Detection of Spurious Modes in Resonance Mode Computations - Pole Condition Method

Dissertation

Benjamin Kettner

Fachbereich Mathematik und Informatik
Freie Universität zu Berlin

März 2012

For Freya & Pia.

Supervisor:	PD Dr. Frank Schmidt, Zuse Institute Berlin
Second referee:	Prof. Dr. Achim Schädle, Heinrich-Heine-Universität Düsseldorf
Date of disputation:	July 5^{th}, 2012

Acknowledgements

I would like to extend the most sincere thanks to my superviser PD Dr. Frank Schmidt, who did not only encourage and support me in every possible way while writing this thesis but also always had an open door and helpful advice whenever I needed them.

I would also like to thank Prof. Achim Schädle who always showed great interest in my work and was very supportive and also gave me a lot of inspiration and help during my visit in Düsseldorf. Also his help with the $2D$-implementation of the pole condition and the finite element method is priceless.

Many thanks are also in order to Prof. Peter Deuflhard, who was always very supportive and has managed to create a productive and inspiring atmosphere at ZIB.

I would like to thank my colleagues Dr. Lin Zschiedrich, Dr. Sven Burger, Dr. Jan Pomplun, Dr. Kirankumar Hiremath, Therese Pollok, Dr. Mark Blome, Daniel Lockau and Martin Hammerschmidt for countless discussions which were most useful. I could always rely on them to help me master any difficulties I encountered during the creation of this thesis. Therese Pollok and Martin Hammerschmidt I would also like to thank for proofreading my manuscript.

Also I would like to thank my family very much, especially my parents and wife for their support and my father in law Prof. Wulf Diepenbrock for always pushing me, showing interest in my work and supporting me with his great strategic knowledge.

I acknowledge the financial support of this thesis within project D23 of MATHEON, the research center "Mathematik für Schlüsseltechnologien" of the "Deutsche Forschungsgemeinschaft" DFG.

Contents

1 Introduction **1**

2 Helmholtz Resonance Problems **3**
2.1 The Helmholtz Equation 3
2.2 Resonance Problems 8
2.3 Transparent Boundary Conditions 11

3 Pole Condition and Implementation **15**
3.1 Derivation: The Pole Condition in 1D 16
3.2 Formalization 18
3.3 Implementation 23
3.4 Alternative Approach 31
3.5 Generalization to Higher Space Dimensions 33

4 Spurious Solutions **43**
4.1 Spurious Solutions in Closed Cavities 44
4.2 Spurious Solutions in Open Resonators 46

5 Detecting Spurious Solutions **57**
5.1 Generalized Eigenvalue Problems 58
5.2 Use and Limitations of Condition Numbers 67
5.3 Exact Perturbation 78
5.4 Convergence Monitor 86
5.5 Removing Spurious Solutions 101

6 Summary and Outlook **105**

7 Zusammenfassung **111**

Bibliography **113**

Chapter 1

Introduction

> A thousand valleys' rustling pines resound.
> My heart was cleansed, as if in flowing water.
> In bells of frost I heard the resonance die.[1]

In the field of numerics, when solving partial differential equations on an unbounded domain, one typically splits the unbounded domain into a bounded interior and an unbounded exterior domain. Then the equation is solved only on the bounded interior domain and transparent boundary conditions are applied to the interface between the bounded interior and the unbounded exterior domain.

This thesis deals with solving resonance problems of the Helmholtz equation on spatially unbounded domains. Here, the application of transparent boundary conditions causes the existence of solutions that bear no physical meaning but are discretization artifacts. These so called *spurious solutions* are sometimes difficult to distinguish from the physical solutions of the problem without a priori knowledge of the expected eigenvalue spectrum or field distributions. In this thesis we will derive a robust algorithm that allows for the detection and removal of spurious solutions from the computed eigenvalue spectrum.

First we will derive the basic equation used throughout this thesis, Equation (2.17). Then we will briefly highlight the physical and technical background of the problem at hand. Following this introduction, we will derive the central tool calculations, the pole condition [Sch02]. Following the implementations of Hohage, Nannen [NS11, Nan08, HN09], Schädle and Ruprecht [RSSZ08], we will obtain a formulation that allows for implementation in the one- and two-dimensional case and fits well in the finite element context. The implementation of the pole condition will reduce to the implementation of two matrices given in Equation (3.45) in the one-dimensional

[1] Quotation originating from Vikram Seth's translations of Li Bai's poems [Set92], used by D. Bindel and M. Zworski in the introduction of their review on scattering poles for the Schrödinger equation `http://www.cims.nyu.edu/~dbindel/resonant1d/theo2.html`

case and to the implementation of two products of sums of matrices in the two-dimensional case, given in Equations (3.66) and (3.67).

We will then investigate the spurious solutions a bit closer and learn that there exist spurious solutions that are caused by the transparent boundary conditions. We will then derive an algorithm to detect the spurious solutions and remove them from the computed eigenvalue spectrum. This algorithm is the central result of this thesis and given in a flow-chart representation in Figure 5.30. Its central building blocks are a formula to compute the reaction of an eigenvalue to the perturbation of the matrix, given in Lemma 5.3, and a convergence monitor that computes the rate of convergence of the transparent boundary condition for an eigenvalue, given in the one-dimensional case by solving (5.20) for κ and in the two-dimensional case by evaluating (5.36) at ω.

The perturbations we will use as input to compute the perturbation of the eigenvalues are perturbations that affect only the exterior part of the problem. These perturbations will be derived in Equations (5.12) and (5.13) for the one-dimensional case and in Equations (5.14) and (5.15) for the two-dimensional case.

Chapter 2

Helmholtz Resonance Problems: Background and Basic Equations

In this introductory chapter we will set the stage for the investigations within this thesis. The chapter is organized as follows: In Section 2.1 we will introduce the Helmholtz equation that is the basic equation we wish to solve and give some physical context. Section 2.2 will then give a problem statement of the resonance mode setting of the Helmholtz equation. Also we will present in short the problem of spurious solutions whose solution is the main result of this thesis. Finally, Section 2.3 will give a brief introduction of the finite element method and an overview of different types of transparent boundary conditions used in the approximation of resonance problems for open resonators, thus introducing the numerical basics for this work. The sections of this chapter bear many references to later chapters of this work as they are supposed to serve as introduction rather than as in-depth explanations or investigations.

2.1 The Helmholtz Equation

The Helmholtz equation is the governing equation of time-harmonic wave propagation. It therefore serves as a simplified model equation for many applications and will be the model equation we will use throughout this thesis. Dealing with the time harmonic case has the premise that our solutions are in steady-state and we can neglect switching and transient effects. Using the circular frequency ω, we can assume that in the steady-state of the time-dependent scalar field $F(x,t)$ the time-dependence can be separated as $F(x,t) = f(x)e^{-i\omega t}$ where f is a stationary function. There are several application fields where the Helmholtz equation is the governing equation [Ihl98]. We will now highlight two of these fields, acoustic waves and electromagnetic

waves. Both of these fields will appear in the examples of this thesis and we will derive the Helmholtz equation for both cases from physical principles in the following sections.

Acoustic Waves

Sound is created by small oscillations of pressure in an acoustic medium (acoustic waves). These oscillations cause energy to be propagated through the medium and using the fundamental laws for compressible fluids, the governing equations can be derived. Let $p(x,t)$ be the pressure, $\rho(x,t)$ be the density and $v(x,t)$ be the velocity of the particles in the fluid, let V be a volume element and ∂V be its boundary, let $n(x)$ be the exterior normal unit vector of V at $x \in \partial V$. The velocity of the normal flux through ∂V is $v(x,t) \cdot n(x)$. First we will derive two basic laws, from which we can derive the desired equations. First, the law of conservation of mass given by

$$-\frac{\partial}{\partial t}\int_V \rho(x,t)dV = \oint_{\partial V} \rho(x,t)(v(x,t)\cdot n(x))dS.$$

Using the Gauss theorem, we can transform the surface integral into a volume integral and thus obtain

$$\int_V \left(\frac{\partial \rho(x,t)}{\partial t} + \operatorname{div}(\rho(x,t)v(x,t))\right) dV = 0,$$

from which we can derive

$$\frac{\partial \rho(x,t)}{\partial t} + \operatorname{div}(\rho(x,t)v(x,t)) = 0, \tag{2.1}$$

the continuity equation describing the conservation of mass. Next we will take into account the linearized equations of motion. Assume that V is subject to pressure $p(x,t)$. Then the total force along ∂V is $F = -\oint p(x,t)n(x)dS$. From the second Newtonian law $F = ma$ we get

$$-\oint_{\partial V} p(x,t)n(x)dS = \int_V \rho \frac{dv(x,t)}{dt} dV.$$

Expanding the total differential dV/dt, we get the nonlinear expression $dV/dt = \partial V/\partial t + (V \cdot \nabla)V$ (cf. [LL87]). Under the assumption of small oscillations, we can linearize this expression and replace the total differential with the partial differential: $dv(x,t)/dt \approx \partial v(x,t)/\partial t$. Applying the Gauss theorem, we arrive at the linearized equation of motion, also called the Euler equation

$$\rho \frac{\partial v(x,t)}{\partial t} = -\nabla p(x,t). \tag{2.2}$$

Returning to the mathematical question of sound waves, we recall that acoustic waves are small oscillations of pressure in a compressible fluid. Such an oscillation can be seen as a small perturbation in pressure and density $(p(x,t), \rho(x,t))$ of a steady state (p_0, ρ_0). At any point x, the functions $p(x,t)$ and $\rho(x,t)$ represent vibrations with a small amplitude. From the Euler equation it follows that the velocities are also small. If we assume a linear material law with the material constant c, we have

$$p(x,t) = c^2 \rho(x,t).$$

Deriving twice partially by t, assuming that the perturbations of ρ and $p(x,t)$ are small, hence that $\rho(x,t) \approx \rho_0$, and using the linearized Equations (2.1) and (2.2), we obtain

$$\begin{aligned} \frac{\partial^2}{\partial t^2} p(x,t) = c^2 \frac{\partial^2}{\partial t^2} \rho(x,t) &= -c^2 \rho_0 \operatorname{div}\left(\frac{\partial}{\partial t}\left(v(x,t)\right)\right) \\ &= c^2 \operatorname{div}(\nabla p(x,t)) = c^2 \Delta p(x,t). \end{aligned} \tag{2.3}$$

Making a time-harmonic ansatz $p(x,t) = p(x)e^{-i\omega t}$, we obtain from Equation (2.3) the Helmholtz equation:

$$\Delta p(x) + k^2 p(x) = 0 \tag{2.4}$$

with $k = \frac{\omega}{c}$ called the *wave number*.

Electromagnetic Waves

We will now deduct the electromagnetic wave equations from Maxwell's equations, the equations that describe the interaction between the electric and the magnetic field. Charges generate electric fields which, in conducting media, enforce currents. The interaction of currents in turn, generates the magnetic force field. We will see in this paragraph that both the electric and the magnetic field satisfy a vector wave equation which under certain assumptions yield a Helmholtz equation.

In order to derive Maxwell's equations, we will first have to establish the following constitutive relations connecting the electric field $\mathbf{E}$ with the conductive current $\mathbf{J}$ and the electric displacement $\mathbf{D}$ and the magnetic field H with the magnetic induction $\mathbf{B}$. These relationships are

$$\mathbf{J} = \sigma \mathbf{E}, \tag{2.5a}$$

$$\mathbf{D} = \varepsilon \mathbf{E}, \tag{2.5b}$$

$$\mathbf{B} = \mu \mathbf{H}. \tag{2.5c}$$

In order to derive Maxwell's equations from physical principles, we first state the dynamic Ampères law including Maxwell's correction term:

$$\oint_{\partial S} \mathbf{H} dl = \int_S \mathbf{J} ds + \int_S \frac{\partial}{\partial t} \mathbf{D} ds. \tag{2.6}$$

Here S is a surface and ∂S is its boundary. Ampères static law then states that the integral of the magnetic field intensity around a closed path is equal to the total current enclosed by the path, Maxwell's correction term includes the temporal change of the electric displacement through the surface which adds to the total current. Furthermore, we require Faradays law of induction:

$$\oint_{\partial S} \mathbf{E} dl = -\frac{\partial}{\partial t} \int_S \mathbf{B} ds. \tag{2.7}$$

It describes the interaction of a time-varying magnetic field that produces an electric field. Further we will have to include Gauss' law into our considerations which states that the electric flux through the surface of a volume V is equal to the total electric charge in the volume:

$$\oint_{\partial V} \varepsilon \mathbf{E} ds = \int_V \rho dV. \tag{2.8}$$

Finally we have to state that there are no magnetic monopoles, hence the integral of the magnetic flux over any closed surface is zero:

$$\oint_{\partial V} \mathbf{B} ds = 0. \tag{2.9}$$

Equations (2.6) to (2.9) together with the relations defined in Equations (2.5a) to (2.5c) yield the full system of Maxwell's equations in integral form:

$$\oint_{\partial V} \mathbf{E} ds = \frac{1}{\varepsilon} \int_V \rho dV, \tag{2.10a}$$

$$\oint_{\partial V} \mathbf{H} ds = 0, \tag{2.10b}$$

$$\oint_{\partial S} \mathbf{E} dl = -\mu \frac{\partial}{\partial t} \int_S \mathbf{H} ds, \tag{2.10c}$$

$$\oint_{\partial S} \mathbf{H} dl = \int_S \mathbf{J} ds + \varepsilon \frac{\partial}{\partial t} \int_S \mathbf{E} ds. \tag{2.10d}$$

By applying the divergence theorem $\oint_{\partial V} \mathbf{F} ds = \int_V \nabla \cdot \mathbf{F} dV$ and Stokes' theorem $\oint_{\partial S} F dl = \int_S (\nabla \times \mathbf{F}) ds$, we can transform the macroscopic Maxwell's Equations (2.10a)-(2.10d) to differential form

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\varepsilon}, \tag{2.11a}$$

$$\nabla \cdot \mathbf{H} = 0, \tag{2.11b}$$

$$\nabla \times \mathbf{E} = -\mu \frac{\partial}{\partial t} \mathbf{H}, \tag{2.11c}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \varepsilon \frac{\partial}{\partial t} \mathbf{E}. \tag{2.11d}$$

Assuming that the electric field is free of charges, that is $\rho \equiv 0$ and that there are no free currents, that ist $\mathbf{J} = 0$, Equations (2.11a) - (2.11d) reduce to the two coupled equations

$$\nabla \times \mathbf{E} + \mu \frac{\partial}{\partial t}\mathbf{H} = 0, \tag{2.12a}$$

$$\nabla \times \mathbf{H} - \varepsilon \frac{\partial}{\partial t}\mathbf{E} = 0, \tag{2.12b}$$

which are complemented by the condition that the vector fields have to be divergence-free. Making a time-harmonic ansatz $\mathbf{E}(\mathbf{x}, t) = \mathbf{E}(\mathbf{x}) \exp(i\omega t)$ and $\mathbf{H}(\mathbf{x}, t) = \mathbf{H}(\mathbf{x}) \exp(i\omega t)$, we obtain from Equations (2.12a) and (2.12b):

$$\nabla \times \mathbf{E} + i\mu\omega\mathbf{H} = 0, \tag{2.13a}$$

$$\nabla \times \mathbf{H} - i\omega\varepsilon\mathbf{E} = 0. \tag{2.13b}$$

By substitution, we have the curl-curl-equations

$$\nabla \times \varepsilon^{-1} \nabla \times \mathbf{H} - \omega^2 \mu \mathbf{H} = 0 \tag{2.14a}$$

$$\nabla \times \mu^{-1} \nabla \times \mathbf{E} - (\omega^2 \varepsilon - i\omega\sigma)\mathbf{E} = 0 \tag{2.14b}$$

Assuming that the electric or magnetic fields are directed along the invariant z-direction, the curl-curl-equations simplify to the following scalar Helmholtz equations for the fields $E_z(x, y)$ and $H_z(x, y)$ in two space dimensions:

$$\Delta E_z(x, y) - \omega^2 n(x, y)^2 E_z(x, y) = 0 \quad \text{and} \tag{2.15a}$$

$$\Delta H_z(x, y) - \omega^2 n(x, y)^2 H_z(x, y) = 0, \tag{2.15b}$$

where $n(x, y) = \sqrt{\varepsilon(x, y)\mu(x, y)}$ is the refractive index.

Radiation Conditions

To motivate the introduction of radiation conditions we will have a look at the one-dimensional wave equation

$$\Delta P(x, t) - \frac{1}{c^2}\frac{\partial^2}{\partial t^2} P(x, t) = 0, \text{ with } x \in \mathbb{R}. \tag{2.16}$$

It is easy to see that $P(x, t) = f(kx - \omega t)$ solves Equation (2.16) if $k = \frac{\omega}{c}$, since $\Delta f(kx - \omega t) = k^2 f''(kx - \omega t)$ and $\frac{\partial^2}{\partial t^2} f(kx - \omega t) = \omega^2 f''(kx - \omega t)$. The value of f does not change if the differential $d(kx - \omega t) = 0$ or, equivalently $\frac{dx}{dt} = \frac{\omega}{k}$. The expression $\frac{dx}{dt}$ is called the *phase velocity* of the solution $f(kx - \omega t)$ and it depends solely on material properties.

The connection between the phase velocity and the wave number can be illustrated by considering time harmonic solutions $P(x, t) = p(x)e^{-i\omega t}$.

Recalling that $k = \omega/c$, the stationary part $p(x)$ satisfies the Helmholtz equation

$$p(x)'' + k^2 p(x) = 0.$$

Its solutions are known to be periodic with $p(x+\lambda) = p(x)$ for all $x \in \mathbb{R}$ with $\lambda = \frac{2\pi}{k}$ and have the general form $p(x) = Ae^{ikx} + Be^{-ikx}$. The corresponding time-dependent solution is $P(x,t) = p(x)e^{-i\omega t}$, hence

$$P(x,t) = Ae^{i(kx-\omega t)} + Be^{-i(kx-\omega t)}.$$

The phase velocities of the two terms of this solution can be evaluated to be c for the first term and $-c$ for the second term. Looking at an arbitrary point x_0, the first term therefore represents a wave traveling towards infinity, hence called outgoing solution, while the second term represents a wave approaching x_0 from infinity, hence called incoming solution.

Considering wave propagation in free space, we postulate that no waves are reflected from infinity. That means that we seek for a condition to suppress incoming solutions. Such a condition, the *Sommerfeld radiation condition* was defined by Sommerfeld [Som49]. Deriving it requires taking into account the free space Green's function in an exterior domain and formulating an integral equation for $u(r)$ where r is in the exterior domain. When truncating the exterior domain by a circle and taking the circles radius to infinity, one obtains a condition ensuring that no incoming solutions exist. This condition will be explicitly given in Chapter 3 when we derive transparent boundary conditions for solving the Helmholtz equation numerically.

2.2 Resonance Problems

The Helmholtz equation

$$\Delta \mathbf{u} + n^2\omega^2 \mathbf{u} = 0, x \in \mathbb{R}^d, d \in \{1,2,3\} \tag{2.17}$$

where n is a material-dependent property and ω the frequency contained in the exponential part of the time-harmonic ansatz, can be solved in different ways yielding two basic problem classes:

- *Scattering problems*: The typical problem setting for scattering problems is sketched in Figure 2.1. In this setting, an incoming wave u_{in} is scattered off an obstacle S within a domain of interest Ω. This generates scattered waves u_{sc} that satisfy a radiation condition and leave the area of interest $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$. The total field of such a problem then is $u_{\text{tot}} = u_{\text{in}} + u_{\text{sc}}$. By the introduction of the

artificial domain of interest Ω with the boundary $\partial\Omega$, we can rename the solution within Ω to $u_{\text{int}}(x) := u_{\text{tot}}(x)$ for $x \in \Omega$ and reformulate

$$\Delta u_{\text{sc}}(x) + k^2 u_{\text{sc}}(x) = 0 \qquad \text{for } x \in \mathbb{R}^d \backslash \Omega, \tag{2.18e}$$

$$\Delta u_{\text{int}}(x) + k^2 u_{\text{int}}(x) = 0 \qquad \text{for } x \in \Omega, \tag{2.18f}$$

$$u_{\text{sc}}(x) + u_{\text{in}}(x) = u_{\text{int}}(x) \qquad \text{for } x \in \partial\Omega, \tag{2.18g}$$

$$\partial_n u_{\text{sc}}(x) + \partial_n u_{\text{in}}(x) = \partial_n u_{\text{int}}(x) \qquad \text{for } x \in \partial\Omega. \tag{2.18h}$$

where ∂_n denotes the derivative in the outward normal direction on $\partial\Omega$. Here, the material parameter n and the frequency of the fields, ω are known and the resulting inner and scattered fields are computed.

- *Resonance problems*: For resonance problems the setting is different since no sources are given, hence Equation (2.17) is reinterpreted as eigenvalue problem where the eigenvector (eigenmode) u and the eigenvalue (eigenfrequency) ω of a structure are sought simultaneously. This problem type is the basis of our considerations and will be investigated in more detail in the subsequent chapters, focusing on transparent boundary conditions and the numerical solution of the problem.

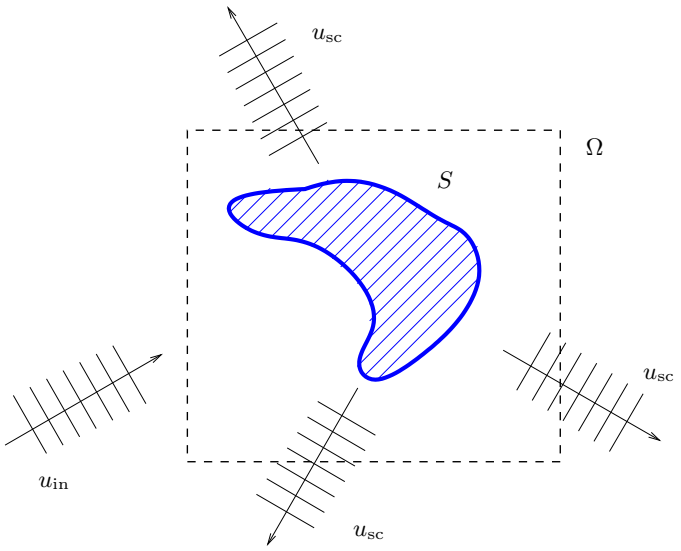

Figure 2.1: Scattering problem. The incoming wave u_{in} is scattered off the obstacle S, resulting in scattered waves u_{sc}.

Now we will first comment on the technological impact of resonance problems before giving the outline of a justification for this problem type. Computing the eigenfrequencies and eigenmodes of cavity resonators is an important technological task with first attempts for solutions dating back almost to the beginning of the computer age [Hoy65, HSR66]. Resonance

problems are of central importance for a wide area of applications. In modern optical applications for example, the functionality of many devices built depends strongly on their ability to respond stronger to incident light at certain frequencies. The impact of these devices is manifold and they affect a wide area of applications:

- In *photovoltaics*, optical resonators are used to achieve light trapping in order to increase the efficiency of solar cells. The aim is to focus the light within certain regions of the solar cell in order to increase the number of free electrons produced in that region. A profound knowledge of the resonances of a structure is required in order to optimize the shape and size of the microstructures that are embedded in the cells.

- In *medicine*, optical resonators can be used for detection of molecules ("lab on a chip") [Flö11, GSH$^+$10]. Here, resonators are coated with a functionalized surface that certain molecules can attach to. When molecules attach to this surface, the volume of the resonator is increased and the resonance frequency changes. Utilizing this effect for molecule detection requires prediction of the fields and frequencies of the resonant states in the highest possible quality-

- In *lasers*, resonators are often used for monochromatic light generation, especially the functionality of vertical cavity surface emitting lasers (VCSELs) depends centrally on the devices acting as optical cavities.

- In *information technology*, resonators are expected to become of central importance for designing and building the envisioned quantum computers. Their ability to store the information inherent in light quanta is required to implement bits required for information processing.

The wide spread of this problem type and significance for its applications make for a high level of interest in solving this problem type numerically in order to optimize structures, understand their physical properties and tailor them to the requirements of the individual applications. However, we will see that the numerical solution of resonance problems is a delicate matter since it introduces unwanted solutions, also called *spurious solutions*, that bear no physical meaning and pollute the solution spectra. We will discuss their origin in Chapter 4 and devote this work to their detection in typical simulation settings arising in modern applications.

