead of the Lanczos iteration an Arnoldi algorithm, Alg. 1, was employed. As a result of the single sided-process, only the first q moments of the resulting transfer function are matched in this process.

Algorithm 1 Arnoldi algorithm

```
$\mathbf{v}_1 = \mathbf{x}/\|\mathbf{x}\|_2$
for $n = 1$ to $q-1$ do
    $\mathbf{v} = \mathbf{A}\mathbf{v}_n$
    // Modified Gram-Schmidt process
    for $j = 1$ to $n-1$ do
        $h_{j,n} = \mathbf{v}_j^H \mathbf{v}$
        $\mathbf{v} = \mathbf{v} - h_{j,n}\mathbf{v}_j$
    end for
    $h_{n+1,n} = \|\mathbf{v}\|_2$
    $\mathbf{v}_{n+1} = \mathbf{v}/h_{n+1,n}$
end for
```

The Krylov vectors $\{\mathbf{v}_0, \ldots, \mathbf{v}_{q-1}\}$ span the qth Krylov subspace, defined as

$$\mathcal{K}_q(\mathbf{P}, \mathbf{u}) = \operatorname{span}\{\mathbf{u}, \mathbf{P}\mathbf{u}, \mathbf{P}^2\mathbf{u}, \ldots, \mathbf{P}^{q-1}\mathbf{u}\}. \tag{3.16}$$

This space can be computed in a numerically stable way by the above introduced Arnoldi algorithm. Projecting the original system (3.1) to the Krylov subspace $\mathcal{K}_q(\mathbf{P}, \mathbf{u})$, with $q \ll N$, results in a ROM, which implicitly matches moments. In particular, building the projection matrix $\mathbf{V} \in \mathbb{C}^{N \times q}$, which spans the qth Krylov subspace

$$\operatorname{span}(\mathbf{V}) = \mathcal{K}_q(\mathbf{P}, \mathbf{u}), \tag{3.17}$$

allows us to compute the low-dimensional system

$$(\tilde{\mathbf{A}}_0 + s\tilde{\mathbf{A}}_1)\tilde{\mathbf{x}}(s) = \tilde{\mathbf{b}}u(s), \qquad \tilde{\mathbf{A}}_0, \tilde{\mathbf{A}}_1 \in \mathbb{C}^{q \times q} \tag{3.18a}$$

$$y(s) = \tilde{\mathbf{c}}^T \tilde{\mathbf{x}}(s), \qquad \tilde{\mathbf{b}}, \tilde{\mathbf{c}}, \tilde{\mathbf{x}} \in \mathbb{C}^q \tag{3.18b}$$

with

$$\tilde{\mathbf{A}}_0 = \mathbf{V}^T\mathbf{A}_0\mathbf{V},\ \tilde{\mathbf{A}}_1 = \mathbf{V}^T\mathbf{A}_1\mathbf{V},\ \tilde{\mathbf{b}} = \mathbf{V}^T\mathbf{b} \text{ and } \tilde{\mathbf{c}} = \mathbf{V}^T\mathbf{c}.$$

As the reduced system (3.18) takes the same structure as the original system (3.1), the transfer function yields

$$\tilde{H}(s) = \tilde{\mathbf{c}}^T(\tilde{\mathbf{A}}_0 + s\tilde{\mathbf{A}}_1)^{-1}\tilde{\mathbf{b}}, \tag{3.19}$$

and its Taylor expansion is derived equivalently as

$$\tilde{H}(s) = \sum_{i=0}^{\infty} \tilde{\mu}_i s^i, \tag{3.20}$$

where the moments μ_i in the series are

$$\tilde{\mu}_i = \tilde{\mathbf{c}}^T(-\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_1)^i\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{b}}. \tag{3.21}$$

The first q moments of the original system (3.1) and the ROM (3.18) match

$$\tilde{\mu}_i = \mu_i,\ \forall\ 0 \leq i < q. \tag{3.22}$$

Proof for Moment Matching

The representation above allows us to proof moment matching by induction [32]. For $i = 0$ we know

$$\exists\ \mathbf{r}_0 \in \mathbb{C}^q :\ \mathbf{A}_0^{-1}\mathbf{b} = \mathbf{V}\mathbf{r}_0, \tag{3.23}$$

and thus

$$\begin{aligned}
\tilde{\mu}_0 &= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b} && (3.24)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T(\mathbf{A}_0\mathbf{A}_0^{-1})\mathbf{b} && (3.25)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{r}_0 && (3.26)\\
&= \mathbf{c}^T\mathbf{V}\mathbf{r}_0 = \mathbf{c}^T\mathbf{A}_0^{-1}\mathbf{b} = \mu_0. && (3.27)
\end{aligned}$$

For $i = 1$ we have

$$\exists\ \mathbf{r}_1 \in \mathbb{C}^q :\ -\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{b} = \mathbf{V}\mathbf{r}_1, \tag{3.28}$$

but we know

$$\mathbf{r}_0 = \left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b}. \tag{3.29}$$

Thus,

$$\begin{aligned}
\tilde{\mu}_1 &= \tilde{\mathbf{c}}^T(-\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_1)\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{b}} && (3.30)\\
&= -\mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b} && (3.31)\\
&= -\mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_1\mathbf{V}\mathbf{r}_0 && (3.32)\\
&= -\mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T(\mathbf{A}_0\mathbf{A}_0^{-1})\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{b} && (3.33)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{r}_1 && (3.34)\\
&= \mathbf{c}^T\mathbf{V}\mathbf{r}_1 = \mathbf{c}^T(-\mathbf{A}_0^{-1}\mathbf{A}_1)\mathbf{A}_0^{-1}\mathbf{b} = \mu_1. && (3.35)
\end{aligned}$$

Note that

$$\mathbf{r}_1 = -\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b}. \tag{3.36}$$

For $2 \le i \le q-1$ we have

$$\exists\ \mathbf{r}_i \in \mathbb{C}^q : (-\mathbf{A}_0^{-1}\mathbf{A}_1)^i\mathbf{A}_0^{-1}\mathbf{b} = \mathbf{V}\mathbf{r}_i. \tag{3.37}$$

Thus, with the hypothesis

$$\mathbf{r}_{i-1} = \left(-\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\right)^{i-1}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b}, \tag{3.38}$$

the moments $2 \le i \le q-1$ of the reduced model match to the original system

$$\begin{aligned}
\tilde{\mu_i} &= \tilde{\mathbf{c}}^T(-\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_1)^i\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{b}} && (3.39)\\
&= \mathbf{c}^T\mathbf{V}\left(-\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\right)^i\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b} && (3.40)\\
&= -\mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_1\mathbf{V}\mathbf{r}_{i-1} && (3.41)\\
&= -\mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T(\mathbf{A}_0\mathbf{A}_0^{-1})\mathbf{A}_1(-\mathbf{A}_0^{-1}\mathbf{A}_1)^{i-1}\mathbf{A}_0^{-1}\mathbf{b} && (3.42)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{r}_i && (3.43)\\
&= \mathbf{c}^T\mathbf{V}\mathbf{r}_i = \mathbf{c}^T(-\mathbf{A}_0^{-1}\mathbf{A}_1)^{-1}\mathbf{A}_0^{-1}\mathbf{b} = \mu_i, && (3.44)
\end{aligned}$$

with

$$\mathbf{r}_i = \left(-\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\right)^i\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{b}. \tag{3.45}$$

3.2.3 Systems with Polynomial Parameterized System Matrix

Higher-Order Systems

In literature, systems with a polynomial parameterized system matrix are often referred to higher-order systems [13], [33]. Second-order systems are discussed more frequently, as damping terms appear in modeling many classes of physical systems [34], [35]. However, in this section, the focus is on general higher-order single-input single-output systems with a polynomially parameterized system matrix

$$\begin{aligned}
(\sum_{i=0}^{a}\mathbf{A}_i s^i)\mathbf{x}(s) &= \mathbf{b}u(s), && (3.46a)\\
y(s) &= \mathbf{c}^T\mathbf{x}(s), && (3.46b)
\end{aligned}$$

where a is an arbitrary finite integer. The transfer function for this system yields

$$H(s) = \mathbf{c}^T(\sum_{i=0}^{a}\mathbf{A}_i s^i)^{-1}\mathbf{b}, \tag{3.47}$$

which can be written as

$$H(s) = \mathbf{c}^T(\mathbf{I} - \sum_{i=1}^{a}\mathbf{D}_i s^i)^{-1}\mathbf{u}, \tag{3.48}$$

with $\mathbf{D}_i = -\mathbf{A}_0^{-1}\mathbf{A}_i$ and $\mathbf{u} = \mathbf{A}_0^{-1}\mathbf{b}$.

Similar to the Neumann series, the matrix inversion can be expanded around $s = 0$ as

$$(\mathbf{I} - \sum_{i=1}^{a} \mathbf{D}_i s^i)^{-1} = \sum_{i=0}^{\infty} \mathbf{P}_i s^i, \tag{3.49}$$

where

$$\mathbf{P}_0 = \mathbf{I} \tag{3.50}$$

$$\mathbf{P}_k = \sum_{i=1}^{\min(k,a)} \mathbf{P}_{k-i}\mathbf{D}_i. \tag{3.51}$$

This is proved in [33] and can be verified by explicitly writing the expansion. Thus, as in the first-order system, the transfer function can be expressed by a Taylor expansion

$$H(s) = \sum_{i=0}^{\infty} \mathbf{c}^T \mathbf{P}_i \mathbf{u} s^i = \sum_{i=0}^{\infty} \mu_i s^i, \tag{3.52}$$

where the ith moment is given as

$$\mu_i = \mathbf{c}^T \mathbf{P}_i \mathbf{u}. \tag{3.53}$$

Higher-Order Krylov Subspaces

For the sequence of matrices $\{\mathbf{D}_i\}_{i=1}^{a}$, $\mathbf{D}_i \in \mathbb{C}^{N\times N}$, and the non-zero vector $\mathbf{u} \in \mathbb{C}^N$, the qth Krylov subspace of ath-order is defined as

$$\mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^{a}; \mathbf{u}) = \operatorname{span}\{\mathbf{w}_0, \mathbf{w}_1, \ldots, \mathbf{w}_{q-1}\}, \tag{3.54}$$

where the recursive definition of the vectors $\mathbf{w}_l$ yield

$$\mathbf{w}_0 = \mathbf{u}, \tag{3.55}$$

$$\mathbf{w}_l = \sum_{i=1}^{\min(l,a)} \mathbf{D}_i \mathbf{w}_{l-i}. \tag{3.56}$$

This definition is from [33], but is in accordance with definitions for second-order Krylov subspaces introduced in [34], [35]. Furthermore, employing the definitions above for $\mathbf{P}_i$ allows an alternative recursive description for the higher-order Krylov vectors

$$\mathbf{w}_i = \mathbf{P}_i \mathbf{u}. \tag{3.57}$$

This approach is also in accordance with the properties introduced for higher-order systems. Hence, the moments can now be written as

$$\mu_i = \mathbf{c}^T \mathbf{w}_i, \tag{3.58}$$

with the Krylov vectors $\mathbf{w}_i$.

Another interesting approach to higher-order Krylov subspaces is to employ the AWE technique [29], [31]. Here, the solution vector is expanded in a Taylor series

$$\mathbf{x} = \sum_{i=0}^{\infty} \tilde{\mathbf{w}}_i s^i \tag{3.59}$$

and plugged into the system (3.46b). Then, by equating the coefficients, the vectors $\tilde{\mathbf{w}}_i$ yield

$$\tilde{\mathbf{w}}_0 = \mathbf{A}_0^{-1}\mathbf{b}, \tag{3.60}$$

$$\tilde{\mathbf{w}}_1 = -\mathbf{A}_0^{-1}\mathbf{A}_1\tilde{\mathbf{w}}_0, \tag{3.61}$$

$$\tilde{\mathbf{w}}_2 = \mathbf{A}_0^{-1}(-\mathbf{A}_1\tilde{\mathbf{w}}_1 - \mathbf{A}_2\tilde{\mathbf{w}}_0), \tag{3.62}$$

$$\vdots \tag{3.63}$$

$$\tilde{\mathbf{w}}_i = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,i)}\mathbf{A}_m\tilde{\mathbf{w}}_{i-m}). \tag{3.64}$$

It is straightforward to see that $\mathbf{w}_i = \tilde{\mathbf{w}}_i$ for $\mathbf{D}_i = -\mathbf{A}_0^{-1}\mathbf{A}_i$ and $\mathbf{u} = \mathbf{A}_0^{-1}\mathbf{b}$, as defined above.

A final approach to higher-order Krylov subspaces is by rewriting the higher-order system as an equivalent first-order system. This allows us to show the connection of the resulting first-order Krylov subspace of the linearized system and the higher-order subspace [24], [36].

Reducing the Higher-Order System

Employing for the reduction of the system (3.46) the projection matrix $\mathbf{V} \in \mathbb{C}^{N\times q}$, which spans the qth Krylov subspace of ath-order

$$\mathrm{span}(\mathbf{V}) = \mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a; \mathbf{u}), \tag{3.65}$$

the low-dimensional system yields

$$(\sum_{i=0}^{a}\tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{x}}(s) = \tilde{\mathbf{b}}u(s), \tag{3.66a}$$

$$\tilde{y}(s) = \tilde{\mathbf{c}}^T\tilde{\mathbf{x}}(s), \tag{3.66b}$$

where

$$\tilde{\mathbf{A}}_i = \mathbf{V}^T\mathbf{A}_i\mathbf{V},\ \tilde{\mathbf{b}} = \mathbf{V}^T\mathbf{b}, \tilde{\mathbf{c}} = \mathbf{V}^T\mathbf{c}. \tag{3.67}$$

The transfer function for the ROM results in

$$\tilde{H}(s) = \tilde{\mathbf{c}}^T(\sum_{i=0}^{a}\tilde{\mathbf{A}}_i s^i)^{-1}\tilde{\mathbf{b}}, \tag{3.68}$$

which is

$$\tilde{H}(s) = \tilde{\mathbf{c}}^T(\mathbf{I} - \sum_{i=1}^{a}\tilde{\mathbf{D}}_i s^i)^{-1}\tilde{\mathbf{u}}, \tag{3.69}$$

with $\tilde{\mathbf{D}}_i = -\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_i$ and $\tilde{\mathbf{u}} = \tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{b}}$. Again the property (3.49) is employed to write

$$(\mathbf{I} - \sum_{i=1}^{a}\tilde{\mathbf{D}}_i s^i)^{-1} = \sum_{i=0}^{\infty}\tilde{\mathbf{P}}_i s^i, \tag{3.70}$$

where

$$\tilde{\mathbf{P}}_0 = \mathbf{I}, \tag{3.71}$$

$$\tilde{\mathbf{P}}_k = \sum_{i=1}^{\min(k,a)} \tilde{\mathbf{P}}_{k-i}\tilde{\mathbf{D}}_i. \tag{3.72}$$

Thus, the transfer function expressed with moments yields

$$\tilde{H}(s) = \sum_{i=0}^{\infty} \tilde{\mathbf{c}}^T \tilde{\mathbf{P}}_i \tilde{\mathbf{u}} s^i = \sum_{i=0}^{\infty} \tilde{\mu}_i s^i, \tag{3.73}$$

where

$$\tilde{\mu}_i = \tilde{\mathbf{c}}^T \tilde{\mathbf{P}}_i \tilde{\mathbf{u}}. \tag{3.74}$$

The first q moments of the original system (3.46) and the ROM (3.66) match

$$\tilde{\mu}_i = \mu_i, \ \forall\ 0 \leq i < q. \tag{3.75}$$

Proof for Moment Matching

The proof is for a fixed integer $a > 1$. The moment matching for $\tilde{\mu}_0 = \mu_0$ and $\tilde{\mu}_1 = \mu_1$ is identical to Section 3.2.2, respectively, because $\mathbf{P}_0 = \mathbf{I}$ and $\mathbf{P}_1 = \mathbf{D}_1 = -\mathbf{A}_0^{-1}\mathbf{A}_1$.

For $i = 2$ we have

$$\exists\ \mathbf{r}_2 \in \mathbb{C}^q :\ \mathbf{P}_2\mathbf{u} = ((-\mathbf{A}_0^{-1}\mathbf{A}_1)^2 - \mathbf{A}_0^{-1}\mathbf{A}_2)\mathbf{A}_0^{-1}\mathbf{b} = \mathbf{V}\mathbf{r}_2, \tag{3.76}$$

Thus,

$$\tilde{\mu}_2 = \tilde{\mathbf{c}}^T \left(\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_1\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_1 - \tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{A}}_2\right)\tilde{\mathbf{A}}_0^{-1}\tilde{\mathbf{b}} \tag{3.77}$$

$$= \mathbf{c}^T\mathbf{V}\left(-\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{V}\right)\mathbf{r}_1 - \left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_2\mathbf{V}\right)\mathbf{r}_0\right) \tag{3.78}$$

$$= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{b} - \mathbf{V}^T\mathbf{A}_2\mathbf{A}_0^{-1}\mathbf{b}\right) \tag{3.79}$$

$$= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{A}_0^{-1}\left(\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{b} - \mathbf{A}_2\mathbf{A}_0^{-1}\mathbf{b}\right) \tag{3.80}$$

$$= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\left(\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{A}_1\mathbf{A}_0^{-1}\mathbf{b} - \mathbf{A}_0^{-1}\mathbf{A}_2\mathbf{A}_0^{-1}\mathbf{b}\right) \tag{3.81}$$

$$= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{P}_2\mathbf{u} \tag{3.82}$$

$$= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{r}_2 \tag{3.83}$$

$$= \mathbf{c}^T\mathbf{V}\mathbf{r}_2 = \mathbf{c}^T\mathbf{P}_2\mathbf{u} = \mu_2 \tag{3.84}$$

Note that

$$\mathbf{r}_2 = \tilde{\mathbf{P}}_2\tilde{\mathbf{u}}. \tag{3.85}$$

For $2 < i < q$ we know

$$\exists\ \mathbf{r}_i \in \mathbb{C}^q :\ \mathbf{P}_i\mathbf{u} = \mathbf{V}\mathbf{r}_i. \tag{3.86}$$

With the hypothesis

$$\mathbf{r}_{i-k} = \tilde{\mathbf{P}}_{i-k}\tilde{\mathbf{u}},\ \forall\, 0 < k < \min(i,a) \tag{3.87}$$

we have

$$\begin{aligned}
\tilde{\mu}_i &= \tilde{\mathbf{c}}^T\tilde{\mathbf{P}}_i\tilde{\mathbf{u}} &(3.88)\\
&= \tilde{\mathbf{c}}^T\left(\sum_{k=1}^{\min(i,a)}\tilde{\mathbf{P}}_{i-k}\tilde{\mathbf{D}}_k\right)\tilde{\mathbf{u}} &(3.89)\\
&= \tilde{\mathbf{c}}^T\left(\sum_{k=1}^{\min(i,a)}\tilde{\mathbf{D}}_k\tilde{\mathbf{P}}_{i-k}\right)\tilde{\mathbf{u}} &(3.90)\\
&= \tilde{\mathbf{c}}^T\left(\sum_{k=1}^{\min(i,a)}\tilde{\mathbf{D}}_k\mathbf{r}_{i-k}\right) &(3.91)\\
&= \mathbf{c}^T\mathbf{V}\left(\sum_{k=1}^{\min(i,a)} -\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\left(\mathbf{V}^T\mathbf{A}_k\mathbf{V}\right)\mathbf{r}_{i-k}\right) &(3.92)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\left(\sum_{k=1}^{\min(i,a)} -\mathbf{A}_k\mathbf{V}\mathbf{r}_{i-k}\right) &(3.93)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\left(\sum_{k=1}^{\min(i,a)} -\mathbf{A}_0^{-1}\mathbf{A}_k\mathbf{V}\mathbf{r}_{i-k}\right) &(3.94)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\left(\sum_{k=1}^{\min(i,a)} -\mathbf{A}_0^{-1}\mathbf{A}_k\mathbf{P}_{i-k}\right)\mathbf{u} &(3.95)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{P}_i\mathbf{u} &(3.96)\\
&= \mathbf{c}^T\mathbf{V}\left(\mathbf{V}^T\mathbf{A}_0\mathbf{V}\right)^{-1}\mathbf{V}^T\mathbf{A}_0\mathbf{V}\mathbf{r}_i &(3.97)\\
&= \mathbf{c}^T\mathbf{V}\mathbf{r}_i = \mathbf{c}^T\mathbf{P}_i\mathbf{u} = \mu_i, &(3.98)
\end{aligned}$$

and thus

$$\mathbf{r}_i = \tilde{\mathbf{P}}_i\tilde{\mathbf{u}}. \tag{3.99}$$

Well-Conditioned AWE

Although the AWE was shown to generate higher-order Krylov vectors, and thus span higher-order Krylov subspaces, the technique exhibits similar numerical deficits as explained for first-order systems. This limits the numerically precise matched moments in the reduction process to a low number. However, the WCAWE process, introduced in [24], provides a means to span higher-order Krylov subspaces in a numerically stable way. The iterative process is given in Alg. 2 and refers to a polynomial parameterized system of linear equations of the form

$$\left(\sum_{i=0}^{a}\mathbf{A}_i s^i\right)\mathbf{x}(s) = \sum_{i=0}^{a_1}\mathbf{b}_i s^i. \tag{3.100}$$

The WCAWE algorithm needs the following additional definitions. The columns of the matrix $\tilde{\mathbf{V}}_q$ are the iteratively computed vectors

$$\tilde{\mathbf{V}}_q = [\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2, \ldots, \tilde{\mathbf{v}}_q]. \tag{3.101}$$

The matrix $\mathbf{U}$ is a non-singular upper triangular matrix which relates the matrices $\tilde{\mathbf{V}}_q$ and $\mathbf{V}_q$ by

$$\tilde{\mathbf{V}}_q \mathbf{U}^{-1} = \mathbf{V}_q,\ \mathbf{U} \in \mathbb{C}^{q \times q}, \tag{3.102}$$

where

$$\mathbf{V}_q = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_q]. \tag{3.103}$$

In the algorithm, the correction matrix $\mathbf{P}_{\mathbf{U}_w}(n,m)$, is computed as

$$\mathbf{P}_{\mathbf{U}_w}(n,m) = \prod_{t=w}^{m} \mathbf{U}^{-1}_{[t:n-m+t-1,t:n-m+t-1]},\ \mathbf{P}_{\mathbf{U}_w}(n,m) \in \mathbb{C}^{n-m \times n-m}. \tag{3.104}$$

The vectors $\mathbf{e}_i$ are defined as unity vectors with the ith entry set to one and all others equal to zero. The lengths of $\mathbf{e}_i$ conforms on the matrix that operates on it.

The coefficients in the matrix $\mathbf{U}$, i.e. the relation of $\tilde{\mathbf{V}}_q$ and $\mathbf{V}_q$, is an important part of the method. Employing the modified Gram-Schmidt process to evaluate (3.102) is claimed to result in high accuracy [9], where the resulting columns of $\mathbf{V}_q$ are orthonormal. The properties of other choices for (3.102) can be found in the referred paper.

The proof for the WCAWE algorithm in [24] shows, that each new computed vector $\mathbf{v}_q$ is a superposition of the higher-order Krylov vectors

$$\mathbf{v}_q = \sum_{i=0}^{q-1} \mathbf{w}_i x_i, \tag{3.105}$$

where the coefficients x_i result from the algorithm.

Algorithm 2 Well-conditioned AWE

```
$\tilde{\mathbf{v}}_1 = \mathbf{A}_0^{-1}\mathbf{b}_0$
$\tilde{\mathbf{v}}_2 = \mathbf{A}_0^{-1}(\mathbf{b}_1\mathbf{e}_1^T\mathbf{P}_{\mathbf{U}_1}(2,1)\mathbf{e}_1 - \mathbf{A}_1\mathbf{v}_1)$
$\tilde{\mathbf{v}}_3 = \mathbf{A}_0^{-1}(\mathbf{b}_1\mathbf{e}_1^T\mathbf{P}_{\mathbf{U}_1}(3,1)\mathbf{e}_2 + \mathbf{b}_2\mathbf{e}_1^T\mathbf{P}_{\mathbf{U}_1}(3,2)\mathbf{e}_1 - \mathbf{A}_1\mathbf{v}_2 - \mathbf{A}_2\mathbf{V}_1\mathbf{P}_{\mathbf{U}_2}(3,2)\mathbf{e}_1)$
    $\vdots$
$\tilde{\mathbf{v}}_q = \mathbf{A}_0^{-1}(\sum_{m=1}^{\min(a_1,q-1)} \mathbf{b}_m\mathbf{e}_1^T\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{e}_{q-m} - \mathbf{A}_1\mathbf{v}_{q-1} - \sum_{m=2}^{\min(a,q-1)}\mathbf{A}_m\mathbf{V}_{q-m}\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{e}_{q-m})$
```

3.3 Adaptivity and Error Measures

Projection-Based Model Order Reduction

To keep the basic idea of the introduced method as general as possible, the considered system takes the form

$$(\sum_{i=0}^{a_0} s^i \mathbf{A}_i)\mathbf{X}(s) = \sum_{i=0}^{a_1} s^i \mathbf{B}_i, \qquad \mathbf{A}_i \in \mathbb{C}^{N\times N}, \tag{3.106a}$$

$$\mathbf{Y} = (\sum_{i=0}^{a_2} s^i \mathbf{C}_i)^T \mathbf{X}(s), \qquad \mathbf{B}_i \in \mathbb{C}^{N\times b}, \mathbf{C}_i \in \mathbb{C}^{N\times c}, \tag{3.106b}$$

where N, b and c denote the number of unknowns, the number of input vectors and output vectors, respectively. The numbers a_0, a_1 and a_2 give the polynomial degrees of the parameterization. The parameter s stands for the wavenumber k, or a substitution as explained in Section 2.4.

The projection-based MORe methods considered in this thesis have in common that they seek an approximation $\mathbf{Q}\tilde{\mathbf{X}}$ to the full solution $\mathbf{X}$ in the range of a low-dimensional unitary matrix $\mathbf{Q} \in \mathbb{C}^{N\times u}$ with $u \ll N$. Furthermore, they employ a Bubnov-Galerkin process to reduce the original system (3.106) to the ROM of the form

$$(\sum_{i=0}^{a_0} s^i \tilde{\mathbf{A}}_i)\tilde{\mathbf{X}}(s) = \sum_{i=0}^{a_1} s^i \tilde{\mathbf{B}}_i, \tag{3.107a}$$

$$\tilde{\mathbf{Y}} = (\sum_{i=0}^{a_2} s^i \tilde{\mathbf{C}}_i)^T \tilde{\mathbf{X}}(s), \tag{3.107b}$$

where

$$\tilde{\mathbf{A}}_i = \mathbf{Q}^T \mathbf{A}_i \mathbf{Q} \in \mathbb{C}^{u\times u}, \tag{3.108}$$

$$\tilde{\mathbf{B}}_i = \mathbf{Q}^T \mathbf{B}_i \in \mathbb{C}^{u\times b}, \tag{3.109}$$

$$\tilde{\mathbf{C}}_i = \mathbf{Q}^T \mathbf{C}_i \in \mathbb{C}^{u\times c}, \tag{3.110}$$

which can be solved at very low cost.

Adaptive Process

In the following, we focus on the iterative enlargement of the projection matrix, as well as on the corresponding enlargement of the ROM. The processes will be employed in the adaptive algorithms introduced later. Hence, assume that the matrix $\mathbf{Q} = \mathbf{Q}_{n-1}$ is employed to compute the ROM of the form (3.107). In the next iteration, a matrix $\mathbf{V}_{new} \in \mathbb{C}^{N\times b}$ is generated by the MORe technique, to enlarge the subspace in which the ROM is solved. Note that the columns of $\mathbf{V}_{new}$ probably neither provide an orthonormal basis nor do they have to be orthonormal to $\mathbf{Q}_n$. Therefore, the modified Gram-Schmidt orthonormalization process is employed for

$$[\mathbf{Q}_{n-1}, \mathbf{Q}_{new}]\mathbf{U} = [\mathbf{V}_{n-1}, \mathbf{V}_{new}], \tag{3.111}$$

where the columns of matrix $\mathbf{V}_{n-1}$ are the previously constructed vectors from the MORe technique. Only the new added vectors need to be orthonormalized. However, the projection matrix will be enlarged by

$$\mathbf{Q}_n = [\mathbf{Q}_{n-1}, \mathbf{Q}_{new}]. \tag{3.112}$$

The subspace projection can be an expensive process in the adaptive algorithm. Thus, to improve the performance, the ROM from the previous iteration is enlarged in each adaption step

$$\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \mathbf{Q}_{n-1}^T \mathbf{A}_i \mathbf{Q}_{new} \\ \mathbf{Q}_{new}^T \mathbf{A}_i \mathbf{Q}_{n-1} & \mathbf{Q}_{new}^T \mathbf{A}_i \mathbf{Q}_{new} \end{bmatrix}, \tag{3.113}$$

$$\tilde{\mathbf{B}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{B}}_i \\ \mathbf{Q}_{new}^T \mathbf{B}_i \end{bmatrix}, \tag{3.114}$$

$$\tilde{\mathbf{C}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{C}}_i \\ \mathbf{Q}_{new}^T \mathbf{C}_i \end{bmatrix}. \tag{3.115}$$

This results in a ROM of the form (3.107) of higher-dimension. Employing the proposed ROM enlargement procedure above reduces the MORe process runtime by not projecting the columns $\mathbf{Q}_{n-1}$ of the ROM, which are already available from the previous iterations.

Error Measure

The performance of the reduction methods as well as the accuracy of the generated ROMs is an important issue in this thesis. Therefore, error measures are introduced to make the quality of the reduced systems comparable, but are also employed in the later introduced adaptive MORe techniques.

We define the set of L equidistant evaluation points $\mathcal{B} = \{s_1, s_2, \ldots, s_L\}$, within the bounds $s_1 = s_{min}$ and $s_L = s_{max}$. For this set $\mathcal{B}$, a sequence of matrices $\{\mathbf{M}\}_{\mathcal{B}} = \{\mathbf{M}(s_1), \mathbf{M}(s_2), \ldots, \mathbf{M}(s_L)\}$ is defined. Hence, for two sequences $\{\mathbf{M}\}_{\mathcal{B}}$ and $\{\mathbf{N}\}_{\mathcal{B}}$, the error measure is defined as

$$E_2(\{\mathbf{M}\}_{\mathcal{B}}, \{\mathbf{N}\}_{\mathcal{B}}) = \sqrt{\frac{1}{N_f N_t^2} \sum_{n=1}^{L} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} |M_{ij}(s_n) - N_{ij}(s_n)|^2}. \tag{3.116}$$

In this thesis, we chose the scattering matrix at the evaluation point s_i as matrix $\mathbf{M}(s_i)$. Thus, the true error $E_2(\{\mathbf{S}\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_n\}_{\mathcal{B}})$, where $\{\mathbf{S}\}_{\mathcal{B}}$ are the scattering parameters computed by the full FE system and $\{\tilde{\mathbf{S}}_n\}_{\mathcal{B}}$ denotes the sweep of the ROM, allows us to evaluate the accuracy of the ROM. The subscript n stands for the number of iterations with which the adaptive MORe process was run. The error measure $E_2(\{\tilde{\mathbf{S}}_n\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{n-1}\}_{\mathcal{B}})$ therefore gives the differences of ROMs of nth and $(n-1)$th iteration. This definition is employed as termination criterion in the later provided adaptive MORe algorithms. Furthermore, the error measure $E_2(\{\tilde{\mathbf{S}}_n\}_{\mathcal{B}_q}, \{\tilde{\mathbf{S}}_{n-1}\}_{\mathcal{B}_q})$ allows us to define an error measure on the subinterval $\mathcal{B}_q \subseteq \mathcal{B}$, which is used to find the next expansion point in the adaptive multi-point method. The efficiency and reliability of the introduced error measure are shown in [28] and [37], where numerical experiments in the later reference also compare the new error measure to alternative approaches.

For some numerical experiments, additional error measures are defined as

$$E_1(\{\mathbf{M}\}_{\mathcal{B}}, \{\mathbf{N}\}_{\mathcal{B}}) = \frac{1}{N_f N_t^2} \sum_{n=1}^{L} \sum_{i=1}^{N_t} \sum_{j=1}^{N_t} \left| M_{ij}(s_n) - N_{ij}(s_n) \right|, \tag{3.117a}$$

$$E_\infty(\{\mathbf{M}\}_{\mathcal{B}}, \{\mathbf{N}\}_{\mathcal{B}}) = \max_{i,j,n} \left| M_{ij}(s_n) - N_{ij}(s_n) \right|. \tag{3.117b}$$

3.4 A Basic Adaptive Single-Point Method

3.4.1 Broadband FE Simulation of Electromagnetic Structures

Several approaches exist, where MORe methods are employed for the broadband FE simulation of electromagnetic structures, based on explicit moment matching [31], [38], [39] or implicit moment matching [25], [9]. In this section, an adaptive fast frequency sweep technique for the simulation of passive microwave structures is provided. Therefore, the FE system (2.85) is rewritten as

$$(\sum_{i=0}^{a} s^i \mathbf{A}_i)\mathbf{X}(s) = s^t\mathbf{B}, \qquad \mathbf{A}_i \in \mathbb{C}^{N\times N}, \mathbf{A}_i = \mathbf{A}_i^T \tag{3.118a}$$

$$\mathbf{Y} = \mathbf{B}^T\mathbf{X}(s), \qquad \mathbf{B} \in \mathbb{C}^{N\times b}, \mathbf{B} = [\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^b], \tag{3.118b}$$

where N and b denote the number of unknowns and the number of input/output vectors, respectively. A main property of the system (3.118) is its symmetry. Employing the projection matrix $\mathbf{Q} \in \mathbb{C}^{N\times q}$, the reduced system takes the form

$$(\sum_{i=0}^{a} s^i \tilde{\mathbf{A}}_i)\tilde{\mathbf{X}}(s) = s^t\tilde{\mathbf{B}}, \tag{3.119a}$$

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}}(s), \tag{3.119b}$$

where

$$\tilde{\mathbf{A}}_i = \mathbf{Q}^T\mathbf{M}_i\mathbf{Q}, \tag{3.120}$$

$$\tilde{\mathbf{B}} = \mathbf{Q}^T\mathbf{B}. \tag{3.121}$$

As the system is driven by b excitations, which are the columns $\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^b$, the Krylov subspace associated to each excitation has to be computed and plugged into the projection matrix $\mathbf{Q}$. Hence, the subspace spanned by $\mathbf{Q}$ is

$$\text{span}(\mathbf{Q}) = \mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a; \mathbf{u}^1) \cup \mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a; \mathbf{u}^2) \cup \ldots \cup \mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a; \mathbf{u}^b), \tag{3.122}$$

where $\mathbf{u}^j = \mathbf{A}_0^{-1}\mathbf{b}^j$ and again $\mathbf{D}_i = -\mathbf{A}_0^{-1}\mathbf{A}_i$.

Due to the symmetry of the system, it is straightforward to prove that the reduced model (3.119) matches in the first $2q$ moments to the full system (3.118), although only a single-sided projection is employed [40].

3.4.2 A First Adaptive Approach

Alg. 3 gives a basic adaptive single-point MORe method for the broadband FE simulation of electromagnetic structures. For the sake of simplicity, the system matrices $\mathbf{A}_i$, for $i = 0, \ldots, a$, represent the shifted matrices in the expansion point $\check{s}$. The method employs the WCAWE process, combined with a modified Gram-Schmidt orthonormalization. The computed WCAWE vectors for each right-hand side have to be orthonormalized against each other in Alg. 3, Line 5. Later in the adaptive loop, the vectors are also orthonormalized against the columns of the previous projection matrix in Alg. 3, Line 13. This results in the updated projection matrix $\mathbf{Q}_q$ which spans the subspace (3.122).

The right-hand side in the system has a purely linear wavenumber dependency, which only causes an index shift in the WCAWE process. This simplifies the WCAWE process, and only the correction matrices for the polynomial parametrized system matrices are required in Alg. 3, Line 10.

In each adaptive iteration, the projection matrix as well as the ROM are enlarged in Alg. 3, Line 13 and 14, respectively, as described in Section 3.3. The process stops as converged, if the error indicator in Alg. 3, Line 16 $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}})$ is below the threshold value $E_{2,tol}$. Otherwise, if $q = q_{max}$, the algorithm aborts with status 'not converged'.