The numerical simulation of resonance problems requires solving eigenvalue problems on unbounded domains. This implies that the eigenvalue, that is the resonance frequency ω, is a complex number even for real-valued coefficient matrices, which was early found to be useful since it enables taking into account radiation and decay properties [XZX89]. Yet the notion

of a complex frequency is one that is not valid within the physical model so a justification is required. Numerical justification was already given in several papers by comparing the results obtained by solving the scattering problem for a frequency range with the resonance frequencies computed by solving the resonance problem [BPSZ11, BZS10, BSZ10]. From a practical point of view, the complex frequencies can be interpreted as follows: the real part $\Re(\omega)$ describes the rest energy of a state, that is the wavelength of the resonance while the imaginary part $\Im(\omega)$ describes the rate of decay. Consequently they should be understood in terms of long time behavior of solutions to the wave equation. However, there is also a mathematical justification for computing complex resonances. Using a "black box" formalism, Zworski and Tang showed that solutions to the wave equation in $\mathbb{R}^d$ can be expanded in resonances [TZ00]. They stated that since an expansion in eigenfunctions is possible on bounded domains and resonances are the equivalent to eigenfunctions when considering problems posed on unbounded domains, an expansion involving resonances was to be expected as anticipated by Lax and Phillips [LP89]. They showed that for odd space dimensions n the solution $u(x,t)$ to the wave equation can be expanded as

$$u(x,t) = \sum_{\Im(\omega_l)\leq C} \sum_{k=1}^{m(\omega_l)} u_k(x) e^{it\omega_l} t^{k-1} + \mathcal{O}(e^{-(C-\varepsilon)t}).,\tag{2.19}$$

for x in the unbounded exterior domain where $m(\omega_l)$ is the multiplicity of ω_l. The expansion given here and in [WMS88, MS84, Zwo99] shows that even though the complex resonance frequencies and the associated field distributions seem unphysical and unintuitive at first they still bear a physical meaning by acting as a basis that can be used for the approximation of solutions to scattering problems.

2.3 Transparent Boundary Conditions

Many books have been written on the solution of partial differential equations using the finite element method. For in-depth literature on this topic we refer to [Ihl98, Mon03, Bra10] and many more. We will however present the most important implementations of transparent boundary conditions in this section since our implementation of transparent boundary conditions, the pole condition will be central for our work.

The necessity to formulate transparent boundary conditions stems from numerics. In order to numerically solve a problem posed on an unbounded domain, the objects of interest (scatterers or resonators) are enclosed in a bounded domain Ω_{int} and an unbounded exterior domain Ω_{ext} which acts as computational domain. To account for light that radiates out of this computational domain to infinity, transparent boundary conditions have to

be imposed on the boundary $\partial\Omega$. The main different concepts for the realization of transparent boundary conditions are

- *Green's tensor methods:* This class of methods relies on the availability of fundamental solutions to the equations that are investigated in Ω_{ext}. A class of examples applying Green's tensor methods are classical boundary integral methods [CK98, Néd01], techniques for certain special cases such as layered materials are available [MP98, PM01]. All these methods however share the downside, that for many applications the Green's tensor is not numerically feasible.

- *Perfectly matched layers:* The general idea behind the perfectly matched layers is to surround Ω_{int} with a layer of finite thickness with a material that is specially designed to damp or slow down the radiating solutions within the finite thickness so that they do not reflect back when a zero boundary condition is applied to the end of the layer. This method is very elaborate and can deal with all geometries arising in current applications [Ber94, Zsc09]. The perfectly matched layers were shown to be equivalent to an analytic continuation of the equation to complex coordinates which replaces the propagating oscillating waves by exponentially decaying waves.

- *Infinite elements:* The basic idea of the infinite element method is to decompose Ω_{ext} into disjoint infinite patches similar to the patches with ansatz functions similar to the finite patches and ansatz functions used by the finite element method in the interior Ω_{int} [Lei86, CRD03]. These ansatz functions then span a subspace of the underlying weighted Sobolev space. The drawback here is that the construction of conformal ansatz functions requires knowledge on the asymptotic behavior of the solution, thus rendering a treatment of heterogeneous exterior domains impossible.

- *Mode matching methods:* For mode matching methods, the solution is expanded into resonances on infinite patches in the exterior. Then matching conditions are enforced on the interfaces between two infinite patches and on $\partial\Omega$ to ensure the field continuity. This works well for many applications including complex heterogeneous exterior domains [Che95, Ham07]. However, the number of modes required for an accurate approximation is quite high for some examples, giving large dense blocks in the system matrices which drastically increase the computational cost.

- *Pole condition method:* The pole condition method which we will use in this thesis, will be explained in detail in Chapter 3. It can be seen as a generalization of the infinite element method that makes the consideration of heterogeneous exterior domains possible.

For more in-depth discussion of the different types of transparent boundary conditions, we refer to review papers such as [Tsy98, Hag99, Hag03].

Chapter 3

Pole Condition and Implementation

In this chapter, we will present the transparent boundary condition that will be used in this thesis. It is a Hardy Space Infinite Element implementation of the pole condition.
The central condition underlying this implementation, the pole condition, was developed by F. Schmidt [Sch98, Sch02]. It defines outward radiating solutions by the location of the poles of their transform with respect to a generalized distance variable in the complex plane. It is equivalent to the PML [Ber94, HSZ03b] and its correspondence with the Sommerfeld radiation condition for homogeneous exterior domains was shown by Hohage et al. [HSZ03a]. In their review paper [AAB$^+$07] Antoine et al. discuss its relation with other concepts for transparent boundary conditions for the Schrödinger equation. In contrast to earlier implementations of the pole condition that were based on BDF and Runge-Kutta methods [HSZ02], our approach implements the Hardy Space Infinite Element approach [NS11, Nan08, HN09]. It is based on a Galerkin method in the Hardy space $H^+(D)$ of the complex unit disk. As we will see later in this chapter, the approach chosen here does not change the structure of the underlying eigenvalue problem since it is linear in the resonance frequency ω^2. Therefore the eigenvalue problems that will occur in our implementation can be solved with standard sparse eigenvalue solvers.

In Section 3.1 we will motivate the pole condition by formulating it in a one-dimensional resonance mode setting. It should be stressed here, that this section serves to depict the method used and to ensure the confidence in the method but is not required for the derivation of the method. Sections 3.2 and 3.3 will formalize this approach and give details on the implementation we will use. Section 3.4 will briefly outline an alternative approach that gives almost the same discretization but will be useful in Section 3.5 when we deal with the implementation for two space dimensions.

3.1 Derivation: The Pole Condition in 1D

In this section we aim at deriving the condition we will later refer to as the pole condition in a simple one-dimensional setting. This section serves to depict the method used. In order to derive the pole condition we will first consider the Helmholtz resonance problem on an unbounded one-dimensional domain. In order to obtain numerical solutions with the finite element method, we will split that domain into a bounded interior domain Ω_{int} than can be dealt with by standard finite elements and two unbounded exterior domains $\Omega_{\text{ext},l}$ and $\Omega_{\text{ext},r}$. Then we will use the simplicity of the one-dimensional case to derive special solutions for our equation in $\Omega_{\text{ext},l}$ and $\Omega_{\text{ext},r}$. Based on the Sommerfeld radiation condition, we can deduct the explicit forms of these special solutions. By observing their behavior under the Laplace transform, we can formulate a condition on the Laplace transform that guarantees that the solutions satisfy the Sommerfeld radiation condition and thus are outgoing.

In a one dimensional setting, we want to solve the scalar Helmholtz equation on an unbounded domain $\Omega \subseteq \mathbb{R}$

$$\partial_{xx} u(x) + n(x)^2 \omega^2 u(x) = 0 \text{ for } x \in \Omega. \tag{3.1}$$

In a resonance mode setting, we search for a pair $(u(x), \omega)$ with $u(x) \in C^2(\Omega)$ and $\omega \in \mathbb{C}$. To tackle this task numerically, we divide the unbounded domain of interest Ω into a bounded interior domain Ω_{int} and an unbounded exterior domain $\Omega_{\text{ext}} := \Omega \backslash \Omega_{\text{int}}$. In the one-dimensional case, Ω_{ext} typically consists of two components. We will choose the division such that the refractive index $n(x)$ is constant in each component of Ω_{ext}. In the one-dimensional case we can think of Ω_{int} as a finite interval, $\Omega_{\text{int}} = [x_l, x_r]$, and of Ω_{ext} as the domain

$$\Omega_{\text{ext}} := \mathbb{R} \backslash [x_l, x_r] = \{x \in \mathbb{R} : x < x_l\} \cup \{x \in \mathbb{R} : x > x_r\}.$$

First we do a change of variables and use distance variables ξ_l and ξ_r instead of the variable x in both parts of the exterior domain. We refer to the solution in the right hand component of the exterior domain as $u_{\text{ext},r}(\xi_r)$ with $\xi_r = x - x_r > 0$ for $x > x_r$ and in the left hand component of the exterior domain as $u_{\text{ext},l}(\xi_l)$ with $\xi_l = x_l - x > 0$ for $x < x_l$. Let the solution inside the computational domain be $u_{\text{int}}(x)$ and without restriction of generality let $x_l < 0$ and $x_r > 0$. Since we chose x_l and x_r, the boundary points of Ω_{int}, so that $n(x)$ is constant in each of the components of the exterior domain, $n(\xi_l) \equiv n_l$ for $\xi_l > 0$ and $n(\xi_r) \equiv n_r$ for $\xi_r > 0$, the solutions in these components are superpositions of complex exponential functions

$$u_{\text{ext},l}(\xi_l) = a_l \exp(-i n_l \omega \xi_l) + b_l \exp(i n_l \omega \xi_l) \text{ and} \tag{3.2}$$

$$u_{\text{ext},r}(\xi_r) = a_r \exp(i n_r \omega \xi_r) + b_r \exp(-i n_r \omega \xi_r). \tag{3.3}$$

Of the two exponential functions, $\exp(-i\xi_{l,r}n_{l,r})$ is a wave traveling in negative $\xi_{l,r}$-direction while $\exp(i\xi_{l,r}n_{l,r})$ is a wave traveling in positive $\xi_{l,r}$-direction. Since we defined ξ_l and ξ_r to be distance variables, the first of the summands in each of the given superpositions is a wave traveling to the right and the second is a wave traveling to the left.

In order to derive a physical solution of our problem, we have to ensure, that there are no sources outside Ω_{int}. That is we need to ensure, that our solution in Ω_{ext} consists only of waves that are leaving Ω_{int}. We refer to such solutions as *outgoing* solutions and define them as follows. A solution is called outgoing if it satisfies the *Sommerfeld radiation condition*. The formal definition of the Sommerfeld radiation condition is given in Definition 3.1.

Definition 3.1.
For $n(x) \equiv n$ constant, a solution to $(\Delta + n^2\omega^2)u(x) = 0, x \in \mathbb{R}^d$ is called *radiating* if it satisfies the *Sommerfeld radiation condition*

$$\lim_{|\xi|\to\infty} |\xi|^{\frac{d-1}{2}} \left(\frac{\partial}{\partial|\xi|} - in^2\omega^2\right) u(\xi) = 0$$

uniformly for all directions $\frac{\xi}{|\xi|}$.

As a direct consequence of the Sommerfeld radiation condition, we can determine the coefficients a_l and b_r in Equations (3.2) and (3.3) to be zero. In order to determine the missing coefficients a_r and b_l, we have to make the further assumption that the solution u is continuous at x_l and x_r, the boundary points of Ω_{int}. This assumption is reasonable if $n(x) = n_l$ in a neighborhood of x_l and $n(x) = n_r$ in a neighborhood of x_r. This assumption is no restriction to generality since we are free to move the artificial boundary points x_l and x_r to a suitable position. Taking the continuity into account, we can derive the missing coefficients a_r and b_l:

$$a_l = 0, \qquad b_l = u_{\text{int}}(x_l) \quad \text{and}$$
$$a_r = u_{\text{int}}(x_r), \qquad b_r = 0.$$

We have now explicitly derived $u_{\text{ext},l}(\xi)$ and $u_{\text{ext},r}(\xi)$ for the special one-dimensional case. Next we wish to investigate them in order to find a property that ensures that a solution satisfies the Sommerfeld radiation condition. For this we will use the Laplace transform, a widely used integral transformation that maps a function $f(t)$ with a real argument t onto a function $\mathcal{L}\{f(t)\}(s)$ with a complex argument.

Definition 3.2.
The Laplace transform $\mathcal{L}\{f(t)\}(s)$ of a function $f(t) : \mathbb{R}^+ \to \mathbb{C}$ is

$$\mathcal{L}\{f(t)\}(s) := \int_0^\infty e^{-st} f(t) dt.$$

We now form $\mathcal{L}\{u_{\text{ext},l}(\xi)\}$ and $\mathcal{L}\{u_{\text{ext},r}(\xi)\}$, the Laplace transforms of $u_{\text{ext},l}(\xi)$ and $u_{\text{ext},r}(\xi)$. Before inserting the coefficients a_l, b_l, a_r and b_r that we derived before, they read

$$\begin{aligned}\mathcal{L}\{u_{\text{ext},l}(\xi)\}(s) &= \int_0^\infty e^{-s\xi}\left(a_l e^{-in_l\omega\xi} + b_l e^{in_l\omega\xi}\right) d\xi \\ &= \frac{a_l}{s+in_l\omega} + \frac{b_l}{s-in_l\omega} \quad \text{and} \\ \mathcal{L}\{u_{\text{ext},r}(\xi)\}(s) &= \int_0^\infty e^{-s\xi}\left(a_r e^{in_r\omega\xi} + b_r e^{-in_r\omega\xi}\right) d\xi \\ &= \frac{a_r}{s-in_r\omega} + \frac{b_r}{s+in_r\omega}.\end{aligned}$$

Both Laplace transforms have holomorphic extensions to $\mathbb{C}$ except for two poles at $s = \pm in_l\omega$ and $s = \pm in_r\omega$ respectively. Now we insert the coefficients $a_{l,r}, b_{l,r}$ derived for our specific case. Then $\mathcal{L}\{u_{\text{ext},l}\}(s)$ and $\mathcal{L}\{u_{\text{ext},r}\}(s)$ simplify to

$$\mathcal{L}\{u_{\text{ext},l}(\xi)\}(s) = \frac{u_{\text{int}}(x_l)}{s-in_l\omega} \quad \text{and} \tag{3.4a}$$

$$\mathcal{L}\{u_{\text{ext},r}(\xi)\}(s) = \frac{u_{\text{int}}(x_r)}{s-in_r\omega}. \tag{3.4b}$$

Now we can see, that $u_{\text{ext},l}$ and $u_{\text{ext},r}$ satisfy the Sommerfeld radiation condition if $a_l = b_r = 0$, which is equivalent to the absence of the poles $s = -in_l\omega$ of $\mathcal{L}\{u_{\text{ext},l}(\xi)\}(s)$ and $s = -in_r\omega$ of $\mathcal{L}\{u_{\text{ext},r}(\xi)\}(s)$. Hence we have derived the equivalence that both exterior solutions are outgoing if the holomorphic extensions of their Laplace transforms to $\mathbb{C}$ have no poles with negative imaginary part for our special problem setting. The case of poles with zero imaginary part are special cases that correspond to solutions that are non-oscillatory and are just exponentially increasing or decaying in ξ-direction. These solutions are neither incoming nor outgoing.

We have thus split $\mathbb{C}$ in two parts $\mathbb{C}_{in} := \{z \in \mathbb{C} : \Im(z) \leq 0\}$ and $\mathbb{C}_{out} := \{z \in \mathbb{C} : \Im(z) > 0\}$ for each subset of Ω_{ext} and reformulated the fact that a solution is outward radiating to the non-existence of poles of the holomorphic extension of the Laplace transform in the subset $\mathbb{C}_{in}$. We will formalize this approach in the following section.

3.2 Formalization

We will now derive a formal formulation of the condition that was intuitively derived in the previous section and bring it to the formal context required for an implementation. We will start off with a variational formulation

of the one-dimensional Helmholtz equation on an unbounded domain. By dividing the domain into a bounded interior and an unbounded exterior and restricting the test functions to a suitable set, we will obtain a formulation that contains the Laplace transforms of the solutions in the exterior parts.

We will then give a more general derivation of the condition on the poles of the Laplace transform in terms of Cauchy's integral formula and a representation of the resulting path integral by Riemann sums. This will yield our first formal definition of the pole condition. Further we will verify that $\mathbb{C}_{in}$ and $\mathbb{C}_{out}$ from the previous section are valid for our equation in this more general framework.

Our starting point is the resonance mode setting of the one-dimensional Helmholtz equation (3.1) on a possibly unbounded domain $\Omega \subseteq \mathbb{R}$. Multiplying with a test function $v \in H^1_{\text{loc}}(\Omega)$, the space of functions that restricted to a compact subset $\Omega_F \subset \Omega$ are in $H^1(\Omega_F)$, we obtain a variational formulation: Find $u \in H^1_{\text{loc}}(\Omega)$, such that

$$\int_\Omega \partial_{xx} u(x)v(x) + n(x)^2\omega^2 u(x)v(x)dx = 0 \tag{3.5}$$

for all $v \in H^1_{\text{loc}}(\Omega)$. Splitting Ω into Ω_{int} and Ω_{ext} yields a splitting of the integral and we obtain

$$\begin{aligned} &\int_{\Omega_{\text{int}}} \partial_{xx} u(x)v(x) + n(x)^2\omega^2 u(x)v(x)dx + \\ &\int_{\Omega_{\text{ext}}} \partial_{xx} u(x)v(x) + n_{l,r}^2\omega^2 u(x)v(x)dx = 0 \qquad \forall v \in H^1_{\text{loc}}(\Omega). \end{aligned} \tag{3.6}$$

Integrating the exterior part of Equation (3.6) by parts, we have

$$\begin{aligned} &\int_{\Omega_{\text{int}}} \partial_{xx} u(x)v(x) + n(x)^2\omega^2 u(x)v(x)dx + \\ &\int_{\Omega_{\text{ext}}} -\partial_x u(x)\partial_x v(x) + n_{l,r}^2\omega^2 u(x)v(x)dx \pm u'(x_{l,r})v(x_{l,r}). \end{aligned} \tag{3.7}$$

Next we insert the special forms $\Omega_{\text{int}} = [x_l, x_r]$ and $\Omega_{\text{ext}} = (-\infty, x_l) \cup (x_r, \infty)$ in the one-dimensional case into Equation (3.7) and have

$$\begin{aligned} &\int_{x_l}^{x_r} -\partial_x u(x)\partial_x v(x) + n(x)^2\omega^2 u(x)v(x)dx + \\ &\quad u'(x_r)v(x_r) - u'(x_l)v(x_l) + \\ &\int_{x<x_l} -\partial_x u(x)\partial_x v(x) + n_l^2\omega^2 u(x)v(x)dx + u'(x_l)v(x_l) + \\ &\int_{x>x_r} -\partial_x u(x)\partial_x v(x) + n_r^2\omega^2 u(x)v(x)dx - u'(x_r)v(x_r) = 0 \\ &\quad \forall v \in H^1_{\text{loc}}(\Omega). \end{aligned} \tag{3.8}$$

Since it is sufficient to test against all functions v in a dense subset of $H^1_{\text{loc}}(\Omega)$, we will restrict the set of test functions to the exponentially decaying functions in $H^1_{\text{loc}}(\Omega)$:

$$v_l \in H^1_{\text{loc}}(\Omega): \quad v_l(x) = ce^{-s(x_l - x)} \quad \text{for } x < x_l, \Re(s) > 0, c \in \mathbb{C}, \tag{3.9}$$

$$v_r \in H^1_{\text{loc}}(\Omega): \quad v_r(x) = ce^{-s(x - x_r)} \quad \text{for } x > x_r, \Re(s) > 0, c \in \mathbb{C}. \tag{3.10}$$

We can further restrict ourselves to using only functions with $c = 1$. Now we insert the test functions from the set of functions defined in Equations (3.9) and (3.10) into Equation (3.8). After a coordinate transform, the two infinite integrals yield the Laplace transforms of $u_{\text{ext},l,r}$ in the two disjoint subsets of Ω_{ext}:

$$\begin{aligned}
0 = &\int_{x_l}^{x_r} -\partial_x u(x)\partial_x v(x) + n(x)^2\omega^2 u(x)v(x)dx + u'(x_r) - u'(x_l) \\
&+ \int_{x<x_l} -\partial_x u(x)\partial_x e^{s(x-x_l)} + n_l^2\omega^2 u(x)e^{s(x-x_l)}dx + u'(x_l) \\
&+ \int_{x>x_r} -\partial_x u(x)\partial_x e^{-s(x-x_r)} + n_r^2\omega^2 u(x)e^{-s(x-x_r)}dx - u'(x_r) \\
\Leftrightarrow 0 = &\int_{x_l}^{x_r} -\partial_x u(x)\partial_x v(x) + n(x)^2\omega^2 u(x)v(x)dx + u'(x_r) - u'(x_l) \\
&+ \int_{x>0} \partial_x u(-x + x_l)se^{-sx} + n_l^2\omega^2 u(-x + x_l)e^{-sx}dx + u'(x_l) \\
&+ \int_{x>0} \partial_x u(x + x_r)se^{-sx} + n_r^2\omega^2 u(x + x_r)e^{-sx}dx + u'(x_r).
\end{aligned}$$

The last two integrals correspond to the Laplace transform of the Helmholtz equation in the left and right exterior domains so we can rewrite the above equations in terms of $\mathcal{L}\{\partial_x u_{\text{ext},l}\}, \mathcal{L}\{\partial_x u_{\text{ext},r}\}, \mathcal{L}\{u_{\text{ext},l}\}$ and $\mathcal{L}\{u_{\text{ext},r}\}$:

$$0 = \int_{x_l}^{x_r} -\partial_x u(x)\partial_x v(x) + n(x)^2\omega^2 u(x)v(x)dx \tag{3.11a}$$

$$+ s\mathcal{L}\{\partial_x u_{\text{ext},l}\}(s) + n_l^2\omega^2\mathcal{L}\{u_{\text{ext},l}\}(s) + u'(x_l) \tag{3.11b}$$

$$+ s\mathcal{L}\{\partial_x u_{\text{ext},r}\}(s) + n_r^2\omega^2\mathcal{L}\{u_{\text{ext},r}\}(s) + u'(x_r). \tag{3.11c}$$

A central result of the Laplace transform is the Laplace transform of $\partial_x u(x)$. It can be obtained by integration by parts and states that

$$\mathcal{L}\{\partial_x u\}(s) = s\mathcal{L}\{u\}(s) - u(0). \tag{3.12}$$

We can now rewrite Equations (3.11a)-(3.11c) using (3.12):

$$0 = \int_{x_l}^{x_r} -\partial_x u(x)\partial_x v(x) + n(x)^2\omega^2 u(x)v(x)dx \tag{3.13a}$$

$$+ s^2\mathcal{L}\{u_{\text{ext},l}\}(s) - su_{\text{ext},l}(x_l) + n_l^2\omega^2\mathcal{L}\{u_{\text{ext},l}\}(s) + u'(x_l) \tag{3.13b}$$

$$+ s^2\mathcal{L}\{u_{\text{ext},r}\}(s) - su_{\text{ext},r}(x_r) + n_r^2\omega^2\mathcal{L}\{u_{\text{ext},r}\}(s) + u'(x_r). \tag{3.13c}$$

While equation (3.13a) is the variational formulation for the solution in Ω_{int}, Equations (3.13b) and (3.13c) are the Laplace transforms of the original Helmholtz equation in the two sub-domains of Ω_{ext}.

We will now construct transparent boundary conditions by imposing conditions on these Laplace transforms $\mathcal{L}\{u_{\text{ext},l,r}\}(s)$. For an arbitrary function $u(x)$, its Laplace transform $\mathcal{L}\{u\}(s)$ as a function of s has some singularities in the complex plane. By Cauchy's integral formula

$$\mathcal{L}\{u\}(s) = \frac{1}{2\pi i} \oint_{\gamma} \frac{\mathcal{L}\{u\}(\tau)}{\tau - s} d\tau, \tag{3.14}$$

where γ is a path enclosing the singularities of $\mathcal{L}\{u\}$. Now we parametrize γ with a bijective function $r : [a, b] \to \gamma$. We obtain a partitioning of γ by partitioning the interval $[a, b]$ into n sub-intervals $[\tau_{i-1}, \tau_i]$, each of the length $\Delta\tau = (b - a)/n$. We can use the set $\{r(\tau_i), i = 1, ..., n\}$ of points on γ to approximate γ with a polygonal path by introducing straight lines between $r(\tau_{i-1})$ and $r(\tau_i)$. We denote the distance between the sample points on the curve by Δs_i. Now we can insert the Riemann sum for the path integral in (3.14):

$$\begin{aligned} \mathcal{L}\{u\}(s) &= \frac{1}{2\pi i} \lim_{\Delta s_i \to 0} \sum_{j=1}^{N} \mathcal{L}\{u\}(r(\tau_j)) \Delta s_i \frac{1}{\tau_j - s} \\ &= \lim_{N \to \infty} \sum_{j=1}^{N} \alpha_j(N, \mathcal{L}\{u\}(\tau_j)) \frac{1}{\tau_j - s}. \end{aligned} \tag{3.15}$$

In Equation (3.15), $\alpha_j(N, \mathcal{L}\{u\}(\tau_j))$ can be seen as weights and the entire sum may thus be reinterpreted as superposition of $(\tau - s)^{-1}$. Transforming these summands back into the space domain, we get the correspondence

$$\frac{1}{\tau - s} \leftrightarrow -e^{\tau x}.$$

Depending on the location of τ in the complex plane $\mathbb{C}$, $-\exp(\tau x)$ is moving to the left/exponentially increasing or moving to the right/exponentially decreasing. So for each disjoint subset of Ω_{ext}, the complex plane can be divided into the two regions

$$\begin{aligned} \mathbb{C}_{in} &:= \{\tau \in \mathbb{C} : -\exp(\tau x) \text{ is incoming or not oscillating}\} \text{ and} \\ \mathbb{C}_{out} &:= \{\tau \in \mathbb{C} : -\exp(\tau x) \text{ is outward radiating}\}. \end{aligned}$$

This enables us to deform the path γ from (3.14) and then split it into two paths $\gamma_{in} \subset \mathbb{C}_{in}$ and $\gamma_{out} \subset \mathbb{C}_{out}$ that each enclose all the singularities of $\mathcal{L}\{u\}(s)$ in the respective region. Equation (3.14) then decomposes as follows:

$$\mathcal{L}\{u\}(s) = \oint_{\gamma_{in}} \frac{\mathcal{L}\{u\}(\tau)}{\tau - s} d\tau + \oint_{\gamma_{out}} \frac{\mathcal{L}\{u\}(\tau)}{\tau - s} d\tau. \tag{3.16}$$

Requiring that u is outward radiating then implies requiring that $\oint_{\gamma_{in}} \frac{\mathcal{L}\{u\}(\tau)}{\tau - s} d\tau$ is zero. This corresponds to the condition that $\mathcal{L}\{u\}(s)$ is analytic in $\mathbb{C}_{in}$. The splitting into $\mathbb{C}_{in}$ and $\mathbb{C}_{out}$ is possible whenever it is possible to distinguish between incoming and outgoing solutions, that is when the problem is not degenerated. We are now in the position to summarize the method in the following definition:

Definition 3.3.

A function $u \in H^1(\Omega)$ is said to obey the *pole condition* if the complex continuation of its Laplace transform $\mathcal{L}\{u\}(s)$ is analytic in $\mathbb{C}_{in}$.

Now we will apply this to the one-dimensional Helmholtz resonance problem. Suppose that u_{int} is given inside $\Omega_{\text{int}} = [x_l, x_r]$. Since the following considerations are the same for the left and right exterior domains, we will take into account only the right hand component of the exterior domain. Then the equation for $\mathcal{L}\{u_{\text{ext},r}\}(s)$ derived from Equation (3.13c) is then

$$\begin{aligned} 0 &= s^2 \mathcal{L}\{u_{\text{ext},r}\}(s) - s u_{\text{ext},r}(x_r) + n_r^2 \omega^2 \mathcal{L}\{u_{\text{ext},r}\}(s) + u'(x_r) \\ \Leftrightarrow \mathcal{L}\{u_{\text{ext},r}\} &= (s^2 + n_r^2 \omega^2)^{-1} (s u_{\text{ext},r} + u'(x_r)). \end{aligned} \tag{3.17}$$

For fixed ω, $(s^2 + n_r^2 \omega^2)$ has two roots

$$s_{+/-} = \pm i \sqrt{n_r^2 \omega^2}. \tag{3.18}$$

We can use the roots s_+ and s_- defined in Equation (3.18) to obtain a partial fraction decomposition of $\mathcal{L}\{u_{\text{ext},r}\}$ from (3.17):

$$\begin{aligned} \mathcal{L}\{u_{\text{ext},r}\}(s) =& (s + s_+)^{-1} \frac{1}{2} \left(s u_{\text{ext},r}(x_r) + u'(x_r) \right) + \\ & (s + s_-)^{-1} \frac{1}{2} \left(s u_{\text{ext},r}(x_r) + u'(x_r) \right). \end{aligned} \tag{3.19}$$

Transforming the summands in Equation (3.19) back to the space domain, they correspond to

$$\frac{1}{2} e^{-s_+ x} (u'(x_r) - i n_r \omega u_{\text{ext},r}(x_r)) \text{ and } \frac{1}{2} e^{-s_- x} (u'(x_r) - i n_r \omega u_{\text{ext},r}(x_r)).$$

Depending on the location of $s_{+/-}$ in the complex plane, they are incoming, exponentially increasing, outgoing or exponentially decreasing. For the left boundary we can obtain a similar splitting by the same arguments. The frequency ω takes values everywhere in the complex plane: $\omega \in \mathbb{C}$. Thus s_+ and s_-, the roots of $(s^2 + n_{l,r}^2 \omega^2)$, split the complex plane into two parts, $\mathbb{C}_{in} = \{s \in \mathbb{C} : \Im(s) \leq 0\}$ and $\mathbb{C}_{out} = \{s \in \mathbb{C} : \Im(s) > 0\}$. The imaginary axis $\Re(s) = 0$ is mapped to the real axis $\Im(s) = 0$.

3.3 Implementation

We will now develop an implementation of the pole condition. Our implementation follows the lines of [RSSZ08] and [HN09]. In order to reformulate the boundary condition in a way that will allow for an easy implementation within the context of a finite element formulation of the problem, we will first map the domain $\mathbb{C}_{in}$ where $\mathcal{L}\{u\}(s)$ is required to be analytic to the unit disc. The mapping that performs this coordinate change is a Möbius transform that maps a half-space in the complex plane to the unit disc D. On $\mathbb{C}_{in}$ and D we will use certain function spaces between which the Möbius transform also forms a connection. On D we can then reformulate the condition that $\mathcal{L}\{u_{\text{ext}}\}$ is analytic in $\mathbb{C}_{in}$ to the condition that its Möbius transform lies in a certain function space on D, the Hardy space $H^+(D)$. The functions of this space can then be expanded into a power series. In order to obtain a discrete formulation, this power series is approximated by a polynom. In order to obtain a formulation that fits well in the finite element context, we will use well-chosen test functions that when testing against them directly give the Laplace transform of u_{ext} and allow for an easy coupling to the solution in the interior.

So in short the outline of what follows next is:

1. Define a mapping $P_{s_0} := \mathbb{C}_{in} \to D$.
2. From this obtain a mapping of the function spaces $H^-(P_{s_0}) \to H^+(D)$.
3. Reformulate the pole condition in terms of these function spaces.
4. Approximate a function in $H^+(D)$ with a power series to obtain a discrete formulation.
5. Choose a test-function for the exterior domain such that the previous formulation can be embedded within a finite element context.
6. Derive the local element matrices this yields for each component of the exterior domain.

For the first of these steps, we will now give a formal definition of the Möbius transform. In its general formulation it is given as:

$$\mathcal{M}(s) = \widetilde{s} := \frac{as + b}{cs + d}.$$

For mapping the half space $P_{s_0} := \{z \in \mathbb{C} : \Re(z/s_0) \geq 0\}$ below the line connecting 0 and is_0 to the unit disc, we set $a = c = 1$, $b = s_0$ and $c = -s_0$:

$$\mathcal{M}_{s_0}(s) = \widetilde{s} := \frac{s + s_0}{s - s_0}. \tag{3.20}$$

The complex parameter s_0 determines the position of the half-space and acts as a tuning parameter. It will be used to identify the spurious modes of the resonance problem at a later point. The Möbius transform with the parameter s_0 will be noted with a subscript s_0: $\mathcal{M}_{s_0}$, see Fig. 3.1.

Since we require $\mathcal{L}\{u\}(s)$ to be analytic in $\mathbb{C}_{in}$, we can use the property that an analytic function can be expanded into a power series that converges inside some ball to obtain a formulation of the pole condition that can be implemented. $\mathcal{M}_{s_0}$ maps the infinite point to 1 and s_0 to zero, thus an approximation of $\mathcal{L}\{u_{\text{ext}}\} \circ \mathcal{M}_{s_0}$ by an power series expansion will be best near s_0. So choosing s_0 in the region where one expects the resonances of interest to be located is typically a good choice.

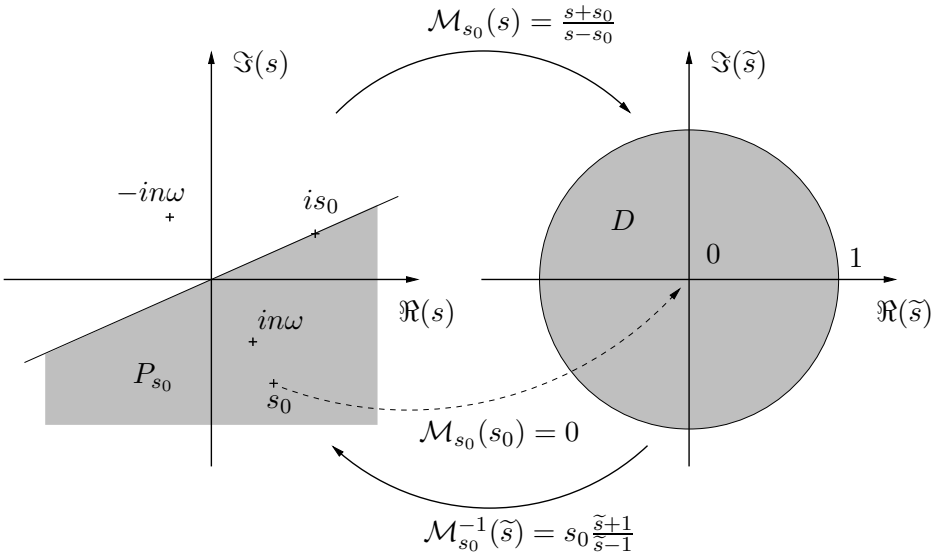

Figure 3.1: The Möbius transform $\mathcal{M}_{s_0}$ and its inverse.

The inverse of the transform (3.20) for our choice of a, b, c and d is

$$\mathcal{M}_{s_0}^{-1} : \widetilde{s} \to s = s_0 \frac{\widetilde{s}+1}{\widetilde{s}-1}. \tag{3.21}$$

In order to be able to give a formal definition of the pole condition in the setting that is fit for implementation, we will have to give some definitions of the function spaces on P_{s_0} and D.

Definition 3.4.

As before let $P_{s_0} := \{z \in \mathbb{C} : \Re(z/s_0) \geq 0\}$ be the half space below the line connecting 0 and is_0. The *Hardy Space* $H^-(P_{s_0})$ is the space of all functions u that are holomorphic in P_{s_0} such that

$$\int_{\mathbb{R}} |u(is_0 x - \epsilon)|^2 dx$$

is uniformly bounded for $\epsilon > 0$.
Let $D = \{z \in \mathbb{C} : |z| < 1\}$ be the open unit disc in $\mathbb{C}$. The *Hardy Space* $H^+(D)$ is the space of all functions u that are holomorphic in D such that

$$\int_0^{2\pi} |u(re^{it})|^2 dt$$

is uniformly bounded for $r \in [0, 1)$.

According to [Har15, Dur70, Hof62], for $\Re(s_0) = 0$, i.e. for P_{s_0} being the half-space below the real axis, functions in $H^-(P_{s_0})$ can be regarded as L^2 boundary functions of functions that are holomorphic in P_{s_0}. Furthermore, in [Nan08] Nannen showed that for $s_0 \in \mathbb{C}, \Re(s_0) > 0$ the Möbius transform is an isomorphism connecting the function space $H^-(P_{s_0})$ and $H^+(D)$.

These results allow us to identify functions in the Hardy Spaces $H^-(P_{s_0})$ and $H^+(D)$ with their boundary functions in $L^2(P_{s_0})$ or $L^2(S^1)$ respectively. The connection that the Möbius transform forms between the function spaces $H^-(P_{s_0})$ and $H^+(D)$ can be phrased as follows:

$$f \in H^-(P_{s_0}) \to H^+(D) \ni (\mathcal{M}_{s_0} f)(\widetilde{s}) := f\left(\mathcal{M}_{s_0}^{-1}(\widetilde{s})\right) \frac{1}{\widetilde{s} - 1}. \tag{3.22}$$

We can now reformulate the pole condition from Definition 3.3 in terms of the function spaces from Definition 3.4:

Definition 3.5.

Let $s_0 \in \mathbb{C}$ with $\Re(s_0) > 0$. Then a solution to (3.1) is said to obey the *pole condition* and is called outgoing if $\mathcal{M}_{s_0}\mathcal{L}\{u_{\text{ext}}\}(\widetilde{s})$, the Möbius transform $\mathcal{M}_{s_0}$ of the holomorphic extension of the Laplace transform of the exterior part, lies in $H^+(D)$.

We can use Equation (3.22) to define $\mathcal{L}_D$, the Laplace transform on the unit disc D. Since this can be done in the same way for both $u_{\text{ext},l}$ and $u_{\text{ext},r}$ we will give the formulation only for the right exterior domain.

$$\begin{aligned} \mathcal{L}_D\{u_{\text{ext},r}\}(\widetilde{s}) &:= (\mathcal{M}_{s_0}\mathcal{L}\{u_{\text{ext},r}\})(\widetilde{s}) \\ &= \mathcal{L}\{u_{\text{ext},r}\}\left(\mathcal{M}_{s_0}^{-1}(\widetilde{s})\right) \frac{1}{\widetilde{s} - 1}. \end{aligned} \tag{3.23}$$

Since we require the Laplace transform to be analytic in the unit disc, it can be expanded into a power series, $\mathcal{L}_D\{u_{\text{ext}}\}(\widetilde{s}) = \sum_{k=0}^{\infty} a_k \widetilde{s}^k$. Hence

$$\mathcal{L}\{u_{\text{ext},r}\}\left(\mathcal{M}_{s_0}^{-1}(\widetilde{s})\right) = (\widetilde{s} - 1) \sum_{k=0}^{\infty} a_k \widetilde{s}^k. \tag{3.24}$$

The coefficients a_k in Equation (3.24) contain the dependence on the parameter s_0.

We will now describe the implementation of a transparent boundary condition based on the pole condition within a finite element context. We will describe the situation for the one-dimensional problem in some detail. The discretization for higher space dimensions is typically done via Cartesian products and will be detailed in the next sections.

As described in the derivation of Equations (3.13a)-(3.13c), we will use ansatz functions that directly yield the Laplace transform in the exterior domain Ω_{ext}. These ansatz functions are called "boundary exp-elements" and consist of the standard interior element coupled with a complex exponential function that will result in a formulation of the Laplace transform in the exterior. Such an element is sketched in Fig. 3.2 for a linear discretization in the interior.

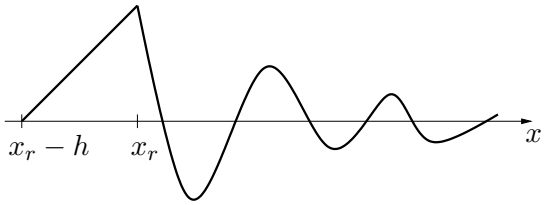

Figure 3.2: Exp-element for the right hand side boundary with first order discretization in the interior.

The boundary exp-element test function for polynomial degree $p = 1$ in the interior at the right artificial boundary x_r is given by

$$\psi_s^{(r)}(x) = \begin{cases} e^{-s(x-x_r)} : x \geq x_r, \\ \frac{x-(x_r-h)}{h} : x_r - h \leq x \leq x_r. \end{cases} \tag{3.25}$$

The first line in Equations (3.25) is the exponential part, that for fixed s gives the Laplace transform of $u_{\text{ext},r}$. The second line is the interior part, h is the mesh width in the interior.

The exp-element $\psi_s^{(r)}(x)$ is not one function but a family of functions parametrized by $s \in \mathbb{C}^+$. They are globally continuous by definition and their support is infinite: $\operatorname{supp}\left(\psi_x^{(r)}\right) = [x_r - h, \infty)$. By using $\psi_s(x)$ as test function in the variational formulation (3.5), we obtain

$$0 = \int_{\mathbb{R}} \partial_{xx} u(x)\psi_s(x) + \omega^2 n(x)^2 u(x)\psi_s(x) dx$$

Due to the definition of $\psi_s(x)$ and assuming that $n(x) \equiv n_i$ for $x \in [x_r - h, x_r]$ and $n(x) \equiv n_r$ for $x > x_r$, after integration by parts and neglecting zero

boundary terms we obtain:

$$\begin{aligned}
0 &= \int_{x_r-h}^{x_r} -\partial_x u_{\text{int}}(x)\partial_x\psi_s(x) + \omega^2 n_i^2 u_{\text{int}}(x)\psi_s(x)dx \\
&\quad + \int_{x_r}^{\infty} -\partial_x u_{\text{ext},r}(x)\partial_x\psi_s(x) + \omega^2 n_r^2 u_{\text{ext},r}(x)\psi_s(x)dx \\
\Leftrightarrow 0 &= -\int_{x_r-h}^{x_r} \partial_x u_{\text{int}}(x)\partial_x\psi_s(x) + \omega^2 n_i^2 u_{\text{int}}(x)\psi_s(x)dx \\
&\quad + s\mathcal{L}\{\partial_x u_{\text{ext},r}\}(s) + \omega^2 n_r^2 \mathcal{L}\{u_{\text{ext},r}\}(s) \\
\Leftrightarrow 0 &= -\int_{x_r-h}^{x_r} \partial_x u_{\text{int}}(x)\frac{1}{h}dx + \omega^2 n_i^2 \int_{x_r-h}^{x_r} u_{\text{int}}(x)\frac{x-(x_r-h)}{h}dx \\
&\quad + s\left(s\mathcal{L}\{u_{\text{ext},r}\}(s) - u_{\text{ext},r}(x_r)\right) + \omega^2 n_r^2 \mathcal{L}\{u_{\text{ext},r}\}(s). \qquad (3.26)
\end{aligned}$$

The boundary Neumann terms occurring due to the integration by parts are here given in weak form in the first integrals. We have now obtained $\mathcal{L}\{u_{\text{ext},r}\}$ but this does not yield a discrete formulation since we lack a convenient orthonormal basis for $H^-(P_{s_0})$. To remedy this deficit, we will make use of the continuity at x_r and then transform Equation (3.26) to $H^+(D)$ by applying (3.23) and inserting $\widetilde{s}$ as defined in Equation (3.20):

$$\begin{aligned}
0 &= -\int_{x_r-h}^{x_r} \partial_x u_{\text{int}}(x)\frac{1}{h}dx + \omega^2 n_i^2 \int_{x_r-h}^{x_r} u_{\text{int}}(x)\frac{x-(x_r-h)}{h}dx \\
&\quad + s_0\frac{\widetilde{s}+1}{\widetilde{s}-1}\left(s_0\frac{\widetilde{s}+1}{\widetilde{s}-1}(\widetilde{s}-1)\mathcal{L}_D\{u_{\text{ext},r}\}(\widetilde{s}) - u_{\text{ext},r}(x_r)\right) \\
&\quad + \omega^2 n_r^2(\widetilde{s}-1)\mathcal{L}_D\{u_{\text{ext},r}\}(\widetilde{s}). \qquad (3.27)
\end{aligned}$$

The next step is to insert a series expansion for $\mathcal{L}_D\{u_{\text{ext},r}\}(\widetilde{s})$. However, for ease of implementation we will not use the direct power series approximation from Equation (3.24) but reformulate it. To obtain an easy formulation for the coupling of the transformed exterior to the interior problem, we take a closer look at $\mathcal{L}\{u_{\text{ext},l}\}$ and $\mathcal{L}\{u_{\text{ext},r}\}$ by inserting their known form (3.4a) and (3.4b) respectively:

$$\begin{aligned}
\mathcal{L}\{u_{\text{ext},l,r}\}(s) &= \frac{u_{\text{int}}(x_{l,r})}{s - in_{l,r}\omega} \\
\xrightarrow{\mathcal{M}_{s_0}} \mathcal{L}_D\{u_{\text{ext},l,r}\}(\widetilde{s}) &= \frac{u_{\text{int}}(x_{l,r})}{s_0(\widetilde{s}+1) - in_{l,r}\omega(\widetilde{s}-1)}.
\end{aligned} \qquad (3.28)$$

If we would attempt to do a power series approximation of Equation (3.28), the boundary degree of freedom $u_{\text{int}}(x_{l,r})$ would couple with each degree of freedom in the exterior. In order to obtain a local coupling, we note that $\mathcal{L}_D\{u_{\text{ext},l,r}\}(1) = u_{\text{int}}(x_{l,r})/2s_0$. To take advantage of this fact, we now decompose

$$\mathcal{L}_D\{u_{\text{ext},l,r}\}(\widetilde{s}) = \frac{1}{2s_0}\left(u_{\text{int}}(x_{l,r}) + (\widetilde{s}-1)\frac{2s_0\mathcal{L}_D\{u_{\text{ext}}\}(\widetilde{s}) - u_{\text{int}}(x_{l,r})}{\widetilde{s}-1}\right).$$

(3.29)

Inserting the series representation (3.24) and rescaling its coefficients

$$\widetilde{a}_k = \begin{cases} 2a_0 - u_{\text{int}}(x_{l,r}) : k = 0 \\ 2s_0 a_k : k \geq 1 \end{cases}$$

we have

$$\mathcal{L}_D\{u_{\text{ext},l,r}\}(\widetilde{s}) = \frac{u_{\text{int}}(x_{l,r})}{2s_0} + (\widetilde{s}-1)\frac{1}{2s_0}\sum_{k=0}^{\infty}\widetilde{a}_k\widetilde{s}^k. \tag{3.30}$$

Inserting (3.30) into Equation (3.27), we get for the right hand side boundary

$$0 = -\int_{x_r-h}^{x_r}\partial_x u_{\text{int}}(x)\frac{1}{h}dx + \int_{x_r-h}^{x_r}u_{\text{int}}(x)\frac{x-(x_r-h)}{h}dx \tag{3.31a}$$

$$+ u_{\text{int}}(x_r)\frac{s_0}{2}(\widetilde{s}+1) + \frac{s_0}{2}(\widetilde{s}+1)^2\sum_{k=0}^{\infty}\widetilde{a}_k\widetilde{s}^k \tag{3.31b}$$

$$+ \omega^2 n_r^2\left(u_{\text{int}}(x_r)\frac{1}{2s_0}(\widetilde{s}-1) + \frac{1}{2s_0}(\widetilde{s}-1)^2\sum_{k=0}^{\infty}\widetilde{a}_k\widetilde{s}^k\right). \tag{3.31c}$$

Equation (3.31a) is the weak form of the Neumann data at the right hand side boundary of Ω_{int}. Equation (3.31b) is the term the exp-element adds to the mass matrix for the right hand side boundary of Ω_{int} and Equation (3.31c) is the stiffness-term for the exp-element at the right hand side boundary of Ω_{int}. The interior and exterior solutions are coupled at the right hand side boundary via $u_{\text{int}}(x_r)$.