Algorithm 3 Self-Adaptive Single-Point Model Order Reduction

```
for $\beta = 1$ to $b$ do
    Initial WCAWE: $\tilde{\mathbf{v}}_1^\beta = \tilde{\mathbf{V}}_1^\beta = \mathbf{A}_0^{-1}\mathbf{b}^\beta$,
    Normalize: $\mathbf{v}_1^\beta = \mathbf{V}_1^\beta = \tilde{\mathbf{V}}_1^\beta / |\tilde{\mathbf{V}}_1^\beta|$, $U^\beta = |\tilde{\mathbf{V}}_1^\beta|$
end for
Modified Gram-Schmidt: $\mathbf{Q}_1\mathbf{U} = [\mathbf{V}_1^1, \dots, \mathbf{V}_1^b]$
Initial subspace projection: $\tilde{\mathbf{A}}_i = \mathbf{Q}_1^T\mathbf{A}_i\mathbf{Q}_1$, $\tilde{\mathbf{B}} = \mathbf{Q}_1^T\mathbf{B}$
Solve frequency sweep:
$(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_1\}_{\mathcal{B}}$
for $q = 2$ to $q_{max}$ do
    for $\beta = 1$ to $b$ do
        WCAWE process:
        $\tilde{\mathbf{v}}_q^\beta = \mathbf{A}_0^{-1}(-\mathbf{A}_1\mathbf{v}_{q-1}^\beta - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m\mathbf{V}_{q-m}^\beta\mathbf{P}_{\mathbf{U}_2}^\beta(q,m)\mathbf{e}_{q-m})$
        Modified Gram-Schmidt: $[\mathbf{V}_{q-1}^\beta, \mathbf{v}_q^\beta]\mathbf{U}^\beta = [\tilde{\mathbf{V}}_{q-1}^\beta, \tilde{\mathbf{v}}_q^\beta]$
    end for
    Apply modified Gram-Schmidt only to new generated vectors:
    $[\mathbf{Q}_{q-1}, \mathbf{Q}_{new}]\mathbf{U} = [\mathbf{V}_{q-1}^1, \dots, \mathbf{V}_{q-1}^b, \mathbf{v}_q^1, \dots, \mathbf{v}_q^b]$ ,   $\mathbf{Q}_q = [\mathbf{Q}_{q-1}, \mathbf{Q}_{new}]$
    Enlarge ROM:
    $\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \mathbf{Q}_{q-1}^T\mathbf{A}_i\mathbf{Q}_{new} \\ \mathbf{Q}_{new}^T\mathbf{A}_i\mathbf{Q}_{q-1} & \mathbf{Q}_{new}^T\mathbf{A}_i\mathbf{Q}_{new} \end{bmatrix}$, $\tilde{\mathbf{B}} \leftarrow \begin{bmatrix} \tilde{\mathbf{B}} \\ \mathbf{Q}_{new}^T\mathbf{B} \end{bmatrix}$.
    Solve frequency sweep:
    $(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}$
    if $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}}) < E_{2,tol}$ then
        return(converged)
    end if
end for
```

3.5 An Adaptive Multi-Point Method

3.5.1 Projection-Based Model Order Reduction

In contrast to single-point approaches, multi-point methods employ system solutions at a set of expansion points $\{\check{s}_1, \ldots, \check{s}_M\}$ to construct the projection matrix $\mathbf{Q}$. Multi-point methods offer flexibility in choosing the expansion points and enjoy great numerical robustness. While it is possible to include higher-order Krylov vectors at each expansion point, as in [30] and [41], the algorithm proposed in the following only employs the system solutions themselves. This technique is called a rational Krylov method of lowest-order [42]. Each solution $\mathbf{X}(\check{s}_m)$ of the considered FE system of the form

$$(\sum_{i=0}^{a} s^i \mathbf{A}_i)\mathbf{X}(s) = s^t \mathbf{B}, \qquad \mathbf{A}_i \in \mathbb{C}^{N\times N}, \mathbf{A}_i = \mathbf{A}_i^T \tag{3.123a}$$

$$\mathbf{Y} = \mathbf{B}^T \mathbf{X}(s), \qquad \mathbf{B}_i \in \mathbb{C}^{N\times b}, \tag{3.123b}$$

spans the ath Krylov subspace of first-order. Thus, employing $\mathbf{X}(\check{s}_m)$ as projection matrix for the symmetric system above generates a ROM which matches in the first and second moment. In this thesis, an orthonormal projection matrix $\mathbf{Q}$, with

$$\mathrm{span}\,(\mathbf{Q}) = \mathrm{span}(\mathbf{X}(\check{s}_1), \mathbf{X}(\check{s}_2), \ldots, \mathbf{X}(\check{s}_M)), \tag{3.124}$$

is employed for the subspace projection, which results in the ROM

$$(\sum_{i=0}^{a} s^i \tilde{\mathbf{A}}_i)\tilde{\mathbf{X}}(s) = s^t \tilde{\mathbf{B}}, \tag{3.125a}$$

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T \tilde{\mathbf{X}}(s), \tag{3.125b}$$

where

$$\tilde{\mathbf{A}}_i = \mathbf{Q}^T \mathbf{A}_i \mathbf{Q}, \tag{3.126}$$

$$\tilde{\mathbf{B}}_i = \mathbf{Q}^T \mathbf{B}_i. \tag{3.127}$$

The employed projection matrix spans the first Krylov subspace at each of the M expansion points $\check{s}_m$. Thus, the first and second moment of the ROM and the original system match at each expansion point $\check{s}_m$.

3.5.2 Proposed Adaptive Algorithm

This thesis provides an adaptive multi-point algorithm, which employs the projection-based approach above and is listed in Alg. 4. The general strategy of the algorithm is to divide the bandwidth of interest, i.e. the set of evaluation points $\mathcal{B}$ within the bandwidth, into subintervals $\mathcal{B}_q \subseteq \mathcal{B}$. At each adaptive step, the interval of largest error $\mathcal{B}_{\check{e}}$ is identified with the help of the error indicator $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}_i}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}_i})$ in Line 18, which evaluates the differences between the scattering parameters obtained from the current and the previous ROM. The next adaptive expansion point $\check{s}_p$ is chosen at the center of $\mathcal{B}_{\check{e}}$, Line 9, and the interval is separated into two new subintervals in Line 10. This procedure is repeated until the error indicator $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}})$, is below the chosen threshold value $E_{2,tol}$.

Algorithm 4 Self-Adaptive Multi-Point Model Order Reduction

```
Solve system at $\check{s}_1 = \min(\mathcal{B})$:
$(\sum_{i=1}^{a} \mathbf{A}_i \check{s}_1^i)\mathbf{X}(\check{s}_1) = \check{s}_1^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_1)$
Solve system at $\check{s}_2 = \max(\mathcal{B})$:
$(\sum_{i=1}^{a} \mathbf{A}_i \check{s}_2^i)\mathbf{X}(\check{s}_2) = \check{s}_2^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_2)$
Initial modified Gram-Schmidt: $\mathbf{Q}_2\mathbf{U} = [\mathbf{X}(\check{s}_1), \mathbf{X}(\check{s}_2)]$
Initial subspace projection: $\tilde{\mathbf{A}}_i = \mathbf{Q}_2^T\mathbf{A}_i\mathbf{Q}_2$, $\tilde{\mathbf{B}} = \mathbf{Q_2}^T\mathbf{B}$
Solve frequency sweep for $s \in \mathcal{B}$:
$(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_2\}_{\mathcal{B}}$
Set interval of worst error: $\check{e} = 1$
Initialize first interval: $\mathcal{B}_{\check{e}} = \mathcal{B}$
for $q = 3$ to $q_{max}$ do
    Find next expansion point: $\check{s}_q = \arg\min_{s\in\mathcal{B}} \left|s - \frac{\min\mathcal{B}_{\check{e}}+\max\mathcal{B}_{\check{e}}}{2}\right|$
    Split interval: $\mathcal{B}_{q-1} = [\check{s}_q, \max(\mathcal{B}_{\check{e}})]$, $\mathcal{B}_{\check{e}} \leftarrow [\min(\mathcal{B}_{\check{e}}), \check{s}_q]$
    Solve system at $\check{s}_q$:
    $(\sum_{i=1}^{a} \mathbf{A}_i \check{s}_q^i)\mathbf{X}(\check{s}_q) = \check{s}_q^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_q)$
    Apply modified Gram-Schmidt only to new generated vectors:
    $[\mathbf{Q}_{q-1}, \mathbf{Q}_{new}]\mathbf{U} = [\mathbf{X}(\check{s}_1), \dots, \mathbf{X}(\check{s}_{q-1}), \mathbf{X}(\check{s}_q)]$,
    $\mathbf{Q}_q = [\mathbf{Q}_{q-1}, \mathbf{Q}_{new}]$
    Enlarge ROM:
    $\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \mathbf{Q}_{q-1}^T\mathbf{A}_i\mathbf{Q}_{new} \\ \mathbf{Q}_{new}^T\mathbf{A}_i\mathbf{Q}_{q-1} & \mathbf{Q}_{new}^T\mathbf{A}_i\mathbf{Q}_{new} \end{bmatrix}, \tilde{\mathbf{B}} \leftarrow \begin{bmatrix} \tilde{\mathbf{B}} \\ \mathbf{Q}_{new}^T\mathbf{B} \end{bmatrix}.$
    Solve frequency sweep:
    $(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}$
    if $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}}) < E_{2,tol}$ then
        return(converged)
    end if
    Find interval of worst error: $\check{e} = \arg\max_{i=1,\dots,q-1} E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}_i}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}_i})$
end for
```

3.6 Single-Point Methods and Multi-Point Methods in Comparison

As a practical example for the comparison of the MORe approaches of this chapter, the bandpass filter in Fig. 3.1 is considered, which is taken from [43]. To obtain reference results for the frequency response of the filter at high spectral resolution, individual FE solutions were computed at $N = 2001$ equidistant frequency points in the range from 4 to 12 GHz. This set of evaluation frequencies is denoted by $\mathcal{B}$. The FE model is based on basis functions of second-order, and the matrix dimension is $N = 103848$. Fig. 3.2 gives the magnitudes of the reflection and transmission coefficients s_{11} and s_{12} versus frequency. As a first numerical experiment, the true error of the adaptive multi-point method for ROMs up to iteration $q = 50$ is evaluated for the norms $E_1(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}})$, $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}})$ and $E_\infty(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}})$. Fig. 3.3 shows a steep descent around the iteration $q = 35$ for all norms employed. At $q = 37$, the ROM exhibits errors lower than 10^{-8} in all norms, which shows the high accuracy of the scattering parameters on the evaluation points in $\mathcal{B}$, compared to the large-scale FE model.

Next it is shown that the proposed adaptive multi-point technique needs lower dimension for high accuracy than single-point methods, even if the expansion point for the moment matching process is chosen at its optimum. To find the best available expansion frequency, which is a priori not known, ROMs with expansion frequencies in the range from 9 to 11 GHz are generated and the true error is evaluated in the same frequency range. The results of this process are presented in Fig. 3.4 and Fig. 3.5, respectively, which show the errors $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}})$ and $E_\infty(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}})$ for ROMs, build at iterations $q = 39$ to $q = 51$.

It can be seen that, even when the optimum expansion frequency $\check{f} = 10.15$ GHz is chosen for the single-point method, a ROM build at iteration $q = 51$ is required to yield results of similar error as the adaptive multi-point approach. Specifically, the single-point errors at $q = 51$ are $E_\infty(\{\tilde{\mathbf{S}}_n\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}}) \approx 6 \cdot 10^{-6}$ and $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\mathbf{S}\}_{\mathcal{B}}) \approx 5 \cdot 10^{-9}$; still worse than for the adaptive multi-point method with $q = 37$. When the iteration numbers of the single-point and adaptive multi-point models are both taken to be $q = 37$, the single-point method is clearly inferior, as can be seen from the errors $e_{11} = |\tilde{s}_{11} - s_{11}|$ and $e_{12} = |\tilde{s}_{12} - s_{12}|$, which are plotted versus frequency in Fig. 3.6. In addition, Fig. 3.6 shows the locations of the expansion frequencies for the adaptive multi-point method, indicated by the symbol ▼. Fig. 3.7 presents different norms of the single-point error for $q = 37$ and 401 different locations of the expansion frequency. This confirms that no single-point method can produce a ROM of similar accuracy as the adaptive multi-point model of same order.

Table 3.1 gives computer runtimes for ROM generation and evaluation. When ROMs of similar quality i.e., adaptive multi-point models with $q = 37$ and single-point models with $q = 51$, are compared, adaptive multi-point evaluation times are 39% shorter. On the other hand, for original system dimensions where the matrix factorizations are efficiently evaluated, model generation is faster with single-point methods, even for a ROM generated at $q = 51$.

Table 3.1: Bandpass filter [43]: Computational data.

		Generation	Evaluation
MOR method	Number of iterations q	LU factorization $N = 103\ 848$	L=2001
Single-point	37	67 s	1.1 s
Single-point	51	101 s	1.8 s
Multi-point	37	694 s	1.1 s

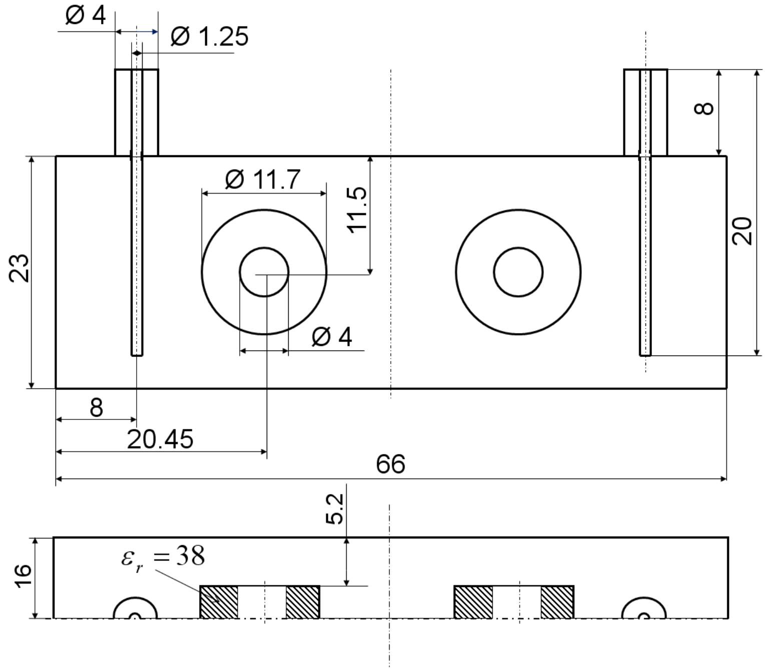


Figure 3.1: Bandpass filter [43]: Geometric dimensions in mm.

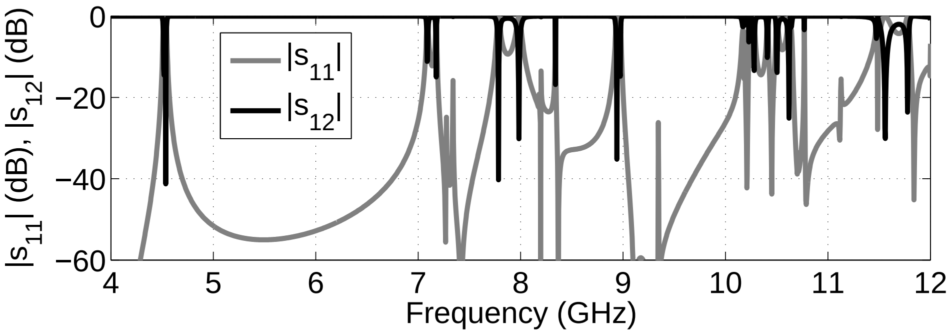


Figure 3.2: Bandpass filter [43]: magnitudes of s_{11} and s_{12} versus frequency.

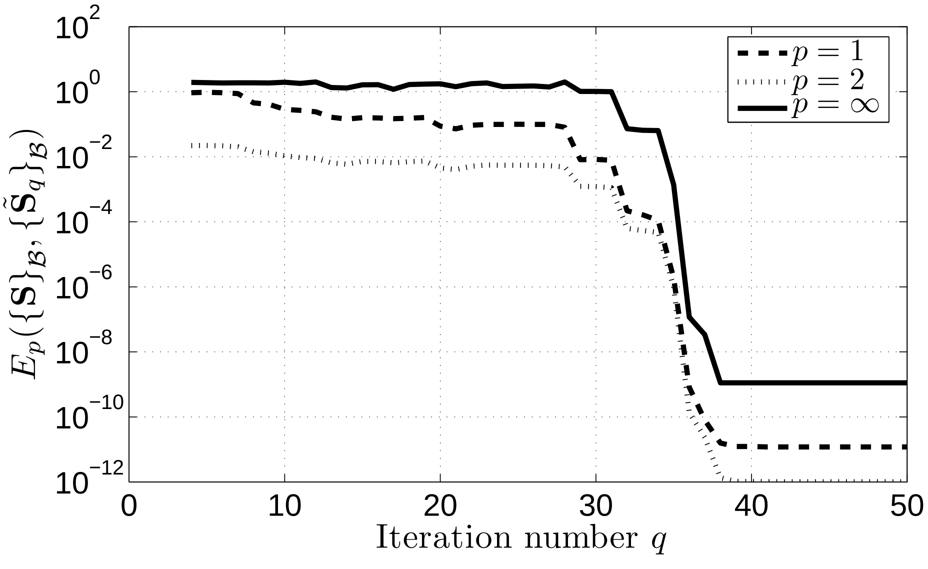


Figure 3.3: Bandpass filter [43]: True errors E_1, E_2 and E_∞ versus iteration number, multi-point method.

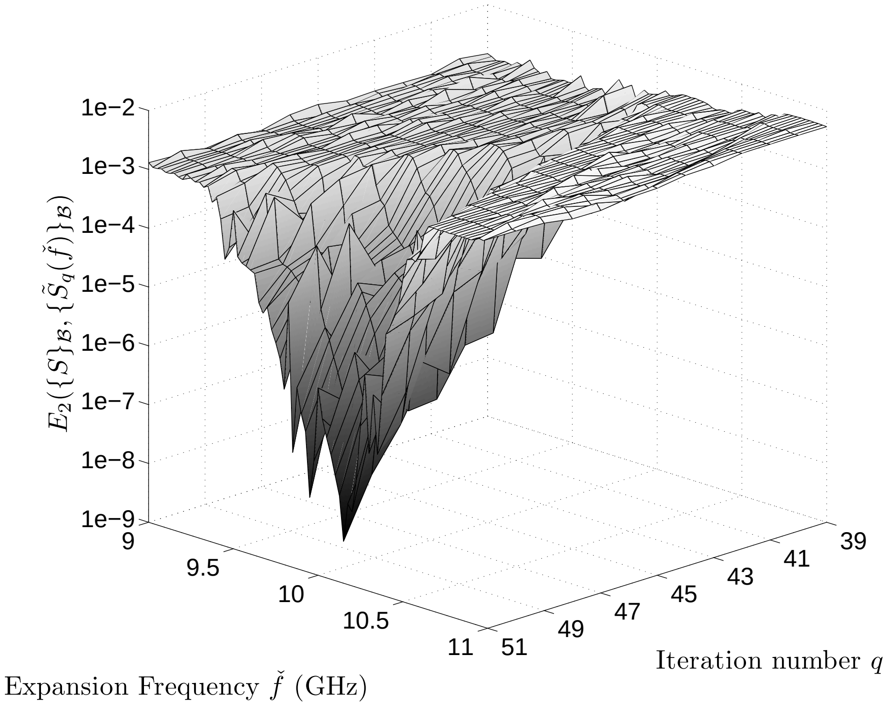


Figure 3.4: Bandpass filter [43]: True error E_2 versus iteration number and expansion frequency, single-point method.

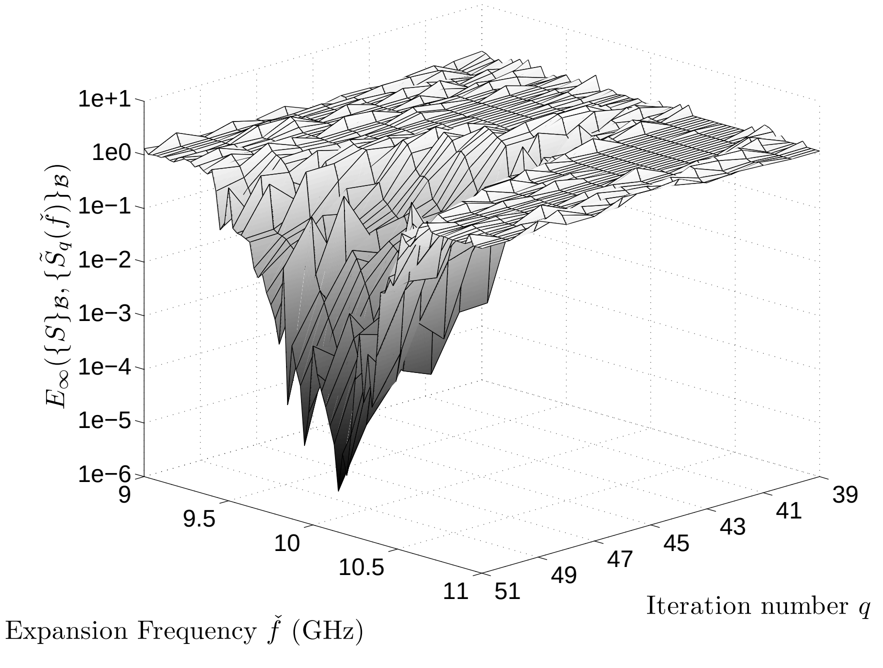


Figure 3.5: Bandpass filter [43]: True error E_∞ versus iteration number and expansion frequency, single-point method.

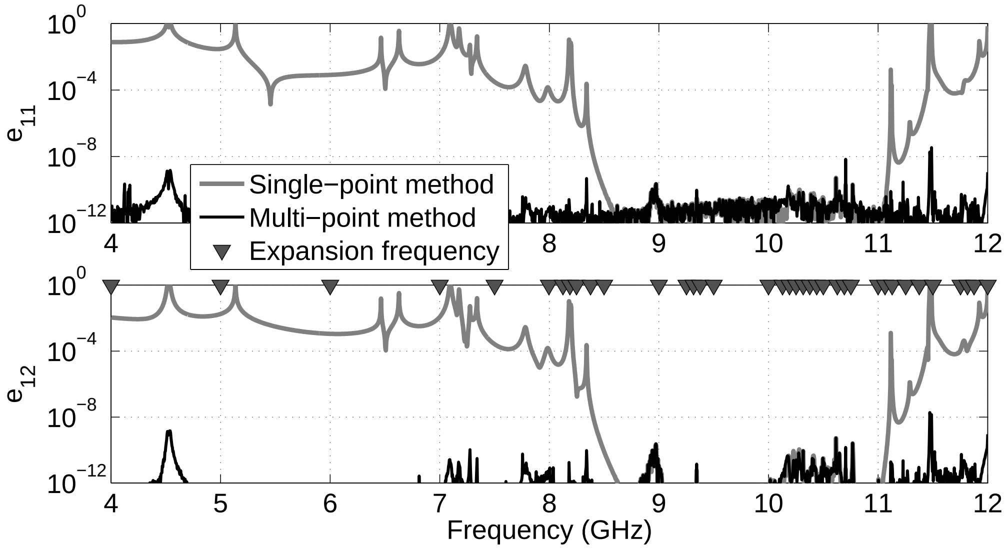


Figure 3.6: Bandpass filter [43]: True errors in scattering parameters versus frequency, ROMs with $q = 37$ of single-point and multi-point method in comparison.

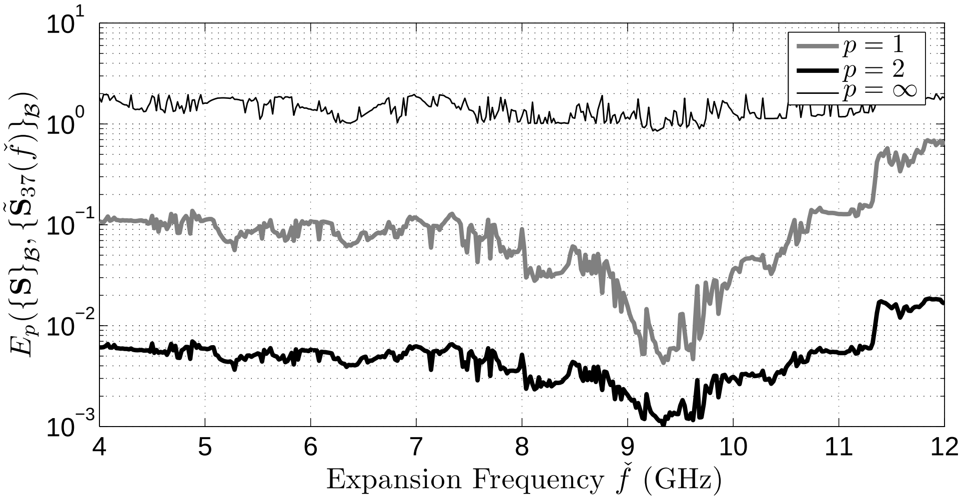


Figure 3.7: Bandpass filter [43]: True errors E_1, E_2 and E_∞ versus expansion frequency for single-point ROMs with $q = 37$.

Chapter 4

Out-of-Core Model Order Reduction Methods

4.1 Systems with a Large Number of Right-Hand Sides

4.1.1 Problem Statement

Technological advances in computer hardware allow the FE simulations of more and more complex electromagnetic structures. Today's simulations include complete printed circuits boards, whole integrated circuit packages, but also interconnectors with large numbers of pins. With the increasing complexity of the simulated structures, also the number of considered excitations became larger. This means for the numerical simulation of such complex structures that, on one hand the systems of linear equations may result in higher dimensions, on the other hand the systems may need to be solved for a larger number of right-hand sides.

Linear equation systems with a large number of right-hand sides is an area of extensive research, which also includes Krylov subspace methods for multiple starting vectors [44], [45]. This research is closely related to the Krylov subspace methods applied for MORe techniques which consider multiple input and output vectors, e.g. [46].

The focus of this chapter is on MORe techniques for high-dimensional systems with a large number of right-hand sides, where the system matrices are polynomially parameterized in the frequency. These are the properties of a system of linear equations, which results from the FE discretization of the considered complex electromagnetic structures.

The adaptive MORe techniques need to increase the projections matrix in each iteration, which requires more and more memory capacity. Although the computer operating system may start a swapping process to store Random Access Memory (RAM) data on the hard disk, the complete MORe process becomes inefficient and the ROM generation may become very slow. However, if the MORe process needs to be aborted, the projection matrix does not span a sufficient subspace and the generated ROM is not accurate within the considered bandwidth. To overcome this limitation, this thesis provides algorithms that swap carefully chosen computation data to the hard disk. These out-of-core approaches keep the RAM requirements for the projection matrix data on a constant low level. For both single-point and multi-point techniques out-of-core algorithms are presented. While the memory swapping can be accomplished easily in the multi-point algorithm, the single-point approach needs major structural changes to remain efficient.

The operation system as well as the compiler provide highly optimized data buffering techniques. Thus, the hard disk access time for the data swapping is not transparent for high level language programming. However, thanks to this intelligent buffering system, the resulting data access turned out to be very fast and does not cut down the efficiency of the MORe techniques.

As a final remark it should be added, that the complete swapping process uses binary data. This reduces the data amount, which reduces access times and requires less hard disk space.

4.1.2 Memory Considerations

Finite Element System

To explain the need for data swapping in the broadband FE simulation of electromagnetic structures, we consider the system (3.118). Looking back to the theory of Section 2.3, each column $\mathbf{b}^i$ of the block right-hand side $\mathbf{B}$ contains one entry and thus is extremely sparse. The sparsity of the system matrices $\mathbf{A}_i$ depends on the FE formulation, the basis functions, the FE mesh of the considered structure as well as on the imposed boundary conditions. Furthermore, it is worth mentioning here, that for systems considered in single-point methods the sparsity pattern may change and the matrices may have more entries, if the expansion point is not chosen at frequency zero.

The FE simulation of complex structures, which results in systems with high-dimensional matrices, may struggle on the memory limitations, as the requirements for the solution/factorization of the system may need a lot of memory. Software packages as [47] therefore provide out-of-core techniques to make the factorization of larger systems available.

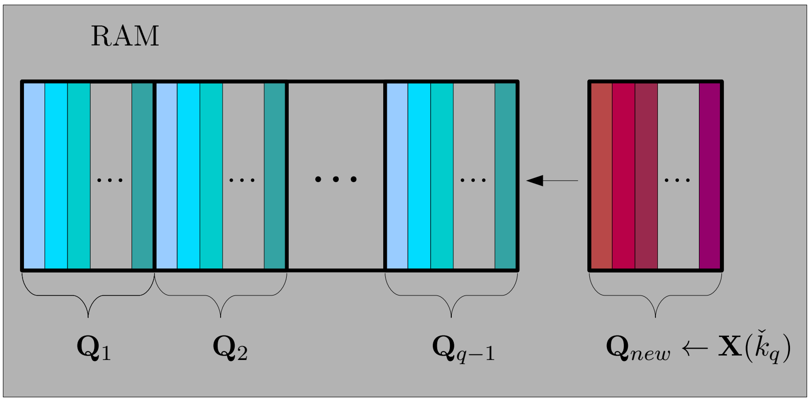
The memory for the system and its solution is one major part of the used memory, but is not subject of this thesis and is not further discussed. Nevertheless, memory requirement for system matrices and their factorization for the simulated structures will be provided in the numerical results.

In contrary to the full FE system, the low-dimensional ROM is irrelevant for the memory considerations. As for the reduced model itself, the memory requirements for the solution of the ROM are negligible.

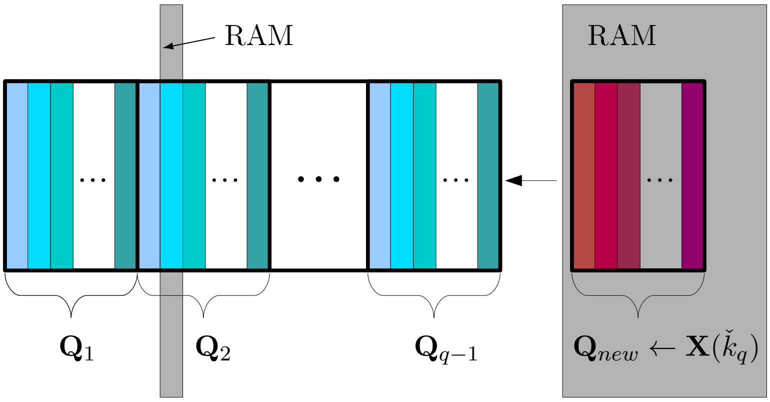
Multi-Point Methods

The adaptive multi-point method, see Alg. 4, represents the more simple technique, also from the memory handling point of view. For each expansion point, one block of b columns and N rows is added, which is the matrix $\mathbf{Q}_{new}$. This is the dimension of the block right-hand side vector $\mathbf{B}$. In contrast to the block right-hand side, the matrix $\mathbf{Q}_{new}$ is dense. Therefore, the algorithm enlarges the projection matrix in each iteration, and the full projection matrix $\mathbf{Q}_q$ of dimension $N \times bq$ fills more and more the RAM, see Fig. 4.1(a).

To improve the memory performance of the adaptive multi-point algorithm, only the matrix $\mathbf{Q}_{new}$ is kept in the memory, which is the orthonormalized solution block vector $\mathbf{X}(\check{k}_q)$ of the current iteration. The complete projection matrix $\mathbf{Q}_{q-1}$ of the previous iterations is swapped to the hard disk. The columns of $\mathbf{Q}_{q-1}$ are loaded one after another from the hard disk to the RAM for the orthonormalization process of the new computed block $\mathbf{X}(\check{k}_q)$ as well as for the subspace projection, see Fig. 4.1(b). Through this process, the required RAM capacity for the projection matrix is constant and does not increase with newly added expansion points. A generalization of the proposed approach is to allow only a predefined number of columns of $\mathbf{Q}_{new}$ to be kept in the memory. This number could even be automatically adapted to available memory.



(a) Multi-point method memory requirements.



(b) Out-of-core multi-point method memory requirements.

Figure 4.1: Multi-point method memory requirements.

Single-Point Methods

The situation in the case of single-point methods is rather difficult. The major drawback of the introduced adaptive single-point approach in Alg. 3 is that, for each right-hand side, the projection data needs to be kept twice in the memory. The WCAWE process of each right-hand side performs its own orthonormalization process. This data has to be kept in the memory additional to the projection matrix $\mathbf{Q}_q$, see Fig. 4.2.

Therefore, a blocked WCAWE approach will be employed, which operates only on the projection matrix itself. Hence, the memory requirements are reduced as Fig. 4.3(a) shows. The blocked process generates in each iteration a non-orthonormalized block WCAWE vector $\tilde{\mathbf{V}}_q = [\tilde{\mathbf{v}}_q^1, \tilde{\mathbf{v}}_q^2, \ldots, \tilde{\mathbf{v}}_q^b]$. This block vector is orthonormalized to the projection matrix and within its columns. The orthonormalized block $\mathbf{V}_q$ is finally added to the projection matrix. Thus, this approach will provide the same memory usage situation as the multi-point approach in Alg. 4.

The block algorithm is also the basis for the development of an out-of-core single-point method. The goal is again to swap as much projection matrix data as possible to the hard disk. This thesis provides a technique that keeps only two blocks of the dimension of $\mathbf{B}$ in the RAM, see Fig. 4.3(b). One block is the newly generated block vector, on which the algorithm operates. The second block is used as a data buffer and is needed to keep the blocked WCAWE process efficient. Additionally one projection matrix column after another is loaded from the hard disk to the RAM for the reduction process operations. Hence, the memory requirements for the projection matrix in the RAM is on a constant low level and does not increase in the iterative process. However, the single-point out-of-core technique needs more RAM capacity to stay efficient compared to the multi-point approach.

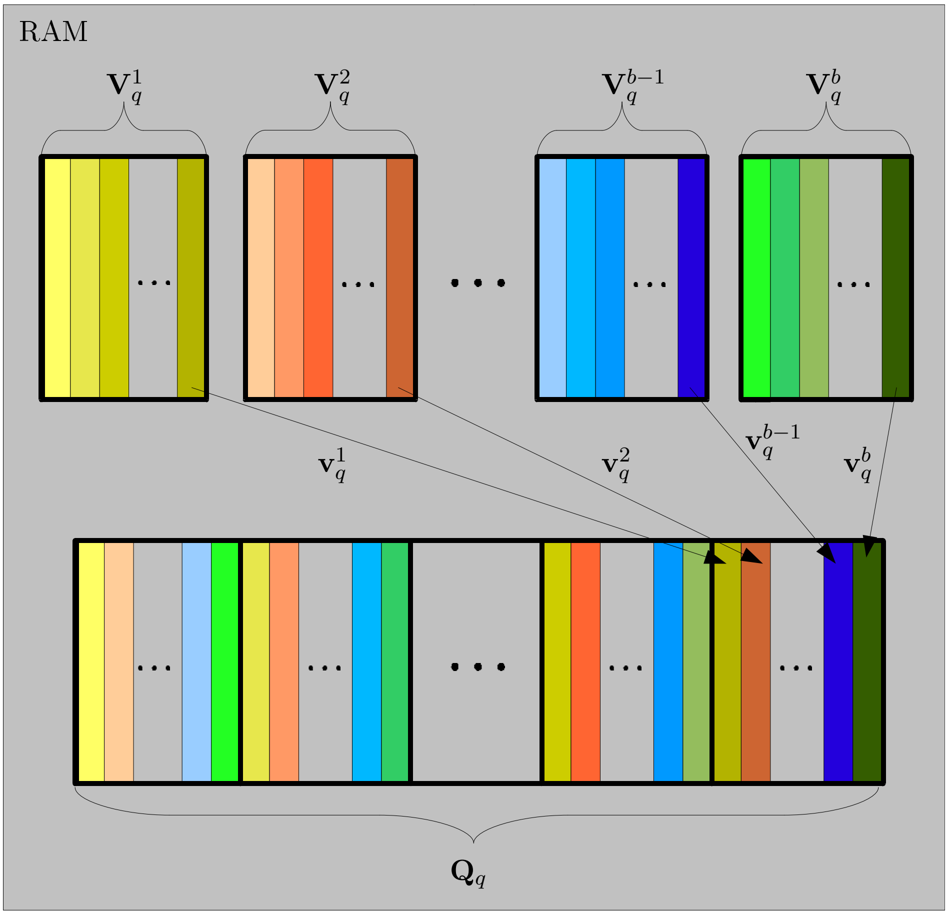
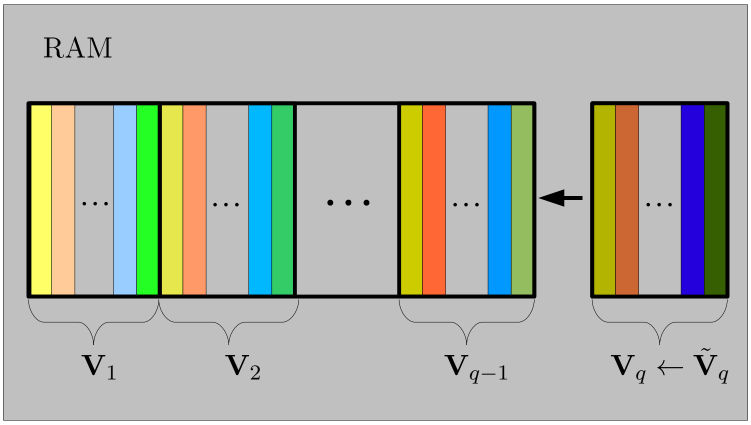
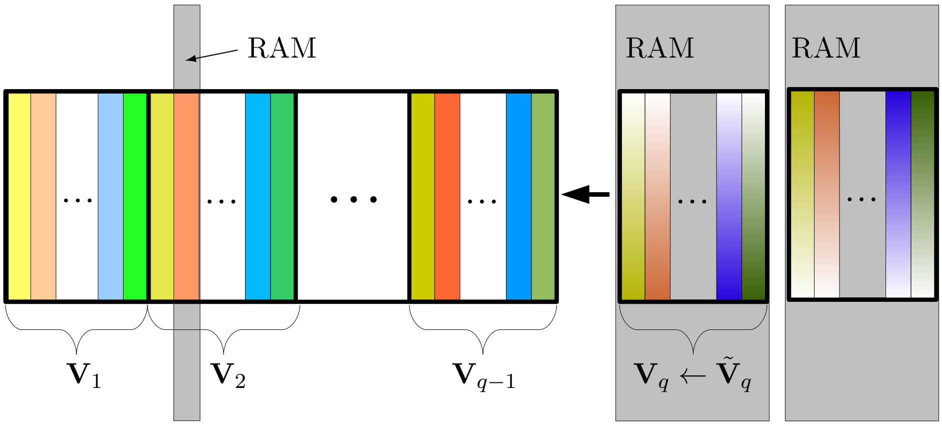


Figure 4.2: Single-point method memory requirements.