In order to obtain the local element matrix for the exp-element, we sort (3.31a)-(3.31c) by powers of $\widetilde{s}$ and compare coefficients:

$$\widetilde{s}^0 :\ 0 = -\int_{x_r-h}^{x_r}\partial_x u_{\text{int}}(x)\frac{1}{h}dx + \omega^2 n_i^2\int_{x_r-h}^{x_r}u_{\text{int}}(x)\frac{x-(x_r-h)}{h}dx$$

$$+ \frac{s_0}{2}\left(u_{\text{int}}(x_r) + \widetilde{a}_0\right) - \omega^2 n_r^2\frac{1}{2s_0}\left(u_{\text{int}}(x_r) - \widetilde{a}_0\right) \tag{3.32a}$$

$$\widetilde{s}^1 :\ 0 = \frac{s_0}{2}\left(u_{\text{int}}(x_r) + 2\widetilde{a}_0 + \widetilde{a}_1\right) - \omega^2 n_r^2\frac{1}{2s_0}\left(-u_{\text{int}}(x_r) + 2\widetilde{a}_0 - \widetilde{a}_1\right) \tag{3.32b}$$

$$\widetilde{s}^k :\ 0 = \frac{s_0}{2}\left(\widetilde{a}_{k-2} + 2\widetilde{a}_{k-1} + \widetilde{a}_k\right) - \omega^2 n_r^2\frac{1}{2s_0}\left(-\widetilde{a}_{k-2} + 2\widetilde{a}_{k-1} - \widetilde{a}_k\right)$$

$$k \geq 2. \tag{3.32c}$$

Collecting these degrees of freedom, we get the following local element matrices for the exterior part of the infinite exp-element:

$$A_{\text{ext}}^{\text{loc}} = s_0\frac{1}{2}\begin{pmatrix} 1 & 1 & 0 & \cdots & & \\ 1 & 2 & 1 & 0 & \cdots & \\ 0 & 1 & 2 & 1 & 0 & \cdots \\ & \ddots & \ddots & \ddots & \ddots & \end{pmatrix} \tag{3.33}$$

and

$$B_{\text{ext}}^{\text{loc}} = n_{l,r}^2 \frac{1}{2s_0} \begin{pmatrix} 1 & -1 & 0 & \cdots & & \\ -1 & 2 & -1 & 0 & \cdots & \\ 0 & -1 & 2 & -1 & 0 & \cdots \\ & \ddots & \ddots & \ddots & \ddots & \end{pmatrix}. \tag{3.34}$$

The first degree of freedom in these local element matrices is $u_{\text{int}}(x_{l,r})$ that is common to the interior and the exterior solution and thus provides the coupling. The integral terms occurring for $\widetilde{s}^0$ are the weak formulation of the Neumann data and can be assembled together with the interior degrees of freedom. For implementation we truncate the series approximation of $\mathcal{L}_D\{u_{\text{ext}}\}(\widetilde{s})$ by setting $\widetilde{a}_k = 0$ for $k \geq L$, making $A_{\text{ext}}^{\text{loc}}$ and $B_{\text{ext}}^{\text{loc}}$ finite:

$$\mathcal{L}_D\{u_{\text{ext},l,r}\}(\widetilde{s}) \approx \frac{u_{\text{int}}(x_{l,r})}{2s_0} + (\widetilde{s}-1)\frac{1}{2s_0}\sum_{k=0}^{L} \widetilde{a}_k \widetilde{s}^k.$$

Discretization of Ω_{int} with normal finite elements yields a sparse eigenvalue problem with stiffness matrix A_{int}, mass matrix B_{int} and vector of unknowns $\mathbf{u}_{\text{int}}$:

$$\left(A_{\text{int}} - \omega^2 B_{\text{int}}\right)\mathbf{u}_{\text{int}} = 0. \tag{3.35}$$

Choosing linear ansatz and test functions for an element with length h_k and refractive index $n(x) \equiv n_k$ yields the well known local element matrices

$$A_{\text{int}}^{\text{loc}} = \frac{1}{h_k}\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \text{and} \quad B_{\text{int}}^{\text{loc}} = \frac{n_k^2 h_k}{6}\begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}.$$

For higher order ansatz functions $\Phi_i, i = 1, ..., p$ of degree $p-1$, the entries in the $p \times p$ local element matrices are

$$A_{\text{int},(i,j)}^{\text{loc}} = \int_I \partial_x \Phi_i(x) \partial_x \Phi_j(x) dx \quad \text{and} \quad B_{\text{int},(i,j)}^{\text{loc}} = \int_I n(x)^2 \Phi_j(x) \Phi_i(x) dx.$$

Combining the degrees of freedom $\widetilde{a}_k$ in the exterior with the degrees of freedom $\mathbf{u}_{\text{int}}$ in the interior and collecting them into one vector of unknowns $\mathbf{u}$, we arrive at a generalized sparse eigenvalue problem

$$(A - \omega^2 B)\mathbf{u} = 0 \tag{3.36}$$

where the matrices A and B have the structure as sketched in Fig. 3.3.

In order to insert the exterior degrees of freedom into the global matrices A and B, we need a mapping P that maps the local degrees of freedom of a component of the exterior domain to the global degrees of freedom. Let $\mathbf{u}_1, ...\mathbf{u}_n$ be the interior degrees of freedom and $N = n + 2L$ be the total

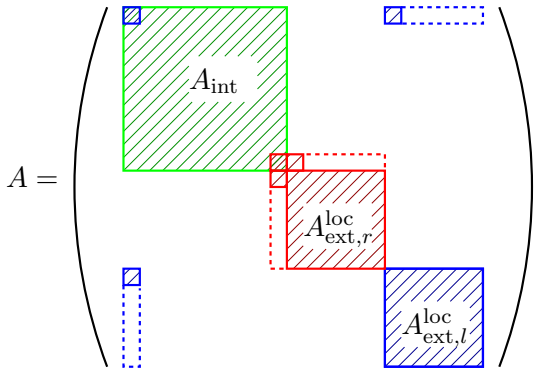

Figure 3.3: Structure for sparse matrix A for a one-dimensional Helmholtz resonance problem discretized with linear finite elements in the interior and the pole condition implemented with first order exp-elements on both boundaries. B has the same structure with coupling in only one interior degree of freedom.

number of degrees of freedom. Then P is a $L \times N$ matrix that has an entry $P_{j,k} = 1$ if the j-th local degree of freedom is mapped to the k-th global degree of freedom. The $N \times N$ matrices $P^\top A^{\text{loc}}_{\text{ext},l} P$, $P^\top A^{\text{loc}}_{\text{ext},r} P$, $P^\top B^{\text{loc}}_{\text{ext},l} P$ and $P^\top B^{\text{loc}}_{\text{ext},r} P$ are the contributions of the left and right exterior domain to the global system matrices A and B. If we call the mapping for the left hand exterior domain degrees of freedom P_l and the mapping for the right hand exterior domain degrees of freedom P_r, we can collect only the exterior degrees of freedom we get the exterior parts A_{ext} and B_{ext} of A and B by:

$$
\begin{aligned}
A_{\text{ext}} &= (P_l^\top A^{\text{loc}}_{\text{ext},l} P_l + P_r^\top A^{\text{loc}}_{\text{ext},r} P_r) \qquad (3.37)\\
&= \frac{s_0}{2} P_l^\top \frac{s_0}{2} \begin{pmatrix} 1 & 1 & 0 & \dots \\ 1 & 2 & 1 & \dots \\ 0 & 1 & 2 & 1 \\ & \ddots & \ddots & \\ & & 0 & 1 & 1 \end{pmatrix} P_l \\
&\quad + P_r^\top \frac{s_0}{2} \begin{pmatrix} 1 & 1 & 0 & \dots \\ 1 & 2 & 1 & \dots \\ 0 & 1 & 2 & 1 \\ & \ddots & \ddots & \\ & & 0 & 1 & 1 \end{pmatrix} P_r \text{ and}
\end{aligned}
$$

$$
\begin{aligned}
B_{\text{ext}} &= (n_l^2 P_l^\top B_{\text{ext},l}^{\text{loc}} P_l + n_r^2 P_r^\top A_{\text{ext},r}^{\text{loc}} P_r) \\
&= P_l^\top \frac{n_l^2}{2s_0} \begin{pmatrix} 1 & -1 & 0 & ... \\ -1 & 2 & -1 & ... \\ 0 & -1 & 2 & -1 \\ & \ddots & \ddots & \\ & & 0 & -1 \quad 1 \end{pmatrix} P_l \\
&\quad + P_r^\top \frac{n_r^2}{2s_0} \begin{pmatrix} 1 & -1 & 0 & ... \\ -1 & 2 & -1 & ... \\ 0 & -1 & 2 & -1 \\ & \ddots & \ddots & \\ & & 0 & -1 \quad 1 \end{pmatrix} P_r.
\end{aligned} \tag{3.38}
$$

3.4 Alternative Approach: Variational Formulation in $H^1(\Omega_{\text{int}}) \times H^+(D)$

An alternative approach to derive almost the same matrices was presented in [HN09, NS11]. Since it will prove to be useful when generalizing to higher space dimensions we will sketch its outlines here. In this approach, we aim at achieving local coupling in Equation (3.28) in a different way than before. This will not only yield a different approach to the same implementation but also give a more formal derivation of the exterior local element matrices and a variational formulation of the coupled problem including the interior and the exterior part.

The first step in this ansatz is to formalize the decomposition used in Equation (3.29) in order to obtain local coupling between the boundary degrees of freedom and the exterior degrees of freedom. For this we define an operator $\mathbf{T}^{(-)}$ that carries out the decomposition $\mathcal{L}_D\{f\}(\widetilde{s}) = \frac{1}{s_0}\mathbf{T}^{(-)}(f_0, F)^\top$ with $F(\widetilde{s}) = \frac{2s_0\mathcal{L}_D\{f\}(\widetilde{s}) - f_0}{\widetilde{s}-1}$ and

$$
\mathbf{T}^{(-)} \begin{pmatrix} f_0 \\ F \end{pmatrix} := \frac{1}{2}(f_0 + (\widetilde{s}-1)F(\widetilde{s})) \tag{3.39}
$$

for $(f_0, F)^\top \in \mathbb{C} \times H^+(D)$ where f_0 is the boundary degree of freedom.

Next we will make use of the identity

$$
\int_0^\infty f(\tau)g(\tau)d\tau = \frac{-s_0}{\pi} \int_{S^1} \mathcal{L}_D\{f\}(\overline{z})\mathcal{L}_D\{g\}(z)|dz|. \tag{3.40}
$$

To prove this equality, f and g are extended by zero to $f^0, g^0 : \mathbb{R} \to \mathbb{C}$. Then the left hand side integral is rewritten as Fourier transform. Using the properties of the Fourier transform and a suitable integration path, it is transformed to an integral along $s_0\mathbb{R}$, the line connecting 0 and is_0. Then this transformed integral is mapped to S^1 using the Möbius transform. A

full formal proof can be found at [Nan08, Lemma 5.3] and [HN09, Lemma A.1]. Substituting $A(F,G) := \frac{1}{2\pi}\int_{S^1} F(\overline{z})G(z)|dz|$ for $F, G \in H^+(D)$ for brevity, Equation (3.40) reads

$$\int_0^\infty f(\tau)g(\tau)d\tau = -2s_0 A(\mathcal{L}_D\{f\}, \mathcal{L}_D\{g\}). \tag{3.41}$$

This holds for u_{ext} and suitable test functions v_{ext} as well as for the derivatives u'_{ext} and v'_{ext}. In order to obtain simple formulas for the derivatives u'_{ext} and v'_{ext}, we again use the basic property of the Laplace transform

$$\mathcal{L}\{f'\}(s) = s\mathcal{L}\{f\}(s) - f_0. \tag{3.42}$$

By applying the Möbius transform to Equation (3.42), we have

$$\begin{aligned}
\mathcal{M}_{s_0}\mathcal{L}\{f'\}(\widetilde{s}) &= s_0\frac{\widetilde{s}+1}{\widetilde{s}-1}\mathcal{L}_D\{f\}(\widetilde{s}) - \frac{f_0}{\widetilde{s}-1} \\
&= \frac{1}{2}(f_0 + (\widetilde{s}+1)F(\widetilde{s})) \text{ with } F(\widetilde{s}) = \frac{2s_0\mathcal{L}_D\{f\}(\widetilde{s}) - f_0}{\widetilde{s}-1} \\
&=: \mathbf{T}^{(+)}\begin{pmatrix} f_0 \\ F \end{pmatrix}
\end{aligned} \tag{3.43}$$

Now we are able to deduce from the variational formulation (3.5) and the identity (3.41) a variational formulation in $H^1(\Omega_{\text{int}}) \times H^+(D)$:

$$B\left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix}\right) = 0 \tag{3.44}$$

with

$$\begin{aligned}
B\left(\begin{pmatrix} u_{\text{int}} \\ U \end{pmatrix}, \begin{pmatrix} v_{\text{int}} \\ V \end{pmatrix}\right) &:= \int_{\Omega_{\text{int}}} u'_{\text{int}}(x)v'_{\text{int}}(x) - n(x)^2\omega^2 u_{\text{int}}(x)v_{\text{int}}(x)dx \\
-2s_0 A\left(\mathbf{T}^{(+)}\begin{pmatrix} u_0 \\ U \end{pmatrix}, \mathbf{T}^{(+)}\begin{pmatrix} v_0 \\ V \end{pmatrix}\right) &- \frac{2n^2\omega^2}{s_0} A\left(\mathbf{T}^{(-)}\begin{pmatrix} u_0 \\ U \end{pmatrix}, \mathbf{T}^{(-)}\begin{pmatrix} v_0 \\ V \end{pmatrix}\right).
\end{aligned}$$

Equation (3.44) is a variational formulation for $(u_{\text{int}}, U)^\top \in H^1(\Omega_{\text{int}}) \times H^+(D)$ where $H^1(\Omega_{\text{int}})$ is the Sobolev space of weakly differentiable functions in Ω_{int}. For the trigonometric monomials $t^k(z) := \exp(ikz)$, $A(t^j, t^k) = \delta_{j,k}$. Thus, the implementation of the exterior part of B is reduced to the implementation of the two operators $\mathbf{T}^+$ and $\mathbf{T}^{(-)} : \mathbb{C} \times H^+(D) \to H^+(D)$.

If the ansatz space $\{t^0, t^1, ..., t^L\}$ is used for $H^+(D)$, these operators can be discretized by two matrices:

$$\mathcal{T}_L^{\pm} := \frac{1}{2}\begin{pmatrix} 1 & \pm 1 & & & \\ & 1 & \pm 1 & & \\ & & \ddots & \ddots & \\ & & & 1 & \pm 1 \\ & & & & 1 \end{pmatrix}. \tag{3.45}$$

The implementation of $\int_{x_r}^{\infty} \partial_x u_{\text{ext}}(x)\partial_x v_{\text{ext}}(x)dx$ is then done by the simple matrix matrix multiplication $2s_0\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)}$ and the implementation of $\int_{x_r}^{\infty} u_{\text{ext}}(x)v_{\text{ext}}(x)dx$ can be rephrased as $\frac{2}{s_0}\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)}$. These matrices are easily verified to correspond to $A_{\text{ext}}^{\text{loc}}$ and $B_{\text{ext}}^{\text{loc}}$ that were derived in the previous section. However it will be useful in the next section to have the discrete correspondence of $\mathcal{T}_L^{(-)}$ to $\mathcal{L}_D\{u_{\text{ext}}\}(\widetilde{s})$ and of $\mathcal{T}_L^{(+)}$ to $\mathcal{L}_D\{\partial_x f\}(\widetilde{s})$.

In terms of $\mathcal{T}_L^{\pm}$, we can rewrite the equations for the exterior degrees of freedom, Equation (3.37) and Equation (3.38) as

$$A_{\text{ext}} = 2s_0 P_l^{\top}(\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_l + 2s_0 P_r^{\top}(\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_r \text{ and} \tag{3.46}$$

$$B_{\text{ext}} = \frac{2n_l^2}{s_0}P_l^{\top}(\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_l + \frac{2n_r^2}{s_0}P_r^{\top}(\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_r. \tag{3.47}$$

As before P_l and P_r are $L \times N$ matrices mapping the local degrees of freedom for the exterior domain to the global degrees of freedom.

3.5 Generalization to Higher Space Dimensions

In the following section we aim at generalizing the concept we derived and implemented for the one-dimensional case so far in higher space dimensions. We will derive the implementation for the two dimensional case. The steps towards such an implementation are:

1. Subdivide the exterior domain Ω_{ext} into trapezoids that have one edge on Γ, the boundary of Ω_{int} and one edge infinitely far from it,
2. Map these trapezoids onto a reference strip, to obtain a coordinate transform,
3. Transform the variational formulation into the new coordinates,
4. Decouple the equations on the reference strip to obtain bounded integrals in the coordinate alongside the boundary of Ω_{int} and infinite integrals in the normal direction,
5. Treat bounded integrals with standard quadrature formulas,
6. Transform infinite integrals to $H^+(D)$ and use same discretization as in the one-dimensional case.

The mapping onto the reference rectangle in step 2 will give us coordinates (ξ, η), where ξ acts as a distance variable that measures the distance in the outward normal direction of Ω_{int}. Using this mapping we can transform the integrals in the variational formulation of our equation on a trapezoid onto a semi-infinite reference rectangle $[0,1] \times [0,\infty)$. Using Fubini's theorem, we can decouple the integrals on the reference strip in step 4. However, after

decoupling the infinite integrals will contain multiplication with the integration variable ξ alongside the test and ansatz functions and their gradients in the integrand. This is a situation that was not covered before and makes step 6 more involved than the one-dimensional equivalent. It necessitates the definition of a new operator $\mathbf{D}$ which we will derive at the end of this section. Together with the operators $\mathbf{T}^{(+)}$ and $\mathbf{T}^{(-)}$ from the previous section, we will then be able to express all the integral expressions that appear in our formulation. We will also give the discrete form $\mathcal{D}_L$ of $\mathbf{D}$ when using the trigonometric monomials as basis for $H^+(D)$. Together with $\mathcal{T}_L^{(+)}$ and $\mathcal{T}_L^{(-)}$ defined in the previous section, this will enable us to discretize all the integrals occurring in the two-dimensional implementation of the pole condition.

As described by Ruprecht et al. [RSSZ08] and Nannen and Schädle [NS11], the basic idea for an implementation in higher space dimensions is to use tensor product elements. Equation (3.1) for higher space dimensions takes the form

$$\Delta u(x) + n(x)^2\omega^2 u(x) = 0 \quad for \quad x \in \Omega \tag{3.48}$$

where $\Omega \subseteq \mathbb{R}^n$, $n \in \{2,3\}$. Again, Ω is assumed to be unbounded and we divide the domain of interest into a bounded interior Ω_{int} and an unbounded exterior part Ω_{ext}. Our approach is to assume a standard boundary condition at $\partial\Omega$ and the pole condition as radiation condition for the generalized radial part of u. In [HN09] Hohage and Nannen used a ball B_ρ with radius ρ to split Ω into a bounded interior $\Omega_{\text{int}} := B_\rho \cap \Omega$ and a spherical unbounded exterior $\Omega_{\text{ext}} := \mathbb{R}^n \backslash B_\rho$. By transformation to polar coordinates, they could split the exterior part into the unbounded radial direction and the bounded surface direction. Then they applied the one-dimensional Hardy Space Infinite Element approach in the unbounded direction handled the bounded surface direction with standard finite elements.

The approach used here works without the restriction to spherical exterior domains. Instead an arbitrary convex polygon P is used to split the domain into $\Omega_{\text{int}} := \Omega \cap P$ and $\Omega_{\text{ext}} := \Omega \backslash P$. Ω_{int} and Ω_{ext} share the common boundary $\Gamma := \partial P$. While in the interior, $H^1(\Omega_{\text{int}})$ is treated with standard finite elements, we apply a segmentation of Ω_{ext} into infinite trapezoids in the two-dimensional case and infinite prismatoids in the three-dimensional case, see Fig. 3.4. In order for such a segmentation to be valid, we require $n(x)$ to be constant within each trapezoid or prismatoid. See [Ket06, KS08, Sch02] for details on obtaining such a segmentation. We will stick to the two-dimensional case in the following paragraphs.

For the implementation we first need an affine bilinear mapping between a reference strip and each trapezoid, see Fig. 3.5. This mapping is a composition of three mappings, a transformation $T : (\xi, \eta) \to (x, y)$ that takes the reference strip to the right coordinate system, stretches and distorts

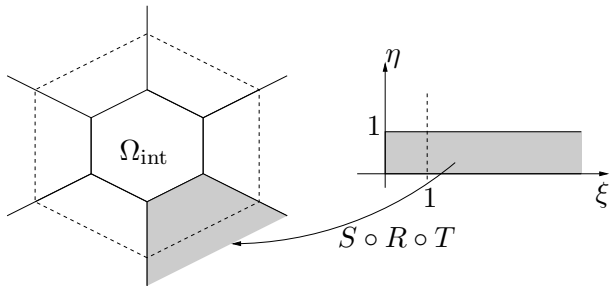

Figure 3.4: Discretization of the exterior with trapezoids. The dashed line is the image of the line $\xi = 1$ under the mapping $(R \circ T)$.

it appropriately. T is followed by a rotation R around $(0, 0)$ and a shift $S : (x, y) \to (x, y) + P_1$.

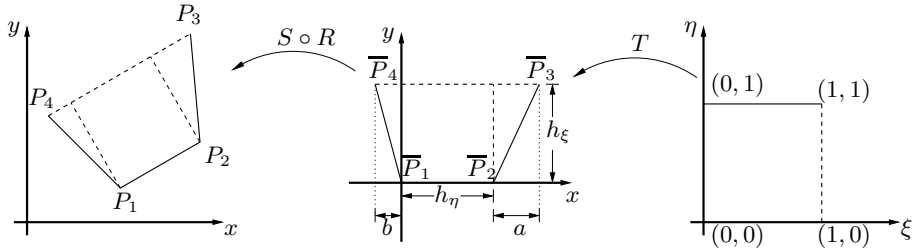

Figure 3.5: Mapping of the reference strip to a trapezoid. We have $T(0,0) = \overline{P}_1$, $T(0,1) = \overline{P}_2$, $T(1,0) = \overline{P}_4$ and $T(1,1) = \overline{P}_3$ and $R(\overline{P}_i) = P_i$ for $i = 1, ..., 4$.

Given the points of a trapezoid T in our discretization $P_i = (x_i, y_i)$ for $i = 1, ..., 4$, we can compute $h_\eta := \|P_2 - P_1\|_2 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$. Geometric calculations yield $a := (P_4 - P_3)(P_2 - P_3)^\top / \|P_4 - P_3\|_2$ and $b := (P_3 - P_4)(P_1 - P_4)^\top / \|P_4 - P_3\|_2$ and $h_\xi = \sqrt{\|P_3 - P_2\|_2^2 - a^2}$. Hence, the first mapping T is given by

$$(x, y) = T(\xi, \eta) = \begin{pmatrix} h_\eta \eta - b\xi + (a + b)\xi\eta \\ h_\xi \xi \end{pmatrix}. \tag{3.49}$$

It is noteworthy that a and b are signed distance variables. ξ will be used as a generalized radial variable whereas η plays the role of the surface variable on Γ. For our implementation to inherit the continuity of the solution in the exterior domain, it is important, that the radial variable ξ on a ray of the segmentation is independent of the neighboring infinite elements. That is why we use a trapezoidal construction so that ξ is constant on parallels to the segments of Γ. The rotation is given by

$$R := \frac{1}{\|P_2 - P_1\|} \begin{pmatrix} x_2 - x_1 & y_2 - y_1 \\ y_2 - y_1 & x_1 - x_2 \end{pmatrix}. \tag{3.50}$$

Hence, the entire mapping is given by

$$\begin{aligned}(x,y) &= (S\circ R\circ T)(\xi,\eta) \\ &= \frac{1}{\|P_2-P_1\|}\begin{pmatrix} x_2-x_1 & y_1-y_2 \\ y_2-y_1 & x_2-x_1\end{pmatrix}\begin{pmatrix} h_\eta\eta - b\xi + (a+b)\xi\eta \\ h_\xi\xi\end{pmatrix} + \begin{pmatrix}x_1\\ y_1\end{pmatrix}.\end{aligned} \tag{3.51}$$

The Jacobi matrix J of (3.51) and its determinant are

$$J = \begin{pmatrix} h_\eta + (a+b)\xi & -b+(a+b)\eta \\ 0 & h_\xi\end{pmatrix},\quad |J| = h_\xi(h_\eta+(a+b)\xi). \tag{3.52}$$

Its inverse is

$$J^{-1} = \begin{pmatrix} \frac{1}{h_\eta+(a+b)\xi} & \frac{b-(a+b)\eta}{h_\xi(h_\eta+\xi(a+b))} \\ 0 & \frac{1}{h_\xi}\end{pmatrix} \tag{3.53}$$

We are now in a position to derive a suitable variational formulation of the exterior part of our problem. First we will transform the integrals over T onto the reference rectangle. On T the integrals read:

$$\int_T (\nabla u(x,y))\cdot(\nabla v(x,y))\,d(x,y) + \omega^2\int_T n(x,y)^2 u(x,y)v(x,y)d(x,y) = 0$$

Transforming to the reference rectangle, the test and ansatz function and refractive index transform as follows: $v(x,y)\to\hat v(\xi,\eta)$, $u(x,y)\to\hat u(\xi,\eta)$ and $n(x,y)\to\hat n(\xi,\eta)$. Thus, the transformed integral terms read

$$\begin{aligned}&\int_T (\nabla u(x,y))\cdot(\nabla v(x,y))\,d(x,y) \qquad (3.54)\\ &= \int_{[0,1]\times[0,\infty)} \left(J^{-\top}\nabla\hat u(\xi,\eta)\right)\cdot\left(J^{-\top}\nabla\hat v(\xi,\eta)\right)|J|d(\xi,\eta) \quad \text{and}\\ &\int_T n(x,y)^2u(x,y)v(x,y)d(x,y) \qquad (3.55)\\ &= \int_{[0,1]\times[0,\infty)} \hat n(\xi,\eta)^2\hat u(\xi,\eta)\hat v(\xi,\eta)|J|d(\xi,\eta).\end{aligned}$$

Since we assume $n(x,y)\equiv n_i$ to be constant in each trapezoid T_i, we can remove $n(x,y)$ and $\hat n(\xi,\eta)$ from the integrals in (3.55):

$$n_i^2\int_T u(x,y)v(x,y)d(x,y) = n_i^2\int_{[0,1]\times[0,\infty)}\hat u(\xi,\eta)\hat v(\xi,\eta)|J(\xi)|d(\xi,\eta). \tag{3.56}$$

If we choose $\hat u(\xi,\eta)$ and $\hat v(\xi,\eta)$ so that we can factorize $\hat u(\xi,\eta)=\hat u_\xi(\xi)\hat u_\eta(\eta)$ and $\hat v(\xi,\eta)=\hat v_\xi(\xi)\hat v_\eta(\eta)$, then the integrals over the trapezoids decouple to independent integrals over ξ and η. Suppose that the interior Ω_{int} is already discretized with standard finite elements. Then the integrals alongside Γ, that is the bounded η-integrals, can be discretized using the traces of the

finite element basis functions in Ω_{int} on Γ as basis functions. If we build the elements for each infinite trapezoid by forming a tensor product of the finite element space formed by the traces of the interior elements and the trigonometric monomials as basis for the Hardy space in ξ-direction, then for the combined basis functions the integrals decouple.

Since the determinant $|J|$ is independent of η, by Fubini's theorem, Equation (3.56) becomes

$$n_i^2 \int_T u(x,y)v(x,y)d(x,y) = n_i^2 \left(\int_0^1 \hat{u}_\eta(\eta)\hat{v}_\eta(\eta)d\eta\right)\left(\int_0^\infty \hat{u}_\xi(\xi)\hat{v}_\xi(\xi)|J|d\xi\right). \tag{3.57}$$

Due to the presence of $J^{-\top}$, the situation for Equation (3.54) is more involved. Inserting the definitions of J and the factorization of u and v yields

$$\begin{aligned}
&\int_T (\nabla u(x,y)) \cdot (\nabla v(x,y))\, d(x,y) \qquad (3.58)\\
&= \left(\int_0^1 \hat{u}'_\eta(\eta)(h_\xi^2 + (b-(a+b)\eta)^2)\hat{v}'_\eta(\eta)d\eta\right)\left(\int_0^\infty \frac{\hat{u}_\xi(\xi)\hat{v}_\xi(\xi)}{h_\xi(h_\eta+(a+b)\xi)}d\xi\right)\\
&\quad + \left(\int_0^1 \hat{u}'_\eta(\eta)\frac{b-(a+b)\eta}{h_\xi}\hat{v}_\eta(\eta)d\eta\right)\left(\int_0^\infty \hat{u}_\xi(\xi)\hat{v}'_\xi(\xi)d\xi\right)\\
&\quad + \left(\int_0^1 \hat{u}_\eta(\eta)\frac{b-(a+b)\eta}{h_\xi}\hat{v}'_\eta(\eta)d\eta\right)\left(\int_0^\infty \hat{u}'_\xi(\xi)\hat{v}_\xi(\xi)d\xi\right)\\
&\quad + \left(\int_0^1 \hat{u}_\eta(\eta)\hat{v}_\eta(\eta)d\eta\right)\left(\int_0^\infty \hat{u}'_\xi(\xi)\frac{h_\eta+(a+b)\xi}{h_\xi}\hat{v}'_\xi(\xi)d\xi\right).
\end{aligned}$$

For the time being we will just take the finite integrals as given and refer to the papers cited in Remark 3.1 (2) for more details on the implementation. So we take the discretization of the finite integrals to be given and result in matrices $T^{\text{ext}}_{\text{loc},i,1}$ to $T^{\text{ext}}_{\text{loc},i,5}$ in order of their appearance in Equations (3.57) and (3.58):

$$\int_0^1 \hat{u}_\eta(\eta)\hat{v}_\eta(\eta)d\eta \approx T^{\text{ext}}_{\text{loc},i,1}, \tag{3.59a}$$

$$\int_0^1 \hat{u}'_\eta(\eta)(h_\xi^2 + (b-(a+b)\eta)^2)\hat{v}'_\eta(\eta)d\eta \approx T^{\text{ext}}_{\text{loc},i,2}, \tag{3.59b}$$

$$\int_0^1 \hat{u}'_\eta(\eta)\frac{b-(a+b)\eta}{h_\xi}\hat{v}_\eta(\eta)d\eta \approx T^{\text{ext}}_{\text{loc},i,3}, \tag{3.59c}$$

$$\int_0^1 \hat{u}_\eta(\eta)\frac{b-(a+b)\eta}{h_\xi}\hat{v}'_\eta(\eta)d\eta \approx T^{\text{ext}}_{\text{loc},i,4} \text{ and} \tag{3.59d}$$

$$\int_0^1 \hat{u}_\eta(\eta)\hat{v}_\eta(\eta)d\eta \approx T^{\text{ext}}_{\text{loc},i,5}. \tag{3.59e}$$

We will now tackle the infinite ξ-integrals by transforming them to the Hardy space $H^+(D)$ using the techniques presented in the previous section. However, two of these integrals contain factors $(\xi + c)$ and $(\xi + c)^{-1}$ for constant $c > 0$ that appear in the integrands. These factors are new in the higher-dimensional implementation and have to be dealt with separately. The next section will sketch a way to discretize the integrals containing these factors.

Including the argument s of $f(s)$ for clarity, we know from the basic properties of the Laplace transform that $\mathcal{L}\{sf(s)\}(\widetilde{s}) = -(\mathcal{L}\{f(s)\})'(\widetilde{s})$ and $\mathcal{L}\left\{\frac{f(s)}{s}\right\}(\widetilde{s}) = \int_0^\infty \mathcal{L}\{f\}(\sigma)d\sigma$. Using these properties, we can derive an operator $\mathbf{D} : H^+(D) \to H^+(D)$ for the factor ξ in Equation (3.58). Taking the equations to $H^+(D)$, we can implicitly define the operator $\mathbf{D}$ by

$$\mathbf{D}\left(\mathcal{M}_{s_0}\mathcal{L}\{f\}\right)(\widetilde{s}) = \mathcal{M}_{s_0}(-(\mathcal{L}\{f\})')(\widetilde{s}) = \mathcal{M}_{s_0}\mathcal{L}\{sf\}(\widetilde{s}). \tag{3.60}$$

For $F \in H^+(D)$, we can compute

$$(\mathbf{D}F)(\widetilde{s}) = \frac{(\widetilde{s}-1)^2}{2s_0}F'(\widetilde{s}) + \frac{\widetilde{s}-1}{2s_0}F(\widetilde{s}). \tag{3.61}$$

As in the one-dimensional case, we use the trigonometric monomials $t^k(z) := \exp(ikz)$ up to order L as basis of $H^+(D)$ then $\mathcal{D}_L$, the discrete form of $\mathbf{D}$ reads $\frac{1}{2s_0}\mathcal{D}_L$ with the matrix

$$\mathcal{D}_L := \begin{pmatrix} -1 & 1 & & & \\ 1 & -3 & 2 & & \\ & 2 & -5 & 3 & \\ & & \ddots & \ddots & \ddots \\ & & & L & -2L-1 \end{pmatrix}. \tag{3.62}$$

A factor $(\xi+c)\hat{f}(\xi)$ in the integrand thus corresponds to the discrete $(\frac{1}{2s_0}\mathcal{D}_L + c\,\mathrm{id})(\mathcal{L}_D\{f\})$. For the factors $(\xi + c)^{-1}$, we use the fact that they are the inverse of $(\xi+c)$, hence they can be discretized as $\left(\frac{1}{2s_0}\mathcal{D}_L + c\,\mathrm{id}\right)^{-1}(\mathcal{L}_D\{f\})$.

We are now in a position to give the matrices that are the discrete implementations of the infinite ξ-integrals in (3.57) and (3.58):

$$\begin{aligned}
\int_0^\infty \hat{u}_\xi(\xi)\hat{v}_\xi(\xi)|J|d\xi &= \int_0^\infty \hat{u}_\xi(\xi)\hat{v}_\xi(\xi)h_\xi(h_\eta+(a+b)\xi)d\xi \\
&\approx -\frac{2h_\xi}{s_0}\mathcal{T}_L^{(-)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)\mathcal{T}_L^{(-)}, \\
\int_0^\infty \frac{\hat{u}_\xi(\xi)\hat{v}_\xi(\xi)}{h_\xi(h_\eta+(a+b)\xi)}d\xi &\approx -\frac{2}{s_0h_\xi}\mathcal{T}_L^{(-)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)^{-1}\mathcal{T}_L^{(-)}, \\
\int_0^\infty \hat{u}_\xi(\xi)\hat{v}'_\xi(\xi)d\xi &\approx -2\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(+)}, \\
\int_0^\infty \hat{u}'_\xi(\xi)\hat{v}_\xi(\xi)d\xi &\approx -2\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(-)} \text{ and} \\
\int_0^\infty \hat{u}'_\xi(\xi)\frac{h_\eta+(a+b)\xi}{h_\xi}\hat{v}'_\xi(\xi)d\xi &\approx -\frac{2s_0}{h_\xi}\mathcal{T}_L^{(+)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)\mathcal{T}_L^{(+)}.
\end{aligned}$$

Thus, on the ith prismatoid T_i, we have the following local stiffness matrix $A^{\mathrm{ext}}_{\mathrm{loc},i}$ and mass matrix $B^{\mathrm{ext}}_{\mathrm{loc},i}$:

$$\begin{aligned}
A^{\mathrm{ext}}_{\mathrm{loc},i} &:= T^{\mathrm{ext}}_{\mathrm{loc},i,2}\otimes\left[\frac{-2h_\xi}{s_0}\mathcal{T}_L^{(-)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)^{-1}\mathcal{T}_L^{(-)}\right] \qquad (3.63)\\
&\quad +T^{\mathrm{ext}}_{\mathrm{loc},i,3}\otimes\left[(-2)\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)}\right] \\
&\quad +T^{\mathrm{ext}}_{\mathrm{loc},i,4}\otimes\left[(-2)\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)}\right] \\
&\quad +T^{\mathrm{ext}}_{\mathrm{loc},i,5}\otimes\left[\frac{-2s_0}{h_\xi}\mathcal{T}_L^{(+)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)\mathcal{T}_L^{(+)}\right] \text{ and} \\
B^{\mathrm{ext}}_{\mathrm{loc},i} &:= n_i^2T^{\mathrm{ext}}_{\mathrm{loc},i,1}\otimes\left[\frac{-2h_\xi}{s_0}\mathcal{T}_L^{(-)\top}\left(h_\eta\,\mathrm{id}+\frac{a+b}{2s_0}\mathcal{D}_L\right)\mathcal{T}_L^{(-)}\right]. \qquad (3.64)
\end{aligned}$$

However, the inverse $(\mathcal{D}_L+c\,\mathrm{id})^{-1}$ gives a full block and thus a full local stiffness matrix. Thus it may be reasonable to avoid inverting $(\mathcal{D}_L+c\,\mathrm{id})$. We can do so by choosing to double the number of unknowns by the following scheme where it is suitable:

$$\begin{aligned}
(M_1+F_1M_2^{-1}F_2)u=0 &\Leftrightarrow F_1^{-1}M_1u+M_2^{-1}F_2u=0 \\
&\Leftrightarrow F_1^{-1}M_1u+w=0\wedge w=M_2^{-1}F_2u \\
&\Leftrightarrow M_1u+F_1w=0\wedge F_2^{-1}M_2w=u \\
&\Leftrightarrow M_1u+F_1w=0\wedge F_2u-M_2w=0 \\
&\Leftrightarrow \begin{pmatrix}M_1 & F_1\\ F_2 & -M_2\end{pmatrix}\begin{pmatrix}u\\ w\end{pmatrix}=0 \qquad (3.65)
\end{aligned}$$

This will result in local element matrices for the infinite Hardy space elements that do not have full blocks but a higher structure instead (see Figure 3.6). Now we rename the components required to give the full local stiffness and mass matrices for the infinite trapezoid arising from the infinite ξ-integrals according to their order or appearance:

$$\begin{aligned}
N^{\text{ext}}_{\text{loc},i,1} &= -\frac{2h_\xi h_\eta}{s_0}\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)} - \frac{h_\xi(a+b)}{s_0^2}\mathcal{T}_L^{(-)\top}\mathcal{D}_L\mathcal{T}_L^{(-)},\\
N^{\text{ext}}_{\text{loc},i,2} &= -h_\eta\,\text{id} + \frac{a+b}{2s_0}\mathcal{D}_L,\\
N^{\text{ext}}_{\text{loc},i,3} &= -2\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(+)},\\
N^{\text{ext}}_{\text{loc},i,4} &= -2\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(-)}, \text{ and}\\
N^{\text{ext}}_{\text{loc},i,5} &= -\frac{2s_0h_\eta}{h_\xi}\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)} - \frac{a+b}{h_\xi}\mathcal{T}_L^{(+)\top}\mathcal{D}_L\mathcal{T}_L^{(+)}.
\end{aligned}$$

Putting these parts together and multiplying with the matrices for the tangential part $T^{\text{ext}}_{\text{loc},i,1}$, ..., $T^{\text{ext}}_{\text{loc},i,5}$, the matrix M_1 in Equation (3.65) becomes $A^{\text{ext}}_{\text{loc},i,1} := T^{\text{ext}}_{\text{loc},i,5} \otimes N^{\text{ext}}_{\text{loc},i,5} + T^{\text{ext}}_{\text{loc},i,4} \otimes N^{\text{ext}}_{\text{loc},i,4} + T^{\text{ext}}_{\text{loc},i,3} \otimes N^{\text{ext}}_{\text{loc},i,3}$ and M_2 corresponds to $2N^{\text{ext}}_{\text{loc},i,2}$ while the factors F_1 and F_2 are $-2T^{\text{ext}}_{\text{loc},i,1} \otimes \mathcal{T}^-$ and $\mathcal{T}^-$. The local element matrices for the infinite part of the i-th trapezoid therefore read

$$\begin{aligned}
A^{\text{ext}}_{\text{loc},i} &= \begin{pmatrix} A^{\text{ext}}_{\text{loc},i,1} & -2T^{\text{ext}}_{\text{loc},i,1} \otimes \mathcal{T}_L^{(-)\top} \\ 2\mathcal{T}_L^{(-)} & -2N^{\text{ext}}_{\text{loc},i,2} \end{pmatrix} \text{ and}\\
B^{ex}_{\text{loc},i} &= n_i^2 \begin{pmatrix} T^{\text{ext}}_{\text{loc}.i,1} \otimes N_{\text{loc},i,1} & 0 \\ 0 & 0 \end{pmatrix}
\end{aligned}$$

where n_i is the refractive index in the i-th trapezoid.

Since the Kronecker product is bilinear and associative, that is $A \otimes (B + kC) = A \otimes B + k(A \otimes C)$, it is possible to sort $A^{\text{ext}}_{\text{loc},i}$ and $B^{\text{ext}}_{\text{loc},i}$ by powers of

s_0.

$$\begin{aligned} A^{\text{ext}}_{\text{loc},i} &= s_0 \frac{2h_\eta}{h_\xi} \begin{pmatrix} (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)}) \otimes T^{\text{ext}}_{\text{loc},i,5} & 0 \\ 0 & 0 \end{pmatrix} \\ &\quad + \frac{1}{s_0}(a+b) \begin{pmatrix} 0 & 0 \\ 0 & -\mathcal{D}_L \end{pmatrix} \\ &\quad + \begin{pmatrix} -2\left((\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(+)}) \otimes T^{\text{ext}}_{\text{loc},i,3} + (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(-)}) \otimes T^{\text{ext}}_{\text{loc},i,4}\right) & -2T^{\text{ext}}_{\text{loc},i,1} \otimes \mathcal{T}_L^{(-)\top} \\ 2\mathcal{T}_L^{(-)} & 2h_\eta \,\mathrm{id} \end{pmatrix} \\ &=: s_0 A^{\text{ext},(1)}_{\text{loc},i} + \frac{1}{s_0} A^{\text{ext},(-1)}_{\text{loc},i} + A^{\text{ext},(0)}_{\text{loc},i} \text{ and} \end{aligned} \tag{3.66}$$

$$\begin{aligned} B^{\text{ext}}_{\text{loc},i} &= n_i^2 \frac{1}{s_0}\frac{2h_\xi}{h_\eta} \begin{pmatrix} \mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)} \otimes T^{\text{ext}}_{\text{loc},i,1} & 0 \\ 0 & 0 \end{pmatrix} \\ &\quad + n_i^2 \frac{1}{s_0^2} h_\xi (a+b) \begin{pmatrix} \mathcal{T}_L^{(-)\top}\mathcal{D}_L\mathcal{T}_L^{(-)} \otimes T^{\text{ext}}_{\text{loc},i,1} & 0 \\ 0 & 0 \end{pmatrix} \\ &=: n_i^2 \frac{1}{s_0} B^{\text{ext},(-1)}_{\text{loc},i} + n_i^2 \frac{1}{s_0^2} B^{\text{ext},(-2)}_{\text{loc},i}. \end{aligned} \tag{3.67}$$

If, as in the previous sections, P_i denotes the L by N matrix mapping the local degrees of freedom to global degrees of freedom, we obtain the exterior part A_{ext} and B_{ext} of the matrices A and B by summing over all trapezoids:

$$A_{\text{ext}} = \sum_{T_i} P_i^\top A^{\text{ext}}_{\text{loc},i} P_i \text{ and} \tag{3.68}$$

$$B_{\text{ext}} = \sum_{T_i} P_i^\top B^{\text{ext}}_{\text{loc},i} P_i. \tag{3.69}$$

Remark 3.1.

1. Since each segment in Ω_{ext} is treated separately, it is possible to account for unbounded inhomogeneities such as waveguides. This does not require any further implementation but can be dealt with by the methods presented here.
2. The exact statement of the tensor product spaces and the right test functions are quite technical and can be found in great detail in [HN09] and [Nan08].

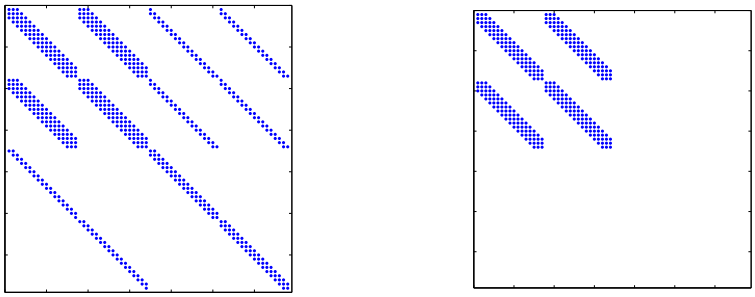

Figure 3.6: Structure of local stiffness matrix $A^{\text{ext}}_{\text{loc},i}$ (left) and local mass matrix $B^{\text{ext}}_{\text{loc},i}$ (right) for one trapezoid with $L = 17$.

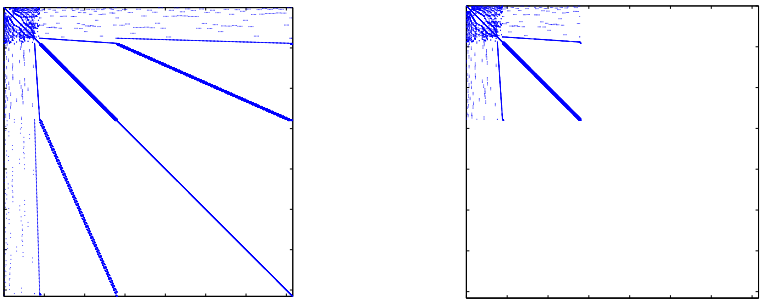

Figure 3.7: Structure of global stiffness matrix A (left) and global mass matrix B (right) for a two-dimensional problem with $N = 3578$ degrees of freedom.

Chapter 4

Spurious Solutions

In this chapter we will discuss the unphysical and thus unwanted *spurious solutions* that pollute the numerical solutions of resonance problems. In the first section we will give a brief overview of different types of spurious solutions and the means that were found useful for avoiding them in the case of resonators with Neumann boundaries. The second section will deal with the issue of spurious solutions occurring when computing the spectra of resonators in open space.

Finite element solutions of resonance problems for various equations have the problem that due to discretization errors there exist solutions in the discrete setting that do not correspond to a physically sensible solution of the problem but are still within the spectral region of interest, sometimes very close to the physically relevant solutions. That these spurious solutions which are frequently branded "notorious", are an issue to be treated can be seen by the fact, that since their first mentioning in the late 1970ies, there have been thousands of papers published mentioning them. [1] It is important to distinguish between two different types of spurious solutions:

1. Spurious solutions in the interior, that also exist when computing closed cavities. These solutions were historically the first to be discovered and can be overcome by using the "correct" finite elements for the discretization of the problem. They are an important issue but to be distinguished from the second type of spurious solutions that we aim at. Section 4.1 gives a brief overview of the first type of spurious solutions and ways to avoid them.

2. Spurious solutions in the exterior. Since we are interested in computing solutions for open resonators, we couple an infinite exterior domain to our interior problem and since we discretize this exterior domain

[1] The ISI-Database finds 1094 papers whose titles mention "spurious solutions" at the beginning of 2012. This does not include equivalent formulations such as "vector parasites", "spectral convergence" etc. so the actual number is even higher.

with a finite number of degrees of freedom, there will always be solutions where the discretization in the exterior is not sufficient which will give rise to unphysical solutions. These unphysical solutions are the spurious solutions, we aim at. We will shed some light on their occurrence in Section 4.2

It should be noted that the problem of the first type of spurious solutions is not exclusive to the finite element method, but was also reported in the context of the finite difference method [CD72, SB84, Su85], the boundary element method [GS77, SSA92] and the spectral method [FA76], however since we apply the finite element method, we will focus on its spurious solutions in the following section.

4.1 Spurious Solutions in Closed Cavities

Literature on spurious solutions of resonance problems is manifold as the problem has always been a serious handicap for the simulation of electromagnetic resonators. The first steps towards the treatment of unwanted spurious solutions dealt with resonance problems with closed resonators [PL91, SMYC95, CHC99, BFGP99, BBG00, CFR01a, FR02b]. That is with problems having a perfectly conducting electric boundary in the electromagnetic case or, mathematically speaking, having Neumann boundary conditions on each boundary.