(a) Block single-point method memory requirement.



(b) Out-of-core single-point memory requirement.

Figure 4.3: New single-point method approaches memory requirements.

4.2 Block Algorithm for Higher-Order Systems

This section is dedicated to single-point methods, in particular, to reduce their memory usage by means of a block algorithm. This thesis provides a proof, that the simple block vector extension of the WCAWE, provided in [13], does match moments. Therefore, the introduced Krylov subspace theory is extended and important properties for blocking the WCAWE techniques are discussed.

4.2.1 Block Krylov Subspaces

Definition and Verification of Higher-Order Block Krylov Subspaces

In this section, the system of the form (3.118) is considered, as the blocked algorithm is sought for the same problem as in the previous chapter. Using the definition of the previous chapter $\mathbf{D}_i = -\mathbf{A}_0^{-1}\mathbf{A}_i$ and extending $\mathbf{u} = \mathbf{A}_0^{-1}\mathbf{b}$ to the block notation $\mathbf{U} = \mathbf{A}_0^{-1}\mathbf{B}$, allows us to introduce block Krylov subspaces.

For the sequence of matrices $\{\mathbf{D}_i\}_{i=1}^{a}$, $\mathbf{D}_i \in \mathbb{C}^{N\times N}$, and the block vector $\mathbf{U} \in \mathbb{C}^{N\times b}$, the qth block Krylov subspace of ath-order is defined as

$$\mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^{a};\mathbf{U}) = \mathrm{colspan}\{\mathbf{W}_0, \mathbf{W}_1, \dots, \mathbf{W}_{q-1}\}, \tag{4.1}$$

where the recursive definition of the block Krylov vectors $\mathbf{W}_l$ yield

$$\mathbf{W}_0 = \mathbf{U}, \tag{4.2}$$

$$\mathbf{W}_l = \sum_{i=1}^{\min(l,a)} \mathbf{D}_i\mathbf{W}_{l-i}. \tag{4.3}$$

This definition is a generalization of the higher-order Krylov subspace definition of Section 3.2.3. Plugging the Taylor expansion

$$\mathbf{X}(s) = \sum_{i=0}^{\infty} \mathbf{W}_i s^i, \tag{4.4}$$

into the system (3.118), allows us to write the block AWE vectors

$$\tilde{\mathbf{W}}_0 = \mathbf{A}_0^{-1}\mathbf{B}, \tag{4.5}$$

$$\tilde{\mathbf{W}}_1 = \mathbf{A}_0^{-1}(-\mathbf{A}_1\tilde{\mathbf{W}}_0), \tag{4.6}$$

$$\tilde{\mathbf{W}}_2 = \mathbf{A}_0^{-1}(-\mathbf{A}_1\tilde{\mathbf{W}}_1 - \mathbf{A}_2\tilde{\mathbf{W}}_0), \tag{4.7}$$

$$\vdots \tag{4.8}$$

$$\tilde{\mathbf{W}}_n = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,n)} \mathbf{A}_m\tilde{\mathbf{W}}_{n-m}). \tag{4.9}$$

With the definition given above, we have $\tilde{\mathbf{W}}_i = \mathbf{W}_i$ and thus, the vectors from the AWE expansion span the Krylov subspace $\mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^{a};\mathbf{U})$.

On each column of a block AWE vector $\mathbf{W}_i$, where as excitation the corresponding right-hand side column of $\mathbf{B}$ is employed, the same operations as for the non-blocked algorithm are performed. Hence, the Krylov subspace $\mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a;\mathbf{U})$ spans the same subspace as the projection matrix employed in (3.122) and thus satisfy

$$\mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a;\mathbf{U}) = \mathcal{K}_q^a(\{\mathbf{D}^i\}_{i=1}^a;\mathbf{u}_1) \cup \mathcal{K}_q^a(\{\mathbf{D}^i\}_{i=1}^a;\mathbf{u}_2) \cup \ldots \cup \mathcal{K}_q^a(\{\mathbf{D}^i\}_{i=1}^a;\mathbf{u}_b). \tag{4.10}$$

Extending the theory for higher-order systems to a block right-hand side excitation, the transfer function can be written as

$$\mathbf{H}(s) = \sum_{i=0}^{\infty} \mathbf{C}^T \mathbf{P}_i \mathbf{U} s^i = \sum_{i=0}^{\infty} \boldsymbol{\mu}_i s^i, \tag{4.11}$$

with the matrix of moments

$$\boldsymbol{\mu}_i = \mathbf{C}^T \mathbf{P}_i \mathbf{U}, \tag{4.12}$$

and a definition for $\mathbf{P}_i$ as in the previous chapter

$$\mathbf{P}_0 = \mathbf{I}, \tag{4.13}$$

$$\mathbf{P}_k = \sum_{i=1}^{\min(k,a)} \mathbf{P}_{k-i} \mathbf{D}_i. \tag{4.14}$$

This definition allows us, similar to the non-blocked case, to write the block Krylov vectors as

$$\mathbf{W}_i = \mathbf{P}_i \mathbf{U}. \tag{4.15}$$

In the literature, e.g. [35], definitions for the first-order block Krylov subspace

$$\mathcal{K}_q(\mathbf{D}_1, \mathbf{U}) = \text{colspan}\{\mathbf{U}, \mathbf{D}_1^1\mathbf{U}, \mathbf{D}_1^2\mathbf{U}, \mathbf{D}_1^3\mathbf{U}, \ldots, \mathbf{D}_1^{q-1}\mathbf{U}\}, \tag{4.16}$$

and the second-order block Krylov subspace

$$\mathcal{K}_q(\mathbf{D}_1, \mathbf{D}_2, \mathbf{U}) = \text{colspan}\{\mathbf{G}_0, \mathbf{G}_1, \ldots, \mathbf{G}_{q-1}\}, \tag{4.17}$$

where

$$\begin{cases} \mathbf{G}_0 = \mathbf{U}, \\ \mathbf{G}_1 = \mathbf{D}_1 \mathbf{G}_0, \\ \mathbf{G}_i = \mathbf{D}_1 \mathbf{G}_{i-1} + \mathbf{D}_2 \mathbf{G}_{i-2}, \end{cases} \tag{4.18}$$

can be found. These definitions are in accordance with the definition in this thesis for $a = 1$ and $a = 2$, respectively.

Moment Matching

For the reduction process (3.119), the projection matrix $\mathbf{Q}$, which spans the block Krylov subspace

$$\text{span}(\mathbf{Q}) = \mathcal{K}_q^a(\{\mathbf{D}_i\}_{i=1}^a;\mathbf{U}), \tag{4.19}$$

is employed. The proof for moment matching is a simple extension of the previous proof for non-blocked right-hand side and is not repeated. Furthermore, also the connections of Krylov subspaces obtained from linearized higher-order systems to first-order systems in [27] and [33], can be extended in a straightforward way to the block right-hand side excitations.

Deflation

In the literature, [45], [48], the definition for the $N \times qb$ block Krylov matrix of the form

$$\mathbf{K} = \begin{bmatrix}\mathbf{U} & \mathbf{D}^1\mathbf{U} & \mathbf{D}^2\mathbf{U} & \mathbf{D}^3\mathbf{U} \dots \mathbf{D}^{q-1}\mathbf{U}\end{bmatrix} \tag{4.20}$$

can be found. The rank of this matrix may be less than qb, if a column $\mathbf{D}^j\mathbf{U}_{[i]}$ is linear dependent on lower-order columns, which implies that all columns $\mathbf{D}^k\mathbf{U}_{[i]}$, with $j < k \leq q-1$ are also linear dependent. This property is called deflation and allows us to additionally define the deflated Krylov matrix

$$\mathbf{K}^{\mathrm{dl}} = [\mathbf{U}_0 \quad \mathbf{D}^1\mathbf{U}_1 \quad \mathbf{D}^2\mathbf{U}_2 \quad \mathbf{D}^3\mathbf{U}_3 \quad \dots \quad \mathbf{D}^{q-1}\mathbf{U}_{q-1}], \tag{4.21}$$

where

$$\mathbf{U}_0 = \mathbf{U}, \qquad \mathbf{U}_0 \in \mathbb{C}^{N \times b_0}, b_0 = b, \tag{4.22a}$$

$$\mathbf{U}_j = \mathbf{U}_{j-1}\mathbf{E}_j,\ 0 < j < q \qquad \mathbf{U}_j \in \mathbb{C}^{N \times b_j},\ b_j \leq b_{j-1} \tag{4.22b}$$

with the deflated identity matrix $\mathbf{E}_k \in \mathbb{C}^{b_{j-1} \times b_j}$, which deletes the linearly dependent columns [48]. For higher-order block Krylov subspaces, similar definitions are available, e.g. [49].

In the block WCAWE algorithm of this thesis, deflation is not considered. Instead, the orthonormalization process excludes numerical deficits, i.e. linear dependent columns, in the resulting projection matrix. However, the algorithm may become more powerful by employing deflation techniques and probably would reduce the projection matrix dimension and thus reduce the resulting ROM size.

4.2.2 Block Well-Conditioned Asymptotic Waveform Evaluation

Notation

The block WCAWE algorithm is a simple extension of the WCAWE algorithm, where scalar values are replaced by a block matrix of the dimension $b \times b$. Some notations need to be introduced for the algorithm and the proof for moment matching.

The block vectors $\mathbf{V}_n$, $\tilde{\mathbf{V}}_n \in \mathbb{C}^{N \times b}$, which are generated in the block algorithm, are gathered in the matrices $\boldsymbol{\mathcal{V}}$, $\tilde{\boldsymbol{\mathcal{V}}}$, with the notation

$$\boldsymbol{\mathcal{V}}_{[1:i]} = \begin{bmatrix}\mathbf{V}_1\,\mathbf{V}_2 \dots \mathbf{V}_i\end{bmatrix} \in \mathbb{C}^{N \times bi}, \tag{4.23}$$

$$\tilde{\boldsymbol{\mathcal{V}}}_{[1:i]} = \begin{bmatrix}\tilde{\mathbf{V}}_1\,\tilde{\mathbf{V}}_2 \dots \tilde{\mathbf{V}}_i\end{bmatrix} \in \mathbb{C}^{N \times bi}. \tag{4.24}$$

The block AWE vectors are collected in $\boldsymbol{\mathcal{W}}$ as

$$\boldsymbol{\mathcal{W}}_{[1:i]} = \begin{bmatrix}\mathbf{W}_0\,\mathbf{W}_1 \dots \mathbf{W}_{i-1}\end{bmatrix} \in \mathbb{C}^{N \times bi}. \tag{4.25}$$

Note the index shift $\mathbf{W}_i = \boldsymbol{\mathcal{W}}_{[i+1]}$, which is performed to conform with the original WCAWE literature [24].

The matrix $\mathbf{U}$ is the upper triangle block matrix

$$\boldsymbol{\mathcal{U}} = \begin{bmatrix} \mathbf{U}_{1,1} & \mathbf{U}_{1,2} & \dots & \mathbf{U}_{1,i} \\ \mathbf{0} & \mathbf{U}_{2,2} & \dots & \mathbf{U}_{2,i} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{U}_{i,i} \end{bmatrix} \in \mathbb{C}^{bi \times bi}, \tag{4.26}$$

with the submatrices

$$\mathbf{U}_{j,k} \in \mathbb{C}^{b \times b}. \tag{4.27}$$

All submatrices $\mathbf{U}_{j,j}$ on the diagonal need to be upper triangular matrices and all $\mathbf{U}_{j,k}$ with $k < j$ need to be null matrices to make $\boldsymbol{\mathcal{U}}$ upper triangular. The subscripts brackets denote the block

$$\boldsymbol{\mathcal{U}}_{[i:k,i:k]} = \begin{bmatrix} \mathbf{U}_{i,i} & \mathbf{U}_{i,i+1} & \dots & \mathbf{U}_{i,k} \\ \mathbf{0} & \mathbf{U}_{i+1,i+1} & \dots & \mathbf{U}_{i+1,k} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{U}_{k,k} \end{bmatrix}. \tag{4.28}$$

The matrix $\mathbf{E}_k \in \mathbb{R}^{M \times b}$ is defined as the block identity matrix

$$\mathbf{E}_k = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \\ \mathbf{0} \end{bmatrix} \begin{array}{l} \} \, (k-1)b \text{ rows} \\ \} \, b \text{ rows} \\ \} \, (q-k)b \text{ rows} \end{array}, \tag{4.29}$$

where k denotes the position of the identity matrix $\mathbf{I} \in \mathbb{R}^{b \times b}$. The number q is defined to be always the number of blocks on which $\mathbf{E}_k$ operates on. Thus, we have $M = qb$.

Block Algorithm

The WCAWE block algorithm provides a means to compute block Krylov subspaces of higher-order for the system (3.118) in a numerically stable way. The WCAWE block vectors are recursively defined as

$$\tilde{\mathbf{V}}_1 = \mathbf{A}_0^{-1}\mathbf{B}, \tag{4.30}$$

$$\tilde{\mathbf{V}}_2 = \mathbf{A}_0^{-1}(-\mathbf{A}_1\mathbf{V}_1), \tag{4.31}$$

$$\tilde{\mathbf{V}}_3 = \mathbf{A}_0^{-1}(-\mathbf{A}_1\mathbf{V}_2 - \mathbf{A}_2\mathbf{V}_1\mathbf{P}_{\mathbf{U}_2}(3,2)\mathbf{E}_1), \tag{4.32}$$

$$\vdots \tag{4.33}$$

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1\mathbf{V}_{q-1} - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m \boldsymbol{\mathcal{V}}_{[1:q-m]}\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{E}_{q-m}), \tag{4.34}$$

where in each iteration the matrix equation

$$\tilde{\boldsymbol{\mathcal{V}}}_{[1:q]} = \boldsymbol{\mathcal{V}}_{[1:q]}\boldsymbol{\mathcal{U}}, \tag{4.35}$$

is updated with the non-singular upper triangle matrix $\boldsymbol{\mathcal{U}} \in \mathbb{C}^{qb \times qb}$. The correction matrix $\mathbf{P}_{\mathbf{U}_w}(n,m)$, which is employed for the computation of the next block WCAWE vector, is defined as

$$\mathbf{P}_{\mathbf{U}_w}(n,m) = \prod_{t=w}^{m} \boldsymbol{\mathcal{U}}^{-1}_{[t:n-m+t-1,t:n-m+t-1]}, \qquad \mathbf{P}_{\mathbf{U}_w}(n,m) \in \mathbb{C}^{(n-m)b \times (n-m)b}, \tag{4.36}$$

where $w \leq m < n < q$. In this thesis, a modified Gram-Schmidt process is chosen to update the relation (4.35) in each iteration. The process orthonormalizes the columns of the matrix $\tilde{\boldsymbol{\mathcal{V}}}_{[1:q]}$, which results in the matrix $\boldsymbol{\mathcal{V}}_{[1:q]}$. Hence, the orthonormalization of each block vector $\tilde{\mathbf{V}}_q$, let each block $\mathbf{U}_{q,q}$ become an upper triangular matrix. Thus, the process generates an upper triangular matrix $\boldsymbol{\mathcal{U}}$ and the columns of $\boldsymbol{\mathcal{V}}_{[1:n]}$ provide a highly accurate basis for the Krylov subspace $\mathcal{K}_n^a(\{\mathbf{D}_i\}_{i=1}^a; \mathbf{U})$.

4.2.3 Proof for Moment Matching

Properties and Interpretations

This section gives a list of properties that helps us to understand better the algorithm and its definitions. The properties are used in particular for the moment matching proof. Detailed proofs for this properties are not given, as they either can be found in standard mathematical literature, e.g. [50], or are simple algebraic considerations. However, mainly for properties which are essential for the proof, some interpretations are provided. Note that some of the properties provided simplify the original proof for the non-blocked WCAWE [24], for the constellation considered in this thesis.

Property 1: The inverse of the upper triangular matrix $\mathbf{U}$, i.e. $\mathbf{U}^{-1}$, is also a upper triangular matrix.

Property 2: The product $\mathbf{U} = \mathbf{U}_1\mathbf{U}_2$ of two upper triangular matrices results in an upper triangular matrix.

Property 3: For the upper triangular matrix $\mathbf{U} \in \mathbb{C}^{n\times n}$, the equality
$\mathbf{U}^{-1}_{[j_1:j_2,j_1:j_2]} = (\mathbf{U}_{[j_1:j_2,j_1:j_2]})^{-1}$ holds for any integers j_1 and j_2 such that $1 \leq j_1 \leq j_2 \leq n$.

Property 4: We have: $\mathrm{span}(\tilde{\boldsymbol{\mathcal{V}}}_{[1:n]}) = \mathrm{span}(\boldsymbol{\mathcal{V}}_{[1:n]})$.

The matrix $\boldsymbol{\mathcal{U}}$, which connects the matrices above by (4.35), is defined to be non-singular.

Property 5: The correction matrix $\mathbf{P}_{\mathbf{U}_w}(n, m)$ is an upper triangular matrix.

The correction matrix $\mathbf{P}_{\mathbf{U}_w}(n, m)$ is a product of upper triangular submatrices of $\boldsymbol{\mathcal{U}}$, see Fig. 4.4(a). Hence, because of Property 2, the resulting correction matrix is upper triangular itself.

Property 6: The equality $\boldsymbol{\mathcal{U}}^{-1}_{[1:n-m,1:n-m]}\mathbf{P}_{\mathbf{U}_2}(n, m) = \mathbf{P}_{\mathbf{U}_1}(n, m)$ holds.

This is the simple multiplication of $\mathbf{P}_{\mathbf{U}_2}(n, m)$ with the first upper triangular matrix. The examples $\mathbf{P}_{\mathbf{U}_1}(n, m)$ and $\mathbf{P}_{\mathbf{U}_2}(n, m)$ in Fig. 4.4(b) and 4.4(c), respectively, show this Property.

Property 7: Assume the integers α, $\bar{\alpha}$ and γ satisfy $1 \leq \gamma < \min(\alpha, \bar{\alpha})$. Then for all integers j_1 and j_2 which satisfy $1 \leq j_1, j_2 \leq \min(\alpha, \bar{\alpha}) - \gamma$, the equality
$\mathbf{E}^T_{j_1}\mathbf{P}_{\mathbf{U}_1}(\alpha, \gamma)\mathbf{E}_{j_2} = \mathbf{E}^T_{j_1}\mathbf{P}_{\mathbf{U}_1}(\bar{\alpha}, \gamma)\mathbf{E}_{j_2}$ holds.

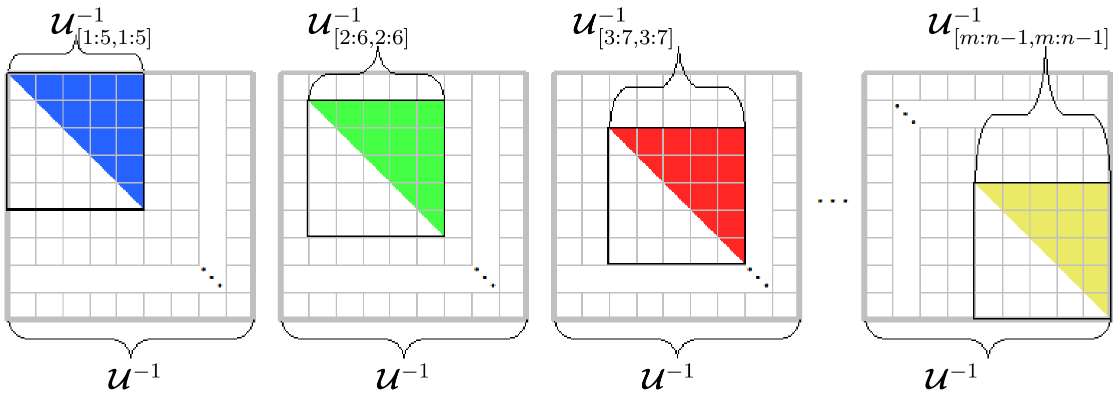
Through modifying the value α to $\bar{\alpha}$, the dimension of the shifted matrix is changed, whereas the number of shifts γ is constant, see Fig. 4.4(d). As only upper triangular matrices are in the products of $\mathbf{P}_{\mathbf{U}_1}(\alpha, \gamma)$ and $\mathbf{P}_{\mathbf{U}_1}(\bar{\alpha}, \gamma)$, the resulting correction matrix is the same, as long as the block unit vectors $\mathbf{E}_{j_1}$ and $\mathbf{E}_{j_2}$ select a block within the smaller dimension.

PROPERTY 8: Let $\boldsymbol{\mathcal{U}} \in \mathbb{C}^{qb\times qb}$ be a non-singular upper triangular matrix, and let β, m and n be integers $1 \leq m < n \leq q$ and $1 \leq \beta \leq n-1$. Then for all integers j_1 and j_2 such that $\beta \leq j_1, j_2 \leq n-m$, the equality $\mathbf{E}_{j_1}^T \mathbf{P}_{\mathbf{U}_1}(n,m)\mathbf{E}_{j_2} = \mathbf{E}_{j_1-\beta+1}^T \mathbf{P}_{\mathbf{U}_\beta}(n, m+\beta-1)\mathbf{E}_{j_2-\beta+1}$ holds.

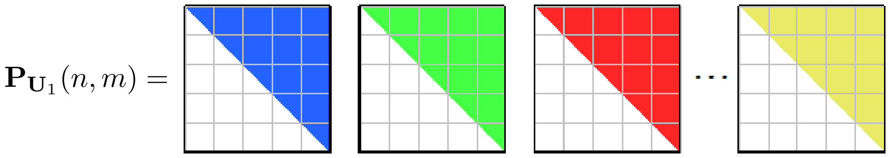
This Property is similar to the Property 7 above. $\mathbf{P}_{\mathbf{U}_\beta}(n, m+\beta-1)$ is the product of upper triangular matrices which are shifted with $\beta-1$ compared to $\mathbf{P}_{\mathbf{U}_1}(n,m)$, where also the dimension of the multiplied matrices is $\beta-1$ smaller. This results in the same matrix product as $\mathbf{P}_{\mathbf{U}_1}(n,m)$, as long as the block unit vectors $\mathbf{E}_{j_1}$ and $\mathbf{E}_{j_2}$ select a block within the smaller dimension. The Property is shown in Fig. 4.4(e).

PROPERTY 9: Let $\boldsymbol{\mathcal{U}} \in \mathbb{C}^{qb\times qb}$ be a non-singular upper triangular matrix, and let n, m and β be integers such that $1 \leq m < n \leq q$. Then for $1 < \beta \leq n-m$ the equality $\mathbf{P}_{\mathbf{U}_1}(n-m, \beta-1)\mathbf{P}_{\mathbf{U}_\beta}(n, m+\beta-1) = \mathbf{P}_{\mathbf{U}_1}(n, m+\beta-1)$ holds.

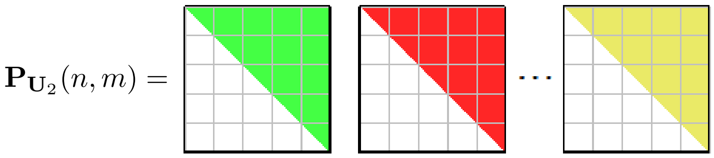
This Property can be interpreted in Fig. 4.4(a). $\mathbf{P}_{\mathbf{U}_1}(n-m, \beta-1)$ is the multiplication of the first $\beta-1$ shifted triangular matrices. The second matrix is the product of the shifted triangular matrices from β to $m+\beta-1$. Thus, the result is the product of all β to $m+\beta-1$ shifted matrices, which is $\mathbf{P}_{\mathbf{U}_1}(n, m+\beta-1)$.



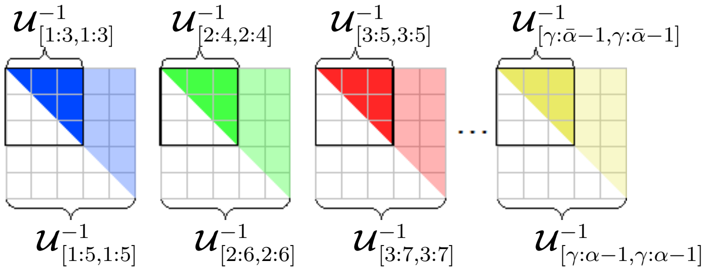
(a) Upper triangular matrices in correction matrix $\mathbf{P}_{\mathbf{U}_w}(n,m)$ for $n-m=5$.



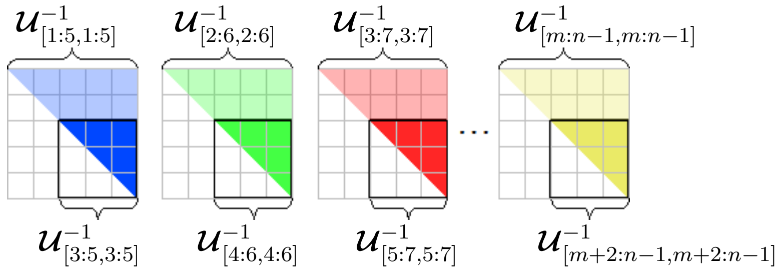
(b) Correction matrix $\mathbf{P}_{\mathbf{U}_1}(n,m)$ for $n-m=5$.



(c) Correction matrix $\mathbf{P}_{\mathbf{U}_2}(n,m)$ for $n-m=5$.



(d) Submatrices for $\mathbf{P}_{\mathbf{U}_1}(\alpha,\gamma)$ and $\mathbf{P}_{\mathbf{U}_1}(\bar{\alpha},\gamma)$ computation, $\alpha-\gamma=5$, $\bar{\alpha}-\gamma=3$.



(e) Submatrices for $\mathbf{P}_{\mathbf{U}_1}(n,m)$ and $\mathbf{P}_{\mathbf{U}_\beta}(n,m+\beta-1)$ computation, $n-m=5$, $\beta=3$.

Figure 4.4: Properties of the correction matrix $\mathbf{P}_{\mathbf{U}_w}(n,m)$, with $\mathcal{U} \in \mathbb{C}^{(n-1)b\times(n-1)b}$.

Proof

We define the block matrix

$$\boldsymbol{\mathcal{X}}_{[1:q,1:q]} = \begin{bmatrix} \mathbf{X}_{1,1} & \mathbf{X}_{1,2} & \dots & \mathbf{X}_{1,q-1} & \mathbf{X}_{1,q} \\ \mathbf{0} & \mathbf{X}_{2,2} & \dots & \mathbf{X}_{2,q-1} & \mathbf{X}_{1,q} \\ \mathbf{0} & \mathbf{0} & & & \mathbf{X}_{1,q} \\ \vdots & \vdots & \ddots & & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{X}_{q,q} \end{bmatrix} \in \mathbb{C}^{bq \times bq}, \tag{4.37}$$

where each block $\mathbf{X}_{j_1,j_2} \in \mathbb{C}^{b \times b}$ is defined as

$$\mathbf{X}_{j_1,j_2} = \begin{cases} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(j_2, j_1 - 1)\mathbf{E}_{j_2 - j_1 + 1} & \text{for } 2 \le j_1 \le j_2 \le q \\ \mathbf{I} & \text{for } j_1 = j_2 = 1 \\ \mathbf{0} & \text{otherwise} \end{cases} . \tag{4.38}$$

Note that $\boldsymbol{\mathcal{X}}$ is an upper triangular non-singular matrix. All diagonal values are 1.

The inductive proof shows that

$$\tilde{\boldsymbol{\mathcal{V}}}_{[1:q]} = \boldsymbol{\mathcal{W}}_{[1:q]} \boldsymbol{\mathcal{X}}_{[1:q,1:q]}. \tag{4.39}$$

and thus $\boldsymbol{\mathcal{V}}$ spans the required space for moment matching. For the induction basis we have $q = 1$

$$\tilde{\mathbf{V}}_1 = \boldsymbol{\mathcal{W}}_{[1]} \mathbf{X}_{1,1} \tag{4.40}$$

and for $q = 2$

$$\begin{aligned} \tilde{\mathbf{V}}_2 = \mathbf{A}_0^{-1}(-\mathbf{A}_1 \mathbf{V}_1) = \mathbf{A}_0^{-1}(-\tilde{\mathbf{V}}_1 \mathbf{U}_{1,1}^{-1}) &= \mathbf{A}_0^{-1}(-\mathbf{A}_1 \boldsymbol{\mathcal{W}}_{[1]} \mathbf{U}_{1,1}^{-1}) & (4.41) \\ = \boldsymbol{\mathcal{W}}_{[2]} \mathbf{U}_{1,1}^{-1} = \boldsymbol{\mathcal{W}}_{[2]} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(2,1) \mathbf{E}_1 &= \boldsymbol{\mathcal{W}}_{[2]} \mathbf{X}_{2,2}. & (4.42) \end{aligned}$$

Note that $\mathbf{X}_{1,2} = \mathbf{0}$. Therefore,

$$\operatorname{span}(\tilde{\boldsymbol{\mathcal{V}}}_{[1:2]}) = \operatorname{span}(\boldsymbol{\mathcal{W}}_{[1:2]}). \tag{4.43}$$

The induction hypothesis states

$$\begin{aligned} \tilde{\boldsymbol{\mathcal{V}}}_{[1:q-1]} &= \boldsymbol{\mathcal{W}}_{[1:q-1]} \mathbf{X}_{[1:q-1,1:q-1]}, & (4.44) \\ \operatorname{span}(\tilde{\boldsymbol{\mathcal{V}}}_{[1:q-1]}) &= \operatorname{span}(\boldsymbol{\mathcal{W}}_{[1:q-1]}). & (4.45) \end{aligned}$$

By induction we will see

$$\begin{aligned} \tilde{\boldsymbol{\mathcal{V}}}_{[1:q]} &= \boldsymbol{\mathcal{W}}_{[1:q]} \mathbf{X}_{[1:q,1:q]}, & (4.46) \\ \operatorname{span}(\tilde{\boldsymbol{\mathcal{V}}}_{[1:q]}) &= \operatorname{span}(\boldsymbol{\mathcal{W}}_{[1:q]}). & (4.47) \end{aligned}$$

For $n \ge 2$ the algorithm is defined as

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1 \mathbf{V}_{q-1} - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m \boldsymbol{\mathcal{V}}_{[1:q-m]} \mathbf{P}_{\mathbf{U}_2}(q,m) \mathbf{E}_{q-m}). \tag{4.48}$$

We use (4.35) to write

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1\tilde{\mathbf{V}}_{q-1}\boldsymbol{\mathcal{U}}^{-1}_{[1:q-1,1:q-1]}\mathbf{E}_{q-1} \\ - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m\tilde{\boldsymbol{\mathcal{V}}}_{[1:q-m]}\boldsymbol{\mathcal{U}}^{-1}_{[1:q-m,1:q-m]}\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{E}_{q-m}), \tag{4.49}$$

and plug in the induction hypothesis (4.44)

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1\boldsymbol{\mathcal{W}}_{[1:q-1]}\boldsymbol{\mathcal{X}}_{[1:q-1,1:q-1]}\boldsymbol{\mathcal{U}}^{-1}_{[1:q-1,1:q-1]}\mathbf{E}_{q-1} \\ - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m\boldsymbol{\mathcal{W}}_{[1:q-m]}\boldsymbol{\mathcal{X}}_{[1:q-m,1:q-m]}\boldsymbol{\mathcal{U}}^{-1}_{[1:q-m,1:q-m]}\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{E}_{q-m}). \tag{4.50}$$

Use Property 6 to write

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1\boldsymbol{\mathcal{W}}_{[1:q-1]}\boldsymbol{\mathcal{X}}_{[1:q-1,1:q-1]}\boldsymbol{\mathcal{U}}^{-1}_{[1:q-1,1:q-1]}\mathbf{E}_{q-1} \\ - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m\boldsymbol{\mathcal{W}}_{[1:q-m]}\boldsymbol{\mathcal{X}}_{[1:q-m,1:q-m]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}). \tag{4.51}$$

Now, we use $\boldsymbol{\mathcal{U}}^{-1}_{[1:n-1,1:n-1]} = \mathbf{P}_{\mathbf{U}_1}(n,1)$

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m\boldsymbol{\mathcal{W}}_{[1:q-m]}\boldsymbol{\mathcal{X}}_{[1:q-m,1:q-m]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}). \tag{4.52}$$

Now use $\boldsymbol{\mathcal{W}}_{[1:n-m]} = \sum_{\beta=1}^{n-m}\boldsymbol{\mathcal{W}}_{[\beta]}\mathbf{E}_\beta^T$

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\sum_{\beta=1}^{q-m}\boldsymbol{\mathcal{W}}_{[\beta]}\mathbf{E}_\beta^T)\boldsymbol{\mathcal{X}}_{[1:q-m,1:q-m]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}). \tag{4.53}$$

Contract now $\mathbf{E}_\beta^T\boldsymbol{\mathcal{X}}_{[1:n-m,1:n-m]} = \boldsymbol{\mathcal{X}}_{[\beta,1:n-m]}$ to have

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\sum_{\beta=1}^{q-m}\boldsymbol{\mathcal{W}}_{[\beta]}\boldsymbol{\mathcal{X}}_{[\beta,1:q-m]})\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}). \tag{4.54}$$

Writing $\boldsymbol{\mathcal{X}}_{[\beta,1:q-m]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}$ as a sum and using the property $\mathbf{X}_{\beta,r} = \mathbf{0}\ \forall\ r < \beta$, we have

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m\sum_{\beta=1}^{q-m}\boldsymbol{\mathcal{W}}_{[\beta]}\sum_{r=\beta}^{q-m}\mathbf{X}_{\beta,r}\mathbf{E}_r^T\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m}). \tag{4.55}$$

From the definition of $\boldsymbol{\mathcal{X}}$ in (4.38) we can write now

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\boldsymbol{\mathcal{W}}_{[1]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m} \\ + \sum_{\beta=2}^{q-m}\boldsymbol{\mathcal{W}}_{[\beta]}\sum_{r=\beta}^{q-m}\mathbf{E}_1^T\mathbf{P}_{\mathbf{U}_1}(r,\beta-1)\mathbf{E}_{r-\beta+1}\mathbf{E}_r^T\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m})), \tag{4.56}$$

where we use $\mathbf{X}_{1,1} = \mathbf{I}$ and $\mathbf{X}_{1,r} = \mathbf{0}$ for $1 < r < q$. Now use Property 7, with $\alpha = r$, $\bar{\alpha} = q - m$, $\gamma = \beta - 1$, $j_1 = 1$ and $j_2 = r - \beta + 1$ to obtain

$$\begin{aligned}\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- &\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\boldsymbol{\mathcal{W}}_{[1]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m} \\ &+ \sum_{\beta=2}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \sum_{r=\beta}^{q-m} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(q-m,\beta-1)\mathbf{E}_{r-\beta+1}\mathbf{E}_r^T\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m})).\end{aligned} \tag{4.57}$$

Now, use Property 8 with $j_1 = r$ and $j_2 = q - m$

$$\begin{aligned}\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- &\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\boldsymbol{\mathcal{W}}_{[1]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m} \\ &+ \sum_{\beta=2}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \sum_{r=\beta}^{q-m} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(q-m,\beta-1)\mathbf{E}_{r-\beta+1}\mathbf{E}_{r-\beta+1}^T\mathbf{P}_{\mathbf{U}_\beta}(q,m+\beta-1)\mathbf{E}_{q-m-\beta+1})).\end{aligned} \tag{4.58}$$

Now, only the product $\mathbf{E}_{r-\beta+1}\mathbf{E}_{r-\beta+1}^T$ depends on r and the sum results in the identity matrix. Thus,

$$\begin{aligned}\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- &\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\boldsymbol{\mathcal{W}}_{[1]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m} \\ &+ \sum_{\beta=2}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(q-m,\beta-1)\mathbf{P}_{\mathbf{U}_\beta}(q,m+\beta-1)\mathbf{E}_{q-m-\beta+1})).\end{aligned} \tag{4.59}$$

The next step is to employ Property 9

$$\begin{aligned}\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- &\sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m(\boldsymbol{\mathcal{W}}_{[1]}\mathbf{P}_{\mathbf{U}_1}(q,m)\mathbf{E}_{q-m} \\ &+ \sum_{\beta=2}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(q,m+\beta-1)\mathbf{E}_{q-m-\beta+1})).\end{aligned} \tag{4.60}$$

Contracting the sum results in

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- \sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m \sum_{\beta=1}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \mathbf{E}_1^T \mathbf{P}_{\mathbf{U}_1}(q,m+\beta-1)\mathbf{E}_{q-m-\beta+1}). \tag{4.61}$$

In this expression, we insert the definition of $\mathbf{X}_{j_1,j_2}$ with $j_1 = m + \beta$ and $j_2 = q$, therefore

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- \sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m \sum_{\beta=1}^{q-m} \boldsymbol{\mathcal{W}}_{[\beta]} \mathbf{X}_{m+\beta,q}). \tag{4.62}$$

This is

$$\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(- \sum_{m=1}^{\min(a,q-1)} \mathbf{A}_m \boldsymbol{\mathcal{W}}_{[1:q-m]} \boldsymbol{\mathcal{X}}_{[m+1:q,q]}). \tag{4.63}$$

After all these algebraic modifications, the following part of the proof helps us to understand why the WCAWE algorithm matches moments. The equation (4.63) allows to write

$$
\begin{aligned}
\tilde{\mathbf{V}}_q &= \Big(\sum_{m=1}^{\min(a,q-1)} -\mathbf{A}_0^{-1}\mathbf{A}_m \boldsymbol{\mathcal{W}}_{[1:q-m]} \boldsymbol{\mathcal{X}}_{[m+1:q,q]} \Big) \qquad (4.64) \\
&= \sum_{n=2}^{q} \Big(- \sum_{m=1}^{\min(a,n-1)} \mathbf{A}_0^{-1}\mathbf{A}_m \boldsymbol{\mathcal{W}}_{[n-m]} \mathbf{X}_{n,q} \Big) \qquad (4.65) \\
&= \sum_{n=2}^{q} \underbrace{\Big(- \sum_{m=1}^{\min(a,n-1)} \mathbf{A}_0^{-1}\mathbf{A}_m \boldsymbol{\mathcal{W}}_{[n-m]} \Big)}_{\boldsymbol{\mathcal{W}}_{[n]}} \mathbf{X}_{n,q} \qquad (4.66) \\
&= \sum_{n=2}^{q} \boldsymbol{\mathcal{W}}_{[n]} \mathbf{X}_{n,q}. \qquad (4.67)
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\tilde{\boldsymbol{\mathcal{V}}}_{[1:n]} &= \boldsymbol{\mathcal{W}}_{[1:n]} \mathbf{X}_{[1:n,1:n]}, \qquad (4.68) \\
\operatorname{span}(\tilde{\boldsymbol{\mathcal{V}}}_{[1:n]}) &= \operatorname{span}(\boldsymbol{\mathcal{W}}_{[1:n]}). \qquad (4.69)
\end{aligned}
$$

and thus the induction hypothesis holds.

4.3 Adaptive Algorithms

4.3.1 Blocking the Single-Point Algorithm

Employing the blocked WCAWE process allows us to introduce the more efficient blocked adaptive single-point method in Alg. 5. Algorithm 5, Line 1 and Line 6, with the following modified Gram-Schmidt orthonormalization, provide a very efficient way to compute a stable basis for the sought Krylov subspace for a block right-hand side excitation. Compared to the non-blocked Alg. 3, the blocked approach does not need separated WCAWE processes for each excitation anymore. Furthermore, the additional orthonormalization process to generate the projection matrix from the separated WCAWE processes is thus dispensable.

This algorithm provides a solid basis for efficiently swapping the projection matrix to the hard disk by some modifications in the WCAWE process, which will be explained in the following.

Algorithm 5 Self-Adaptive Block Single-Point Model Order Reduction

```
Initial WCAWE: $\tilde{\mathbf{V}}_1 = \mathbf{A}_0^{-1}\mathbf{B}$
Initial modified Gram-Schmidt: $\tilde{\mathbf{V}}_1 = \mathbf{V}_1\mathbf{U}$
Initial subspace projection: $\tilde{\mathbf{A}}_i = \mathbf{V}_1^T\mathbf{A}_i\mathbf{V}_1$, $\tilde{\mathbf{B}} = \mathbf{V}_1^T\mathbf{B}$
Solve frequency sweep:
$(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_1\}_{\mathcal{B}}$
for $q = 2$ to $q_{max}$ do
    Blocked WCAWE process:
    $\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}(-\mathbf{A}_1\mathbf{V}_{q-1} - \sum_{m=2}^{\min(a,q-1)} \mathbf{A}_m\boldsymbol{\mathcal{V}}_{[1:q-m]}\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{E}_{q-m})$
    Apply Modified Gram-Schmidt to new generated block vector:
    $[\boldsymbol{\mathcal{V}}_{[1:q-1]}, \mathbf{V}_q]\mathbf{U} = [\tilde{\boldsymbol{\mathcal{V}}}_{[1:q-1]}, \tilde{\mathbf{V}}_q]$
    Enlarge ROM:
    $\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \boldsymbol{\mathcal{V}}_{[1:q-1]}^T\mathbf{A}_i\mathbf{V}_q \\ \mathbf{V}_q^T\mathbf{A}_i\boldsymbol{\mathcal{V}}_{[1:q-1]} & \mathbf{V}_q^T\mathbf{A}_i\mathbf{V}_q \end{bmatrix}$, $\tilde{\mathbf{B}} \leftarrow \begin{bmatrix} \tilde{\mathbf{B}} \\ \mathbf{V}_q^T\mathbf{B} \end{bmatrix}$.
    Solve frequency sweep:
    $(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}$
    if $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}}) < E_{2,tol}$ then
        return(converged)
    end if
end for
```

4.3.2 Adaptive Out-of-Core Algorithm for Single-Point Method

Algorithm 6 presents the adaptive out-of-core MORe single-point method. For the memory management, additional operations need to be introduced. The operator SWAP($\mathbf{V}_i$) stands for saving the block matrix $\mathbf{V}_i$ to hard disk and free the main memory. LOAD($\mathbf{V}_i$) means to load the previously saved block $\mathbf{V}_i$ from the hard disk to the RAM and CLEAR($\mathbf{V}_i$) is written for free the memory used for the block $\mathbf{V}_i$. For a better algorithm illustration, matrices which are swapped to the hard disk are colored in gray.

The operations in Alg. 5 where the projection matrix $\boldsymbol{\mathcal{V}}_{[1:q-1]}$ appears are of special interested for the memory management. These are Alg. 5, Lines 6, 7 and 8. The out-of-core operations for these lines are discussed in the following. However, to keep the WCAWE process in the out-of-core technique efficient, two block vectors of size $\mathbf{V}_q \in \mathbb{C}^{N \times b}$ are needed to operate on. Otherwise a large number of read/write processes would dramatically slow down the algorithm.

Furthermore, it is worth commenting on the WCAWE block vector generation in the out-of-core algorithm, Alg. 6, Lines 8 to 15. In the algorithm development process, the matrix read/write processes were expected to be very time-consuming. Thus, the first approach to generate the WCAWE block vector was to perform all operations on a loaded block:

```
1: Initialize: LOAD($\mathbf{V}_{q-1}$) ;   $\tilde{\mathbf{V}}_q = -\mathbf{A}_1 \mathbf{V}_{q-1}$ ;   CLEAR($\mathbf{V}_{q-1}$) ;
2: for $p = 1$ to $q - 2$ do
3:    LOAD($\mathbf{V}_p$)
4:    for $m = 2$ to $\min(a, q - p)$ do
5:       $\tilde{\mathbf{V}}_q = \tilde{\mathbf{V}}_q - \mathbf{A}_m \mathbf{V}_p \mathbf{E}_p \mathbf{P}_{\mathbf{U}_2}(q, m) \mathbf{E}_{q-m})$
6:    end for
7:    CLEAR($\mathbf{V}_p$)
8: end for
9: $\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1} \tilde{\mathbf{V}}_q$
```

However, it turned out that for increasing q the additionally performed matrix-vector multiplications in Line 5 are more time-consuming than the chosen operations in Alg. 6. Thanks to the highly optimized reading operation, provided by the compiler and operating system, the time overhead of the swapping mechanism is very low as numerical results will show.

Alg. 6, Lines 17 to 22 introduce an approach for the orthonormalization process of the projection matrix with low memory requirements. The modified Gram-Schmidt algorithm is rearranged such that each column of the new block vector $\mathbf{V}_q$ is orthonormalized against each column of a single loaded block vector $\mathbf{V}_p$. Thus, in the loop for $k = 1 \, \ldots \, q - 1$, each block vector needs to be loaded only once. After this loop, the columns of the block vector $\mathbf{V}_q$ are orthonormalized against each other in Line 22, which completes the orthonormalization process.

Finally, for the projection process the projection matrix data must be read again from the hard disk. This is performed by loading each block vector $\mathbf{V}_p$ once and evaluate the reduced blocks for the ROM enlargement. This is shown in Alg. 6, Lines 23 to 27.

Algorithm 6 Self-Adaptive Out-of-Core Single-Point Model Order Reduction

```
Initial WCAWE: $\tilde{\mathbf{V}}_1 = \mathbf{A}_0^{-1}\mathbf{B}$
Initial modified Gram-Schmidt: $\tilde{\mathbf{V}}_1 = \mathbf{V}_1\mathbf{U}$
Initial subspace projection: $\tilde{\mathbf{A}}_i = \mathbf{V}_1^T\mathbf{A}_i\mathbf{V}_1$, $\tilde{\mathbf{B}} = \mathbf{V}_1^T\mathbf{B}$
Free memory: SWAP($\mathbf{V}_1$),
Solve frequency sweep:
$(\sum_{i=0}^{a}\tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_1\}_{\mathcal{B}}$
for $q = 2$ to $q_{max}$ do
    Initialize: LOAD($\mathbf{V}_{q-1}$) ;  $\tilde{\mathbf{V}}_q = -\mathbf{A}_1\mathbf{V}_{q-1}$ ;  CLEAR($\mathbf{V}_{q-1}$) ;
    for $m = 2$ to $\min(a, q-1)$ do
        for $p = 1$ to $q-1$ do
            LOAD($\mathbf{V}_p$)
            $\tilde{\mathbf{V}}_q = \tilde{\mathbf{V}}_q - \mathbf{A}_m\mathbf{V}_p\mathbf{E}_p\mathbf{P}_{\mathbf{U}_2}(q,m)\mathbf{E}_{q-m})$
            CLEAR($\mathbf{V}_p$)
        end for
    end for
    $\tilde{\mathbf{V}}_q = \mathbf{A}_0^{-1}\tilde{\mathbf{V}}_q$
    Modified Gram-Schmidt in out-of-core technique:
    for $p = 1$ to $q-1$ do
        LOAD($\mathbf{V}_p$)
        Orthonormalize only against loaded block vector:
        $[\mathcal{V}_{[1:p-1]}, \mathbf{V}_p, \mathcal{V}_{[p:q-1]}]\mathbf{U} = [\ldots, \tilde{\mathbf{V}}_q]$
        CLEAR($\mathbf{V}_p$)
    end for
    Orthonormalize columns of current block vector: $[\mathcal{V}_{[1:q-1]}, \mathbf{V}_q]\mathbf{U} = [\ldots, \tilde{\mathbf{V}}_q]$
    Enlarge ROM in out-of-core technique:
    for $p = 1$ to $q-1$ do
        LOAD($\mathbf{V}_p$) ;  $\tilde{\mathbf{M}}_{p,i} = \mathbf{V}_p^T\mathbf{A}_i\mathbf{V}_q$ ;  CLEAR($\mathbf{V}_p$);
    end for
    $\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \begin{bmatrix}\tilde{\mathbf{M}}_{1,i} \\ \vdots \\ \tilde{\mathbf{M}}_{q-1,i}\end{bmatrix} \\ \begin{bmatrix}\tilde{\mathbf{M}}_{1,i}^T & \ldots & \tilde{\mathbf{M}}_{q-1,i}^T\end{bmatrix} & \mathbf{V}_q^T\mathbf{A}_i\mathbf{V}_q \end{bmatrix}$, $\tilde{\mathbf{B}} \leftarrow \begin{bmatrix}\tilde{\mathbf{B}} \\ \mathbf{V}_q^T\mathbf{B}\end{bmatrix}$.
    Free memory: SWAP($\mathbf{V}_q$),
    Solve frequency sweep:
    $(\sum_{i=0}^{a}\tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}$
    if $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}}) < E_{2,tol}$ then
        return(converged)
    end if
end for
```

4.3.3 Adaptive Out-of-Core Algorithm for Multi-Point Method

Compared to the efforts for single-point methods, swapping the projection matrix to the hard disk is a rather simple task in the multi-point approach. Algorithm 7 shows the proposed modifications, where again the notation SWAP($\mathbf{V}_i$), LOAD($\mathbf{V}_i$) and CLEAR($\mathbf{V}_i$) are used for the swapping operations. In contrast to the single-point method above, the multi-point algorithm keeps only the currently generated block vector and one single column of the projection matrix in the RAM.

The out-of-core extension in the multi-point case is particularly simple, because no data of previous expansion points are required to evaluate the block $\mathbf{X}(\check{s}_q)$. However, the orthonormalization process as well as the reduction of the projection matrix is performed in exactly the same manner as for the single-point method.

Algorithm 7 Self-Adaptive Out-of-Core Multi-Point Model Order Reduction

```
Solve system at $\check{s}_1 = \min(\mathcal{B})$:
$(\sum_{i=1}^{a} \mathbf{A}_i \check{s}_1^i)\mathbf{X}(\check{s}_1) = \check{s}_1^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_1)$
Solve system at $\check{s}_2 = \max(\mathcal{B})$:
$(\sum_{i=0}^{a} \mathbf{A}_i \check{s}_2^i)\mathbf{X}(\check{s}_2) = \check{s}_2^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_2)$
Initial modified Gram-Schmidt: $[\mathbf{X}(\check{s}_1), \mathbf{X}(\check{s}_2)] = \mathbf{Q}_{[1:2]}\mathbf{U}$
Initial subspace projection: $\tilde{\mathbf{A}}_i = \mathbf{Q}_{[1:2]}^T\mathbf{A}_i\mathbf{Q}_{[1:2]}$, $\tilde{\mathbf{B}} = \mathbf{Q}_{[1:2]}^T\mathbf{B}$
SWAP($\mathbf{Q}_{[1:2]}$)
Solve frequency sweep for $s \in \mathcal{B}$:
$(\sum_{i=0}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_2\}_{\mathcal{B}}$
Set interval of worst error: $\check{e} = 1$
Initialize first interval: $\mathcal{B}_{\check{e}} = \mathcal{B}$
for $q = 3$ to $q_{max}$ do
    Find next expansion point: $\check{s}_q = \arg\min_{s\in\mathcal{B}} \left|s - \frac{\min\mathcal{B}_{\check{e}}+\max\mathcal{B}_{\check{e}}}{2}\right|$
    Split interval: $\mathcal{B}_{q-1} = [\check{s}_q, \max(\mathcal{B}_{\check{e}})]$, $\mathcal{B}_{\check{e}} \leftarrow [\min(\mathcal{B}_{\check{e}}), \check{s}_q]$
    Solve system at $\check{s}_q$:
    $(\sum_{i=0}^{a} \mathbf{A}_0 \check{s}_q^i)\mathbf{X}(\check{s}_q) = \check{s}_q^t\mathbf{B}, \Rightarrow \mathbf{X}(\check{s}_q)$
    Modified Gram-Schmidt in out-of-core technique:
    for $p = 1$ to $q-1$ do
        LOAD($\mathbf{Q}_p$)
        Orthonormalize only against loaded block vector:
        $[\mathbf{Q}_{[1:p-1]}, \mathbf{Q}_p, \mathbf{Q}_{[p:q-1]}]\mathbf{U} = [\dots, \mathbf{X}(\check{p}_q)]$
        CLEAR($\mathbf{Q}_p$)
    end for
    Orthonormalize columns of current block vector: $[\mathbf{Q}_{[p:q-1]}, \mathbf{Q}_q]\mathbf{U} = [\dots, \tilde{\mathbf{Q}}_q]$
    Enlarge ROM in out-of-core technique:
    for $p = 1$ to $q-1$ do
        LOAD($\mathbf{Q}_p$) ;   $\tilde{\mathbf{M}}_{p,i} = \mathbf{Q}_p^T\mathbf{A}_i\mathbf{Q}_q$ ;   CLEAR($\mathbf{Q}_p$)
    end for
    $\tilde{\mathbf{A}}_i \leftarrow \begin{bmatrix} \tilde{\mathbf{A}}_i & \begin{bmatrix}\tilde{\mathbf{M}}_{1,i}\\ \vdots \\ \tilde{\mathbf{M}}_{q-1,i}\end{bmatrix} \\ \begin{bmatrix}\tilde{\mathbf{M}}_{1,i}^T & \dots & \tilde{\mathbf{M}}_{q-1,i}^T\end{bmatrix} & \mathbf{Q}_q^T\mathbf{A}_i\mathbf{Q}_q \end{bmatrix}$, $\tilde{\mathbf{B}}_i \leftarrow \begin{bmatrix}\tilde{\mathbf{B}}\\ \mathbf{Q}_q^T\mathbf{B}\end{bmatrix}$.
    Free memory: SWAP($\mathbf{Q}_q$),
    Solve frequency sweep:
    $(\sum_{i=1}^{a} \tilde{\mathbf{A}}_i s^i)\tilde{\mathbf{X}} = s^t\tilde{\mathbf{B}}$, $\tilde{\mathbf{Y}} = \tilde{\mathbf{B}}^T\tilde{\mathbf{X}} \Rightarrow$ Sequence $\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}$
    if $E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}}) < E_{2,tol}$ then
        return(converged)
    end if
    Find interval of worst error: $\check{e} = \arg\max_{i=1,\dots,q-1} E_2(\{\tilde{\mathbf{S}}_q\}_{\mathcal{B}_i}, \{\tilde{\mathbf{S}}_{q-1}\}_{\mathcal{B}_i})$
end for
```

4.4 Numerical Results

The numerical experiments demonstrate the computational improvements as well as the need for the MORe approaches introduced in this thesis. Single-point methods, multi-point methods and full FE simulation sweeps are compared in the broadband analysis of electromagnetic structures. Furthermore, an explicit discussion of the memory usage shows the importance of out-of-core techniques in MORe processes.

All computational work in this section is performed on a single core of an *Intel Xeon* 2.33 GHz quad-core processor. The computer runs on an *openSUSE 11.1* operating system with a RAM size of 16 GByte. The operating system swapping partition size is set to 38.5 GBytes. The introduced broadband simulation framework is implemented in a c++ program, which is linked to the PARDISO [47] solver project for the factorization and solution of the FE systems.

4.4.1 Superiority of Blocking the WCAWE Process

Although the blocked WCAWE process spans the same Krylov subspaces as the non-blocked in theory, different operations are performed in the algorithms. Slight differences in the numerical computations of the Krylov subspaces result in a significantly improved MORe process, using the blocked approach. ROMs which are generated from the blocked approach have turned out to be highly accurate on a set of evaluation points $\mathcal{B}$, i.e. with a low true error $E_2(\{S\}_\mathcal{B}, \{\tilde{S}_q\}_\mathcal{B})$, with lower dimensions than ROMs generated from the non-blocked method. During this thesis, no further investigations on this effect beyond this numerical experiments were done.

Structure Definition

The considered test structure consists of a set of 8 lumped ports which, together with perfect electric conductor (PEC) sheets and lumped elements, form a kind of chain, see Fig. 4.5(a). The chain is surrounded by free space and a box bounds the computational domain. On both ends of the box surface impedance boundary conditions are imposed, which can be seen in Fig. 4.5(b), where $\sigma = 5.8 \cdot 10^5 A/(Vm)$ and $\mu = \mu_0$. All other surfaces of the box are defined as PEC. Detailed geometric dimensions of the structure are given in Fig. 4.6. Lumped element sheets have the values $R = 10k\Omega$, $L = 1mH$ and $C = 1pF$. Note that defining surface impedances and lumped elements results in a system parameterization where the frequency polynomial is of degree $a = 4$.

Numerical Experiments

For the test structure simulation, the set $\mathcal{B}$ is defined as the 1001 equidistant evaluation points in the bandwidth from 20 GHz to 40 GHz. The simulation employs first-order FE basis function and results in a system matrix dimension of $N = 39229$.

The most time-consuming process in the discrete sweep evaluation is the factorization, whose average computation time is about $3.9s$. Thus, with 1001 evaluation points, the complete time is extrapolated to $3904s$. As the dimension of the scattering matrix of this model is 8×8, which makes 64 scattering parameters, the frequency sweep in Fig. 4.7 shows only the eight scattering parameters S_{41} to S_{48} as representative selection.

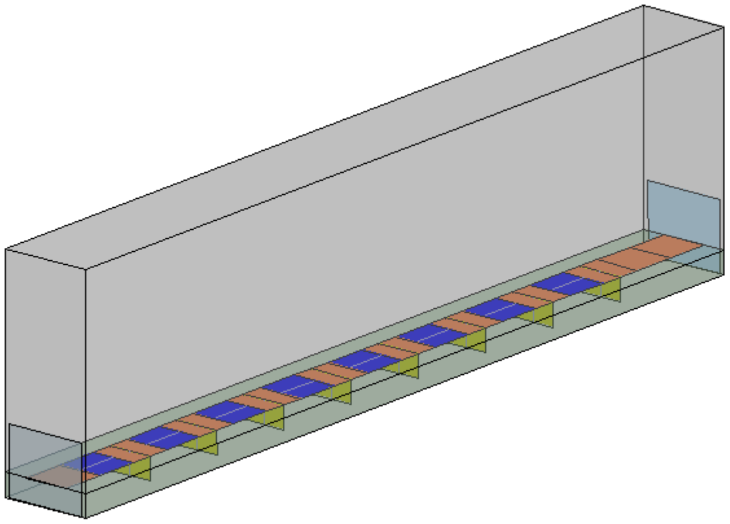
The MORe processes are applied to the FE discretization of the full FE run above, with the same number of degrees of freedom. The frequency of 20 GHz is chosen as the expansion point. This is not the optimal choice and will increase the iteration number, but shows in this numerical experiment the differences between blocked and non-blocked algorithms better.

The error indicator $E_2(\{\tilde{S}_{q-1}\}_{\mathcal{B}}, \{\tilde{S}_q\}_{\mathcal{B}})$, which is evaluated during the adaptive MORe process, is compared for the two approaches in Fig. 4.8(a). Until the error indicator falls under the threshold value of $E_{abort} = 10^{-5}$, the non-blocked approach runs $q = 38$ iterations, while the block process stops at $q = 25$ iterations. More interesting for the convergence, however, is the true error $E_2(\{S\}_{\mathcal{B}}, \{\tilde{S}_q\}_{\mathcal{B}})$, which is shown for the two runs in Fig. 4.8(b). To get a true error with $E_2(\{S\}_{\mathcal{B}}, \{\tilde{S}_q\}_{\mathcal{B}}) < 10^{-5}$, we need $q = 38$ for the non-blocked and $q = 24$ for the blocked WCAWE. In addition to the lower iteration number for the blocked WCAWE, the more smoothly falling error is particularly noticeable in the plots, which may be interpreted as an indicator for improved numerical robustness. Moreover, the error sweeps for both approaches at $q = 25$ in Fig. 4.9, where $e_{ij}(f) = |S_{ij}(f) - \tilde{S}_{ij}(f)|$, confirm the better convergence for the blocked WCAWE. Note that only for visualization reasons not all scattering parameters are shown in this plot again. Unsurprisingly, the blocked approach with a lower number of iterations results in lower computation times, see Table 4.1.

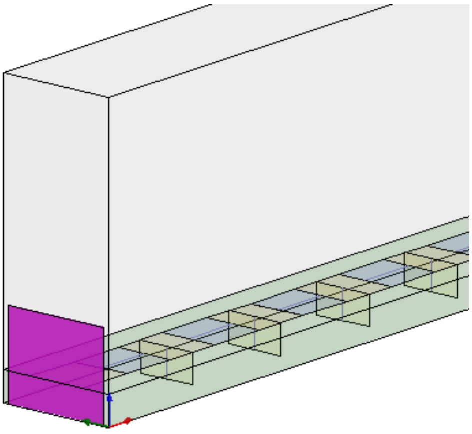
In this numerical experiment the blocked WCAWE approach generated a ROM with lower dimension to reach the same accuracy as the non-blocked approach. Thus, the blocked approach is not only superior in memory requirements, but also the dimension of the resulting ROM is smaller. The author guesses that the orthogonalization process in the blocked approach is more stable and therefore the projection matrix spans the Krylov subspace more properly.

Table 4.1: Port chain: Computational data.

	Number of iterations q	ROM generation time (s)	ROM evaluation time (s)	Sweep time (s)
Full FE run	-	-	-	3904
Non-blocked WCAWE	38	420	16	436
Blocked WCAWE	25	211	6	217



(a) Sheets: Lumped ports (blue), lumped elements (yellow) and PEC (red).



(b) Impedance boundary (purple).

Figure 4.5: Port chain: View of the structure.

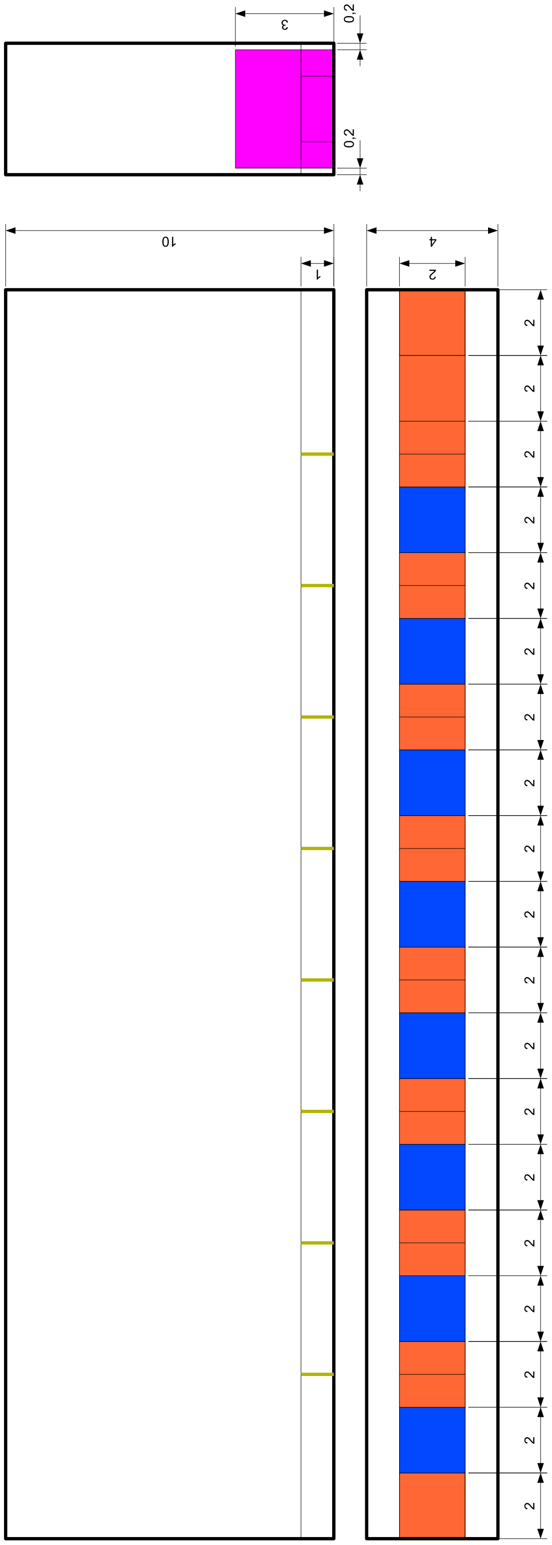


Figure 4.6: Port chain: Geometric dimensions in mm.

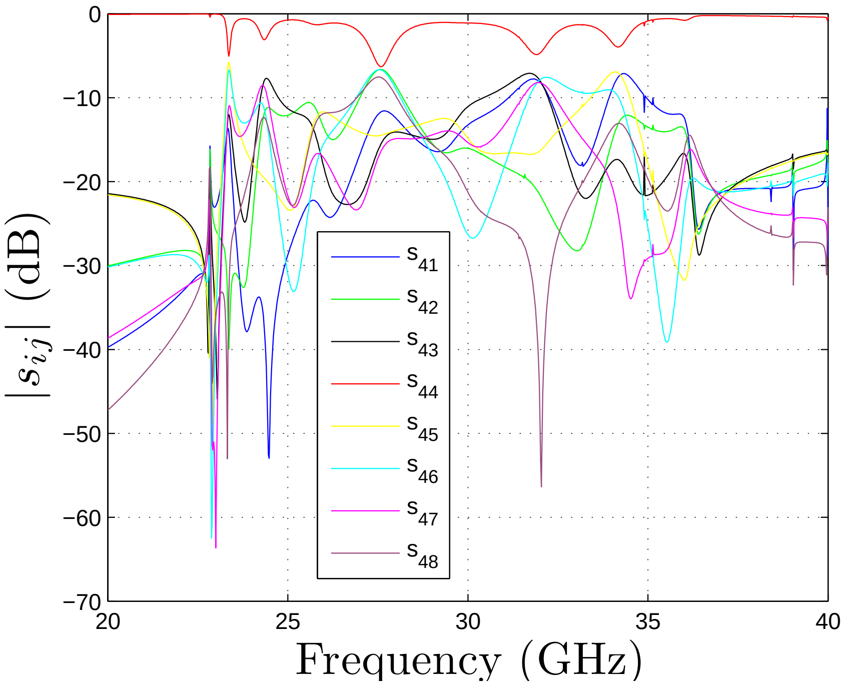
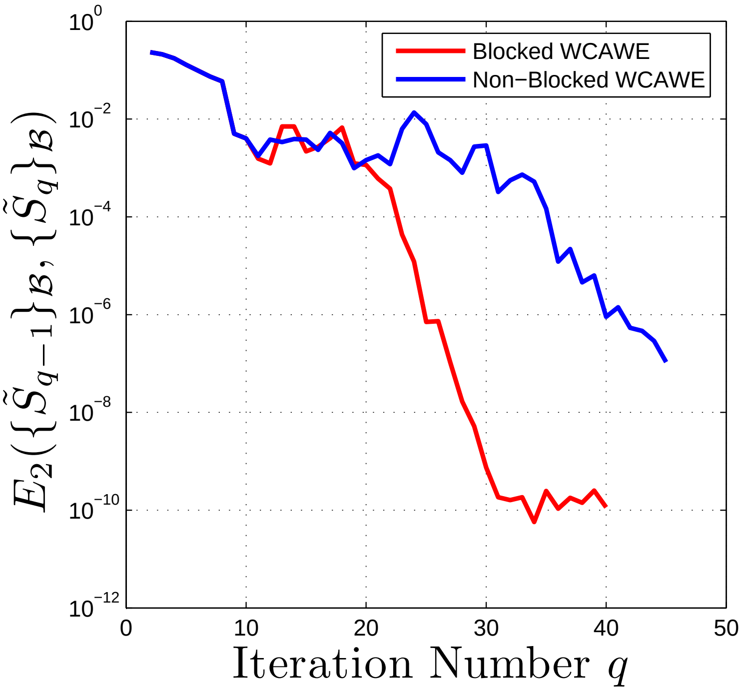
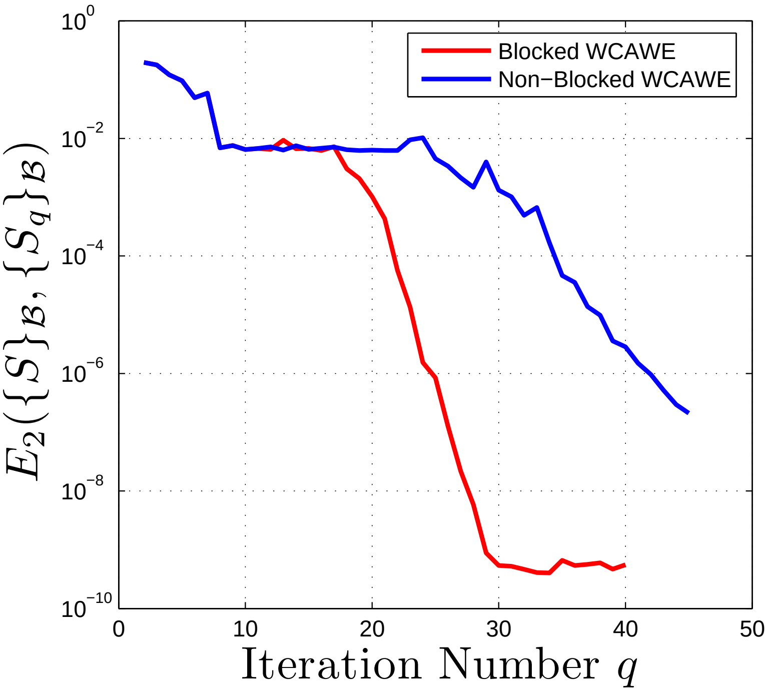


Figure 4.7: Port chain: Scattering parameter versus frequency.

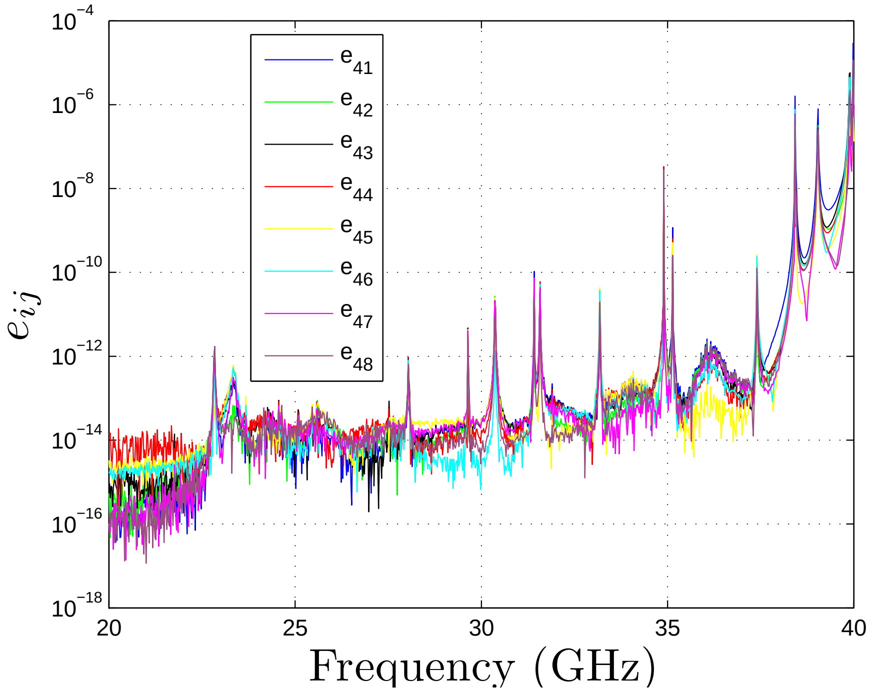


(a) Error indicator versus iteration number.

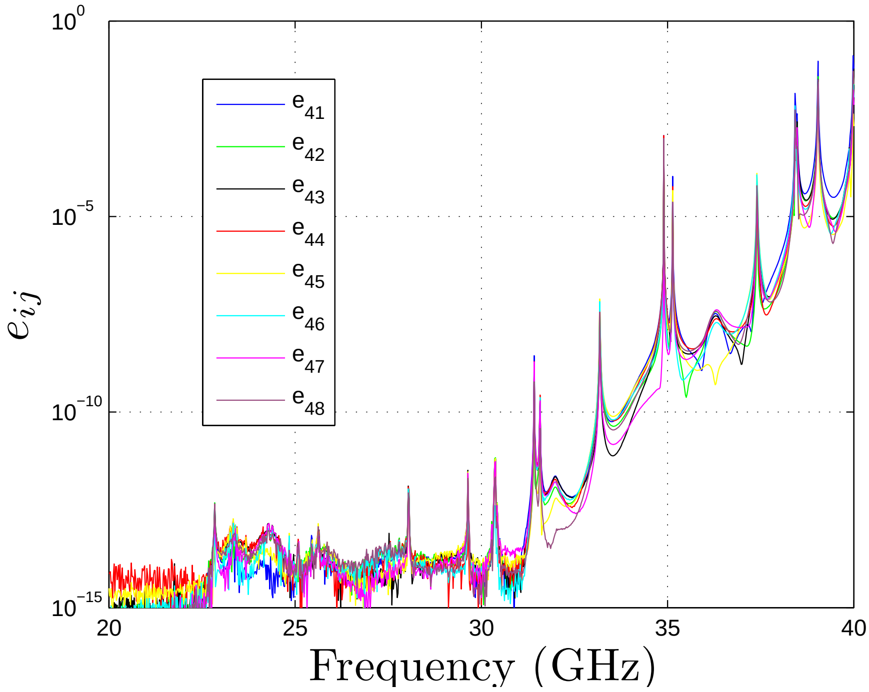


(b) True error versus iteration number.

Figure 4.8: Port chain: Error indicator and true error versus iteration number.



(a) Blocked WCAWE process.



(b) Non-blocked WCAWE process.

Figure 4.9: Port chain: Error in scattering parameters versus frequency for $q = 25$.

4.4.2 Out-of-Core Technique Results

The second part of this section is dedicated to the out-of-core techniques. To give meaningful data and results for the introduced techniques, the chosen test structure needs to satisfy a couple of requirements. The system of linear equations, which results from the FE simulation of the structure, needs to be of large dimension, with many right-hand sides. The frequency parameterization should be polynomial, otherwise the WCAWE process will reduce to an Arnoldi iteration. To increase the number of adaptive iterations in the MORe processes, the test structure for this numerical experiment should exhibit some resonant behavior in the evaluated bandwidth.

Test Structure Definition with Absorbing Boundaries

The considered structure is the cuboid in Fig. 4.10. The structure contains 8 port chains, each with 8 lumped ports, in total 64 lumped ports. The ports are connected by PEC sheets and lumped elements, which have resistance, inductance and capacitance of $R = 10k\Omega$, $L = 1mH$ and $1pF$, respectively. On the top surface of the cuboid an absorbing boundary condition is defined, all other surface are PEC. The chains are surrounded by a vacuum. Fig. 4.11 defines the detailed geometric dimension of the structure.

Discrete Sweep Data

The set of evaluation points $\mathcal{B}$ in the simulation is defined as 2001 equidistant points in the bandwidth from 20 GHz to 30 GHz. The full FE simulation, which is used as reference, employs second-order FE shape functions and results in a sparse system of linear equations with 827666 degrees of freedom. The assembled system matrix counts 16119523 non-zeros, which causes a memory requirement of 257.9 MBytes, if 8 Bytes double precision floating point numbers are taken and all non-zero values are assumed to be complex (thus 16 Bytes per entry). The memory requirement for the block right-hand side, with 64 extremely sparse excitation vectors, may be neglected. However, the factorization of the system matrix counts 236161487 non-zeros, which requires 3.8 GBytes. Note that the peak memory requirement of the factorization is not considered in this analysis.

The typical factorization time for a solution of the system is 186.2s. The additional time for each right-hand side solution is typically 4.3s, which is not negligible in this case of 64 right-hand sides. As a numerical result, Fig. 4.12 shows the scattering parameters $S_{20\,17}$ to $S_{20\,24}$ for the evaluation points in $\mathcal{B}$ as a chosen selection.