Early attempts at computing numerical approximations of resonance problems, that is of computing approximations of frequencies and fields of cavity resonators, showed very soon that the solutions of finite element models that seemed reasonable were affected by solutions that lacked any physical meaning. These solutions were dubbed *spurious solutions* [CS70, DFP82, HWFK83, PCS88].

The severe drawback that the pollution of the eigenvalue spectra with these unphysical spurious solutions presented, was overcome by the introduction of specific finite elements and meshes [Bos90, Bos88, LSC91, Cen91, CSJ88, DLW94, WC88]. From a practical viewpoint these special elements and meshes offered solutions to the problems of spurious solutions in closed cavities, however from a mathematical point of view the reasons for their success was often given wrongly [CSJ88, WC88, FS92, WI91], as proved in [DLW94, FS92, CFR95, CFR96, CFR97, BFLP99].

The mathematical theory for the convergence of finite element approximations of resonance problems is highly involved and beyond the scope of this work (see [Mon03, MD01, CFR01a, BFLP99, Bof01]). However, we will briefly outline the generally accepted reason for spurious solutions in the simulation of closed cavities and ways of avoiding them in the following paragraphs.

It was possible to consider the question of spurious solutions solved from an users point of view when Nédélec introduced a new family of finite elements which also became known as edge elements [Néd80]. Even though there were clear indications that the edge elements resolved the issue of spurious solutions in the simulation of the electromagnetic fields in closed cavities, a correct explanation of that behavior was only given much later [CFR01b, FR02a].

For this explanation, Caorsi et al. defined an approximation to be spurious-free, if it satisfies five conditions:

1. *Completeness of the spectrum*: For any eigenvalue ω of the original problem we can find a sequence of approximated eigenvalues $\{\omega_h\}$ such that $\omega_h \to \omega$ as $h \to 0$.

2. *Non-pollution of the spectrum*: For any bounded sequence $\{\omega_h\}$, the distance of ω_h from the exact spectrum vanishes as $h \to 0$.

3. *Completeness of the eigenspace*: For any eigenvector u of the original problem we can find a sequence of approximated eigenvectors $\{u_h\}$ such that $u_h \to u$ as $h \to 0$.

4. *Non-pollution of the eigenspaces*: No sequence $\{u_h\}$ that consists of normalized numerical eigenvectors corresponding to a bounded sequence of numerical eigenvalues can have a non-vanishing distance from the union of all eigenspaces of the original problem.

5. No sequence of strictly positive eigenvalues $\{\omega_h\}, \quad \omega_h > 0 \quad \forall h$ can converge to $\omega = 0$.

Given the function spaces $V = H_0(\mathrm{curl}, \Omega) := \{v \in L^2(\Omega) : \nabla \times v \in L^2(\Omega), n \times v = 0 \text{ on } \Gamma\}$, $V_0 := \{v \in V : \nabla \times v = 0\}$ and $V_1 := \{v \in V : \nabla \cdot v = 0\}$, they found that there are three conditions on a sequence of finite element spaces $\{V_h\}$ generated on a regular family of triangulations that make for a spurious-free approximation of a resonance:

1. "Completeness of the approximating subspace":

$$\forall v \in V : \lim_{h \to 0} \|v - v_h\|_V = 0,$$

2. "Completeness of the discrete kernel":

$$\forall v \in V_0 : \lim_{h \to 0} \inf_{v_h \in V_{0h}} \|v - v_h\|_V = 0$$

and

3. "Discrete compactness property": Any sequence $\{v_h\}$ such that $v_h \in V_{1h}, \|v_h\|_V \leq C \ \forall h$ contains a sub-sequence $\{u_h\}$ such that

$$\exists v \in L^2(\Omega) : \lim_{h \to 0} \|u_h - v\|_{L^2} = 0.$$

These conditions contain the classical approximation condition 1 and the "divergence free" condition (i.e. condition 2) that was often postulated as sole condition for spurious-free approximation of resonances. However, Caorsi et al. went beyond these classical conditions by introducing the third condition and proved that all three conditions are necessary and sufficient for an approximation of the resonances of a closed cavity to be spurious-free [CFR01a, FR02a]. We will see in the next section that taking into consideration open resonators with non perfectly conducting boundaries introduces a new type of spurious solutions caused by the approximation introduced by the necessary transparent boundary conditions.

4.2 Spurious Solutions in Open Resonators

When open resonators are taken into account, a second type of spurious solution enters the stage. This second type of spurious solutions is caused by the discretization of the exterior domain, as we will see later in this section. The outline of this section is as follows: first we will introduce a simple one-dimensional example that we can tackle analytically. Then we will couple two different types of transparent boundary conditions, the perfectly matched layer method and the pole condition, to the analytic solution in the interior which will show us, that both methods introduce spurious solutions when the infinite exterior domain is discretized with a finite number of degrees of freedom.

The issue of spurious solutions caused by transparent boundary conditions was already addressed in the diploma thesis of M. Rechberger [Rec05]. Based on the observation that there are spurious solutions in open resonators that are caused by the transparent boundary conditions, she aimed at identifying these spurious solutions by computing their sensitivity towards changes of the transparent boundary condition. She found that the sensitivity towards the damping in the perfectly matched layers (PML) is not sufficient as a criterion for the detection of spurious solutions and therefore used a combination of the PML and a wave factorization in the exterior. For the wave factorization, Rechberger factorized the outgoing wave into an exponential term $\exp(i\omega|x|)$ and a term $\widetilde{u}$. She combined both methods by first factorizing the solution and then applying a PML for damping the radial exponential term of the resulting problem. They found that by combining both methods they were able to distinguish physical from spurious solutions for several model problems in acoustics. The downside of this approach

however is, that in the formulation given in [Rec05] it does not account for heterogeneous exterior domains and requires special tuning for each problem. Both of these downsides are not present in the method for detecting spurious solutions that we will present in Chapter 5.

Further attempts towards the detection of spurious solutions caused by the boundary conditions were undertaken by Tischler et. al [Tis03, TH00]. They derived a criterion they called "PPP (Power Part in PML) criterion" that relates the energy flux in the PML layer to the energy flux of the free field. The downside of this criterion is that it will only work if the energy flux of the exterior domain can be directly computed and is very sensitive towards the material composition in Ω_{ext}.

Before deriving our way of detecting this second type of spurious solutions in the next chapter, we will make their existence plausible by investigating thoroughly the following simple example.

Example 4.1.

We want to solve the Helmholtz resonance problem in one space dimension. As layout for our first example, we will use a symmetric one-dimensional cavity layout. The cavity consists of an area with refractive index n_i that stretches from $x_l = -1$ to $x_r = 1$ is embedded in a material with lower refractive index $n_e < n_i$. The layout of this example is sketched in Figure 4.1.

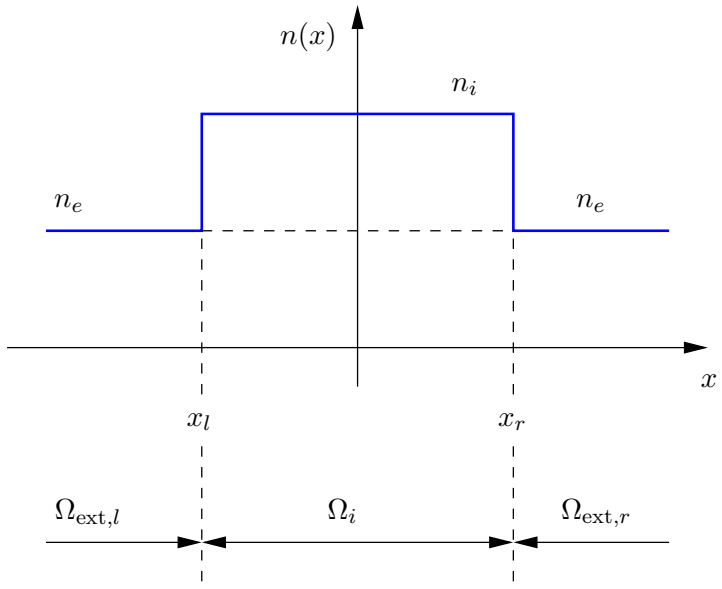

Figure 4.1: One-dimensional cavity layout.

Solution with Special Functions

We will now give an analytic solution for Example 4.1 using our knowledge of the special form of the solution in the interior and the exterior. We know that time-harmonic solutions to Equation (3.1) are superpositions of waves $u_{\text{int},1,2}(x) = c\exp(\pm i\omega n(x)x)$ and that the solutions in the exterior are supposed to be outgoing, hence $u_{\text{ext},l}(x) = \exp(-i\omega n_e x)$ for $x < x_l$ and $u_{\text{ext},r}(x) = \exp(i\omega n_e x)$ for $x > x_r$. Since the solutions themselves as well as their derivatives have to be continuous at x_l and x_r, we can derive the following system of equations which has to hold for a solution:

$$\begin{aligned}
\gamma u_{\text{ext},l}(x_l) &= \alpha u_{\text{int},1}(x_l) + \beta u_{\text{int},2}(x_l) && \text{(4.1a)}\\
\delta u_{\text{ext},r}(x_r) &= \alpha u_{\text{int},1}(x_r) + \beta u_{\text{int},2}(x_r) && \text{(4.1b)}\\
\gamma \frac{d}{dx} u_{\text{ext},l}(x_l) &= \alpha \frac{d}{dx} u_{\text{int},1}(x_l) + \beta \frac{d}{dx} u_{\text{int},2}(x_l) && \text{(4.1c)}\\
\delta \frac{d}{dx} u_{\text{ext},r}(x_r) &= \alpha \frac{d}{dx} u_{\text{int},1}(x_r) + \beta \frac{d}{dx} u_{\text{int},2}(x_r). && \text{(4.1d)}
\end{aligned}$$

Equations (4.1a) and (4.1b) guarantee the continuity of the solution at x_l and x_r and Equations (4.1c) and (4.1d) guarantee the continuity of the derivatives. The parameters α, β, γ and δ need to be determined. Inserting our known solutions $u_{\text{ext},l,r}$ and $u_{\text{int},1,2}$, we have

$$\begin{aligned}
\gamma e^{-i\omega n_e x_l} &= \alpha e^{-i\omega n_i x_l} + \beta e^{i\omega n_i x_l} && \text{(4.2a)}\\
\delta e^{i\omega n_e x_r} &= \alpha e^{-i\omega n_i x_r} + \beta e^{i\omega n_i x_r} && \text{(4.2b)}\\
-i\omega n_e \gamma e^{-i\omega n_e x_l} &= -i\omega n_i \alpha e^{-i\omega n_i x_l} + i\omega n_i \beta e^{i\omega n_i x_l} && \text{(4.2c)}\\
i\omega n_e \delta e^{i\omega n_e x_r} &= -i\omega n_i \alpha e^{-i\omega n_i x_r} + i\omega n_i \beta e^{i\omega n_i x_r}. && \text{(4.2d)}
\end{aligned}$$

We seek non-trivial solutions of the linear system of equations (4.2a)-(4.2d). The existence of these solutions is dependent on the values of ω. Values for which we may find such non-trivial solutions are the resonances of the system. Without restriction of generality we can assume our problem to be axially symmetric with respect to $x = 0$ and $x_l = -x_r$, we also seek for symmetric solutions, that is either $\alpha = \beta$ and $\gamma = \delta$ or $\alpha = -\beta$ and $\gamma = -\delta$. We may tackle this problem by taking the coefficient matrix of the system of equations to be dependent of ω. Non-trivial solutions to the system of equations exist, if the determinant of the coefficient matrix is zero. For the first type of symmetry, the coefficient matrix reads

$$\mathbf{M}^{+}_{sym}(\omega) = \begin{pmatrix} e^{i\omega n_i x_r} + e^{-i\omega n_i x_r} & -e^{i\omega n_e x_r} \\ -i\omega n_i (e^{-i\omega n_i x_r} - e^{i\omega n_i x_r}) & -i\omega n_e e^{i\omega n_e x_r} \end{pmatrix}, \qquad (4.3)$$

in the second case we have

$$\mathbf{M}^{-}_{sym}(\omega) = \begin{pmatrix} e^{i\omega n_i x_r} - e^{-i\omega n_i x_r} & -e^{i\omega n_e x_r} \\ -i\omega n_i (e^{i\omega n_i x_r} + e^{-i\omega n_i x_r}) & i\omega n_e e^{i\omega n_e x_r} \end{pmatrix}. \qquad (4.4)$$

Now we can compute the values ω for which the system has a solution:

Lemma 4.1.
In the symmetric case, the resonance frequencies, that is the values ω for which the system of Equations (4.2a)-(4.2d) has a non-trivial solution, all have the same imaginary part and are equidistantly distributed in the direction of the real axis.

Proof.
Setting $a := n_i + n_e$ and $b := n_i - n_e$ the determinants of (4.3) and (4.4) read:

$$\begin{aligned}\det(\mathbf{M}^{+}_{sym}(\omega)) &= i\omega b e^{i\omega a x_r} - i\omega a e^{-i\omega b x_r}\\ \det(\mathbf{M}^{-}_{sym}(\omega)) &= -i\omega b e^{i\omega a x_r} - i\omega a e^{-i\omega b x_r}.\end{aligned}$$

We want to equate both determinants to zero. Since we are looking for nontrivial solutions, $\omega \neq 0$, we may cancel the $i\omega$ factors and have:

$$\begin{aligned}&\det(\mathbf{M}^{+}_{sym}(\omega)) &&= 0\\ \Leftrightarrow\quad &\tfrac{b}{a} e^{i\Re(\omega)(a+b)x_r - \Im(\omega)(a+b)x_r} &&= 1 \quad \text{and} \qquad (4.5)\\ &\det(\mathbf{M}^{-}_{sym}(\omega)) &&= 0\\ \Leftrightarrow\quad &-\tfrac{b}{a} e^{i\Re(\omega)(a+b)x_r - \Im(\omega)(a+b)x_r} &&= 1. \qquad (4.6)\end{aligned}$$

From $|e^{i\Re(\omega)(a+b)x_r}| = 1$ and (4.5) we can deduct

$$\begin{aligned}&\tfrac{b}{a} e^{-\Im(\omega)(a+b)x_r} &&= 1\\ \Rightarrow\quad &\Im(\omega) &&= -\frac{1}{(a+b)x_r}\ln\left(\frac{a}{b}\right)\\ &&&= -\frac{1}{2n_i x_r}\ln\left(\frac{n_i+n_e}{n_i-n_e}\right). \qquad (4.7)\end{aligned}$$

The same argumentation holds for (4.6). We have shown that all values ω for which (4.2a)-(4.2d) has a non-trivial solution have the same imaginary part. Inserting $\Im(\omega)$ into (4.5), we obtain for $\Re(\omega)$:

$$\begin{aligned}&e^{i\Re(\omega)a} &&= e^{-i\Re(\omega)b}\\ \Rightarrow\quad &\Re(\omega) &&= \frac{2k\pi}{(a+b)x_r}, \quad k \in \mathbb{N}\\ &&&= \frac{2k\pi}{2n_i x_r}. \qquad (4.8)\end{aligned}$$

From (4.6) we obtain by the same argumentation

$$\begin{aligned}\Re(\omega) &= \frac{(2k+1)\pi}{(a+b)x_r}, \quad k \in \mathbb{N}\\ &= \frac{(2k+1)\pi}{2n_i x_r}. \qquad (4.9)\end{aligned}$$

Putting together (4.8) and (4.9) yields

$$\Re(\omega) = \frac{k\pi}{2n_i x_r}, \quad k \in \mathbb{N}. \tag{4.10}$$

□

Remark 4.1.

1. The real part $\Re(\omega)$ is what could be expected from physical reasoning. At resonance we expect an integer number of half wavelengths, $\frac{wl}{2}$, inside a cavity of length l. Hence we have $k\frac{wl}{2} = l$ for $k \in \mathbb{N}$ which means $wl = \frac{2l}{k}$. Inserting $wl = \frac{2\pi}{n\Re(\omega)}$, we can solve for $\Re(\omega)$ which gives $\Re(\omega) = \frac{\pi k}{ln}$ for $k \in \mathbb{N}$.
2. The same result can be reached by computing the determinant of the full 4×4 matrix derived from Equations (4.2a)-(4.2d)

$$\mathbf{M}(\omega) = \begin{pmatrix} e^{-i\omega n_e x_l} & 0 & e^{-i\omega n_i x_l} & e^{i\omega n_i x_l} \\ 0 & e^{i\omega n_e x_r} & e^{-i\omega n_i x_r} & e^{i\omega n_i x_r} \\ -i\omega n_e e^{-i\omega n_e x_l} & 0 & -i\omega n_i e^{-i\omega n_i x_l} & i\omega n_i e^{i\omega n_i x_l} \\ 0 & i\omega n_e e^{i\omega n_e x_r} & -i\omega n_i e^{-i\omega n_i x_r} & i\omega n_i e^{i\omega n_i x_r} \end{pmatrix}.$$

Equating its determinant to zero gives the resonances without the assumption of symmetry:

$$\omega = \frac{k\pi}{(x_r - x_l)n_i} - i\frac{1}{(x_r - x_l)n_i} \ln\left(\frac{n_i + n_e}{n_i - n_e}\right), \quad k \in \mathbb{N}. \tag{4.11}$$

Solution with Perfectly Matched Layers

In this section, we wish to apply the perfectly matched layer (PML) method to our problem [Ber94, Zsc09]. We will give a very brief introduction of the method, which will just serve to justify the equations used in this section. The basic idea is to replace the real variable $x \in \mathbb{R}$ with a complex variable $z(x) \in \mathbb{C}$ and obtain an analytic continuation of the solution along $z(x)$. We choose $z(x)$ such that $z(x) = x$ for $x_l \leq x \leq x_r$ and in the exterior $z(x) = x + i\sigma(x - x_r)$ for $x \geq x_r$ and $z(x) = x + i\sigma(x - x_l)$ for $x \leq x_l$ where $0 < \sigma \equiv \text{const.} \in \mathbb{R}$.

Then our solutions in the interior remain unchanged,

$$u(z(x)) = u(x) = c\exp(\pm i n_i x) \text{ for } x_l \leq x \leq x_r.$$

In the exterior, our solutions become

$$\begin{aligned} u_{\text{ext},l}(z(x)) &= e^{-i\omega n_e(x + i\sigma(x - x_l))} &&= e^{-i\omega n_e[(1+i\sigma)x - i\sigma x_l]} \text{ and} \\ u_{\text{ext},r}(z(x)) &= e^{i\omega n_e(x + i\sigma(x - x_r))} &&= e^{i\omega n_e[(1+i\sigma)x - i\sigma x_r]}. \end{aligned}$$

The derivatives of the exterior solutions then read

$$\begin{aligned}
\frac{d}{dz}u_{\text{ext},l}(z(x)) &= -\frac{1}{1+i\sigma}i\omega n_e(1+i\sigma)e^{-i\omega n_e[(1+i\sigma)x-i\sigma x_l]} \\
&= -i\omega n_e e^{-i\omega n_e[(1+i\sigma)x-i\sigma x_l]} \text{ and} \\
\frac{d}{dz}u_{\text{ext},r}(z(x)) &= \frac{1}{1+i\sigma}i\omega n_e(1+i\sigma)e^{i\omega n_e[(1+i\sigma)x-i\sigma x_r]} \\
&= i\omega n_e e^{i\omega n_e[(1+i\sigma)x-i\sigma x_r]}.
\end{aligned}$$

Substituting these functions and their derivatives into Equations (4.1a)-(4.1d), we obtain a coefficient matrix

$$\mathbf{M}_{PML} = \begin{pmatrix}
e^{-i\omega n_i x_r} & e^{i\omega n_i x_r} & -e^{i\omega n_e x_r} & 0 \\
e^{-i\omega n_i x_l} & e^{i\omega n_i x_l} & 0 & -e^{-i\omega n_e x_l} \\
-i\omega n_i e^{-i\omega n_i x_r} & i\omega n_i e^{i\omega n_i x_r} & -i\omega n_e e^{i\omega n_e x_r} & 0 \\
-i\omega n_i e^{-i\omega n_i x_l} & i\omega n_i e^{i\omega n_i x_l} & 0 & i\omega n_e e^{-i\omega n_e x_l}
\end{pmatrix}.$$

It can be checked with computer algebra systems such as MAPLE, that $\det(\mathbf{M}_{PML})$ is zero for $\omega = \frac{k\pi}{(x_r-x_l)n_i} - i\frac{1}{(x_r-x_l)n_i}\ln\left(\frac{n_i+n_e}{n_i-n_e}\right)$, $k \in \mathbb{N}$, the resonances we derived from the ansatz using special functions in the previous section.

In our next step, we will truncate the PML as it is done in typical algorithms. This means, that we will have two additional equations, truncating the PML at a distance $\rho > 0$ from the boundary of the computational domain. Further, truncating the system means that we will have to allow for back-reflected waves in the exterior, which gives us the following coefficient matrix

$$\mathbf{M}_{PML}^{trunc} = \begin{pmatrix}
M_{\text{int}} & M_{\text{ext}} \\
M_{d,int} & M_{d,ext} \\
0 & M_{trunc}
\end{pmatrix}$$

with the sub-matrices

$$\begin{aligned}
M_{\text{int}} &= \begin{pmatrix} e^{-i\omega n_i x_r} & e^{i\omega n_i x_r} \\ e^{-i\omega n_i x_l} & e^{i\omega n_i x_l} \end{pmatrix}, \\
M_{\text{ext}} &= \begin{pmatrix} -e^{i\omega n_e x_r} & -e^{-i\omega n_e x_r} & 0 & 0 \\ 0 & 0 & -e^{i\omega n_e x_l} & -e^{-i\omega n_e x_l} \end{pmatrix}, \\
M_{d,int} &= \begin{pmatrix} -i\omega n_i e^{-i\omega n_i x_r} & i\omega n_i e^{i\omega n_i x_r} \\ -i\omega n_i e^{-i\omega n_i x_l} & i\omega n_i e^{i\omega n_i x_l} \end{pmatrix} \text{ and} \\
M_{d,ext} &= \begin{pmatrix} -i\omega n_e e^{i\omega n_e x_r} & i\omega n_e e^{-i\omega n_e x_r} & 0 & 0 \\ 0 & 0 & -i\omega n_e e^{i\omega n_e x_l} & i\omega n_e e^{-i\omega n_e x_l} \end{pmatrix}, \\
M_{trunc} &= \begin{pmatrix} e^{i\omega n_e(x_r+(1+i\sigma)\rho)} & -e^{-i\omega n_e(x_r+(1+i\sigma)\rho)} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \\
&\quad + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & e^{i\omega n_e(x_l-(1+i\sigma)\rho)} & -e^{-i\omega n_e(x_l-(1+i\sigma)\rho)} \end{pmatrix}.
\end{aligned}$$

If we add the equations for truncating the PML, the 4×4-matrix $\mathbf{M}_{PML}$ becomes the 6×6-matrix $\mathbf{M}_{PML}^{trunc}$. Its determinant is a polygon whose roots can not be computed exactly any more with the help of MAPLE. Hence, we resort to rasterizing the part of the complex plane containing the resonances of interest with a rectangular mesh and computing the determinant $\det\left(\mathbf{M}_{PML}^{trunc}\right)$ at each mesh point. Figure 4.2 shows a contour plot visualizing the value of $\det(\mathbf{M}_{PML}^{trunc}$ for $0 \leq \Re(\omega) \leq 10$ and $-2 \leq \Im(\omega) \leq 2$. As values for the parameters we chose $x_l = -1, x_r = 1, n_i = 2, n_e = 1, \rho = 1.3$ and $\sigma = 0.6$.