Broadband Simulation Techniques

In the next numerical experiment, the scattering parameter sweep on $\mathcal{B}$ is performed using the proposed MORe techniques, i.e. adaptive single-point and multi-point as in-core as well as out-of-core methods. The MORe is applied to the original FE system above, where in this constellation the system is parameterized in the frequency to the order $a = 2$. Single-point methods use the arbitrarily chosen expansion frequency 25 GHz.

The plot of the employed error indicator $E_2(\{\tilde{S}_{q-1}\}_\mathcal{B}, \{\tilde{S}_q\}_\mathcal{B})$ in Fig. 4.13(a) shows, that the adaptive single-point method needs to run to the iteration $q = 9$, where the multi-point process needs only $q = 8$, to satisfy the threshold value $E_2(\{\tilde{S}_{q-1}\}_\mathcal{B}, \{\tilde{S}_q\}_\mathcal{B}) < 10^{-5}$. The plot for the true error $E_2(\{S\}_\mathcal{B}, \{\tilde{S}_q\}_\mathcal{B})$ confirms this faster convergence, see Fig. 4.13(b). For a true error below 10^{-5}, the multi-point method needs $q = 7$ and the single-point method $q = 8$.

MORe Memory Analysis

A first performance analysis is done on the plots in Fig. 4.14, which show the memory usage versus time for the considered approaches. In the single-point memory plot, see Fig. 4.14(a), it can be seen that the memory requirements increase at each iteration for the in-core algorithm, which results from the projection matrix enlargement. Each enlargement of a full block right-hand side is 64 columns, each with 827666 degrees of freedom and complex double precision floating point numbers (16 Bytes) per degree of freedom, which is a memory size of 848 MByte. However, the out-of-core approach does not increase the projection matrix data for higher iterations. Note that, for each iteration, an additional full right-hand side block is required in the out-of-core block WCAWE process for a short time, as explained in Section 4.3.2. These are the eye-catching jumps in the memory plot in each iteration.

Turning to the multi-point curves in Fig. 4.14(b), a similar behavior can be observed. In each iteration the memory block of 848 MByte is allocated for the new projection matrix columns. While the out-of-core process swaps data from previous iterations to the hard disk and the memory requirements remain on a low level, the memory requirements in the in-core approach increase at each iteration with the size of this block. In each iteration, the factorization memory allocation and deallocation can be noticed, the size of which was evaluated above to be 3.8 GBytes.

In the next experiment, a projection matrix data size is forced that exceeds the available system main memory. For this purpose, all adaptive MORe approaches are run for $q = 18$ iterations. The resulting in-core computations can be seen in 4.15(a), where the memory requirements run into the limit of 16 GBytes and the operating system swapping process starts to work. This makes the process extremely inefficient. While the single-point approach got stuck at the iteration $q = 11$, the multi-point approach at least works at the last iterative step, when the experiment was manually aborted after running $1.6 \cdot 10^5 s$. The out-of-core approach, however, runs without any remarkable additional memory requirements, see Fig. 4.15(b).

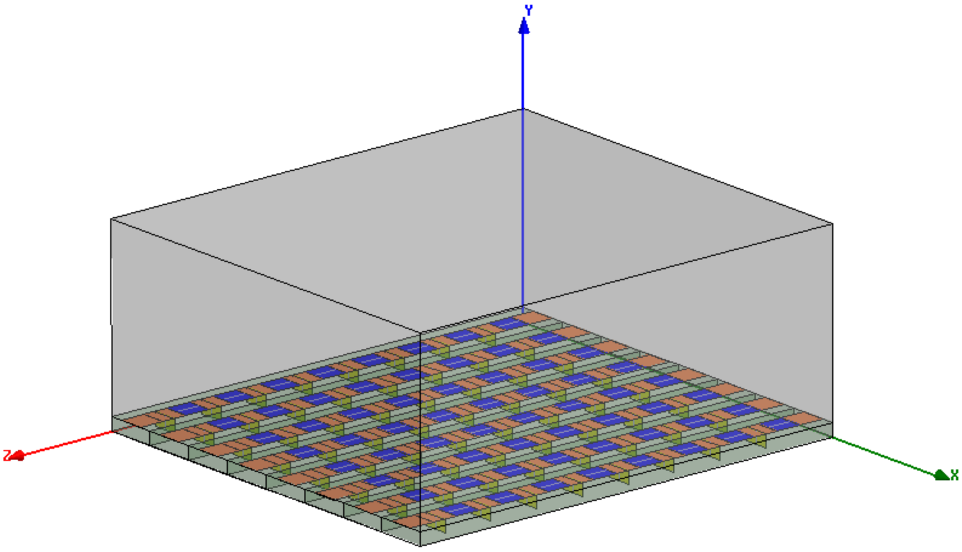
Runtime Discussion

The runtime discussion for the considered MORe techniques starts with Table 4.2. Keeping in mind that the multi-point approach needs one fewer adaptive iteration than the single-point method, i.e. $q = 8$, to generate a ROM which is considered converged, the computational costs are lower, as can be expected. Anyhow, the listed results in the Table 4.2 show clearly that the out-of-core process swapping mechanism is not the limiting operation in the performed experiments. It should also be mentioned that all MORe techniques are far superior to the discrete FE sweep, which is the main purpose of MORe. The full FE computation time only includes the factorization time and the solution of the 64 right-hand sides at each evaluation frequency.

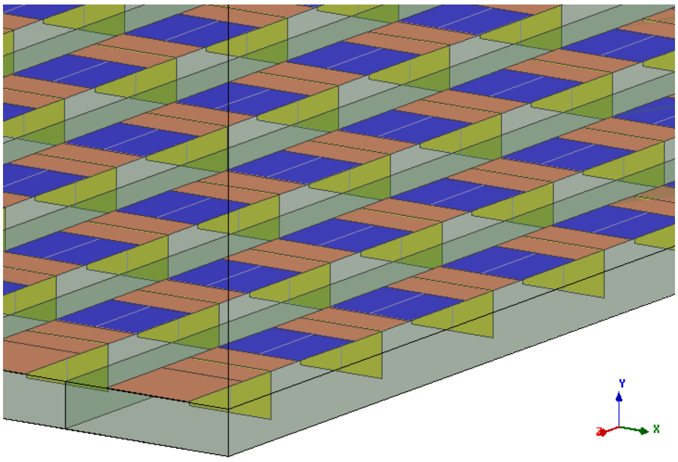
We start the detailed timing analysis by looking at the memory plots in Fig. 4.14(a). Particularly the last iteration of the in-core technique is investigated. This iteration starts before the memory rises the last time (at 10206s) and ends just before the memory is released (at 12748s), which is a time interval of 2542s. The time for this adaptive step results from 941s for the block WCAWE process, 1333s for the subspace projection and 268s for the error measure evaluation. Detailed investigation showed that the current implementation works on non-optimized matrix-vector product and vector-vector product operations. This makes the computation times for the block WCAWE process as well as for the subspace projection rather high compared to the highly optimized operations in the PARDISO solver [47]. This explains also why the multi-point approach is faster than the single-point method in Table 4.2, if both computations run 9 adaptive iterations.

Table 4.2: Port cuboid: Computational data.

	Number of iterations q	ROM generation time (s)	ROM evaluation time (s)	Sweep computation time (s)
Single-point:				
In-core	9	12746	267	13013
Out-of-core	9	12864	268	13132
Multi-point:				
In-core	8	9337	201	9538
Out-of-core	8	9554	201	9755
In-core	9	11451	268	11719
Out-of-core	9	11656	268	11924
Full FE run:	-	-	-	923261



(a) Full cuboid.



(b) Sheets: Lumped ports (blue), lumped elements (yellow) and PEC (red).

Figure 4.10: Port cuboid: View of the structure.

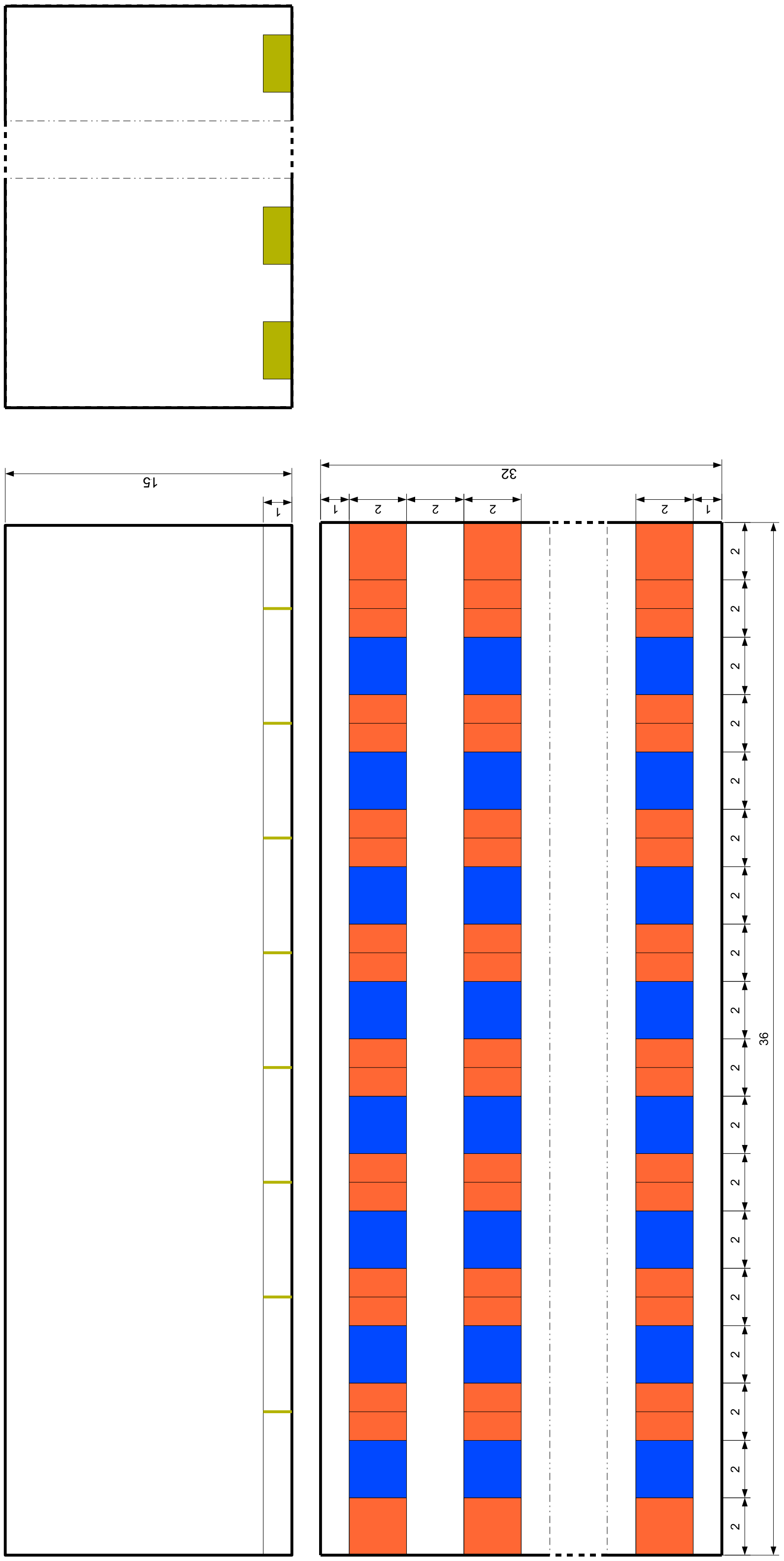


Figure 4.11: Port cuboid: Geometric dimensions in mm.

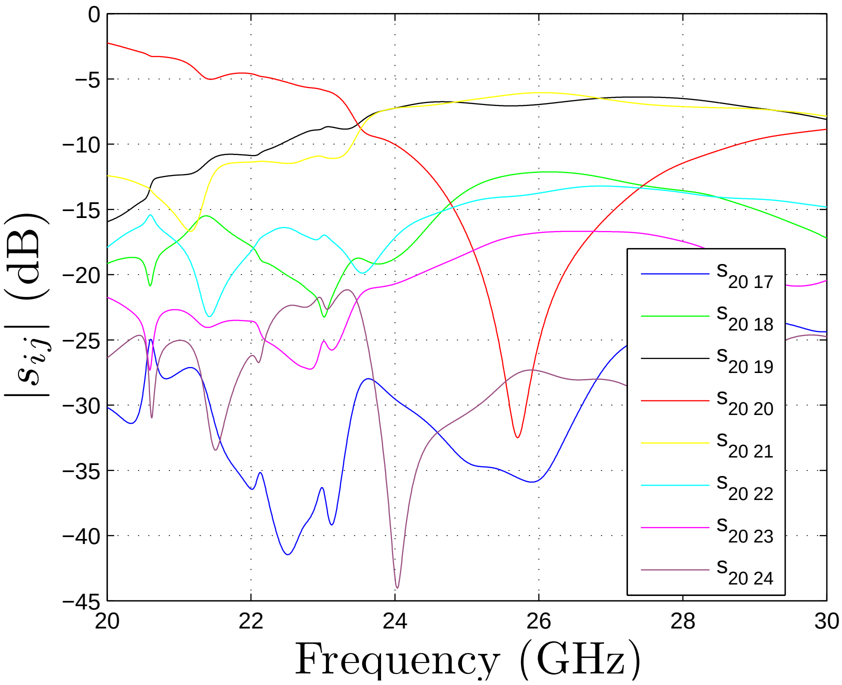
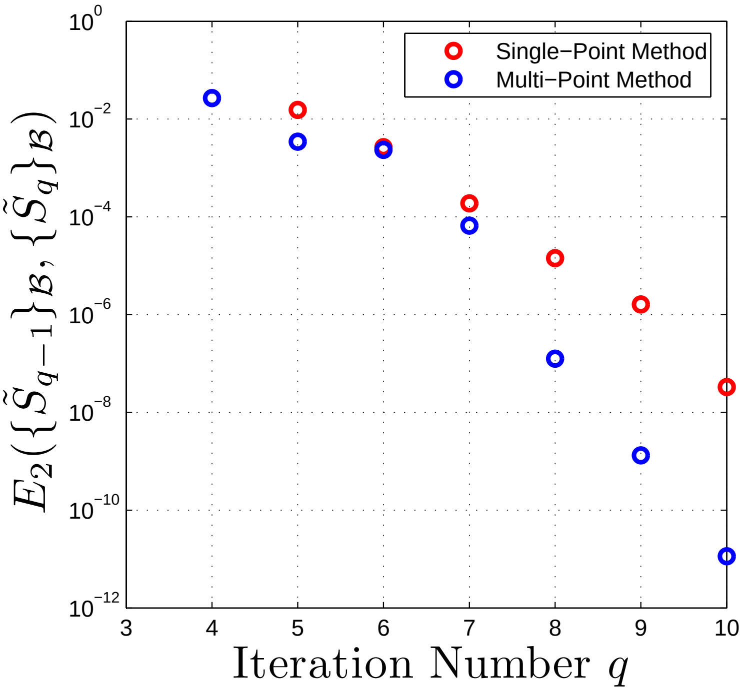
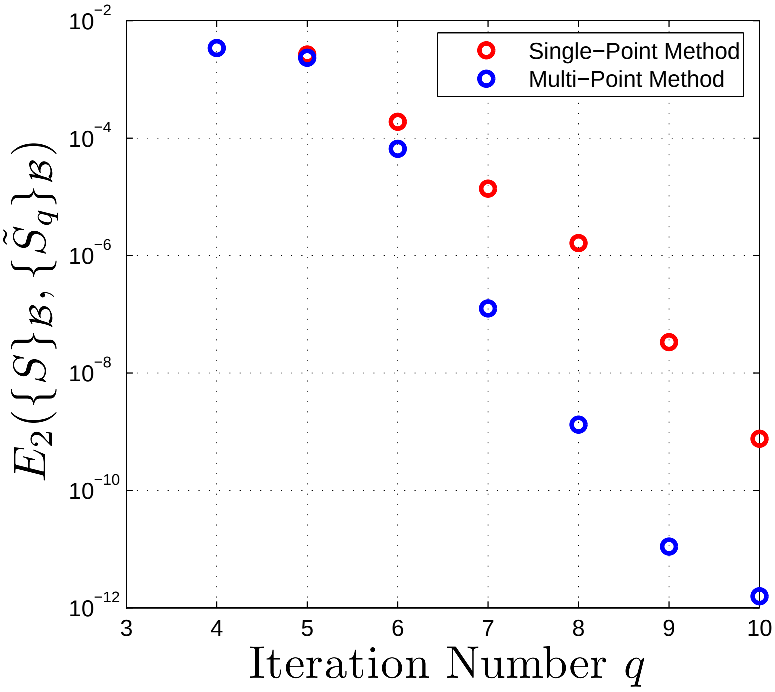


Figure 4.12: Port cuboid: Scattering parameters versus frequency.

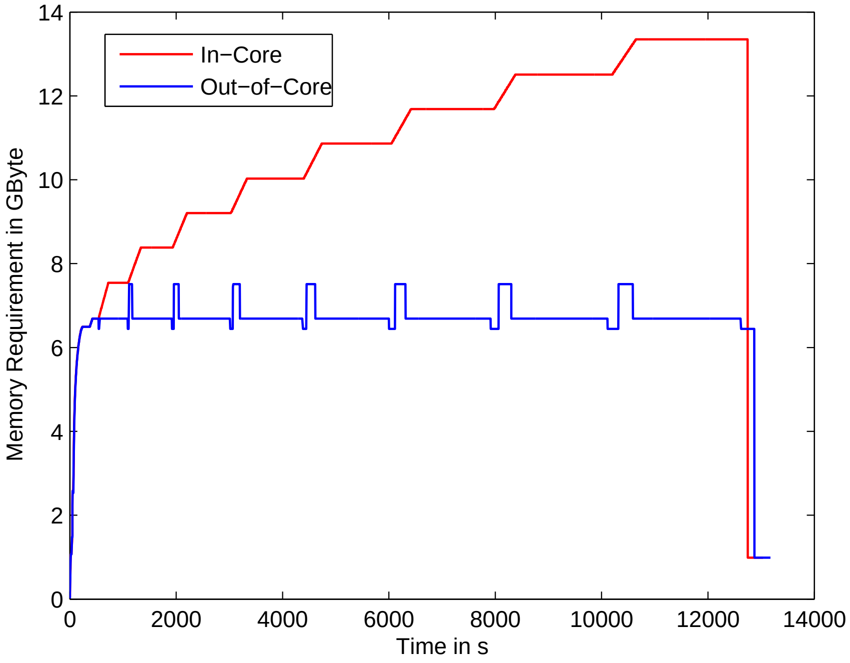


(a) Error indicator versus iteration number.

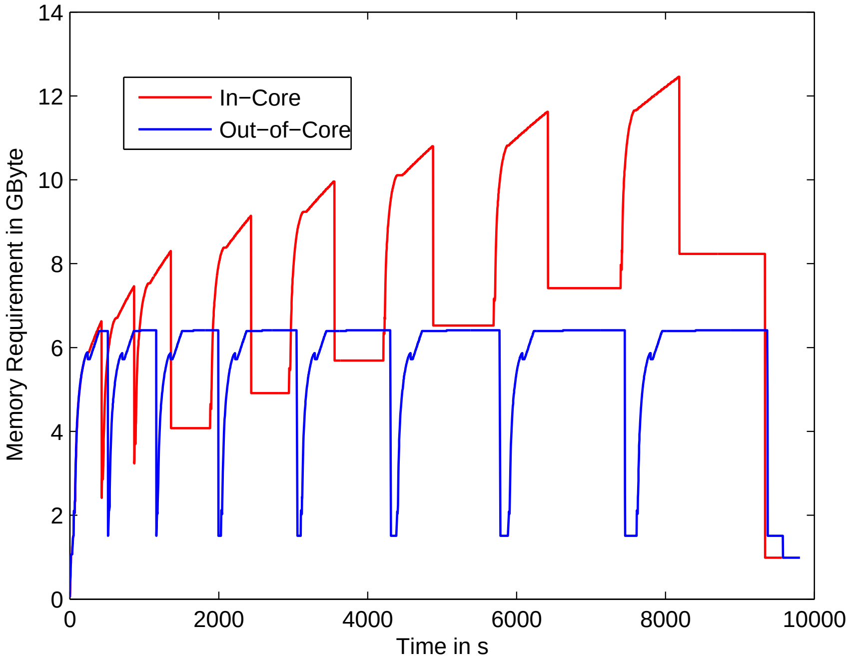


(b) True error versus iteration number.

Figure 4.13: Port cuboid: Error indicator and true error versus iteration number.

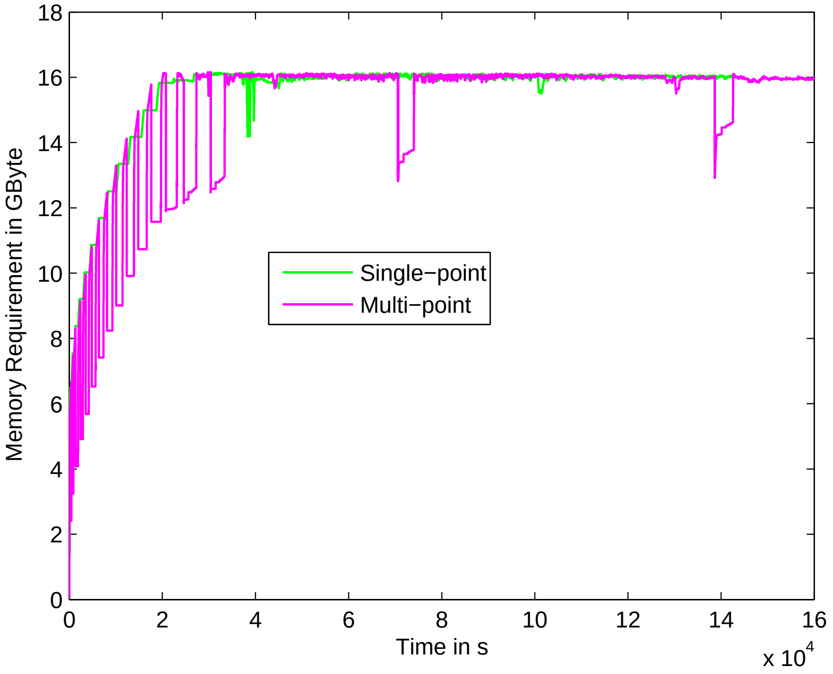


(a) Single-point method, $q = 9$.

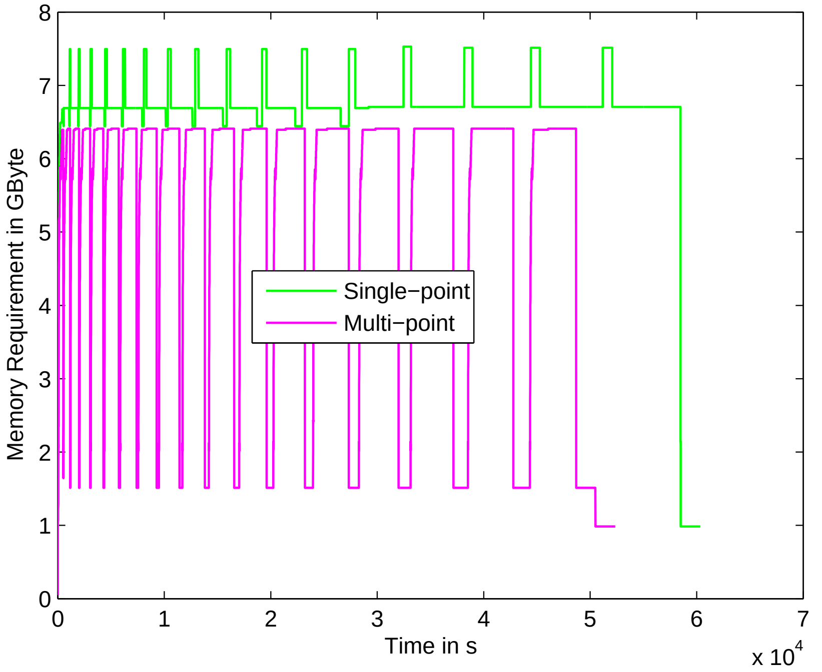


(b) Multi-point method, $q = 8$.

Figure 4.14: Port cuboid: Memory plots for adaptive MORe techniques.



(a) In-core approaches with filled main memory.



(b) Out-of-core approaches.

Figure 4.15: Port cuboid: Memory plots with forced iteration number $q = 18$.

Test Structure with Surface Impedance

For the last numerical experiment in this section the absorbing boundary condition in the structure above is replaced by a surface impedance boundary with a conductivity of $\sigma = 5.8 \cdot 10^5\ A/(Vm)$ and a permeability of $\mu = \mu_0$. Thus, the system matrix is parameterized now to the order $a = 4$ in the frequency. In both constellations the FE simulation employs the same mesh. Therefore, the time and memory requirements in the full FE run do not exhibit any remarkable differences to the run above. However, this modification results in a less smooth behavior in the frequency response of the port cuboid, as can be seen in Fig. 4.16. Both adaptive processes, single-point and multi-point methods, run one adaptive iteration more and converge more slowly as the plots in Fig. 4.17 show. In particular, the multi-point method runs $q = 9$ iterations and the single-point method needs $q = 10$ iterative steps.

The runtime comparison for the multi-point MORe with $q = 9$ in Table 4.3 and Table 4.2 shows that the model modification slowed down the MORe process. The longer computation times can be explained by the higher polynomial degree in the frequency parameterization of the system matrices, which results from the surface impedance boundary condition.

This effect is much more obvious for the single-point method, comparing the memory plots Fig. 4.18(a) and Fig. 4.14(a). The origin for the massive slow down is the same as for the multi-point method, the higher polynomial degree in the system parameterization. The effect, however, is much more pronounced as a result of the expansion point shift, which is performed only in single-point methods. The shift causes that all 5 system matrices have non-zero entries, where some of them may have a much higher density as the non-shifted, which more slows down some operations compared to the multi-point approach. For a detailed timing analysis we choose again the 9th iteration of the in-core technique, which is the same iteration as used in the absorbing boundary condition runtime analysis above. Note that this is not the last adaptive iteration in this run. The interval starts at 14831s and ends at 18712s and consists of 1372s for the block WCAWE, 2226s for the subspace projection and 283s for the error evaluation. Thus it is the subspace projection, which runs almost 15 min longer, that especially slows down the computation.

Table 4.3: Port cuboid with surface impedance: Computational data.

	Number of iterations q	ROM generation time (s)	ROM evaluation time (s)	Sweep computation time (s)
Single-point:				
In-core	10	23112	381	23112
Out-of-core	10	23196	381	23196
Multi-point:				
In-core	9	12731	268	12999
Out-of-core	9	12966	269	13235
Full FE run:	-	-	-	923261

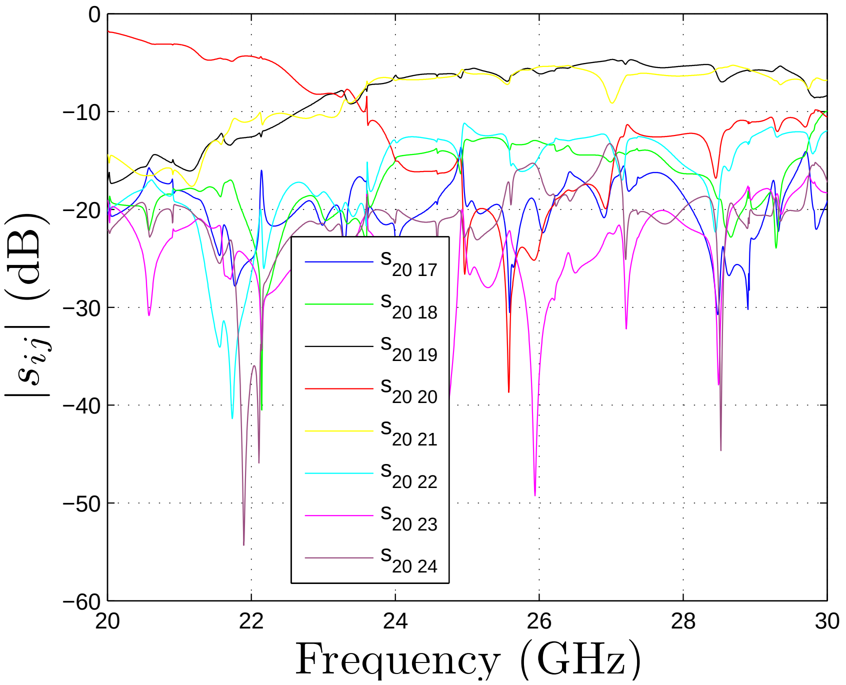
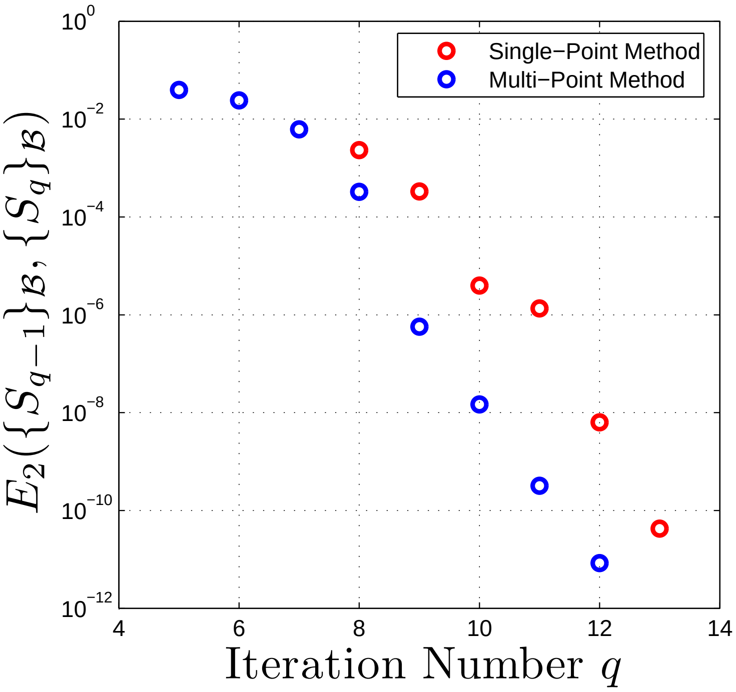
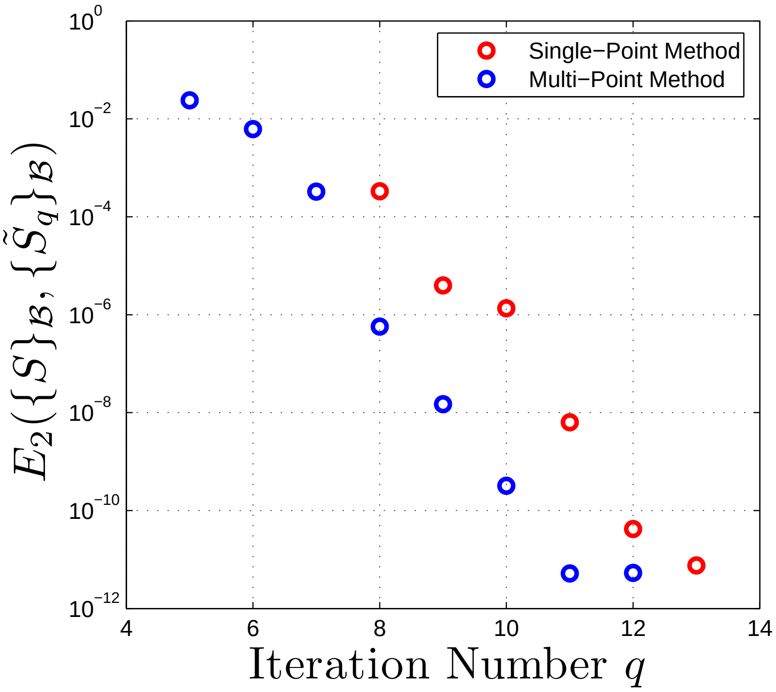


Figure 4.16: Port cuboid with surface impedance: Scattering parameters versus frequency.

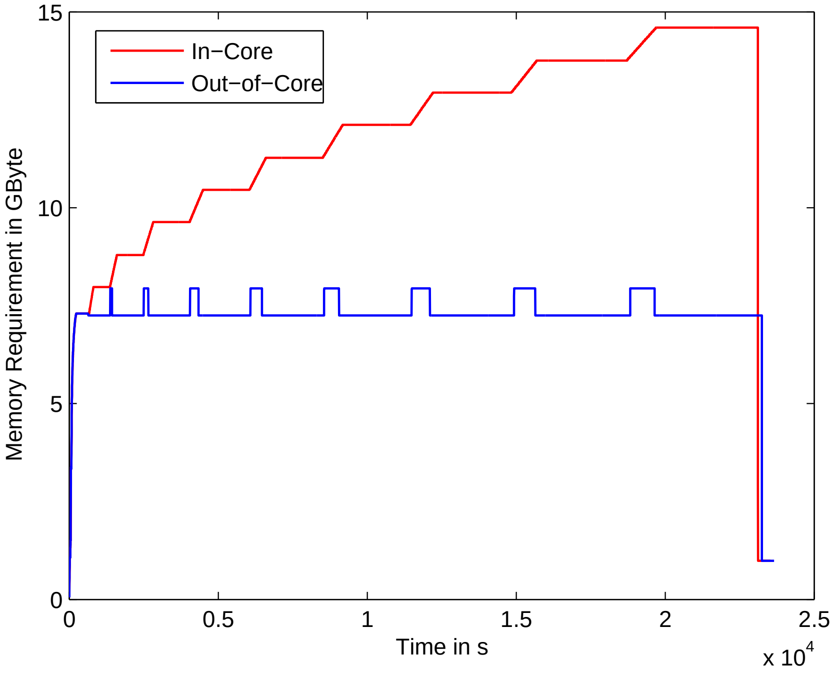


(a) Error indicator versus iteration number.

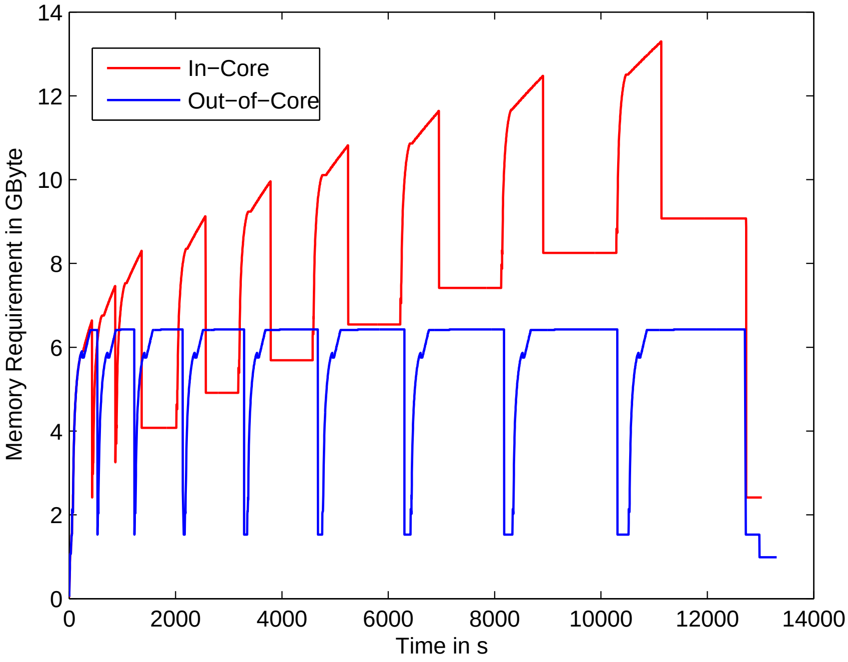


(b) True error versus iteration number.

Figure 4.17: Port cuboid with surface impedance: Error indicator and true error versus iteration number.



(a) Single-point method.



(b) Multi-point method.

Figure 4.18: Port cuboid with surface impedance: Memory plots for adaptive MORe techniques.

4.5 Conclusion

The provided adaptive MORe approaches in combination with the chosen error measure show reliable results and work very efficiently. The memory swapping mechanisms only slightly increase the computation times and therefore do not relevantly reduce the efficiency of the MORe techniques. However, the employed out-of-core approaches may cause an important performance improvement, if the projection matrix memory requirements exceed the available main memory, as may happen in real-world problem broadband simulations. Memory plots show the reduction of the main memory requirement due to the projection matrix data swapping. A high number of adaptive iterations can be performed using the out-of-core approach, without filling the main memory.

To show the practical importance of the introduced numerical techniques in this thesis, numerical computations run on the limit of the system capacity. This showed some unexpected results, which are mainly caused by non-optimized numerical operations in the current implementation. Timing analysis showed that non-optimized matrix-vector and vector-vector products mainly slow down the single-point methods. Although we have this shortcoming in the implemented framework, which affects single-point as well as multi-point methods, all MORe runs are far superior to discrete FE sweeps.

Chapter 5

Broadband Finite Element Simulation for Waveguide Problems

5.1 Model Order Reduction for Waveguide Problems

The FE method is also a very powerful technique for the modal analysis of electromagnetic waveguides. Since it is just the waveguide cross-sections that need to be discretized, single solutions are computationally inexpensive. In many applications, however, the characteristics of waveguide modes are to be determined over wide frequency bands. Since modal field patterns may be frequency-dependent and the corresponding propagation coefficients highly dispersive, and because dispersion curves may feature bifurcations, cross-over points, or coupled-mode sections, the broadband analysis of electromagnetic waveguides typically requires a large number of FE computations, at different frequency points. In such cases, computer runtime is still a limiting factor.

MORe techniques such as [51], [52] and [53], provide a means to speed up the solution times of frequency sweeps very significantly, at little additional error. This chapter introduces a multi-point MORe method employing an adaptive point-placement scheme for controlling the error. An incremental error indicator for the propagation coefficient is provided to guide the adaptive process.

The considered waveguides are assumed to be bounded by electric and magnetic walls, and to possess material properties that are scalar-valued and uniform along the waveguide axis z but non-uniform in the transverse plane t. In consequence, the axial behavior of the modal fields is given by $\exp(-\gamma z)$, wherein γ denotes the propagation coefficient.

5.2 Finite-Element Formulation

For stability reasons, the chosen formulation is based on a magnetic vector potential $\vec{A}$ and a scaled electric scalar potential ϕ, as introduced in [54], [4]. Specifically, we employ the gauge $A_z = 0$ and decompose $\vec{A}_t$ into components of non-vanishing circulation $\vec{A}_t^c$ plus the transverse gradient ∇_t of a scalar field ψ. Hence we have

$$\vec{A} = e^{-\gamma z}(\vec{A}_t^c(x,y) + \nabla_t \psi(x,y)), \tag{5.1}$$

$$\phi = e^{-\gamma z} V(x,y), \tag{5.2}$$

and the electromagnetic fields are represented by

$$\vec{B} = e^{-\gamma z}[\nabla_t \times \vec{A}_t^c - \hat{e}_z \times \gamma(\vec{A}_t^c + \nabla_t \psi)], \tag{5.3}$$

$$\vec{E} = -jc_0 e^{-\gamma z}[\gamma \hat{e}_z(jV) - \nabla_t(jV) + k(\vec{A}_t^c + \nabla_t \psi)]. \tag{5.4}$$

By plugging (5.1) and (5.2) into the time-harmonic Maxwell equations, we arrive at the eigenvalue problem

$$\begin{aligned} \nabla_t \times \mu_r^{-1} \nabla_t \times \vec{A}_t^c - k\varepsilon_r \nabla_t(k\psi - jV) - k^2 \varepsilon_r \vec{A}_t^c = \\ -\gamma^2 \left[\hat{e}_z \times \mu_r^{-1} \hat{e}_z \times \vec{A}_t^c + \hat{e}_z \times \mu_r^{-1} \hat{e}_z \times \nabla_t \psi \right], \end{aligned} \tag{5.5a}$$

$$\nabla_t \cdot \epsilon_r [k\vec{A}_t^c + \nabla_t(k\psi - jV)] = \gamma^2 \epsilon_r jV. \tag{5.5b}$$

FE discretization results in the algebraic EVP

$$(\mathbf{S}_0 + k\mathbf{S}_1 + k^2\mathbf{S}_2)\mathbf{x}_m = \gamma_m^2 \mathbf{T}\mathbf{x}_m \qquad \text{with } \mathbf{x}_m = \begin{bmatrix} \mathbf{x}_A \\ \mathbf{x}_\psi \\ \mathbf{x}_V \end{bmatrix}_m, \tag{5.6}$$

wherein $\mathbf{x}_A$, $\mathbf{x}_\psi$, and $\mathbf{x}_V$ denote the component vectors for $\vec{A}_t^c$, ψ, and (jV), respectively, and $\mathbf{S}_0$, $\mathbf{S}_1$, $\mathbf{S}_2$, and $\mathbf{T}$ are sparse symmetric matrices, whose structure can be found in [4]. Note that (5.5) is satisfied not only by physical waveguide modes but also by a set of null-field solutions, i.e. non-trivial solutions with $\vec{E} = 0$ and $\vec{B} = 0$:

$$\gamma = 0, \tag{5.7}$$

$$\vec{A}_t^c = 0, \tag{5.8}$$

$$jV = k\psi, \tag{5.9}$$

with arbitrary ψ. In the FE context (5.6), the null-field solutions $\mathbf{n}$ read

$$\gamma = 0, \tag{5.10}$$

$$\mathbf{n} = \mathbf{N}\mathbf{x}_\psi, \tag{5.11}$$

$$\mathbf{N}(k) = \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \\ k\mathbf{I} \end{bmatrix}, \tag{5.12}$$

with arbitrary $\mathbf{x}_\psi$. Eq. (5.6) implies the generalized orthogonality equation

$$\mathbf{x}_m^T \mathbf{T} \mathbf{x}_n = 0 \qquad \text{for } m \neq n. \tag{5.13}$$

Hence any superposition of physical modes $\mathbf{p}(k)$ satisfies

$$\mathbf{N}(k)^T \mathbf{T}\mathbf{p}(k) = \mathbf{0}, \tag{5.14}$$

which enables us to reconstruct $\mathbf{p}$ from given components $\mathbf{p}_A$ and $\mathbf{p}_\psi$. The resulting equation takes the form

$$\mathbf{p}(k) = (\mathbf{P}_0 + k\mathbf{P}_1) \begin{bmatrix} \mathbf{p}_A \\ \mathbf{p}_\psi \end{bmatrix}. \tag{5.15}$$

The structure of the matrices $\mathbf{P}_0$ and $\mathbf{P}_1$ can be found in [4].

5.3 Multi-Point Model Order Reduction

The main idea behind the waveguide MORe approach is to restrict the trial and test spaces in the FE system (5.6) to suitable subspaces whose dimension q is much smaller than that of the original system, p. For this purpose, we construct projection matrices $\mathbf{V}(k), \mathbf{W}(k)$ and apply a two-sided projection process to (5.6). The resulting reduced-order eigenvalue problem is of the form

$$(\tilde{\mathbf{S}}_0 + k\tilde{\mathbf{S}}_1 + k^2\tilde{\mathbf{S}}_2)\tilde{\mathbf{x}}_m = \tilde{\gamma}_m^2 \tilde{\mathbf{T}}\tilde{\mathbf{x}}_m \tag{5.16}$$

with

$$\tilde{\mathbf{S}}_i(k) = \mathbf{W}(k)^H \mathbf{S}_i \mathbf{V}(k) \qquad i \in \{0, 1, 2\}, \tag{5.17a}$$

$$\tilde{\mathbf{T}}(k) = \mathbf{W}(k)^H \mathbf{T}\mathbf{V}(k), \tag{5.17b}$$

and the corresponding approximations $\mathbf{x}'_m$ to the eigenvectors of (5.6) are given by

$$\mathbf{x}'_m(k) = \mathbf{V}(k)\tilde{\mathbf{x}}_m(k). \tag{5.18}$$

The matrices $\mathbf{V}(k)$ and $\mathbf{W}(k)$ are constructed as follows: we first compute the dominant M eigenpairs of (5.6) at N expansion wavenumbers k_n and assemble their components in $\vec{A}_t^c$ and ψ to a matrix $\mathbf{\Xi}$:

$$\mathbf{\Xi} = \left[\begin{bmatrix} \mathbf{x}_A^1 \\ \mathbf{x}_\psi^1 \end{bmatrix}_{k_1}, \dots, \begin{bmatrix} \mathbf{x}_A^M \\ \mathbf{x}_\psi^M \end{bmatrix}_{k_1}, \dots\dots, \begin{bmatrix} \mathbf{x}_A^M \\ \mathbf{x}_\psi^M \end{bmatrix}_{k_N} \right]. \tag{5.19}$$

To provide a stable basis, we next compute the **QR** factorization of $\mathbf{\Xi}$. In view of (5.14), we then construct $\mathbf{V}(k)$ by

$$\mathbf{V}(k) = (\mathbf{P}_0 + k\mathbf{P}_1)\mathbf{Q}. \tag{5.20a}$$

Hence the trial space of the ROM, $\operatorname{colsp} \mathbf{V}(k)$, contains superpositions of physical modes only. According to (5.18), the approximate eigenvectors $\mathbf{x}'_m$ satisfy (5.14), which assures that the ROM will not lead to null-space solutions.

Following an idea from [55], $\mathbf{W}(k)$ is taken to be

$$\mathbf{W}(k) = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{I} \end{bmatrix} \mathbf{V}(k). \tag{5.20b}$$

By plugging (5.20) and (5.17) into (5.16) and collecting terms of equal power in k, we arrive at the final form of the ROM:

$$(\tilde{\mathbf{S}}_{S0} + k\tilde{\mathbf{S}}_{S1} + k^2\tilde{\mathbf{S}}_{S2} + k^3\tilde{\mathbf{S}}_{S3} + k^4\tilde{\mathbf{S}}_{S4})\hat{\mathbf{x}}_m = \tilde{\gamma}_m^2(\tilde{\mathbf{T}}_{T0} + k\tilde{\mathbf{T}}_{T1} + k^2\tilde{\mathbf{T}}_{T2})\hat{\mathbf{x}}_m. \tag{5.21}$$

Eq. (5.21) features explicit k dependence, and all matrices are in $\mathbb{C}^{q \times q}$. Since $q \ll p$, the eigenvalue problem (5.21) can be solved much more efficiently than the underlying FE system (5.6).

5.3.1 Homogeneous Material Properties

Waveguides with homogeneous material properties are known to support transverse electric (TE), transverse magnetic (TM), and possibly transverse electromagnetic (TEM) modes. They all have in common that the transverse field patterns are independent of frequency. Eq. (5.3) implies that the modal patterns in terms of $\vec{A}_t^c$ and ψ must be frequency-independent, too. Since the corresponding components in V, which do depend on frequency, are reconstructed via (5.15) and (5.20a), respectively, we conclude that a single expansion point suffices in the present MORe method to fully characterize any TE, TM, or TEM mode over an arbitrary frequency range.

5.4 Self-Adaptive Point Placement Strategy

For practical reasons, the following algorithm is formulated in terms of the operating frequency f rather than the wavenumber. The goal is to compute the dispersion characteristics of M dominant modes on a set $\mathcal{B}_0$ of L equidistant evaluation frequencies f_l within user-defined bounds $f_{\min}$ and $f_{\max}$.

The MORe method of Section 5.3 provides two degrees of freedom to control the error: the number of expansion points and their respective locations; see (5.19). The adaptive strategy we propose is based on successive bisection. It places a new expansion point in the middle of that sub-interval $\mathcal{B}_{\hat{n}}$, for which the error indicator E_∞ is worst. The procedure is repeated until the error indicator on the whole of $\mathcal{B}_0$ falls below a user-defined threshold E_∞^{tol}.

Our error indicator is in terms of $\tilde{\gamma}^2$. It is of incremental type and covers all modes and evaluation points. Specifically, we set

$$E_\infty(\mathcal{B}) = \frac{\max\limits_{f_l \in \mathcal{B}} \max\limits_{m=1\ldots M} \left|\tilde{\gamma}_{m+}^2(f_l) - \tilde{\gamma}_{m-}^2(f_l)\right|}{\max\limits_{m=1\ldots M} \left|\gamma_m^2(f_{\max})\right|}, \tag{5.22}$$

wherein the indices + and - denote the present and preceding iteration.

Algorithm 8 gives the details of the proposed technique. Line 2 and Line 8 show that the first two expansion points are always placed at the boundaries of the frequency range. The main loop starts at Line 10. Note that the dimension of the reduced-order EVP at Line 15 is larger than the number of sought modes, M. Hence there are two classes of eigenpairs: close approximations to the dominant modes and higher-order solutions without any merit. The `for`-loop starting at Line 16 provides a simple filter for the propagation constants of the dominant modes. At Line 26 and Line 27, we detect the interval of worst error indicator $\mathcal{B}_{\hat{k}}$ and set the new expansion point $\hat{f}$ at the evaluation frequency closest to its middle.

Algorithm 8 Waveguide MORe with adaptive point placement

```
PARAMETERS: frequency range $[f_{\min}, f_{\max}]$,
            number of evaluation frequencies $L$,
            max. number of expansion points $j_{\max}$,
            error threshold value $E_{\infty}^{tol}$.
Compute $\mathbf{P}_0$, $\mathbf{P}_1$, $\mathcal{B}_0(f_{\min}, f_{\max}, L)$
Solve $(\sum_i f_{\max}^i \mathbf{S}_i)\mathbf{X} = \mathbf{T}\mathbf{X}\,\mathrm{diag}\,\gamma_m^2$
$\mathbf{Q} \leftarrow$ updateQR($\mathbf{Q} = \mathbf{0}, \mathbf{X}$)
$[\tilde{\mathbf{S}}_{Si}, \tilde{\mathbf{T}}_{Ti}] \leftarrow$ updateROM($\mathbf{Q}; \mathbf{P}_0, \mathbf{P}_1$)
for $l = 1$ to $L$ do
  Solve $(\sum_i f_l^i \tilde{\mathbf{S}}_{Si})\tilde{\mathbf{X}} = (\sum_i f_l^i \tilde{\mathbf{T}}_{Ti})\tilde{\mathbf{X}}\,\mathrm{diag}\,\tilde{\gamma}_{m-}^2(f_l)$
end for
$\hat{f} = f_{\min}$ {Next expansion point}
$\mathcal{B}_1 = \mathcal{B}_0$ {First interval}
for $j = 2$ to $j_{max}$ do
  Solve $(\sum_i \hat{f}^i \mathbf{S}_i)\mathbf{X} = \mathbf{T}\mathbf{X}\,\mathrm{diag}\,\gamma_m^2$
  $\mathbf{Q} \leftarrow$ updateQR($\mathbf{Q}, \mathbf{X}$)
  $[\tilde{\mathbf{S}}_{Si}, \tilde{\mathbf{T}}_{Ti}] \leftarrow$ updateROM($\mathbf{Q}; \mathbf{P}_0, \mathbf{P}_1$)
  for $l = 1$ to $L$ do
    Solve $(\sum_i f_l^i \tilde{\mathbf{S}}_{Si})\tilde{\mathbf{X}} = (\sum_i f_l^i \tilde{\mathbf{T}}_{Ti})\tilde{\mathbf{X}}\,\mathrm{diag}\,\tilde{\gamma}^2$
    for $m = 1$ to $M$ do
      $\tilde{\gamma}_{m+}^2(f_l) = \arg\min_{\tilde{\gamma}^2}(|\tilde{\gamma}^2 - \tilde{\gamma}_{m-}^2(f_l)|)$
    end for
  end for
  for $k = 0$ to $j - 1$ do
    Compute $E_{\infty}(\mathcal{B}_k)$
  end for
  if $E_{\infty}(\mathcal{B}_0) < E_{\infty}^{tol}$ then
    return converged
  end if
  $\hat{k} = \arg\max_{k=1\ldots j} E_{\infty}(\mathcal{B}_k)$ {Interval of worst error}
  $\hat{f} = \arg\min_{f \in \mathcal{B}_{\hat{k}}} \left| f - \frac{\min \mathcal{B}_{\hat{k}} + \max \mathcal{B}_{\hat{k}}}{2} \right|$ {Expansion point}
  $\mathcal{B}_{\hat{k}} \leftarrow [\min \mathcal{B}_{\hat{k}}, \hat{f}]$
  $\mathcal{B}_j = [\hat{f}, \max \mathcal{B}_{\hat{k}}]$
  for all $m, l$ do
    $\tilde{\gamma}_{m-}^2(f_l) \leftarrow \gamma_{m+}^2(f_l)$
  end for
end for
```

5.5 Numerical Examples

In the following, errors in propagation constant are computed with respect to full FE solutions, using the same discretization as the MORe method. The termination criterion for the adaptive process is set to $E_{\infty}^{tol} = 10^{-6}$. An overview of all computational parameters and results is given in Table 5.1. Note that the number of modes computed in the underlying FE model has chosen to be somewhat larger than that displayed in the frequency sweep, to account for the fact that the preconditioned Arnoldi method used for solving the FE system may sometimes produce higher-order modes first [56].

5.5.1 Shielded Microstrip Line

Fig. 5.1(a) shows a shielded microstrip line [52]. The FE model represents one half of the structure and uses a magnetic wall for the middle plane. In our first test, the dielectric substrate is replaced by free space, so that the resulting waveguide has homogeneous material properties. Fig. 5.2 presents the results based on a single expansion point at 25 GHz. (Note that the actual Algorithm 8 will always result in a minimum of two points.) As predicted in Subsection 5.3.1, both the dominant TEM mode as well as the higher-order TE and TM modes are perfectly represented everywhere in the range 0-25 GHz.

Dispersion curves and error plots for the first 10 modes of the inhomogeneous waveguide can be seen in Fig. 5.3. Note the occurrence of bifurcations and complex modes. The adaptive loop finishes after 5 expansion points. Again, the MORe solutions are in excellent agreement with reference results from FE computations. However, Table 5.1 shows that the MORe scheme is 53 times faster.

5.5.2 Dielectric Loaded Waveguide

Our second example is the dielectric loaded waveguide [57] of Fig. 5.1(b). We consider the dominant 19 modes in the range 0-30 GHz. In this case, the method terminates after 7 iterations. Fig. 5.4 presents dispersion curves and error plots for the propagation coefficients. Note the highly non-uniform distribution of the expansion points. It can be seen that the MORe solutions are in excellent agreement with reference results from FE computations. This time, MORe is 12 times faster than conventional FE analysis.

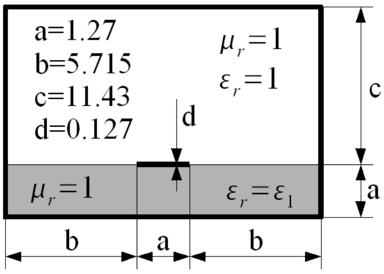
Table 5.1: Waveguide structures: Computational data[1]

	Model	A ($\varepsilon_1 = 1$)	A ($\varepsilon_1 = 8.875$)	B
Sweep	Frequency (GHz)	0 to 25	0 to 25	0 to 30
	Evaluation points	1001	1001	1001
	Number of modes	10	10	19
	Results	Fig. 5.2	Fig. 5.3	Fig. 5.4
FE model	Degrees of freedom	25553	25553	12612
	Order of FE basis	2	2	2
	Number of modes	11	12	23
	Total runtime[2] (s)	**4721**	**5045**	**4827**
ROM	Expansion points	1	5	7
	Error threshold E_∞^{tol}	–	10^{-6}	10^{-6}
	ROM dimension	11	60	161
	Total runtime[3] (s)	**11**	**95**	**402**

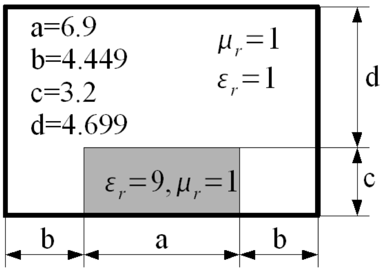
[1] For a single core of the Intel Core 2 Extreme 3 GHz processor.

[2] Including all evaluation points.

[3] Including adaptivity and all evaluation points.



(a) Shielded microstrip line



(b) Dielectric loaded waveguide

Figure 5.1: Waveguide structures. All dimensions are in mm.

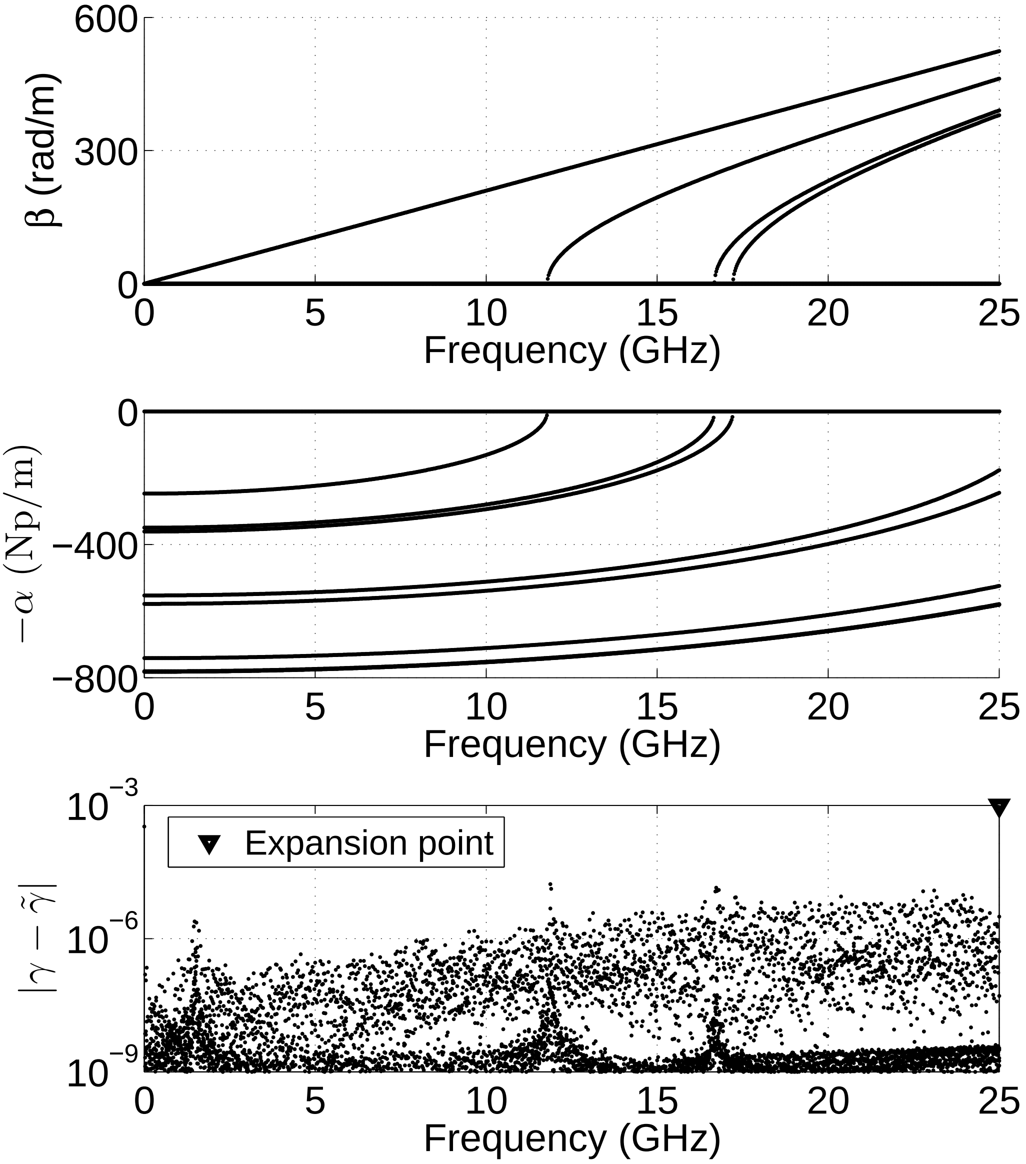


Figure 5.2: Microstrip line with homogeneous material properties ($\varepsilon_1 = 1$): Dispersion curves and error plot.

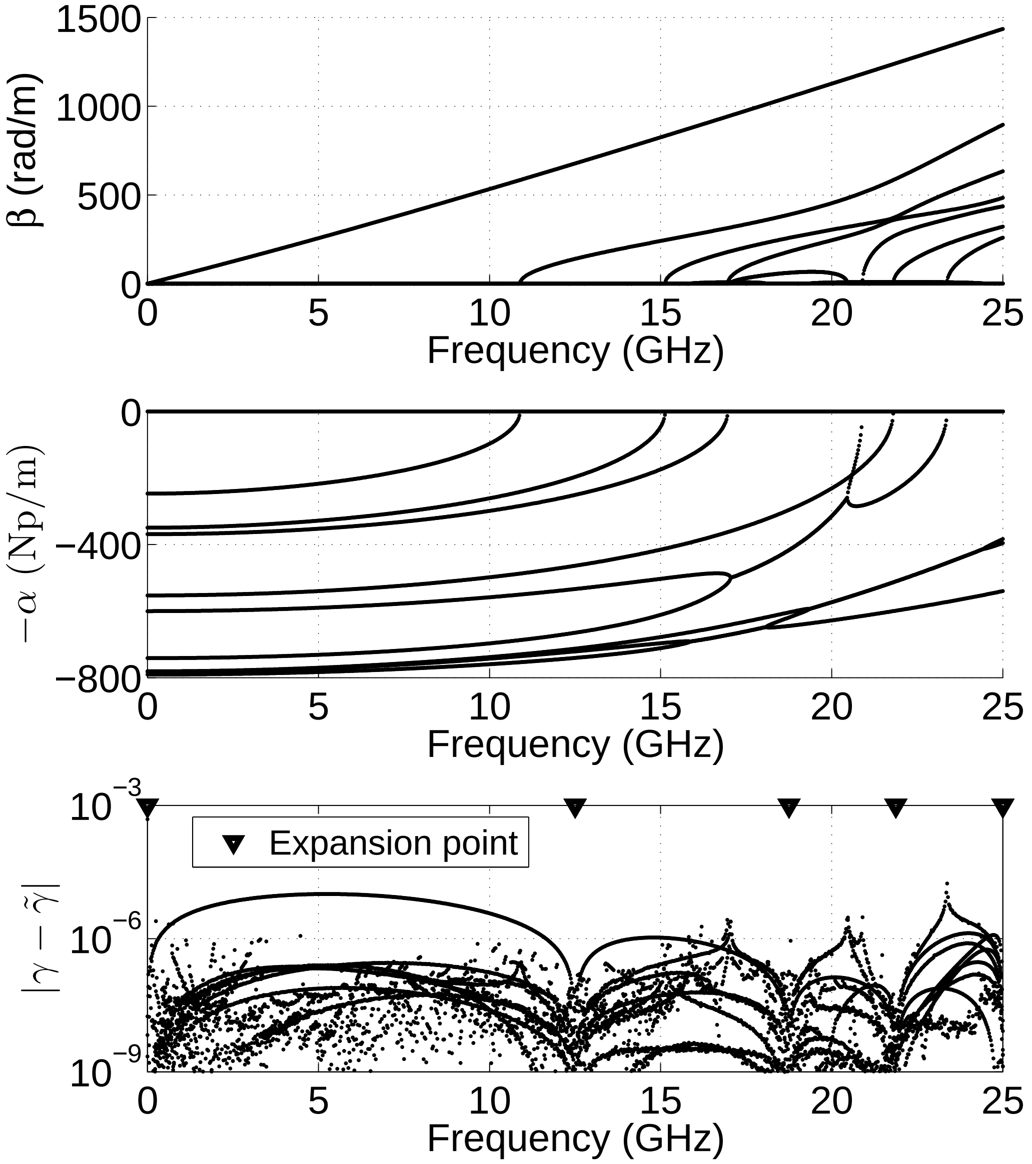


Figure 5.3: Microstrip line with dielectric substrate ($\varepsilon_1 = 8.875$): Dispersion curves and error plot.

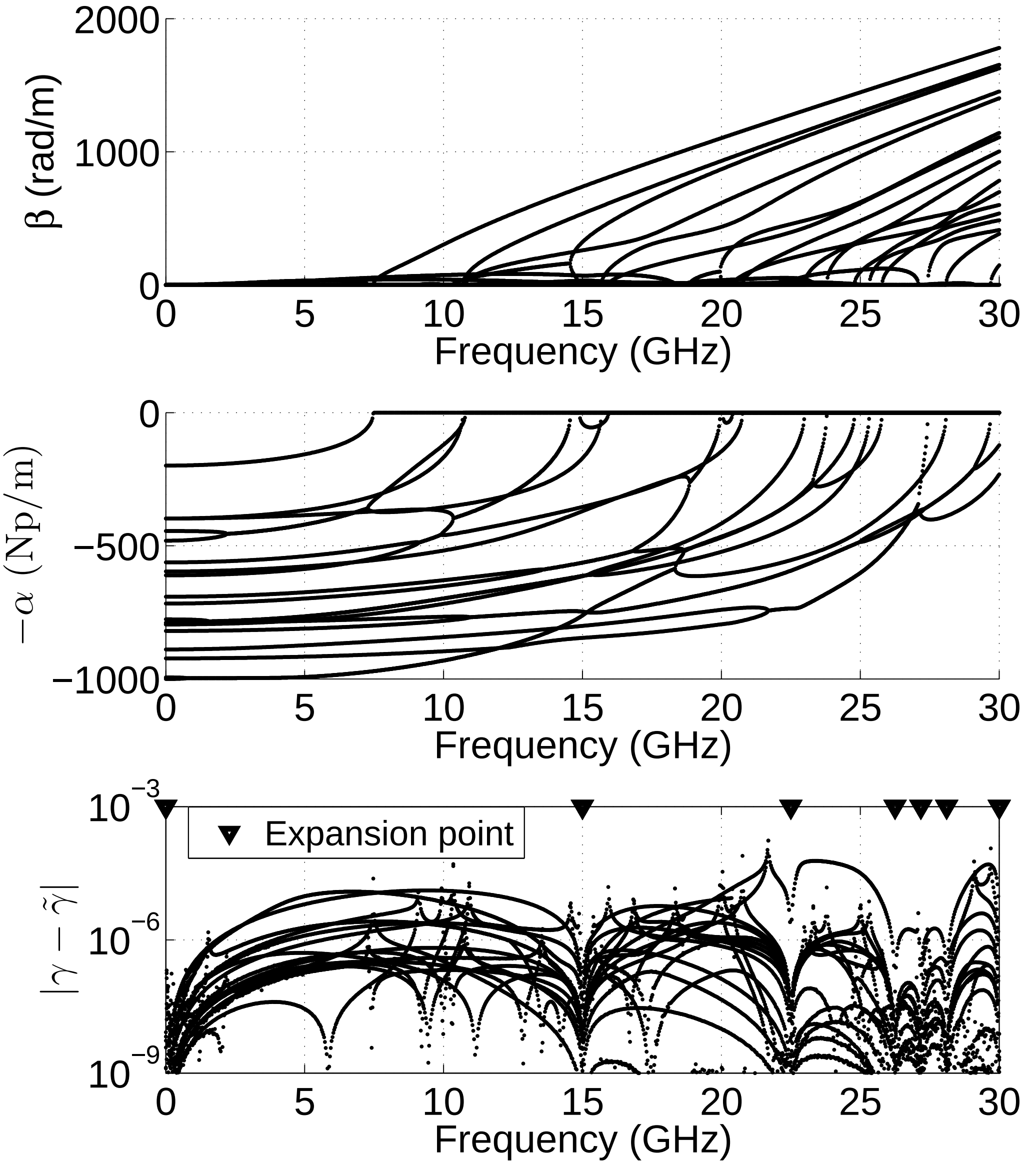


Figure 5.4: Dielectric loaded waveguide: Dispersion curves and error plot.

5.6 Conclusion

In this chapter, a multi-point MORe technique with a self-adaptive point placement strategy for the broadband FE analysis of electromagnetic waveguides has been introduced. The underlying MORe method employs two-sided projections with null-field orthogonalization, and the adaptive scheme is based on successive bisection, guided by an incremental error indicator for the propagation constant.

The numerical tests of Section 5.5 confirm that the number of expansion points required by the adaptive scheme remains very small, even for very wide broadband applications. In consequence, the proposed method is significantly faster than traditional FE analysis. At the same time, errors in propagation coefficient are negligible over the whole frequency band.

Chapter 6

Broadband Sensitivity Analysis

6.1 Introduction

The design process of a microwave device is a complex task and depends on many parameters. Frequency responses for different parameter configurations may be evaluated by means of numerical methods. However, the impact of a variation in the design parameters can be used to systematically improve the performance of a microwave device. Sensitivity analysis provides a powerful means for analyzing small modifications in the design parameters.

Based on 2D-FE methods, [58], [59] introduced automatic microwave device shape optimizations in an iterative process, where optimization algorithms are applied. The introduced techniques define cost functions and employ design sensitivity analysis to evaluate the impact of a design variation. An optimization process founded on a 3D-FE technique, where sensitivity analysis is employed as well, was presented in [60]. It is shown in [61] that the design sensitivities can be evaluated from the solution used for the scattering matrix evaluation and no additional system matrix solutions are required. An AWE technique is employed to compute the sensitivity on a frequency range in [62], which is used in [63] to optimize microwave devices with respect to their frequency response.

This thesis provides broadband sensitivity analysis techniques, which are based on the previously introduced projection-based MORe methods. Particularly the adaptive MORe schemes, developed in the previous chapters, are attractive to employ, as fully automatic analysis runs may be performed. However, the employed WCAWE technique as well as the multi-point approach allows us to perform the sensitivity analysis on a larger frequency range as the AWE process applied in [62]. Hence, the provided methods allow us to perform a reliable optimization process over a large bandwidth. Although the new technique is applied only to material parameters in this thesis, other design parameters, such as geometry variations, as performed in [58], may also be employed.

6.2 Theory

6.2.1 System Design Sensitivity

The FE simulation of a passive microwave structure results in a system of linear equations and an output functional for the network parameter evaluation. For the sensitivity analysis, a design parameter p is defined on which a perturbation is applied to. As the modal field patterns on wave ports result from a separated analysis, design parameters must not affect the wave port areas, which is a limitation in the current implementation. Assuming polynomial dependency in the design parameter p and the wavenumber k, we denote the resulting FE system as

$$\sum_{m=0}^{M}\sum_{n=0}^{N}\left(\mathbf{A}_{mn}p^m k^n\right)\mathbf{x}(k,p) = k\mathbf{b}, \tag{6.1a}$$

$$y(k,p) = \mathbf{c}^T\mathbf{x}(k,p) + d, \tag{6.1b}$$

where M and N designate the highest polynomial dependencies in the parameters. The input vector $\mathbf{b}$, the output functional $\mathbf{c^T}$ and the feed through coefficient d are assumed to be constant

$$\mathbf{b} = \text{const}, \tag{6.2a}$$

$$\mathbf{c}^T = \text{const}, \tag{6.2b}$$

$$d = \text{const}\,. \tag{6.2c}$$

For the sensitivity analysis, the first derivative of the output with respect to p, at the expansion point $p = 0$ is chosen. Hence, deriving both sides of the system (6.1) and setting the parameter $p = 0$, the sensitivity system yields

$$\left(\sum_{n=0}^{N}\mathbf{A}_{1n}k^n\right)\mathbf{x}(k,p=0) = -\left(\sum_{n=0}^{N}\mathbf{A}_{0n}k^n\right)\frac{\partial\mathbf{x}(k,p)}{\partial p}\Big|_{p=0}, \tag{6.3a}$$

$$\frac{\partial y(k,p)}{\partial p}\Big|_{p=0} = \mathbf{c}^T\frac{\partial\mathbf{x}(k,p)}{\partial p}\Big|_{p=0}. \tag{6.3b}$$

The derivative of the system output at the expansion point $p = 0$ is denoted as

$$\delta_p(k) = \frac{\partial y(k,p)}{\partial p}\Big|_{p=0}. \tag{6.4}$$

Solving (6.3) for the derivative results in

$$\delta_p(k) = -\mathbf{c}^T\left(\sum_{n=0}^{N}\mathbf{A}_{0n}k^n\right)^{-1}\left(\sum_{n=0}^{N}\mathbf{A}_{1n}k^n\right)\mathbf{x}(k,p=0) \tag{6.5}$$

$$\overset{(6.1)}{=} -\mathbf{c}^T\left(\sum_{n=0}^{N}\mathbf{A}_{0n}k^n\right)^{-1}\left(\sum_{n=0}^{N}\mathbf{A}_{1n}k^n\right)\left(\sum_{n=0}^{N}\mathbf{A}_{0n}k^n\right)^{-1}k\mathbf{b}. \tag{6.6}$$

The impact of a perturbation in the design parameter p is evaluated from

$$dy(k,p=0) = y(k,p=0) + \delta_p(k)\Delta p, \tag{6.7}$$

where a sweep for different wavenumbers may be evaluated. It is remarkable that for a fixed wavenumber, the system matrix at $p = 0$ needs to be factorized only once to evaluate the system sensitivity [61]. Furthermore, the whole process may be performed for a parameter vector $\mathbf{p}$ instead of a single design parameter, where only one system factorization is required for the sensitivity evaluation for all parameters. The analysis process (6.7) can be performed for any combination of parameters.

6.2.2 Broadband Design Sensitivity Technique

Employing the projection-based MORe techniques that were introduced in the previous chapters, the original system (6.1) is reduced to the polynomial parameterized ROM

$$\sum_{m=0}^{M}\sum_{n=0}^{N}\left(\tilde{\mathbf{A}}_{mn}p^m k^n\right)\tilde{\mathbf{x}}(k,p) = k\tilde{\mathbf{b}}, \tag{6.8a}$$

$$\tilde{y}(k,p) = \tilde{\mathbf{c}}^T\tilde{\mathbf{x}}(k,p) + d, \tag{6.8b}$$

where the the projection matrix $\mathbf{V}$ is employed for the reduction

$$\tilde{\mathbf{A}} = \mathbf{V}^T\mathbf{A}\mathbf{V}, \tag{6.9a}$$

$$\tilde{\mathbf{b}} = \mathbf{V}^T\mathbf{b}, \tag{6.9b}$$

$$\tilde{\mathbf{c}} = \mathbf{V}^T\mathbf{c}. \tag{6.9c}$$

As the reduced system (6.8) has the same structure as the original model (6.1), the same procedure as in the large-scale sensitivity analysis is performed. Thus, equivalent to (6.5) to (6.7), the derivative of the system output at $p = 0$ can be written as

$$\tilde{\delta}_p(k) = \frac{\partial\tilde{y}(k,p)}{\partial p}\Big|_{p=0} = -\tilde{\mathbf{c}}^T\left(\sum_{n=0}^{N}\tilde{\mathbf{A}}_{0n}k^n\right)^{-1}\left(\sum_{n=0}^{N}\tilde{\mathbf{A}}_{1n}k^n\right)\left(\sum_{n=0}^{N}\tilde{\mathbf{A}}_{0n}k^n\right)^{-1}k\tilde{\mathbf{b}}, \tag{6.10}$$

and the perturbation evaluation is performed through

$$d\tilde{y}(k,p=0) = \tilde{y}(k,p=0) + \tilde{\delta}_p(k)\Delta p. \tag{6.11}$$

The computational costs for a sensitivity sweep in the reduced case are very low compared to the large-scale system, as the system matrix solutions are very cheap. However, the extension to a parameter sweep with a sensitivity parameter vector is, as in the full system sensitivity analysis, straightforward. Note that this approach is applicable in theory to scattering as well as impedance formulation modeling. These two formulations are explained in detail in [21].