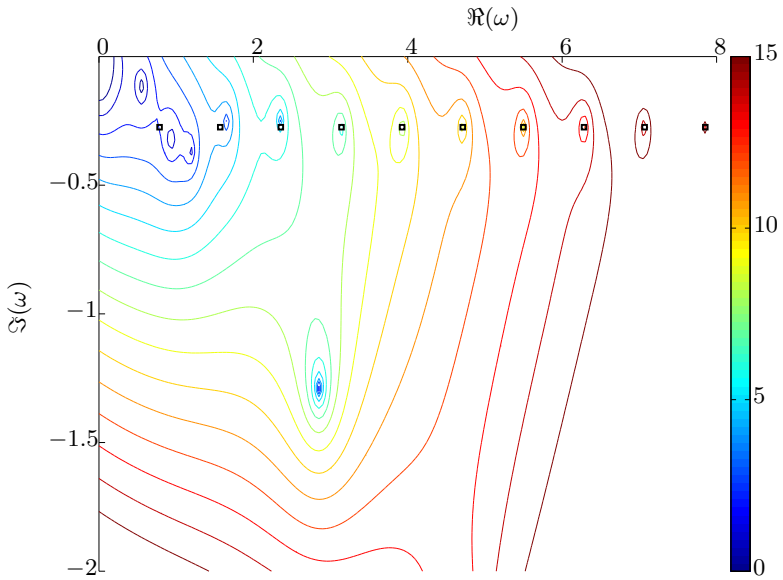

Figure 4.2: Value of $\det(\mathbf{M}_{PML})$ (black squares) and color coded contour of $\log(|\det(\mathbf{M}_{PML}^{trunc})|)$ for different values of ω. We can see dips of $|\det(\mathbf{M}_{PML}^{trunc})|$ at the verified physical resonances of the problem and additional dips, which result in spurious solutions.

Figure 4.2 shows that the truncated PML has resonances not only at the location of the physical solutions of the system, which are marked with black squares, but additional dips not corresponding to any physical solution, e.g. at $\omega \approx 2.7 - 1.3i$. These dips are spurious solutions of the system introduced by truncating the PML. This truncation translates into an approximation of the infinite exterior domain with a finite PML, which in turn introduces discretization errors, that result in spurious solutions that are not due to the discretization of the interior but purely caused by the transparent boundary condition.

Solution with the Pole Condition

Since we have seen in the previous section that the truncation of the PML causes a discretization error that results in spurious solutions, it suggests itself to perform the same heuristic for the pole condition.

For the pole condition, the implementation however is not as straightforward as it was for the PML. However, Equations (3.32a) - (3.32c), show the way to the desired implementation. We recall that the integral terms in Equation (3.32a) are the weak form of the Neumann data on the boundary, and thus combine them to $u'_{\text{int}}(x_r)$ on the right hand side boundary. Inserting this definition, we get the following set of equations for the right hand side boundary:

$$\begin{aligned} 0 &= 2s_0u'(x_r) + s_0^2\left(u(x_r) + \widetilde{a}_0\right) - \omega^2 n_r^2\left(u(x_r) - \widetilde{a}_0\right) && \text{(4.12a)} \\ 0 &= s_0^2\left(u(x_r) + 2\widetilde{a}_0 + \widetilde{a}_1\right) - \omega^2 n_r^2\left(-u(x_r) + 2\widetilde{a}_0 - \widetilde{a}_1\right) && \text{(4.12b)} \\ 0 &= s_0^2\left(\widetilde{a}_{k-2} + 2\widetilde{a}_{k-1} + \widetilde{a}_k\right) - \omega^2 n_r^2\left(-\widetilde{a}_{k-2} + 2\widetilde{a}_{k-1} - \widetilde{a}_k\right), \quad k \geq 2. && \text{(4.12c)} \end{aligned}$$

Making the ansatz $\widetilde{a}_k = z^k$, Equation (4.12c) translates to a second order equation which has two solutions

$$z_1 = \frac{n_r\omega + is_0}{n_r\omega - is_0} \text{ and } z_2 = \frac{n_r\omega - is_0}{n_r\omega + is_0}. \tag{4.13}$$

Next, we solve Equation (4.12b) for $u(x_r)$, and obtain

$$u(x_r) = -\widetilde{a}_0\left(2s_0^2 - 2n_r^2\omega^2\right)\left(s_0^2 + n_r^2\omega^2\right)^{-1}\widetilde{a}_0 - \widetilde{a}^1. \tag{4.14}$$

Inserting Equation (4.14) into Equation (4.12a), we can solve for $u'(x_r)$:

$$u'(x_r) = \frac{1}{2s_0}\left(\left(n_r^2\omega^2 - s_0^2\right)\widetilde{a}_0 - \left(n_r^2\omega^2 + s_0^2\right)\widetilde{a}_1\right). \tag{4.15}$$

The general solution to the recurrence relation in Equation (4.12c) is a superposition ($\alpha z_1 + \beta z_2$) of the roots of its characteristic polynomial given in Equation (4.13). However, since one root corresponds to an outgoing solution and the second one corresponds to an incoming solution, we can set $\alpha = 0$ at the right hand side boundary and $\beta = 0$ at the left hand side boundary. Using Equations (4.14) and (4.15) and the equivalent formulations for the left hand side boundary, we can couple to the Dirichlet and Neumann data of the interior solution, as before and obtain the matrix

$$\mathbf{M}_{pc} = \begin{pmatrix} -e^{-i\omega n_i x_l} & -e^{i\omega n_i x_l} & \frac{n_e\omega - is_0}{n_e\omega + is_0} & 0 \\ -e^{-i\omega n_i x_r} & -e^{i\omega n_i x_r} & 0 & \frac{n_e\omega + is_0}{n_e\omega - is_0} \\ i\omega n_i e^{-i\omega n_i x_l} & -i\omega n_i e^{i\omega n_i x_l} & -in_e\omega & 0 \\ i\omega n_i e^{-i\omega n_i x_r} & -i\omega n_i e^{i\omega n_i x_r} & 0 & in_e\omega \end{pmatrix} \tag{4.16}$$

Again, the resonances of the problem correspond to values of ω for which the determinant of $\mathbf{M}_{pc}$ is zero. We can not compute the determinant of $\mathbf{M}_{pc}$ directly, hence we resort to the same rastering technique we used for $\mathbf{M}_{PML}^{trunc}$. Again we used the values $n_e = 1$, $n_i = 2$, $x_l = -1$ and $x_r = 1$. For the pole condition we used a parameter of $s_0 = 0.1 - 0.3i$. The result is shown in Figure 4.3.

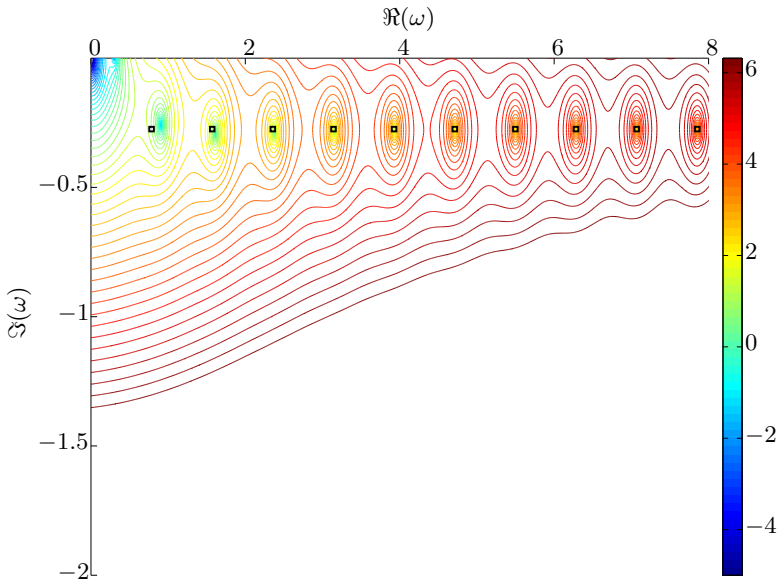

Figure 4.3: Physical resonances of a one-dimensional cavity (black squares) and color coded contour of $\log(|\det(\mathbf{M}_{pc}0)|)$ for different values of ω.

Next we will introduce a cutoff in the equations for the pole condition. This means, that we set $z_1^k = 0$ for $k \geq L$ or $z_2^k = 0$ for $k \geq L$ respectively. In order to achieve this cutoff, we can no longer set $\beta = 0$ or $\alpha = 0$ in the superposition in the exterior and we will have to set $\alpha z_1^L = -\beta z_2^L$. Hence, the matrix coupling the analytic interior to the truncated version of the pole condition reads

$$\mathbf{M}_{pc}^{trunc} = \begin{pmatrix} M_{\text{int}} & M_{\text{ext}} \\ M_{d,int} & M_{d,ext} \\ 0 & M_{trunc} \end{pmatrix}.$$

Again we use the interior sub matrices M_{int} and $M_{d,int}$ from the previous

sections. The remaining sub matrices are

$$
\begin{aligned}
M_{\text{ext}} &= \begin{pmatrix} \frac{n_e\omega - is_0}{n_e\omega + is_0} & \frac{n_e\omega + is_0}{n_e\omega - is_0} & 0 & 0 \\ 0 & 0 & \frac{n_e\omega + is_0}{n_e\omega - is_0} & \frac{n_e\omega - is_0}{n_e\omega + is_0} \end{pmatrix}, \\
M_{d,ext} &= \begin{pmatrix} -in_e\omega & in_e\omega & 0 & 0 \\ 0 & 0 & in_e\omega & -in_e\omega \end{pmatrix} \text{ and} \\
M_{trunc} &= \begin{pmatrix} \left(\frac{n_e\omega - is_0}{n_e\omega + is_0}\right)^L & -\left(\frac{n_e\omega + is_0}{n_e\omega - is_0}\right)^L & 0 & 0 \\ 0 & 0 & \left(\frac{n_e\omega + is_0}{n_e\omega - is_0}\right)^L & -\left(\frac{n_e\omega - is_0}{n_e\omega + is_0}\right)^L \end{pmatrix}.
\end{aligned}
$$

Again, we can not compute the determinant directly, so we resort to the rastering technique again. Figure 4.4 shows a contour plot of $\log(|\det(M_{pc}^{trunc})|)$ using the same parameters as before and $L = 15$ degrees of freedom. We can see again, that truncating the infinite series that implements the pole condition introduces new minima in the contour. These new minima are again eigenvalues of the truncated operator that do not correspond to any physical solutions, which implies that they are spurious solutions.

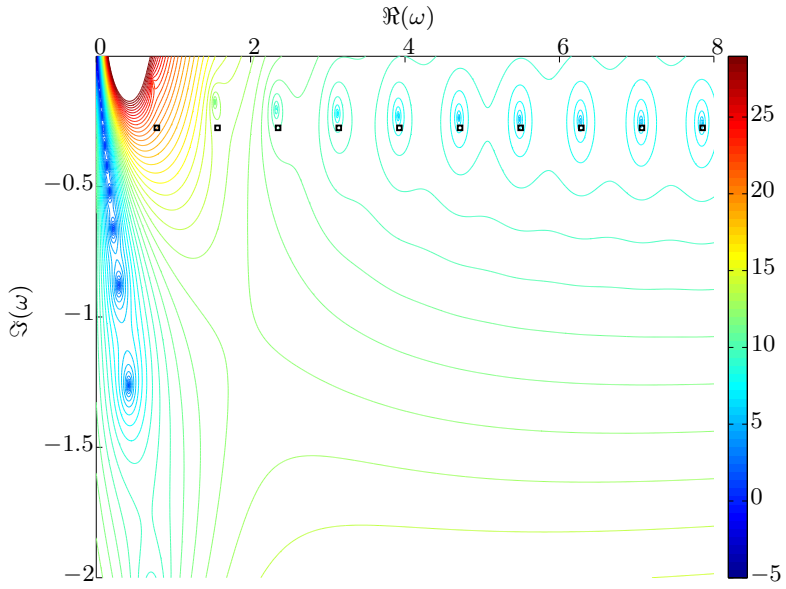

Figure 4.4: Physical resonances of a one-dimensional cavity (black squares) and color coded contour of $\log(|\det(\mathbf{M}_{pc}^{trunc})|)$ for different values of ω.

We have seen in the last sections that independent of the transparent boundary condition we use, whenever we make the infinite exterior domain finite, we introduce a discretization error for our simple one-dimensional problem. This discretization error results in new resonances that lack any

physical meaning and thus are spurious solutions. Given the fact that the cause for spurious solutions caused by the interior discretization has been thoroughly investigated and solved, we will devote the following chapter to distinguishing these spurious solutions from the physical solutions of a problem. For this we will make use of the observation that the spurious solutions are caused by the transparent boundary conditions and the further assumption that they therefore are more sensitive to perturbations of the boundary condition.

Chapter 5

Detecting Spurious Solutions

Since our observation from the previous chapter is that spurious solutions are caused by badly converged solutions in the exterior domain, they respond more strongly to perturbations of this exterior domain than the physical solutions of a problem. We will make use of these results by presenting a method for detecting the spurious solutions within the computed eigenvalue spectrum with a robust algorithm. Our basic idea is to investigate the dependence of the eigenvalues on the pole condition parameter s_0.

Since the reaction of quantities of interest to perturbations is typically determined by condition numbers, we will first give a brief review of perturbation theory for generalized eigenvalue problems. In Section 5.1 we will therefore give the basic definitions and derive the condition numbers of eigenvalues. In the subsequent section, we will explore the usefulness of these condition numbers for the detection of spurious solutions and their limitations. In Section 5.3 we will make use of the fact that we need not deal with an arbitrary perturbation but with a perturbation that is well-defined. This will allow us to directly compute the reaction of the eigenvalues to variations of the pole condition parameter s_0.

Finally we will investigate the domains of convergence for our method that will allow us to implement a convergence monitor that gives us regions where the statements derived in this chapter produce reliable results.

Throughout this chapter we will make extensive use of the terms "eigenvalue" and "spectrum" of a generalized Eigenvalue problem which are defined as follows:

Definition 5.1.

Let $A, B \in \mathbb{C}^{n\times n}$ be complex n by n matrices. We call $\lambda \in \mathbb{C}$ an *eigenvalue*, $\mathbf{u} \in \mathbb{C}^n$ a *(right) eigenvector* and the pair $(\lambda, \mathbf{u})$ a *(right) eigenpair* of the *matrix pair* (A, B) if

- $\mathbf{u} \neq 0$ and
- $(A - \lambda B)\mathbf{u} = 0$.

$\mathbf{v} \neq 0$ is called a *left eigenvector*, if $\mathbf{v}^H(A-\lambda B) = 0$. We will refer to the set of all eigenvalues of a matrix pair as the *spectrum of the pair* (A, B): $\sigma(A, B) := \{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } (A, B)\}$.

5.1 Generalized Eigenvalue Problems

Even though the literature on perturbation of ordinary eigenvalue problems and computing their condition numbers is manifold, the situation for generalized eigenvalue problems of the type $(A - \lambda B)\mathbf{u} = 0$ is more involved and often restricted to special cases (see e.g. [SS90, GL96, Wil88, Ste01]). More general approaches to the problem of the sensitivity of eigenvalues of the generalized eigenvalue problem typically make use of deflating subspaces [Ste72]. However, for our purposes, we may restrict ourselves to *regular matrix pairs* (cf. Definition 5.2). In this section we will give an overview on condition numbers for regular generalized eigenvalue problems. It is mostly based on Stewart and Suns work on the sensitivity of eigenvalue problems [SS90, pp. 271–324].

There are some reasons why generalized eigenvalue problems differ from ordinary eigenvalue problems and why their perturbation theory is more involved. In the first place, it is possible for $\det(A - \lambda B)$ to be identically zero independent of λ. We call such matrix pairs where each scalar λ can be regarded as an eigenvalue *singular matrix pairs*.

Secondly it is possible for B to be singular in which case B has a null vector $\mathbf{u}_0 \neq 0$. Rewriting the problem in the reciprocal form $B\mathbf{u}_0 = \lambda^{-1}A\mathbf{u}_0$, we see that $B\mathbf{u}_0 = 0A\mathbf{u}_0$, hence $\mathbf{u}_0$ is an eigenvector of the reciprocal problem corresponding to the eigenvalue $\lambda^{-1} = 0$, i.e. $\lambda = \infty$.

For a formal definition of the eigenvalue of a generalized eigenvalue problem that also accounts for infinite eigenvalues, we switch from the asymmetric treatment of A and B to an equivalent symmetric formulation by replacing $\lambda = \frac{\alpha}{\beta}$. We then have

$$(A - \lambda B)\mathbf{u} = 0 \Leftrightarrow (A - \frac{\alpha}{\beta}B)\mathbf{u} = 0 \Leftrightarrow (\beta A - \alpha B)\mathbf{u} = 0.$$

Using this symmetric formulation, we can make the following definitions:

Definition 5.2.

Let A and B be square complex matrices of order n.

- The matrix pair (A, B) is called *singular* if for all (α, β), $\det(\beta A - \alpha B) = 0$. Otherwise (A, B) is called *regular matrix pair*.
- If (A, B) is regular and $\beta A\mathbf{u} = \alpha B\mathbf{u}$ for $(\alpha, \beta) \neq (0, 0)$ and $\mathbf{u} \neq 0$ then for $\tau \in \mathbb{C}$, $\tau\beta A\mathbf{u} = \tau\alpha B\mathbf{u}$, hence we refer to the entire

subspace spanned by $(\alpha, \beta)^\top$ as the eigenvalue of (A, B) and write: $\langle \alpha, \beta \rangle := \{\tau(\alpha, \beta)^\top : \tau \in \mathbb{C}\}$.

- To preserve the connection with the ordinary eigenvalue problem, we write $\langle \lambda \rangle$ for $\langle \lambda, 1 \rangle$. Furthermore we define $\langle \infty \rangle := \langle 1, 0 \rangle$.

Remark 5.1.

Two observations follow directly from these definitions:

1. Infinite eigenvalues are not just special cases that can be ignored in perturbation theory. This can be seen by rewriting the eigenvalue problem in the cross-product form $\beta A\mathbf{u} = \alpha B\mathbf{u}$. In this form, an infinite eigenvalue corresponds to a nonzero pair (α, β) with $\beta = 0$ which is not essentially different from the case $\alpha = 0$, i.e. $\lambda = 0$.
2. Since $\mathbf{u} \in \ker A \cap \ker B \Leftrightarrow (\beta A - \alpha B)\mathbf{u} = 0 \quad \forall (\alpha, \beta)$, the matrix pair (A, B) is singular, if and only if $\ker A$ and $\ker B$ have a nonempty intersection.

For a better understanding of the differences between ordinary and generalized eigenvalue problems and of the notion *eigenvalue of the matrix pair* (A, B), we will consider the following comparison: if $(\lambda, \mathbf{u})$ is a (right) eigenpair of the matrix A, then $A\mathbf{u} = \lambda\mathbf{u}$. That means that the direction of $\mathbf{u}$ remains invariant under multiplication by A provided we agree that the direction of the zero vector matches that of any nonzero vector. On the other hand if $(\lambda, \mathbf{u})$ is an eigenpair of the matrix pair (A, B), then $A\mathbf{u} = \lambda B\mathbf{u}$. This means that the direction of $\mathbf{u}$ is not necessarily preserved by multiplication with A and B. Instead, the direction of $A\mathbf{u}$ and $B\mathbf{u}$ are the same. This is illustrated in Figure 5.1.

As for the ordinary eigenvalue problem, we can define the *characteristic equation* of (A, B) as $\det(A - \lambda B) = 0$. The eigenvalues of the pair (A, B) satisfy the characteristic equation. When B is singular, the characteristic equation will have degree less than n. The missing eigenvalues are the infinite ones. Hence, if B is singular, the matrix pair (A, B) has infinite eigenvalues. If B was non-singular, the eigenvalues of (A, B) would behave like the eigenvalues of the ordinary eigenvalue problem $B^{-1}A\mathbf{u} = \lambda\mathbf{u}$ making it possible to apply the perturbation theory known from ordinary eigenvalue problems. In our case, the matrix B that we have to compute is singular or at least very ill-conditioned, thus we cannot resort to this simplification. Consequently we will have to deal with the concept of infinite eigenvalues and develop a perturbation theory for generalized eigenvalue problems.

However the matrix pairs that we will have to deal with are regular. Thus the characteristic polynomial $\det(A - \lambda B)$ is not identically zero. As a

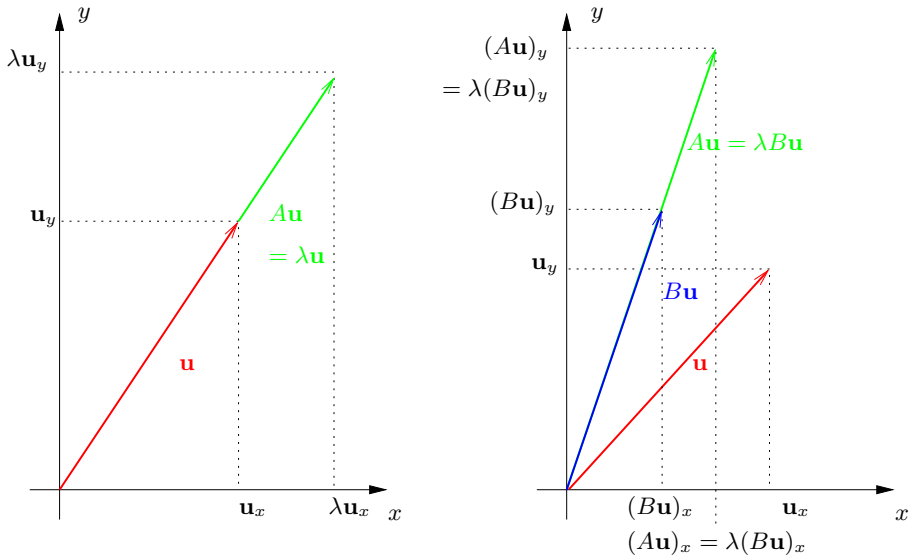

Figure 5.1: Left: For the ordinary eigenvalue problem $A\mathbf{u} = \lambda\mathbf{u}$, the direction of $\mathbf{u}$ is invariant under multiplication by A, only the length is changed. Right: For the generalized eigenvalue problem $A\mathbf{u} = \lambda B\mathbf{u}$, the direction of $A\mathbf{u}$ (green) is equal to the direction of $B\mathbf{u}$ (blue).

consequence, there is an established perturbation theory which is applicable to our problem and which we will present in the following paragraphs.

Since the generalized eigenvalue in the cross-product form is a subspace $\langle \alpha, \beta \rangle$ as mentioned in Definition 5.2, we have to introduce the distance between two such subspaces $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$. The *chordal distance* is such a measure. It is required for a perturbation theory and defined as follows:

Definition 5.3.

The *chordal distance* between $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$ is the number

$$\mathcal{X}(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle) := \rho_g(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle).$$

The function ρ_g is the gap metric defining the distance between the two subspaces $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$:

$$\rho_g(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle) := \max \left\{ \max_{\substack{(\alpha,\beta) \in \langle \alpha_1, \beta_1 \rangle \\ |(\alpha,\beta)|=1}} |(\alpha, \beta) - \langle \alpha_2, \beta_2 \rangle|, \max_{\substack{(\alpha,\beta) \in \langle \alpha_2, \beta_2 \rangle \\ |(\alpha,\beta)|=1}} |\langle \alpha_1, \beta_1 \rangle - (\alpha, \beta)| \right\}.$$

Inserting the definitions for $\langle \alpha_1, \beta_1 \rangle$ and $\langle \alpha_2, \beta_2 \rangle$, we can easily evaluate $\mathcal{X}(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle)$ as

$$\mathcal{X}(\langle \alpha_1, \beta_1 \rangle, \langle \alpha_2, \beta_2 \rangle) = \frac{|\alpha_1 \beta_2 - \beta_1 \alpha_2|}{\sqrt{|\alpha_1|^2 + |\beta_1|^2}\sqrt{|\alpha_2|^2 + |\beta_2|^2}}.$$

In this notation, we can see that

$$\mathcal{X}(\langle \infty \rangle, \langle \lambda \rangle) = \mathcal{X}(\langle 1, 0 \rangle, \langle \lambda, 1 \rangle) = \frac{1}{1 + \sqrt{|\lambda|^2 + 1}}.$$

Thus the chordal metric behaves counter-intuitive by regularizing the point at infinity and making it no more than unit distance from any other point. Returning to the conventional notation by setting $\lambda_1 = \frac{\alpha_1}{\beta_1}$ and $\lambda_2 = \frac{\alpha_2}{\beta_2}$, we have

$$\mathcal{X}(\langle \lambda_1 \rangle, \langle \lambda_2 \rangle) = \frac{|\lambda_1 - \lambda_2|}{\sqrt{1 + |\lambda_1|^2} + \sqrt{1 + |\lambda_2|^2}} \leq 1.$$

The perturbation theory of our problem would be dramatically simplified, if we could rewrite $A\mathbf{u} = \lambda B\mathbf{u}$ in the form $B^{-1}A\mathbf{u} = \lambda \mathbf{u}$. Then we would have reduced the generalized Eigenvalue problem to an ordinary eigenvalue problem, however, since B is singular or ill-conditioned in our application, this reduction is not possible. However, a way to deal with this situation exists in the form of *generalized shifting*. The definition of this shifting in Lemma 5.1 corresponds to that of [SS90, VI, Theorem 1.6].