6.3 Numerical Experiments

The efficiency of the introduced broadband design sensitivity technique is presented by means of some numerical examples. However, this section also discusses the drawback of impedance formulations in combination with the sensitivity analysis. In the analyzed electromagnetic structures, p_ε is chosen as design parameter, which is defined through

$$\varepsilon = (\varepsilon_r + p_\varepsilon)\varepsilon_0. \tag{6.12}$$

All numerical sensitivity computations in this section are based on FE simulations and the developed MORe techniques of this thesis.

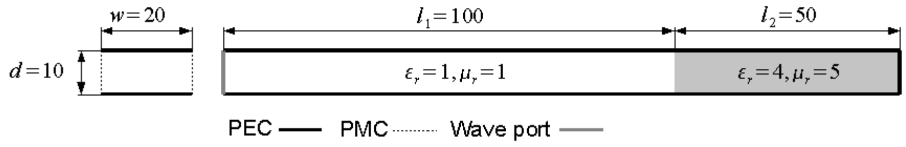
6.3.1 Parallel Plate Waveguide

The first considered structure is the parallel plate waveguide of Fig. 6.1(a), which has only one wave port. The gray part in the waveguide represents a material with properties ε_r and μ_r, whereas the white section is modeled as vacuum. The waveguide is bounded on top and bottom by PEC and side walls are perfect magnetic conductors (PMC).

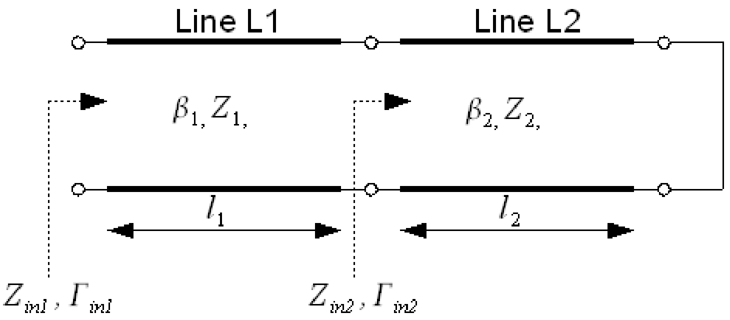
At the end of the parallel plate waveguide a PEC is attached to impose a short circuit. In the following, the permittivity of the material in the dark colored section is chosen as design parameter and is designated with ε_r. The perturbation of this design parameter is written as p_ε.

Transmission Line Model

As the propagating waves in the chosen structure are of TEM type, a transmission line model can be obtained for the waveguide and can be analytically evaluated. In the model of Fig. 6.1(b), the transmission line $L1$ represents the vacuum part, the transmission line $L2$ the gray colored section and the sort-circuit at the end of the line is the PEC boundary at the end of the waveguide. The following formulas for parallel plate waveguides and



(a) Geometric dimensions in mm.



(b) Transmission line model.

Figure 6.1: Parallel plate waveguide: Structure geometry and transmission line model.

transmission line models can be found in many books on microwave theory, e.g. [19]. The characteristic impedance of the transmission line circuit model can be found as

$$Z_0 = \eta \frac{d}{w} = \sqrt{\frac{\mu}{\varepsilon}} \frac{d}{w}, \tag{6.13}$$

which only depends on material coefficients and geometry of the waveguide. The phase velocity in the material medium is given by

$$v_p = \frac{\omega}{\beta} = \frac{1}{\sqrt{\mu\varepsilon}}, \tag{6.14}$$

and only depends on the material. The input impedance Z_{in} of a transmission line with load Z_L can be evaluated from

$$Z_{in} = Z_0 \frac{Z_L + jZ_0 \tan(\beta l)}{Z_0 + jZ_L \tan(\beta l)}, \tag{6.15}$$

where l defines the length of the transmission line. Alternatively, the reflection coefficient Γ can be used to perform the line transformation. At the end of the line, the reflection coefficient is given through

$$\Gamma(0) = \frac{Z_L - Z_0}{Z_L + Z_0}, \tag{6.16}$$

and can be transformed by

$$\Gamma(l) = \Gamma(0) e^{-2j\beta l}, \tag{6.17}$$

to the position l in the transmission line, which of course can be chosen as the input of the line.

Scattering Formulation

In this section, the proposed numerical broadband sensitivity technique is compared to the analytic solution. For the FE simulation and thus for the broadband sensitivity a scattering formulation is employed. The analytic solution is obtained by evaluating the reflection coefficient Γ_{in2} and performing the line transformation in $L1$ through (6.17)

$$Z_{in2} = jZ_2 \tan(\beta_2 l_2), \tag{6.18}$$

$$\Gamma_{in2} = \frac{Z_{in2} - Z_1}{Z_{in2} + Z_1}, \tag{6.19}$$

$$\hat{S}_{11} = \Gamma_{in1} = \Gamma_{in2} e^{-2j\beta_1 l_1}. \tag{6.20}$$

Thus, the analytic solution for the derivative of the design parameter and the perturbed scattering parameter are

$$\hat{\delta}_\varepsilon(k) = \frac{\partial \hat{S}_{11}(k, p_\varepsilon)}{\partial p_\varepsilon}\Big|_{p_\varepsilon = 0}, \tag{6.21}$$

$$d\hat{S}_{11}(k, p_\varepsilon) = \hat{S}_{11}(k, p_\varepsilon = 0) + \hat{\delta}_p(k) p_\varepsilon. \tag{6.22}$$

The numerical simulation data is given in Table 6.1. The ROM generation time includes the time for the FE matrices assembling, the projection matrix generation, and the subspace projection process. The evaluation time represents the time of the evaluation for one sweep, either scattering parameters or the derivatives, as no significant differences are measurable. These definitions are valid for all following simulation data tables in this chapter.

Fig. 6.2 shows the evaluated scattering parameters, their derivatives as well as error plots. The absolute value of S_{11} is constant in the computed bandwidth, while its derivative δ_ε is increasing with higher frequency. Additionally, the phases between parameter and derivative exhibit a shift. Note that the graphs of analytic solution and ROM evaluation are lying upon each other as a result of the low error. The errors e_{S11} and e_{δ_ε} are defined as

$$e_{S11} = |\hat{S}_{11} - \tilde{S}_{11}|, \tag{6.23}$$

$$e_{\delta_\varepsilon} = |\hat{\delta}_\varepsilon - \tilde{\delta}_\varepsilon|. \tag{6.24}$$

The increasing errors with higher frequency originate from the FE simulation. It is a result of shorter wavelengths in the fields on a constant mesh.

The computed sweeps allow us to perform evaluations for a disturbed material parameter ε_r, by applying (6.11) and (6.22). Sweeps for the perturbed analytic solution $d\hat{S}_{11}(p_\varepsilon)$ and the perturbed numerical evaluation $d\tilde{S}_{11}(p_\varepsilon)$ are solved. The results are shown for a perturbation of $p_\varepsilon = 0.004$ in Fig. 6.3 and for a perturbation of $p_\varepsilon = 0.04$ in Fig. 6.4. Additionally, the analytic sweeps are solved for $\varepsilon = \varepsilon_r\varepsilon_0$ and $\varepsilon = (\varepsilon_r + p_\varepsilon)\varepsilon_0$ as references. While the phase shift for the small perturbation can be only noticed in the upper end of the computed bandwidth in Fig. 6.3, the perturbation of $p_\varepsilon = 0.04$ causes a remarkable phase shift on a broadband in the upper end of the sweep in Fig. 6.4. The absolute value of S_{11} increases with frequency into non-physical results. This effect, which becomes more significant with higher perturbation values, results from the sensitivity analysis and is not caused through the numerical broadband approach.

In the error plots, the difference of the analytic solutions at $\varepsilon_r + p_\varepsilon$ to analytic perturbation and to the broadband sensitivity analysis is recorded. The error between perturbed analytic solution and perturbed ROM is evaluated as well as e_{PR}. In particular these errors are defined as

$$e_{AP}(k) = |\hat{S}_{11}(k, \varepsilon_r + p_\varepsilon) - d\hat{S}_{11}(k, p_\varepsilon)|, \tag{6.25}$$

$$e_{AR}(k) = |\hat{S}_{11}(k, \varepsilon_r + p_\varepsilon) - d\tilde{S}_{11}(k, p_\varepsilon)|, \tag{6.26}$$

$$e_{PR}(k) = |d\tilde{S}_{11}(k, p_\varepsilon) - d\hat{S}_{11}(k, p_\varepsilon)|. \tag{6.27}$$

For the very small perturbation of $p_\varepsilon = 0.004$, the error from perturbation to analytical formulas, i.e. e_{AP} and e_{AR}, is very small, see Fig. 6.3. Although the error originating from numerical simulation is dominant at higher frequencies, it is important to notice that the perturbation error is rising with higher frequencies. This effect becomes more pronounced for a higher perturbation, as can be seen in Fig. 6.4, where the error due to perturbation is dominant over the complete evaluated bandwidth (e_{AP} and e_{AR} are lying upon each other).

Specifications:		Simulation data:	
Lowest frequency in Hz	1e7	Original dimension	149787
Highest frequency in Hz	2e9	Number of iterations q	9
Evaluation points, equidistant	501	ROM generation time in s	15.3
FE basis function order	2	Evaluation time in s	0.01
Multi-point method	-		

Table 6.1: Parallel plate waveguide: Scattering formulation simulation data.

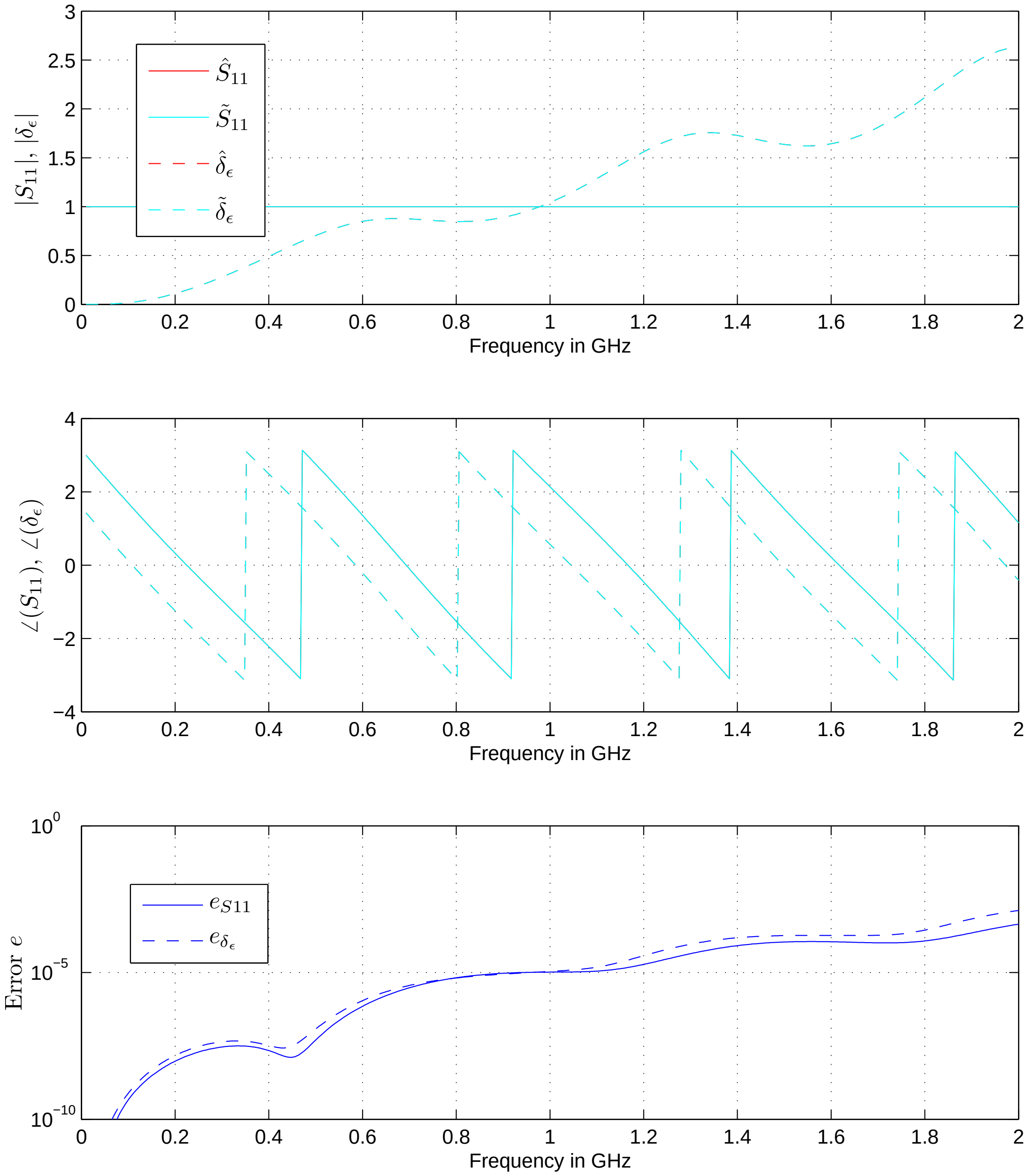


Figure 6.2: Parallel plate waveguide: S_{11} and δ_ε versus frequency using scattering formulation.

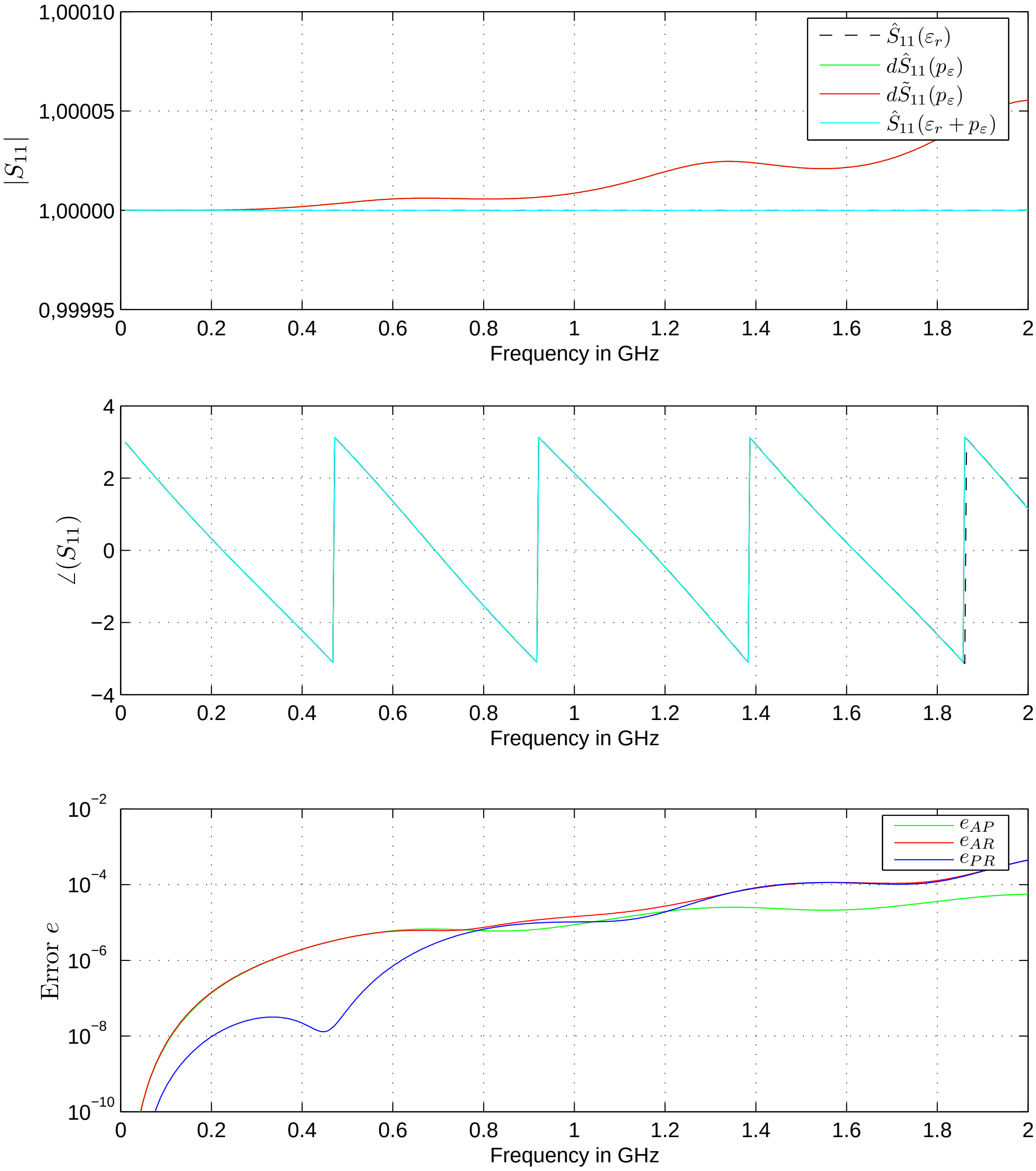


Figure 6.3: Parallel plate waveguide: Material perturbation $p_\varepsilon = 0.004$ versus frequency using scattering formulation.

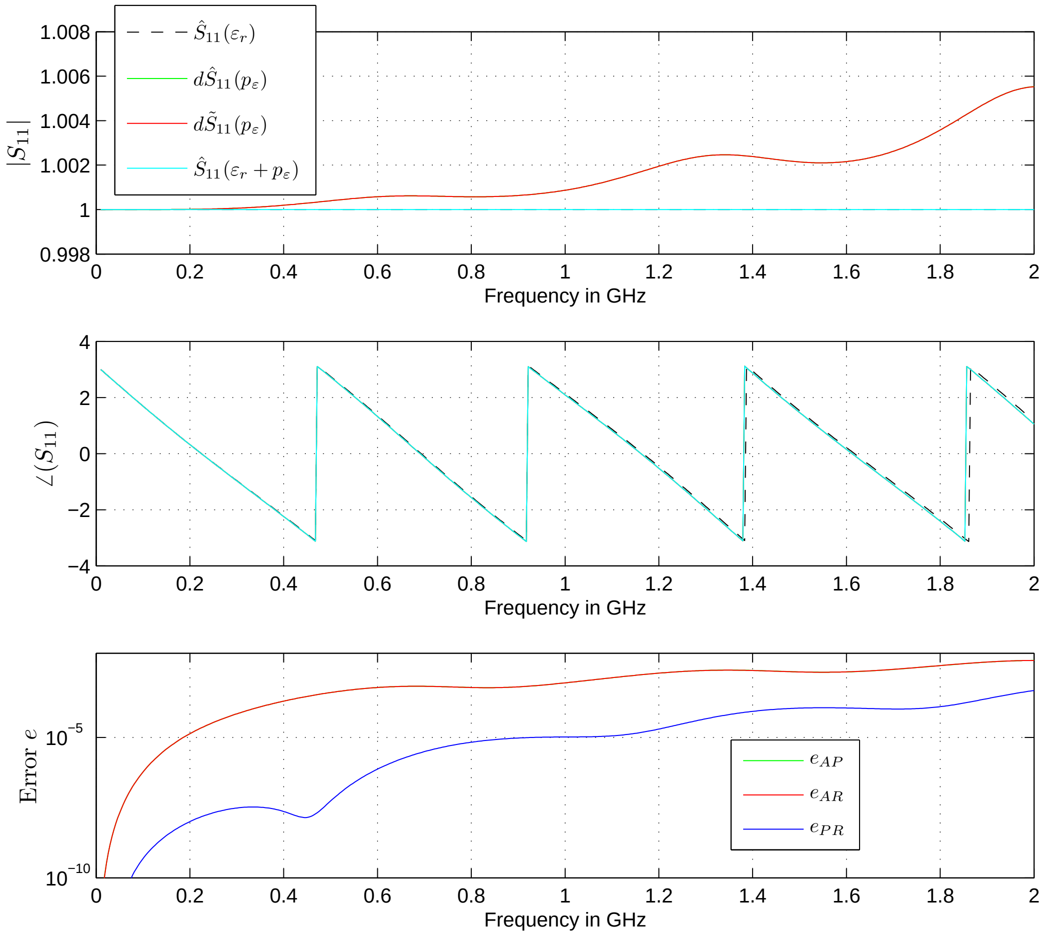


Figure 6.4: Parallel plate waveguide: Material perturbation $p_\varepsilon = 0.04$ versus frequency using scattering formulation.

Impedance Formulation

As for the scattering formulation, a set of analytically formulas for the impedance formulation is provided. Instead of transforming the line $L1$ by (6.17), the transformation is performed by (6.15) and thus we have

$$Z_{in2} = jZ_2 \tan(\beta_2 l_2), \tag{6.28}$$

$$Z_{in1} = Z_1 \frac{Z_{in2} + jZ_1 \tan(\beta_1 l_1)}{Z_1 + jZ_{in2} \tan(\beta_1 l_1)}, \tag{6.29}$$

$$\hat{Z}_{11} = \frac{Z_{in1}}{Z_1}, \tag{6.30}$$

$$\hat{\delta}_\varepsilon = \frac{\partial \hat{Z}_{11}}{\partial p_\varepsilon}\Big|_{p_\varepsilon = 0}. \tag{6.31}$$

As for the analytical solution, an impedance formulation is chosen for the numerical solution as well. The computational data for the simulation can be found in Table 6.2.

Fig. 6.5 shows a sweep for the impedance parameter Z_{11} and its derivative δ_ε. The analyzed waveguide structure exhibits inner resonances which result in a set of singularities in the impedance parameter sweep, in the numerical simulation as well as in the evaluated analytical formulation. The inner resonances cause the same set of singularities in the derivatives of the impedance parameters. The computed errors are defined as

$$e_{Z11} = |\hat{Z}_{11} - \tilde{Z}_{11}|, \tag{6.32}$$

$$e_{\delta_\varepsilon} = |\hat{\delta}_\varepsilon - \tilde{\delta}_\varepsilon|. \tag{6.33}$$

The set of error peaks in Fig. 6.5 is a result of the singularities in the impedance parameters and their derivatives.

In equivalence to (6.16), sweeps for the reflection coefficient S_{11} and the perturbed scattering parameter dS_{11} can be evaluated from the expressions

$$\hat{S}_{11}(k) = \frac{\hat{Z}_{11}(k) - 1}{\hat{Z}_{11}(k) + 1}, \tag{6.34}$$

$$d\hat{S}_{11}(k, p_\varepsilon) = \frac{\hat{Z}_{11}(k) + \hat{\delta}_\varepsilon(k) p_\varepsilon - 1}{\hat{Z}_{11}(k) + \hat{\delta}_\varepsilon(k) p_\varepsilon + 1}. \tag{6.35}$$

Thus, the solutions in impedance formulation allow us to perform the perturbation analysis in the scattering formulation and to evaluate scattering parameters. As in the scattering formulation section, the perturbed sweeps are compared to the analytic solutions at $\varepsilon = \varepsilon_r \varepsilon_0$ and $\varepsilon = (\varepsilon_r + p_\varepsilon)\varepsilon_0$. Again, the perturbations are set to $p_\varepsilon = 0.004$, in Fig. 6.6, and $p_\varepsilon = 0.04$, in Fig. 6.7. The errors are defined as above, i.e. (6.25) to (6.27).

Although non-physical resonances may occur in the impedance formulation, as reported for constructed configurations in [21], the evaluated scattering parameters in this simulation do not exhibit unwanted singularities. However, at the frequencies of the inner resonances, which can be seen in Fig. 6.5, the errors e_{AP} and e_{AR} rise to high values in Fig. 6.6, already for a perturbation of $p_\varepsilon = 0.004$. For a perturbation of $p_\varepsilon = 0.04$, shown in Fig. 6.7, the non-physical resonances cause serious discontinuities in the phase for higher perturbations. It is very important to notice that the analytical perturbation analysis exhibits exactly the same behavior as the numerical evaluation. Therefore, this effect is a result of the formulation and is not an outcome of the numerical simulation.

Specifications:		Simulation data:	
Lowest frequency in Hz	1e7	Original dimension	149787
Highest frequency in Hz	2e9	Number of iterations q	9
Evaluation points	501	ROM generation time in s	8.9
FE basis function order	2	Evaluation time in s	0.01
Multi-point method	-		

Table 6.2: Parallel plate waveguide: Impedance formulation simulation data.

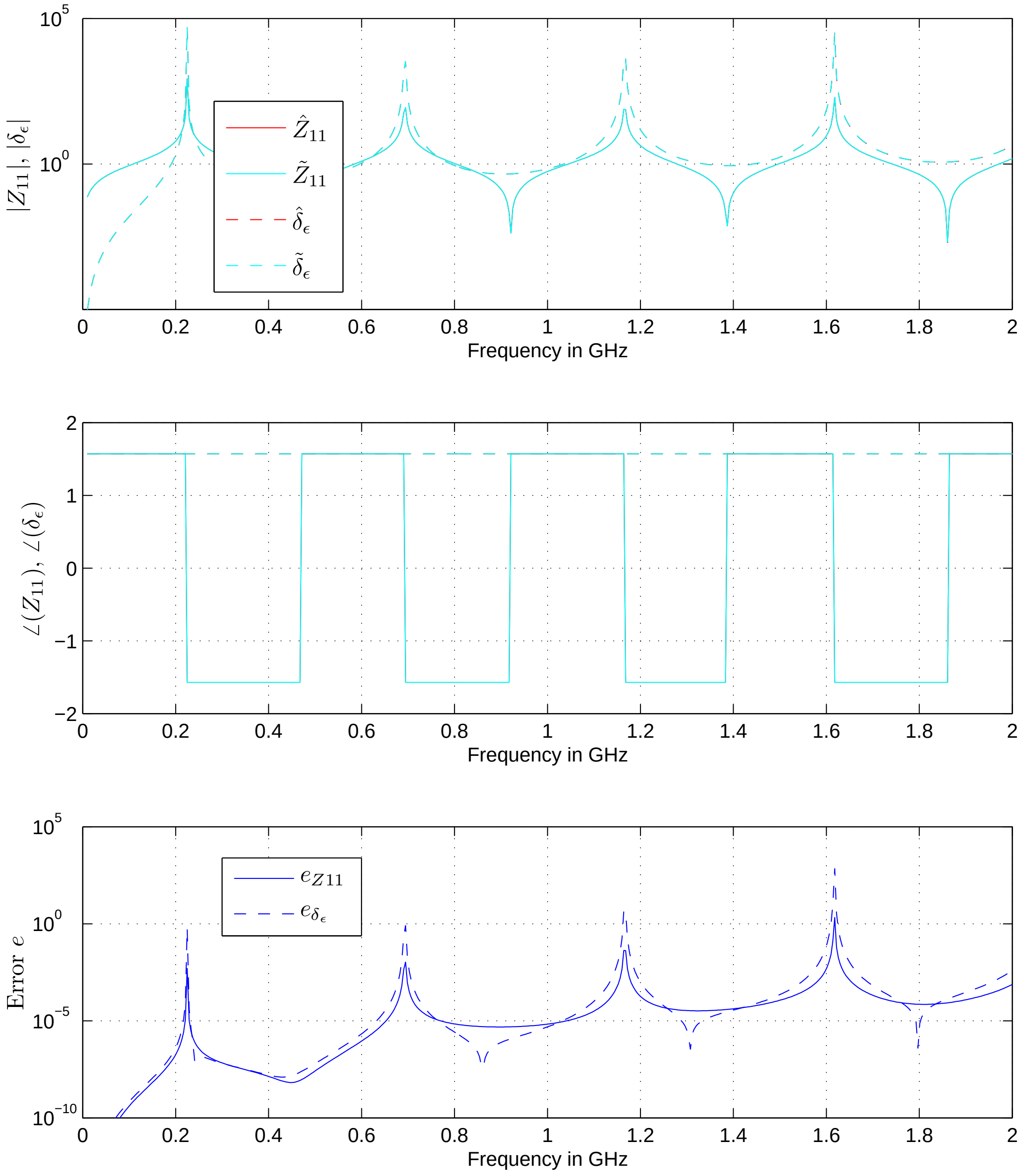


Figure 6.5: Parallel plate waveguide: Z_{11} and δ_ε versus frequency using impedance formulation.

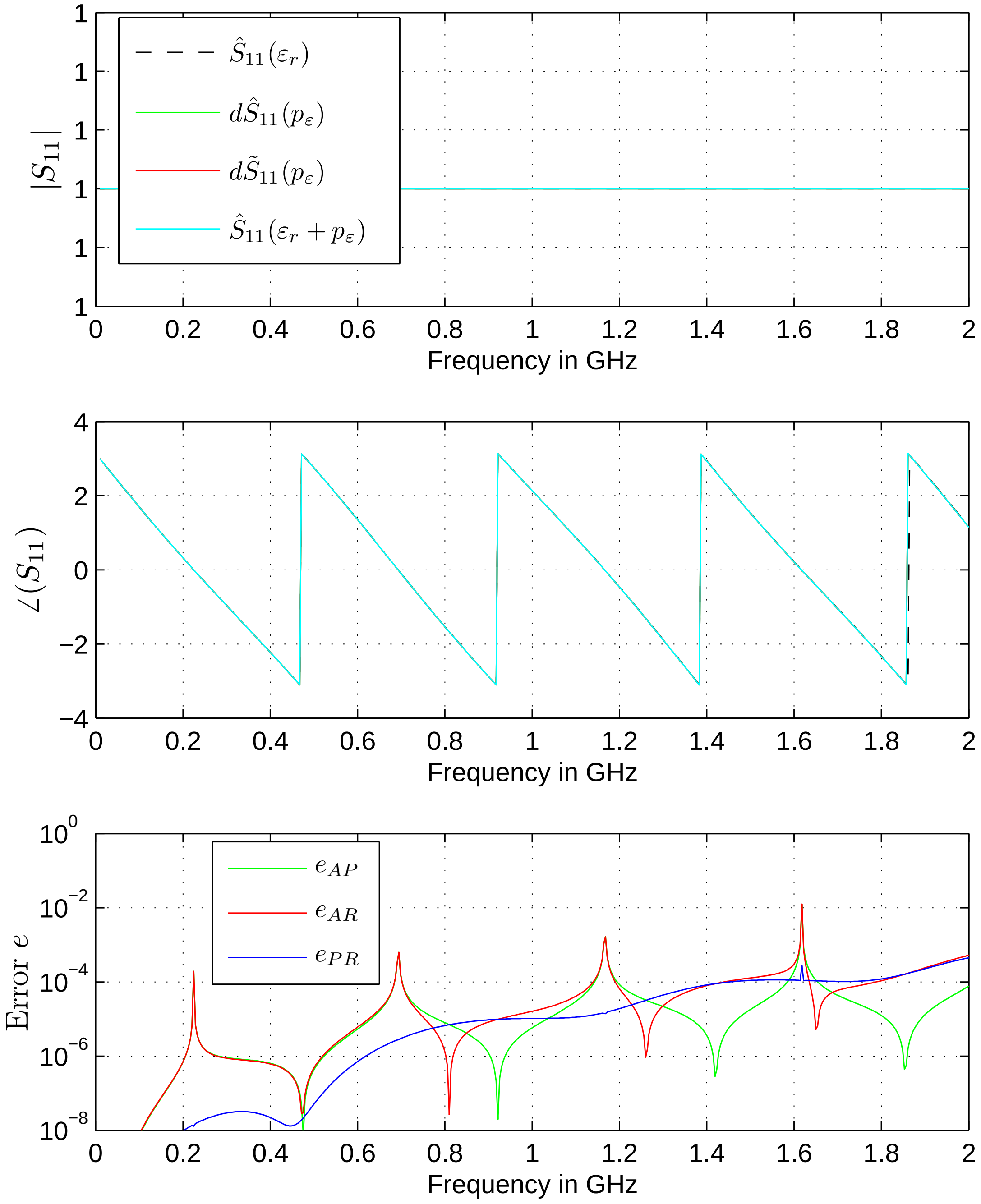


Figure 6.6: Parallel plate waveguide: Material perturbation $p_\varepsilon = 0.004$ versus frequency using impedance formulation.

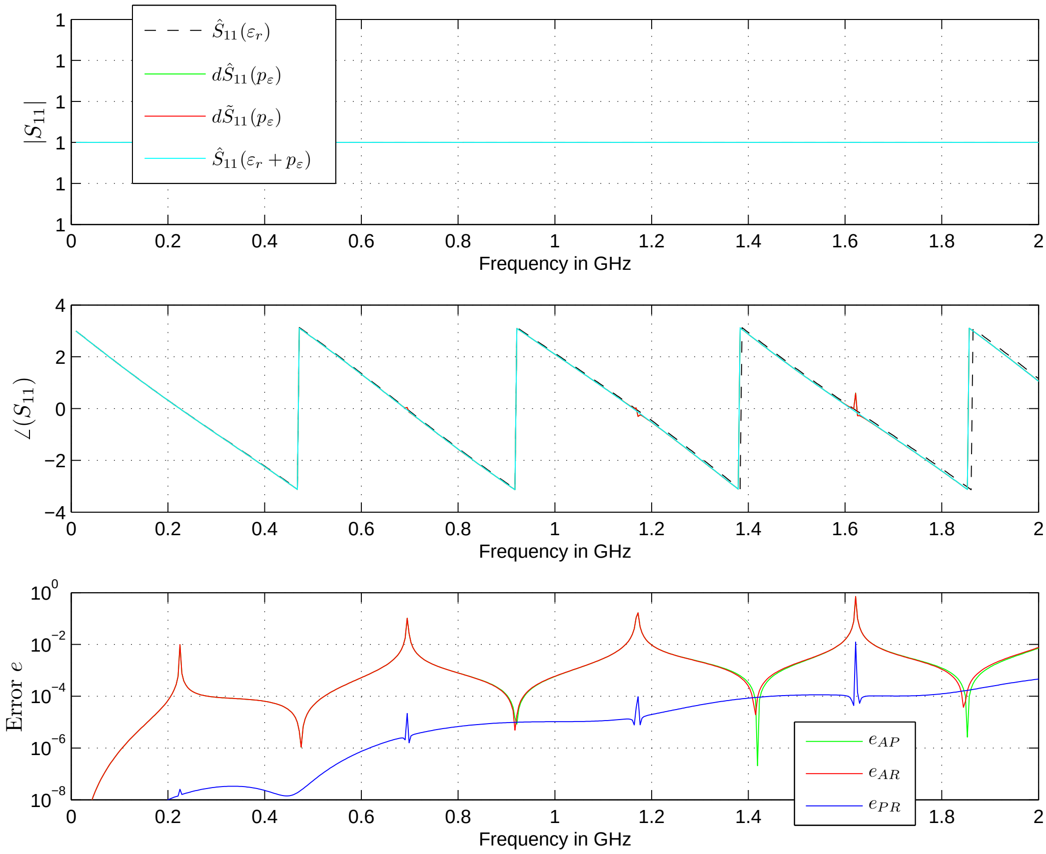


Figure 6.7: Parallel plate waveguide: Material perturbation $p_\varepsilon = 0.04$ versus frequency using impedance formulation.

6.3.2 Dielectric Pole Structure

Structure Definition

The second considered structure is the parallel plate waveguide in Fig. 6.8, where three dielectric poles with $\varepsilon_r = 50$ are inserted. For such kinds of structures it is rather difficult to figure out an analytical solution and even more difficult to find the derivative for the sensitivity analysis. Therefore, for the structure simulation and the sensitivity analysis only numerical methods are employed, i.e. MOR techniques applied to FE simulations and the above introduced broadband sensitivity analysis. As sensitivity parameter, the perturbation p_ε of the permittivity ε_r of the three dielectric poles is chosen.

Simulation and Sensitivity Analysis

In a first numerical experiment, the scattering parameter evaluation and the sensitivity analysis for the three pole structure is performed in scattering formulation. The simulation data is given in Table 6.3, while a frequency sweep is plotted in Fig. 6.9. Because resonances are very interesting from the sensitivity analysis point of view, the focus is on the resonance close to 190 MHz and the set of resonances around 1 GHz. The perturbations for the analysis are set to $p_\varepsilon = 0.2$ and $p_\varepsilon = 1.0$.

Figure 6.10 shows the frequency shift resulting from the material perturbations in the resonance at 190 MHz. In addition to the perturbed numeric evaluation, the scattering parameters are computed as reference solution at the shifted material parameters $\varepsilon_r = 50.2$ and $\varepsilon_r = 51.0$. While for $p_\varepsilon = 0.2$ the perturbed scattering parameter $d\tilde{S}_{11}$ is in good accordance with the shifted material parameter solution, the perturbation $p_\varepsilon = 1.0$ causes a severe difference. This is also noticeable in the evaluated error in Fig. 6.10, which shows the differences between sensitivity analysis and shifted material parameter solution

$$e_{p_\varepsilon} = |\tilde{S}_{11}(\varepsilon_r + p_\varepsilon) - d\tilde{S}_{11}(\varepsilon_r, p_\varepsilon)|. \tag{6.36}$$

For the resonances at 1 GHz, already the smaller chosen perturbation of $p_\varepsilon = 0.2$ leads to differences compared to the solution with shifted parameter $\tilde{S}_{11}(\varepsilon_r = 50.2)$, see Fig. 6.11. These differences can be noticed particularly in the phases of the scattering parameters. The errors are also higher, compared to the perturbation at 190 MHz. For a perturbation of $p_\varepsilon = 1.0$, the errors at 1 GHz resonances are rather high and the scattering parameters evaluated from the sensitivity analysis do not represent the resonances properly. This is a result of setting the perturbation too high, but not a limitation of the provided broadband sensitivity analysis.

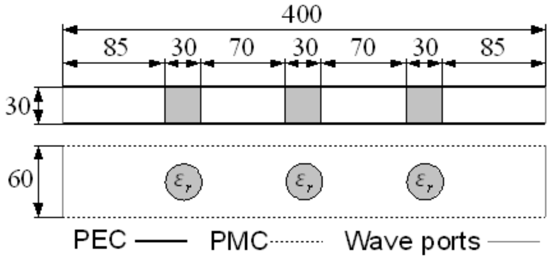


Figure 6.8: Three pole structure: Geometric dimensions in mm.

Specifications:		Simulation data:	
Lowest frequency in Hz	1e7	Original dimension	306768
Highest frequency in Hz	1.2e9	Number of iterations q	8
Evaluation points	10000	ROM generation time in s	907
FE basis function order	2	Evaluation time in s	0.3
Multi-point method	-		

Table 6.3: Three pole structure: Simulation data using scattering formulation.

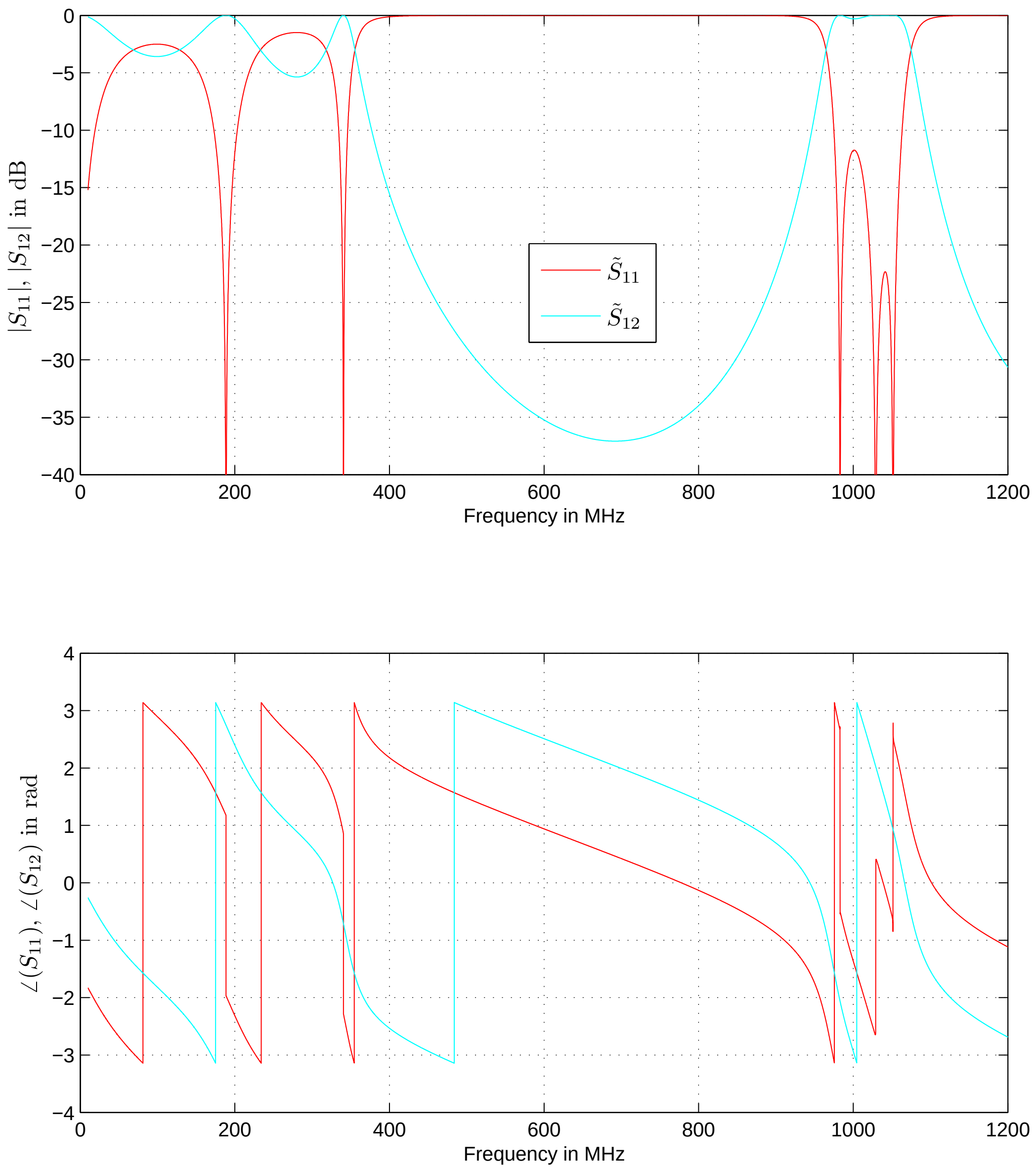


Figure 6.9: Three pole structure: Scattering parameters versus frequency.

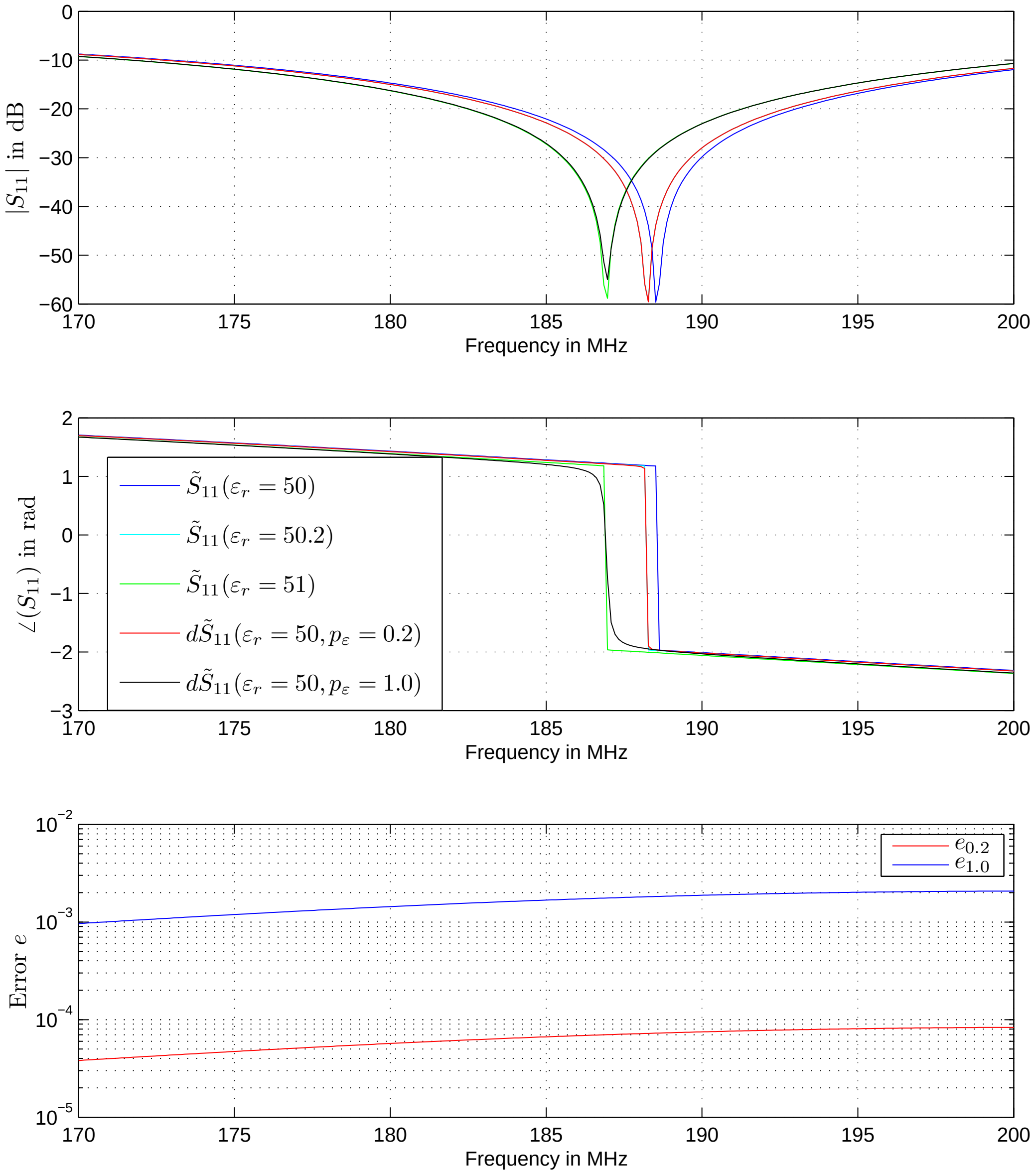


Figure 6.10: Three pole structure: View of the sensitivity analysis at the resonance around 190 MHz using scattering formulation.

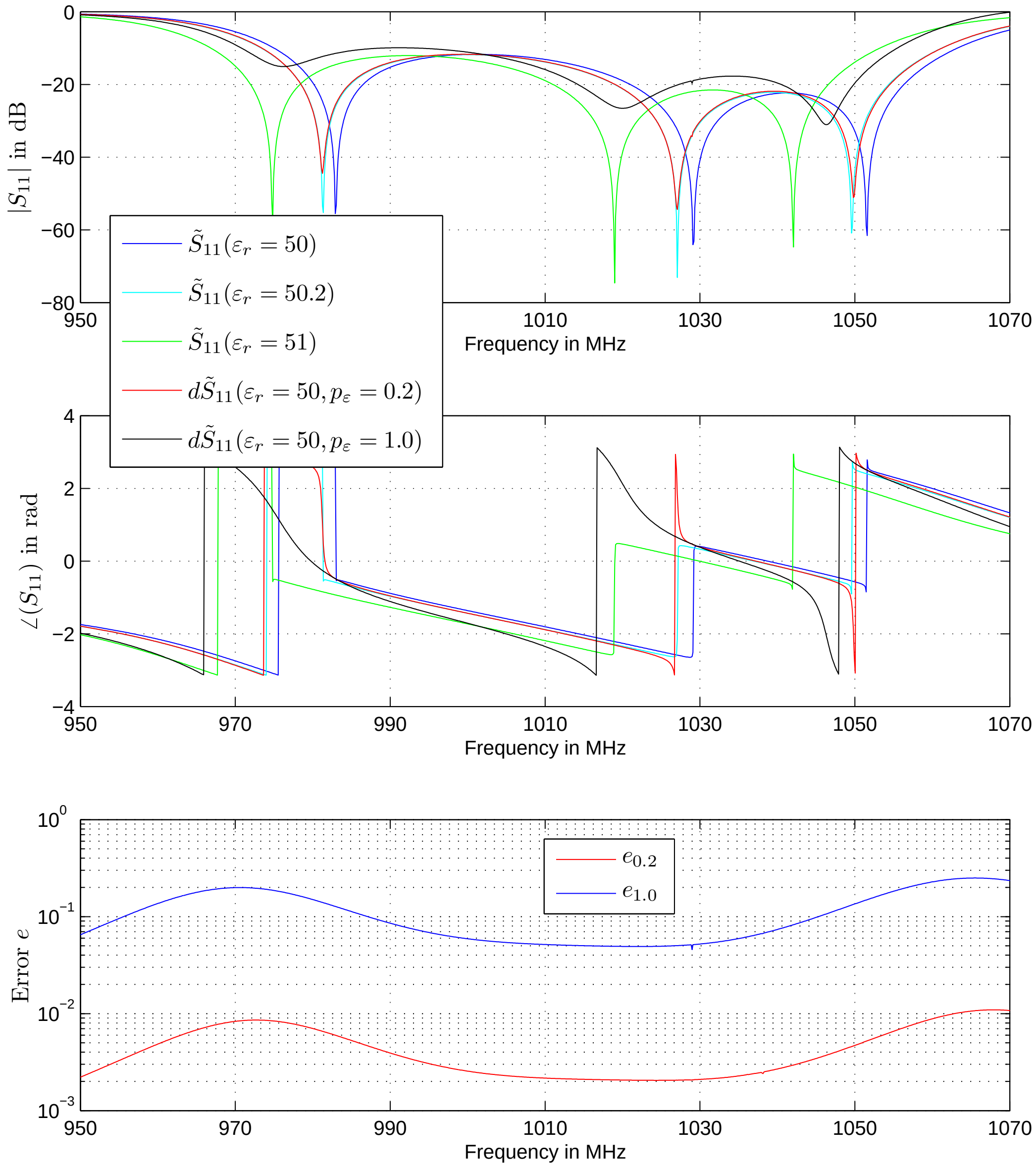


Figure 6.11: Three pole structure: View of the sensitivity analysis at the resonances around 1 GHz using scattering formulation.

Impedance Formulation

In this section, results of the simulation and the broadband sensitivity analysis in the impedance formulation are shown. As seen in the parallel plate waveguide, inner resonances causes high errors in the sensitivity analysis. The same effect can be seen for the dielectric three pole structure, where already for a small perturbation of $p_\varepsilon = 0.2$ the singularities cause high errors. This can be seen in Fig. 6.12, which shows a broadband sensitivity analysis in impedance formulation. The simulation data can be found in Table 6.4. In contrast to the parallel plate waveguide, the discontinuities here can even be found in the absolute values, and not only in the phase. These discontinuities can already be seen at low frequencies, as the extraction of the sweep in Fig. 6.13 shows.

Specifications:		Simulation data:	
Lowest frequency in Hz	1e7	Original dimension	306768
Highest frequency in Hz	1.2e9	Number of iterations q	8
Evaluation points	10000	ROM generation time in s	297
FE basis function order	2	Evaluation time in s	0.3
Multi-point method	-		

Table 6.4: Three pole structure: Simulation data using impedance formulation.

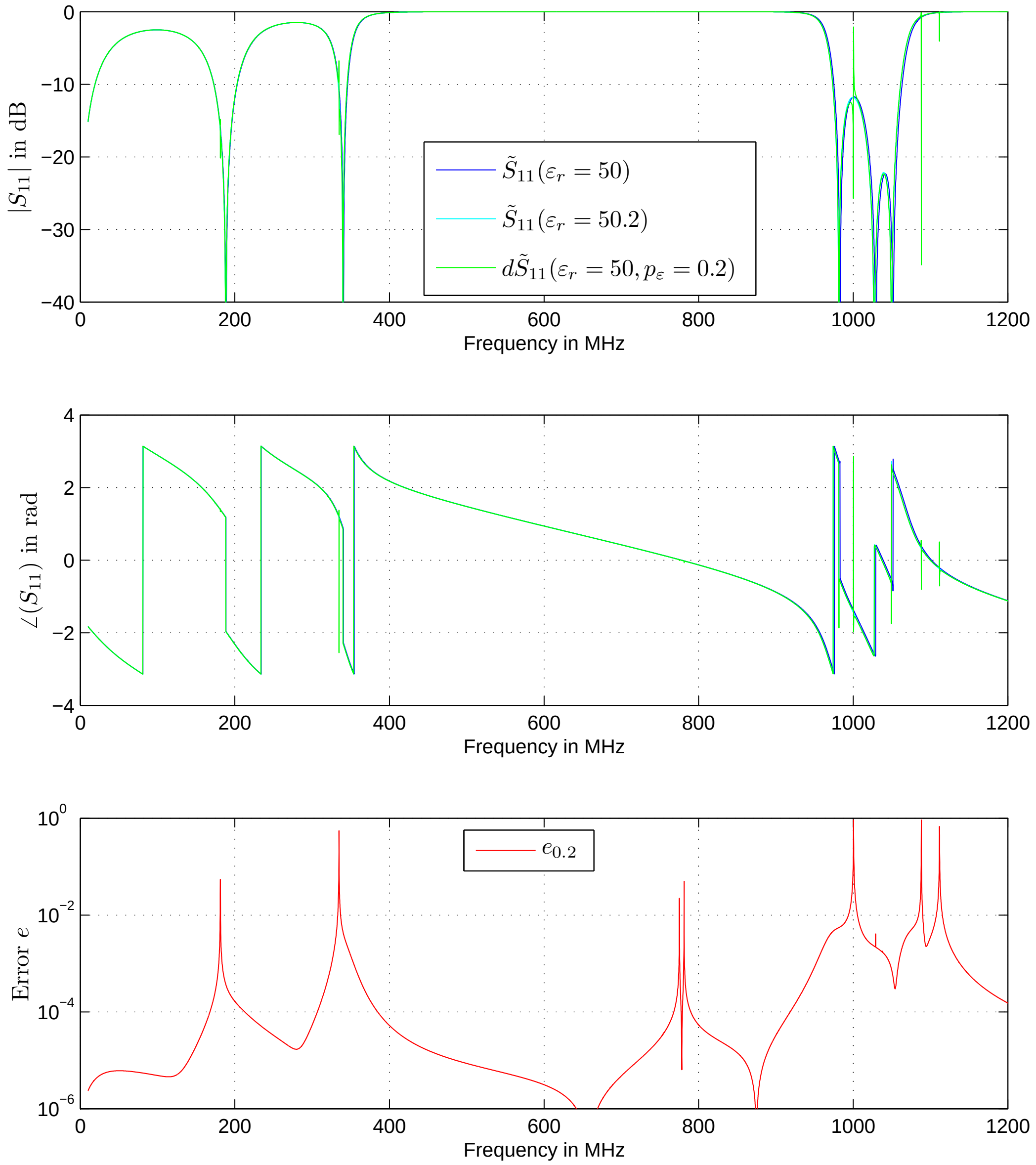


Figure 6.12: Three pole structure: Scattering parameters and sensitivity analysis versus frequency using impedance formulation.

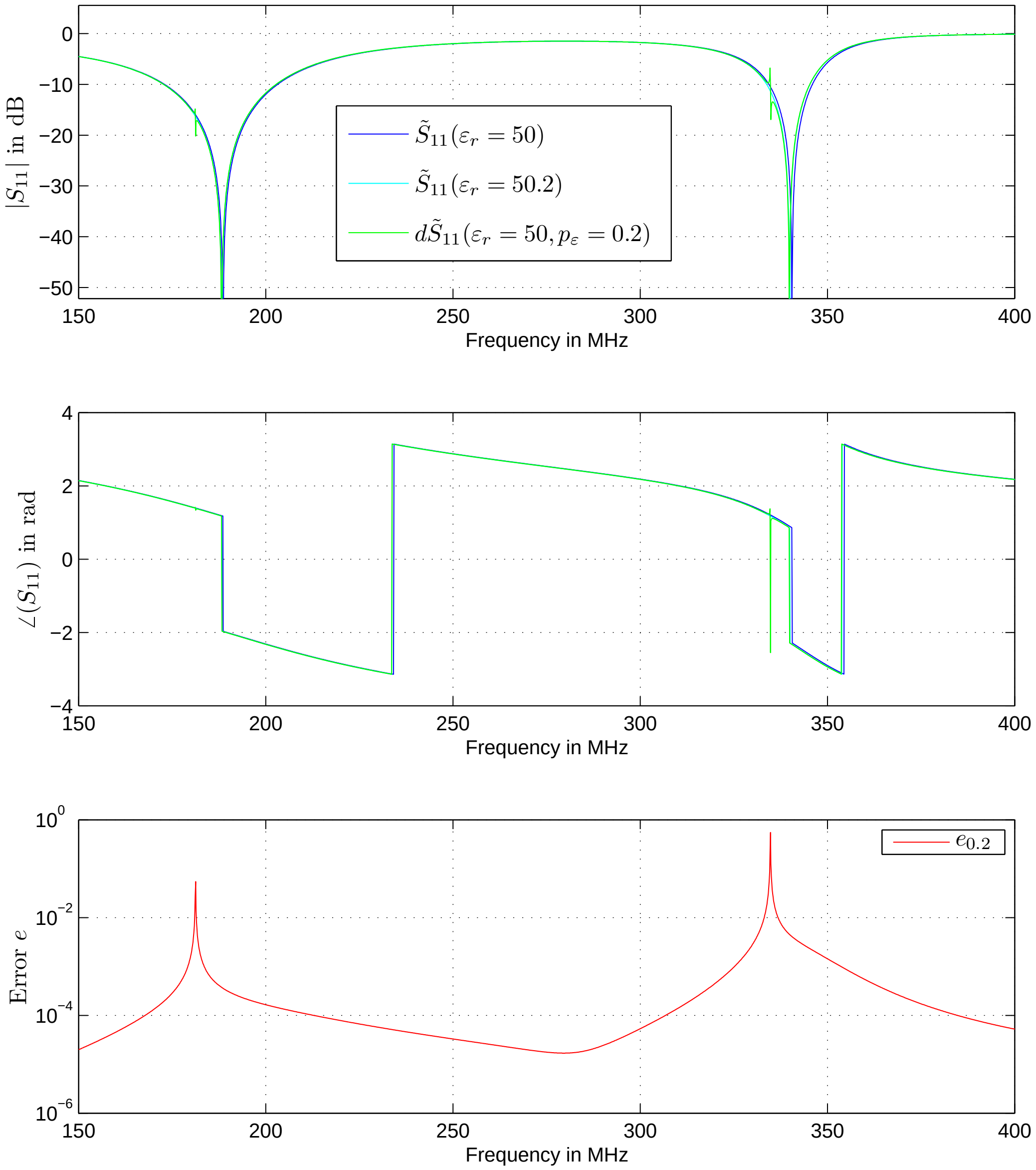


Figure 6.13: Three pole structure: Scattering parameters and sensitivity analysis at low frequencies using impedance formulation.

6.4 Conclusion

This chapter has proposed a powerful tool for analyzing the effect of perturbations in design parameters over a large frequency range. The broadband sensitivity technique is based on the adaptive MORe methods provided in the previous chapters. Thus, the reduction, as well as the sensitivity analysis process runs automatically using the introduced MORe error measures. This technique can be applied to a large set of design parameters and evaluated in arbitrary combinations. Numerical results show the efficiency and reliability of the theory if a scattering formulation is employed.

However, the sensitivity analysis using an impedance formulation exhibits large errors and fails to deliver reliable results. This effect is caused by the inner resonances of the impedance formulation, which is explained by means of analytical formulas in this chapter. Although the scattering parameters derived from the impedance formulation do not exhibit this non-physical behavior, the sensitivity analysis suffers from these singularities.

Chapter 7

Closing Words

This thesis provides adaptive MORe methods for the broadband FE simulation of a large set of real-world problems. In particular, an adaptive multi-point method is proposed and a proof for a blocked WCAWE method is given. Moreover, memory swapping algorithms are presented that allow us to maintain the system memory requirements for the projection matrix at a constant low level. Numerical results show the efficiency and reliability of these techniques. A memory analysis of the simulation runs demonstrates the importance of the swapping mechanisms.

The proposed adaptive multi-point method for broadband waveguide simulations, together with its incremental error measure, is shown to work reliably and efficiently. Additionally, some broadband sensitivity analysis techniques are developed which are based on the introduced MORe methods.

This thesis also provides a basis for future work. The out-of-core concept probably is very attractive for multivariate MORe, where the projection matrix size may become an even more limiting factor. Furthermore, there are some interesting issues concerning the blocked WCAWE algorithm. Deflation and alternative orthogonalization processes in the block algorithm should be investigated, which may reduce the computational efforts. Finally, extending the WCAWE blocking process to more than one parameter would probably improve the efficiency of multivariate single-point methods [40].

List of Abbreviations

AWE	: Asymptotic Waveform Evaluation
FE	: Finite Element
MORe	: Model Order Reduction
PEC	: Perfect Electric Conductor
PMC	: Perfect Magnetic Conductor
RAM	: Random Access Memory
ROM	: Reduced Order Model
TE	: Transverse Electric
TEM	: Transverse Electromagnetic
TM	: Transverse Magnetic
WCAWE	: Well-Conditioned Asymptotic Waveform Evaluation

List of Tables

List of Figures

Bibliography

[1] J. P. Webb and S. Perihar, “Finite element analysis of H-plane rectangular waveguide problems,” *IEE Proceedings H - Microwave, Antennas and Propagation*, vol. 133, no. 2, pp. 91–94, Apr. 1986.

[2] Z. J. Cendes and J.-F. Lee, “The transfinite element method for modeling MMIC devices,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 36, no. 12, pp. 1638–1649, Dec. 1988.

[3] J.-F. Lee, “Analysis of passive microwave devices by using three-dimensional tangential vector finite elements,” *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 3, no. 4, pp. 235–246, Dec. 1990.

[4] O. Farle, V. Hill, and R. Dyczij-Edlinger, “Finite-element waveguide solvers revisited,” *IEEE Transactions on Magnetics*, vol. 40, no. 2, pp. 1468–1471, Mar. 2004.

[5] J.-F. Lee, D.-K. Sun, and Z. J. Cendes, “Full-wave analysis of dielectric waveguides using tangential vector finite elements,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 39, no. 8, pp. 1262–1271, Aug. 1991.

[6] B. C. Moore, “Principal component analysis in linear systems: Controllability, observability, and model reduction,” *IEEE Transactions on Automatic Control*, vol. 26, no. 1, pp. 17–32, Feb. 1981.

[7] L. T. Pillage and R. A. Rohrer, “Asymptotic waveform evaluation for timing analysis,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 9, no. 4, pp. 352–366, Apr. 1990.

[8] P. Feldmann and R. W. Freund, “Efficient linear circuit analysis by Padé approximation via the Lanczos process,” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 5, pp. 639–649, May 1995.

[9] R. D. Slone, R. Lee, and J.-F. Lee, “Well-conditioned asymptotic waveform evaluation for finite elements,” *IEEE Transactions on Antennas and Propagation*, vol. 51, no. 9, pp. 2442–2447, Sep. 2003.

[10] E. Chiprout and M. S. Nakhla, “Analysis of interconnect networks using complex frequency hopping (CFH),” *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 14, no. 2, pp. 186–200, Feb. 1995.

[11] P. D. Ledger, J. Peraire, K. Morgan, O. Hassan, and N. P. Weatherill, “Parameterised electromagnetic scattering solutions for a range of incident wave angles,” *Computer Methods in Applied Mechanics and Engineering*, vol. 193, no. 33-35, pp. 3587–3605, Aug. 2004.

[12] Y. Konkel, O. Farle, and R. Dyczij-Edlinger, “Ein Fehlerschätzer für die Krylov-Unterraum-basierte Ordnungsreduktion zeitharmonischer Anregungsprobleme,” in *Modellbildung, Identifikation und Simulation in der Automatisierungstechnik, Tagungsband GMA Fachausschuss 1.30*, 2008.

[13] T. Wittig, “Zur reduzierung der modellordnung in elektromagnetischen feldsimulationen,” Ph.D. dissertation, Technische Universität Darmstadt, 2003.

[14] P. Monk, *Finite Element Methods for Maxwell's Equations.* Oxford, UK: Oxford University Press, 2003.

[15] A. Bossavit, *Computational Electromagnetism.* Boston, Massachusetts: Academic Press, 1998.

[16] J. D. Jackson, *Klassische Elektrodynamik*, 3rd ed. Berlin, Germany: de Gruyter, 2002.

[17] J.-P. Berenger, “A perfectly matched layer for the absorption of electromagnetic waves,” *Journal of Computational Physics*, vol. 114, no. 2, pp. 185–200, Oct. 1994.

[18] R. E. Collin, *Field Theory of Guided Waves.* Hoboken, New Jersey: John Wiley & Sons Inc., 1991.

[19] D. M. Pozar, *Microwave Engineering*, 3rd ed. Hoboken, New Jersey: John Wiley & Sons Inc., 2005.

[20] P. Ingelström, “A new set of H(curl)-conforming hierarchical basis functions for tetrahedral meshes,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 54, no. 1, pp. 106–114, Jan. 2006.

[21] O. Farle, M. Lösch, and R. Dyczij-Edlinger, “Efficient fast frequency sweep without nonphysical resonances,” *Electromagnetics*, vol. 30, no. 1-2, pp. 51–68, Mar. 2010.

[22] J. Rubio, J. Arroyo, and J. Zapata, “SFELP-an efficient methodology for microwave circuit analysis,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 3, pp. 509–516, Mar. 2001.

[23] M. L. Silveira, M. Kamon, and J. White, “Efficient reduced-order modeling of frequency-dependent coupling inductances associated with 3-D interconnect structures,” *32nd ACM/IEEE Conference on Design Automation*, vol. 14, no. 5, pp. 376–380, Jun. 1995.

[24] R. D. Slone, R. Lee, and J.-F. Lee, “Broadband model order reduction of polynomial matrix equations using single-point well-conditioned asymptotic waveform evaluation: derivations and theory,” *International Journal for Numerical Methods in Engineering*, vol. 58, no. 15, pp. 2325–2342, Dec. 2003.

[25] Y. Zhu and A. C. Cangellaris, “Finite element-based model order reduction of electromagnetic devices,” *International Journal for Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 15, no. 1, pp. 73–92, Jan. 2002.

[26] V. Hill, O. Farle, P. Ingelström, and R. Dyczij-Edlinger, "Efficient implementation of nonuniform refinement levels in a geometric multigrid finite-element method for electromagnetic waves," *IEEE Transactions on Magnetics*, vol. 43, no. 4, pp. 1521–1524, Apr. 2007.

[27] R. D. Slone, J.-F. Lee, and R. Lee, "Automating multipoint Galerkin AWE for a FEM fast frequency sweep," *IEEE Transactions on Magnetics*, vol. 38, no. 2, pp. 637–640, Mar. 2002.

[28] A. Schultschik, O. Farle, and R. Dyczij-Edlinger, "An adaptive multi-point fast frequency sweep for large-scale finite element models," *IEEE Transactions on Magnetics*, vol. 45, no. 3, pp. 1108–1111, Mar. 2009.

[29] O. Farle, V. Hill, and R. Dyczij-Edlinger, "Multi-parameter polynomial order reduction of linear finite element models," *Mathematical and Computer Modeling of Dynamical Systems*, vol. 14, no. 5, pp. 421–434, Oct. 2008.

[30] R. Sanaie, E. Chiprout, M. S. Nakhla, and Q.-J. Zhang, "A fast method for frequency and time domain simulation of high-speed VLSI interconnects," *IEEE Transactions on Microwave Theory and Techniques*, vol. 42, no. 12, pp. 2562–2571, Dec. 1994.

[31] X.-M. Zhang and J.-F. Lee, "Application of the AWE method with the 3-D TVFEM to model spectral responses of passive microwave components," *IEEE Transactions on Microwave Theory and Techniques*, vol. 46, no. 11, pp. 1735–1741, Nov. 1998.

[32] R. W. Freund, "Passive reduced-order models for interconnect simulation and their computation via Krylov-subspace algorithms," in *DAC 99, Proceedings of the 36th annual ACM/IEEE Design Automation Conference, New Orleans, LA*, Jun. 1999.

[33] T. Koyama and S. Govindjee, "Moment matching theorems for dimension reduction of higher-order dynamical systems via higher-order Krylov spaces," Department of Civil and Environment Engineering, University of California, Barkley, http://www.ce.berkeley.edu/~sanjay/, Tech. Rep., Nov. 2008, report No. UCB/SEMM-2008/04.

[34] Z. Bai and Y. Su, "SOAR: A second-order Arnoldi method for the solution of the quadratic eigenvalue problem," *SIAM Journal on Matrix Analysis and Applications*, vol. 26, no. 3, pp. 640–659, 2005.

[35] B. Lohmann and B. Salimbahrami, "Reduction of second order systems using second order Krylov subspaces," in *Proceedings of the IFAC World Congress, Prag, Czech Republic*, Jul. 2005.

[36] T.-J. Su and R. R. Craig, "Model reduction and control of flexible structures using Krylov vectors," *Journal of Guidance, Control and Dynamics*, vol. 14, no. 2, pp. 260–267, Mar. 1991.

[37] Y. Konkel, O. Farle, A. Köhler, A. Schultschik, and R. Dyczij-Edlinger, "Adaptive strategies for fast frequency sweeps," *COMPEL: The International Journal for Computation and Mathematics in Electrical and Electronic Engineering*, vol. 30, no. 6, pp. 1855–1869, 2011.

[38] J. E. Bracken, D.-K. Sun, and Z. J. Cendes, “S-domain methods for simultaneous time and frequency characterization of electromagnetic devices,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 46, no. 9, pp. 1277–1290, Sep. 1998.

[39] D.-K. Sun, Z. J. Cendes, and J.-F. Lee, “ALPS-a new fast frequency-sweep procedure for microwave devices,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 49, no. 2, pp. 398–402, Feb. 2001.

[40] O. Farle and R. Dyczij-Edlinger, “Numerically stable moment matching for linear systems parameterized by polynomials in multiple variables with applications to finite element models of microwave structures,” *IEEE Transactions on Antennas and Propagation*, vol. 58, no. 11, pp. 3675–3684, Nov. 2010.

[41] R. D. Slone, R. Lee, and J.-F. Lee, “Multipoint Galerkin asymptotic waveform evaluation for model order reduction of frequency domain FEM electromagnetic radiation problems,” *IEEE Transactions on Antennas and Propagation*, vol. 49, no. 10, pp. 1504–1513, Oct. 2001.

[42] E. J. Grimme, “Krylov projection methods for model reduction,” Ph.D. dissertation, Coordinated-Science Laboratory, University of Illinois at Urbana-Champaign, 1997.

[43] J. R. Brauer and G. C. Lizalek, “Microwave filter analysis using a new 3-D finite-element modal frequency method,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 5, pp. 810–818, May 1997.

[44] Y. Saad, “On the Lanczos method for solving symmetric linear systems with several right-hand sides,” *Mathematics of Computations*, vol. 48, no. 178, pp. 651–662, Apr. 1987.

[45] J. I. Aliaga, D. L. Boley, R. W. Freund, and V. Hernández, “A Lanczos-type method for multiple starting vectors,” *Mathematics of Computations*, vol. 69, no. 232, pp. 1577–1601, 2000.

[46] R. W. Freund, “Krylov-subspace methods for reduced-order modeling in circuit simulation,” *Journal of Computational and Applied Mathematics*, vol. 123, no. 1-2, pp. 395–421, Nov. 2000.

[47] “Intel Math Kernel Library, Intel corporation,” http://software.intel.com/en-us/articles/intel-mkl/, last visited 15/07/2012.

[48] R. W. Freund, “Krylov subspaces associated with higher-order linear dynamic systems,” *BIT Numerical Mathematics*, vol. 45, no. 3, pp. 495–516, Sep. 2005.

[49] M. H. Gutknecht and T. Schmelzer, “The block grade of a block Krylov space,” *Linear Algebra and its Applications*, vol. 430, no. 1, pp. 147–185, Jan. 2009.

[50] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics.* New York, New York: Springer Verlag, 2000.

[51] S. V. Polstyanko, R. Dyczij-Edlinger, and J.-F. Lee, “Fast frequency sweep technique for the efficient analysis of dielectric waveguides,” *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 7, pp. 1118–1126, Jul. 1997.

[52] F. Bertazzi, O. Peverini, M. Goano, G. Ghione, R. Orta, and R.Tascone, "A fast reduced-order model for the full-wave FEM analysis of lossy inhomogeneous anisotropic waveguides," *IEEE Transactions on Microwave Theory and Techniques*, vol. 50, no. 9, pp. 2108–2114, Sep. 2002.

[53] S.-H. Lee, T.-Y. Huang, and R.-B. Wu, "Fast waveguide eigenanalysis by wide-band finite-element model-order reduction," *IEEE Transactions on Microwave Theory and Techniques*, vol. 53, no. 8, pp. 2552–2558, Aug. 2005.

[54] S.-C. Lee, J.-F. Lee, and R. Lee, "Hierarchical vector finite elements for analyzing waveguiding structures," *IEEE Transactions on Microwave Theory and Techniques*, vol. 51, no. 8, pp. 1897–1905, Aug. 2003.

[55] A. Schultschik, O. Farle, and R. Dyczij-Edlinger, "A model order reduction method for the finite element simulation of inhomogeneous waveguides," *IEEE Transactions on Magnetics*, vol. 44, no. 6, pp. 1394–1397, June 2008.

[56] L. N. Trefethen and D. Bau, *Numerical Linear Algebra.* Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics, 1997.

[57] J. Strube and F. Arndt, "Rigorous hybrid-mode analysis of the transition from rectangular waveguide to shielded dielectric image guide," *IEEE Transactions on Microwave Theory and Techniques*, vol. 33, no. 5, pp. 391–401, May 1985.

[58] H.-B. Lee, H.-K. Jung, S.-Y. Hahn, C. Cheon, and K.-S. Lee, "Shape optimization of H-plane waveguide tee junction using edge finite element method," *IEEE Transactions on Magnetics*, vol. 31, no. 3, pp. 1928–1931, May 1995.

[59] J. Kim, H.-B. Lee, H.-K. Jung, S.-Y. Hahn, C. Cheon, and H. Kim, "Optimal design technique for waveguide device," *IEEE Transactions on Magnetics*, vol. 32, no. 3, pp. 1250–1253, May 1996.

[60] H.-B. Lee and T. Itoh, "A systematic optimum design of waveguide-to-microstrip transition," *IEEE Transactions on Microwave Theory and Techniques*, vol. 45, no. 5, pp. 803–809, May 1997.

[61] H. Akel and J. P. Webb, "Design sensitivities for scattering-matrix calculation with tetrahedral edge elements," *IEEE Transactions on Magnetics*, vol. 36, no. 4, pp. 1043–1046, Jul. 2000.

[62] J. P. Webb, "Design sensitivity of frequency response in 3-D finite-element analysis of microwave devices," *IEEE Transactions on Magnetics*, vol. 38, no. 2, pp. 1109–1112, Mar. 2002.

[63] D. Nair and J. P. Webb, "Optimization of microwave devices using 3-D finite elements and the design sensitivity of the frequency response," *IEEE Transactions on Magnetics*, vol. 39, no. 3, pp. 1325–1328, May 2003.