Lemma 5.1.

Let W be a 2×2 nonsingular matrix $W = \begin{pmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{pmatrix}$. Given the matrix pair (A, B) set

$$(C, D) = (w_{1,1}A + w_{2,1}B, w_{1,2}A + w_{2,2}B) =: (A, B)(W \otimes I). \quad (5.1)$$

Given $(\alpha, \beta) \neq (0, 0)$, define

$$\begin{pmatrix} a \\ -b \end{pmatrix} = W^{-1} \begin{pmatrix} \beta \\ -\alpha \end{pmatrix}. \quad (5.2)$$

Then $\langle \alpha, \beta \rangle$ is an eigenvalue of (A, B) if and only if $\langle a, b \rangle$ is an eigenvalue of (C, D).

Proof.

The proof is purely computational. Using $W = \begin{pmatrix} w_{1,1} & w_{1,2} \\ w_{2,1} & w_{2,2} \end{pmatrix}$ and its inverse $W^{-1} = (w_{1,1}w_{2,2} - w_{2,1}w_{1,2})^{-1} \begin{pmatrix} w_{2,2} & -w_{1,2} \\ -w_{2,1} & w_{1,1} \end{pmatrix}$, we can compute

$$\begin{aligned} a &= \frac{1}{w_{1,1}w_{2,2} - w_{2,1}w_{1,2}}(w_{2,1}\beta + w_{1,1}\alpha) \\ b &= \frac{1}{w_{1,1}w_{2,2} - w_{2,1}w_{1,2}}(w_{2,2}\beta + w_{1,2}\alpha). \end{aligned}$$

Now we can prove the equivalence of $(\beta A - \alpha B)\mathbf{u} = 0$ and $(bC - aD)\mathbf{u} = 0$ by inserting C, D, a and b into the second equation and transforming equivalently to obtain the first equation.

□

From this lemma we can directly deduct the following corollary:

Corollary 5.1.

If (A, B) is a regular pair, there exists a 2×2 matrix W such that for C and D defined as in Lemma 5.1 D is nonsingular.

Proof.

If (A, B) is a regular pair, there are constants σ and τ such that $\det(\tau A - \sigma B) \neq 0$. That means, $\tau A - \sigma B$ is nonsingular. If we set

$$W = \begin{pmatrix} \sigma & \tau \\ \tau & -\sigma \end{pmatrix}, \tag{5.3}$$

then W is nonsingular. That means, we can define (C, D) as in Lemma 5.1 such that the eigenvalues of (A, B) and (C, D) are in one-to-one correspondence and D is nonsingular.

□

We are now in a position to treat the perturbation of the eigenvalues of matrix pairs. Let (A, B) be a complex matrix pair of order n and $(\widetilde{A}, \widetilde{B}) := (A + \Delta A, B + \Delta B)$ be the perturbed pair with perturbations ΔA and ΔB. Our goal is to derive a first order expansion for the eigenvalues of the perturbed system. First we will need a measure for the perturbation of the pair (A, B). As such a measure we will fix $\varepsilon := \sqrt{\|\Delta A\|_2^2 + \|\Delta B\|_2^2}$.

First we will follow [SS90, IV, Theorem 2.1] and show the continuity of the eigenvalues of a regular matrix pair under small perturbations, which will then allow us to investigate the sensitivity of the eigenvalues with respect to a perturbation of the matrix pair.

Theorem 5.1.

Let (A, B) be a regular matrix pair of order n and let $\langle\lambda_1\rangle, ...\langle\lambda_n\rangle$ be its eigenvalues. Then there exists an ordering $\langle\widetilde{\lambda}_1\rangle, ...\langle\widetilde{\lambda}_n\rangle$ of the eigenvalues of the perturbed matrix pair $(\widetilde{A}, \widetilde{B})$ such that

$$\lim_{\epsilon\to 0} \mathcal{X}(\langle\widetilde{\lambda}_i\rangle, \langle\lambda_i\rangle) = 0, i = 1, ..., n.$$

Proof.

By Lemma 5.1 and Corollary 5.1 there is a 2×2 matrix W such that the matrix D in $(C, D) = (A, B)(W \otimes I)$ is nonsingular. Let $\mu_1, ..., \mu_n$ be the eigenvalues of $D^{-1}C$ and let $(\widetilde{C}, \widetilde{D}) = (\widetilde{A}, \widetilde{B})(W \otimes I)$. For ε sufficiently small, $\widetilde{D}$ is nonsingular. By the continuity of the eigenvalues of an ordinary eigenvalue problem (cf. [Ste01, pp. 37–38]) we know that there is an ordering of $\widetilde{\mu}_1, ..., \widetilde{\mu}_n$, the eigenvalues of $\widetilde{D}^{-1}\widetilde{C}$ such that $\lim_{\epsilon\to 0} \widetilde{\mu}_i \to \mu_i$ for $i = 1, ..., n$.

□

Now we can proceed by analyzing the sensitivity of the eigenvalues with respect to a perturbation of (A, B). We will do this by presenting a first order perturbation theory. Let $\langle\alpha, \beta\rangle$ be a simple eigenvalue of (A, B). We will first show that a first order expansion exists. This is evident since by Corollary 5.1 we may assume that B is nonsingular. If ε is sufficiently small, the perturbed matrix $\widetilde{B} = B + \Delta B$ is also nonsingular and there exists an eigenvalue $\widetilde{\lambda} = \lambda + \mathcal{O}(\varepsilon)$ of $\widetilde{B}^{-1}\widetilde{A}$ that corresponds to the eigenvalue λ of $B^{-1}A$. By the theory that holds for ordinary eigenvalue problems (cf. e.g. [Ste01, p. 47]), $\widetilde{\lambda}$ is differentiable in the elements of $\widetilde{B}^{-1}\widetilde{A}$ and thus differentiable in the elements of A and B. It follows that $\langle\widetilde{\lambda}\rangle$ is the required first order expansion. To give a first order expansion for the perturbed eigenvalue we need the following prerequisite:

Theorem 5.2.

Let $\langle\alpha, \beta\rangle$ be a simple eigenvalue of the regular pair (A, B). If $\mathbf{u}$ and $\mathbf{v}$ are the right and left eigenvectors corresponding to $\langle\alpha, \beta\rangle$, then $\langle\alpha, \beta\rangle = \langle\mathbf{v}^H A\mathbf{u}, \mathbf{v}^H B\mathbf{u}\rangle$.

A proof for this theorem can be found e.g. at Stewart [Ste01, 2, Theorem 4.9]. For the desired first order expansion of $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle$ we require another result that describes the effect of an $\mathcal{O}(\varepsilon)$ perturbation of $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle$ in the direction of $\langle\alpha, \beta\rangle$.

Lemma 5.2.

Let $\langle\widetilde{\alpha},\widetilde{\beta}\rangle$ be an $\mathcal{O}(\varepsilon)$ perturbation of $\langle\alpha,\beta\rangle$ in the chordal metric. Let $\phi(\varepsilon)=\mathcal{O}(\varepsilon)$. Then $\mathcal{X}(\langle\widetilde{\alpha}+\phi(\varepsilon)\alpha,\widetilde{\beta}+\phi(\varepsilon)\beta\rangle,\langle\widetilde{\alpha},\widetilde{\beta}\rangle)=\mathcal{O}(\varepsilon^2)$.

Proof.

$$\begin{aligned}
&\mathcal{X}(\langle\widetilde{\alpha},\widetilde{\beta}\rangle,\langle\widetilde{\alpha}+\phi(\varepsilon)\alpha,\widetilde{\beta}+\phi(\varepsilon)\beta\rangle)\\
&=\frac{|\widetilde{\alpha}(\widetilde{\beta}+\phi(\varepsilon)\beta)-\widetilde{\beta}(\widetilde{\alpha}+\phi(\varepsilon)\alpha|}{\sqrt{|\widetilde{\alpha}|^2+|\widetilde{\beta}|^2}\sqrt{|(\widetilde{\alpha}+\phi(\varepsilon)\alpha|^2+|\widetilde{\beta}+\phi(\varepsilon)\beta|^2}}\\
&=\frac{|\phi(\varepsilon)(\widetilde{\alpha}\beta-\widetilde{\beta}\alpha)|}{\sqrt{|\widetilde{\alpha}|^2+|\widetilde{\beta}|^2}\sqrt{|(\widetilde{\alpha}+\phi(\varepsilon)\alpha|^2+|\widetilde{\beta}+\phi(\varepsilon)\beta|^2}},
\end{aligned}$$

the denominator of which is $\mathcal{O}(\varepsilon^2)$.

□

With Lemma 5.2, we have assembled all the preliminaries required to give a first order expansion for the perturbed eigenvalue $\langle\widetilde{\alpha},\widetilde{\beta}\rangle$, which will then in turn yield the relative condition number of an eigenvalue in the following theorem which is a slight modification of [SS90, IV, Theorem 2.2].

Theorem 5.3.

Let $\mathbf{u}$ and $\mathbf{v}$ be the right and left eigenvectors for the simple eigenvalue $\langle\alpha,\beta\rangle=\langle\mathbf{v}^HA\mathbf{u},\mathbf{v}^HB\mathbf{u}\rangle$ of the regular matrix pair (A,B). Let $(\widetilde{A},\widetilde{B})=(A+\Delta A,B+\Delta B)$ be the perturbed pair, $\varepsilon=\sqrt{\|\Delta A\|_2^2+\|\Delta B\|_2^2}$ and $\langle\widetilde{\alpha},\widetilde{\beta}\rangle$ be the perturbed eigenvalue corresponding to $\langle\alpha,\beta\rangle$. Then

$$\langle\widetilde{\alpha},\widetilde{\beta}\rangle=\langle\alpha+\mathbf{v}^H\Delta A\mathbf{u},\beta+\mathbf{v}^H\Delta B\mathbf{u}\rangle+\mathcal{O}(\varepsilon^2). \tag{5.4}$$

Proof.

From [Ste01, Theorem 3.13, Chapter 1] we know that for an eigenvector ξ of the ordinary eigenvalue problem there exists a perturbed eigenvector $\widetilde{\xi}$ of the eigenvalue problem perturbed by $\mathcal{O}(\varepsilon)$ and that $\widetilde{\xi}$ satisfies $\sin(\angle(\xi,\widetilde{\xi}))=\mathcal{O}(\varepsilon)$. Normalizing ξ and $\widetilde{\xi}$, this implies that $\widetilde{\xi}=\xi+\mathcal{O}(\varepsilon)$.

By Corollary 5.1 we may assume that B is nonsingular, hence we can apply these results to $B^{-1}A$ and AB^{-1}. Thus we can take $\widetilde{\mathbf{u}}=\mathbf{u}+\Delta\mathbf{u}$

and $\widetilde{\mathbf{v}} = \mathbf{v} + \Delta\mathbf{v}$ with $\Delta\mathbf{u}, \Delta\mathbf{v} = \mathcal{O}(\varepsilon)$ as left and right eigenvectors corresponding to $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle$. Using Theorem 5.2, we have

$$\begin{aligned}\widetilde{\alpha} &= \widetilde{\mathbf{v}}^H \widetilde{A}\widetilde{\mathbf{x}} \\ &= (\mathbf{v}^H + \Delta\mathbf{v}^H)(A + \Delta A)(\mathbf{u} + \Delta\mathbf{u}) \\ &= \mathbf{v}^H(A + \Delta A)\mathbf{u} + \mathbf{v}^H A\Delta\mathbf{u} + \Delta\mathbf{v}^H A\mathbf{u} + \mathcal{O}(\varepsilon^2)\end{aligned}$$

and

$$\begin{aligned}\widetilde{\beta} &= \widetilde{\mathbf{v}}^H \widetilde{B}\widetilde{\mathbf{u}} \\ &= (\mathbf{v}^H + \Delta\mathbf{v}^H)(B + \Delta B)(\mathbf{u} + \Delta\mathbf{u}) \\ &= \mathbf{v}^H(B + \Delta B)\mathbf{u} + \mathbf{v}^H B\Delta\mathbf{u} + \Delta\mathbf{v}^H B\mathbf{u} + \mathcal{O}(\varepsilon^2).\end{aligned}$$

Since (A, B) is regular, at least one of α or β must be nonzero, say $\beta \neq 0$. Then

$$\mathbf{v}^H B\Delta\mathbf{u} + \Delta\mathbf{v}^H B\mathbf{u} = \beta\frac{\mathbf{v}^H B\Delta\mathbf{u} + \Delta\mathbf{v}^H B\mathbf{u}}{\beta}$$

and because $\mathbf{u}$ and $\mathbf{v}$ are right and left eigenvectors of (A, B), $\beta A\mathbf{u} = \alpha B\mathbf{u}$ and $\mathbf{v}^H\beta A = \mathbf{v}^H\alpha B$, thus

$$\mathbf{v}^H A\Delta\mathbf{u} + \Delta\mathbf{v}^H A\mathbf{u} = \alpha\frac{\mathbf{v}^H B\Delta\mathbf{u} + \Delta\mathbf{v}^H B\mathbf{u}}{\beta}.$$

That means, that $(\mathbf{v}^H A\Delta\mathbf{u} + \Delta\mathbf{v}^H A\mathbf{u}, \mathbf{v}^H B\Delta\mathbf{u} + \Delta\mathbf{v}^H B\mathbf{u})$ is an $\mathcal{O}(\varepsilon)$ perturbation of $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle$. By Lemma 5.2, deleting these terms induces an error of $\mathcal{O}(\varepsilon^2)$ and we end up with $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle = \langle\alpha + \mathbf{v}^H\Delta A\mathbf{u} + \mathcal{O}(\varepsilon^2), \beta + \mathbf{v}^H\Delta B\mathbf{u} + \mathcal{O}(\varepsilon^2)\rangle$ or in a shorter notation $\langle\widetilde{\alpha}, \widetilde{\beta}\rangle = \langle\alpha + \mathbf{v}^H\Delta A\mathbf{u}, \beta + \mathbf{v}^H\Delta B\mathbf{u}\rangle + \mathcal{O}(\varepsilon^2)$.

□

The results we have derived so far allow us to compute the relative condition number of an eigenvalue. First using $\alpha = \mathbf{v}^H A\mathbf{u}$, $\beta = \mathbf{v}^H B\mathbf{u}$, $\widetilde{\alpha} = \alpha + \mathbf{v}^H\Delta A\mathbf{u}$ and $\widetilde{\beta} = \beta + \mathbf{v}^H\Delta B\mathbf{u}$, we may compute

$$\begin{aligned}\mathcal{X}(\langle\alpha, \beta\rangle, \langle\widetilde{\alpha}, \widetilde{\beta}\rangle) &= \frac{|\alpha\widetilde{\beta} - \beta\widetilde{\alpha}|}{\sqrt{|\alpha|^2 + |\beta|^2}\sqrt{|\widetilde{\alpha}|^2 + |\widetilde{\beta}|^2}} \\ &\approx \frac{|\alpha\mathbf{v}^H\Delta B\mathbf{u} - \beta\mathbf{v}^H\Delta Ax|}{|\alpha|^2 + |\beta|^2}\end{aligned} \tag{5.5}$$

The numerator in equation (5.5) can be rewritten as

$$\begin{aligned}|\alpha\mathbf{v}^H\Delta B\mathbf{u} - \beta\mathbf{v}^H\Delta A\mathbf{u}| = &\left|(\alpha, \beta)\begin{pmatrix}\mathbf{v}^H\Delta B\mathbf{u} \\ -\mathbf{v}^H\Delta A\mathbf{u}\end{pmatrix}\right| \\ \leq\ &\varepsilon\|\mathbf{u}\|_2\|\mathbf{v}\|_2\sqrt{|\alpha|^2 + |\beta|^2}.\end{aligned} \tag{5.6}$$

Inserting equation (5.6) into (5.5), we have

$$\mathcal{X}(\langle\alpha,\beta\rangle,\langle\widetilde{\alpha},\widetilde{\beta}\rangle) \lesssim \frac{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}{\sqrt{|\alpha|^2+|\beta|^2}}\varepsilon. \tag{5.7}$$

Now, we will summarize these results in the following theorem giving the relative condition number for a generalized eigenvalue problem, which is for example also stated in [Ste01, 2, Theorem 4.12]:

Theorem 5.4.
Let $\langle\alpha,\beta\rangle$ be a simple eigenvalue of (A,B) and let $\mathbf{u}$ and $\mathbf{v}$ be the right and left eigenvectors corresponding to $\langle\alpha,\beta\rangle$. Let $\widetilde{A}=A+\Delta A,\widetilde{B}=B+\Delta B$ be a perturbation of (A,B) and set $\varepsilon=\sqrt{\|\Delta A\|_F^2+\|\Delta B\|_F^2}$. If ε is sufficiently small, there is an eigenvalue $\langle\widetilde{\alpha},\widetilde{\beta}\rangle$ of $(\widetilde{A},\widetilde{B})$ such that for the chordal distance between $\langle\alpha,\beta\rangle$ and $\langle\widetilde{\alpha},\widetilde{\beta}\rangle$ holds

$$\mathcal{X}(\langle\alpha,\beta\rangle,\langle\widetilde{\alpha},\widetilde{\beta}\rangle) = \kappa_{rel}\left(\langle\alpha,\beta\rangle\right)\varepsilon+\mathcal{O}(\varepsilon^2)$$

with

$$\kappa_{rel}(\langle\alpha,\beta\rangle) = \frac{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}{\sqrt{|\alpha|^2+|\beta|^2}}.$$

Remark 5.2.
We would like to make the following remarks about Theorem 5.4:

1. The number $\kappa_{rel}\left(\langle\alpha,\beta\rangle\right)$ defined in Theorem 5.4 is analogous to the condition number $\kappa_{rel}(\lambda)$ for the eigenvalue λ of the ordinary eigenvalue problem defined e.g. by [SS90, IV.2.8]. It serves the role of a *relative* condition number for its eigenvalue of the generalized eigenvalue problem.
2. Since $\alpha=\mathbf{v}^HA\mathbf{u}$ and $\beta=\mathbf{v}^HB\mathbf{u}$, $\kappa_{rel}\left(\langle\alpha,\beta\rangle\right)$, a rescaling of $\mathbf{u}$ and $\mathbf{v}$ is cancelled. Therefore $\kappa_{rel}\left(\langle\alpha,\beta\rangle\right)$ is independent of the scaling of $\mathbf{u}$ and $\mathbf{v}$.
3. The first order bound on the perturbation of an eigenvalue does not reduce to the first order bound for the ordinary eigenvalue problem for $B=I$ and $\Delta B=0$. However, for a parametrization $A=\tau A,\lambda=\tau\lambda$ and $\Delta A=\tau\Delta A$ it can be shown that the condition number $\kappa_{rel,\tau}(\langle\alpha,\beta\rangle)\to\|\mathbf{u}\|_2\|\mathbf{v}\|_2/|\mathbf{v}^H\mathbf{u}|$ as $\tau\to0$ which gives the usual bound for ordinary eigenvalue problems.
4. If $\|\mathbf{u}\|_2=\|\mathbf{v}\|_2=1$, κ_{rel} is large when α and β are small. This was to be expected from the first order expansion

   $$\langle\widetilde{\alpha},\widetilde{\beta}\rangle\approx\langle\alpha+\mathbf{v}^H\Delta A\mathbf{u},\beta+\mathbf{v}^H\Delta B\mathbf{u}\rangle.$$

   Hence, if both α and β are small, they are sensitive to the perturbation ΔA and ΔB, i.e. ill-conditioned.

5. In order to obtain an implementation of this method, we solve the generalized eigenvalue problem $(A - \lambda B)\mathbf{u} = 0$ with the standard generalized sparse eigenvalue solver in MATLAB, which is an implementation of the Arnoldi method with spectral deformation (see [GL96, Saa80]). This produces the spectrum $\sigma(A, B)$ together with the right eigenvectors. Since a left eigenvector $\mathbf{v}$ satisfies $\mathbf{v}^H(A - \lambda B) = 0$, we can obtain $\mathbf{v}$ by solving the hermitian conjugate of the problem $(A^H - \overline{\lambda} B^H)\mathbf{v} = 0$.

In the next section we will see how the condition number for a generalized eigenvalue problem may be used for detecting spurious solutions and what limitations this method is subjected to.

5.2 Use and Limitations of Condition Numbers

We will now cover the detection of spurious solutions of the Helmholtz resonance problem on unbounded domains via the condition number. We will see in the next paragraphs that this method is error-prone and we will analyze its problems. This sets the stage for the computation of the exact perturbation which we will develop in Section 5.3. This section is based on two one-dimensional examples for the detection of spurious solutions via their condition number. While this works well for the first example, we will run into some difficulties for the more complicated second example.

For our examples we will revert to using ω as eigenvalue instead of λ. This will make the distinction between general statements for eigenvalues, that were given in terms of λ and results for the Helmholtz equation that will be given in terms of ω easier. Since κ_{rel} is a relative condition number, we can expect higher eigenvalues to have lower condition numbers. This may seem unintuitive from a physical point of view since the approximation of higher eigenvalues is typically worse for a fixed grid. However from a purely algebraic viewpoint, this is in good agreement with the notion *relative condition number* since we cannot expect the effect of a small perturbation to increase for bigger eigenvalues. This means that lower physical modes may have higher condition numbers than higher order spurious solutions. The comparison of condition numbers in order to detect spurious solutions is therefore not reliable in a global setting. In order to obtain a global criterion, we have to compare condition numbers of eigenvalues that have a similar distance from the origin. We achieve that by re-scaling our condition numbers with a factor $|\omega|$.

Geometrically speaking we could say that we draw concentric rings around 0 within which we expect all eigenvalues to have similar condition numbers. We will see in the next section that some eigenvalues within a circle have

significantly larger condition numbers than others, these eigenvalues correspond to spurious solutions.

We will start off with the simple one-dimensional Example 4.1 introduced in Chapter 3. It illustrates how condition numbers can be utilized to separate spurious solutions from physical solutions. For this example we will compare the numerical solution with the analytic solution derived in Chapter 3.

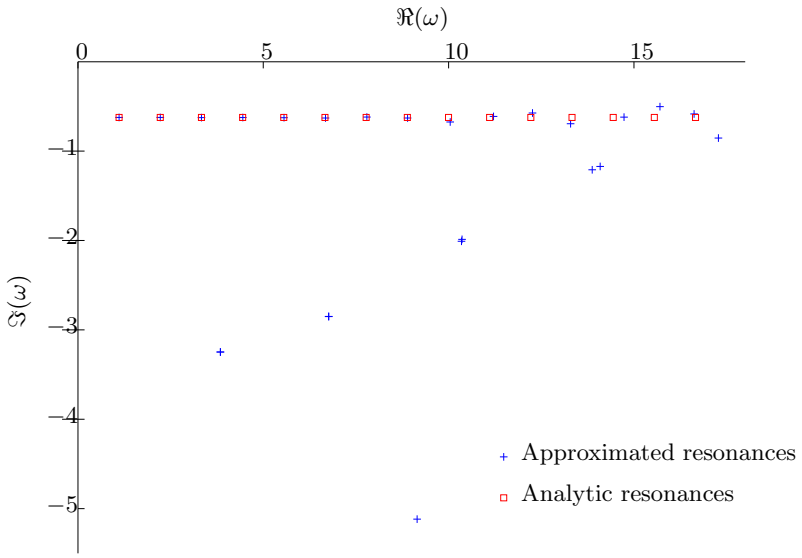

Figure 5.2: Resonances computed for Example 4.1. Red squares mark the analytic resonances computed by Lemma 4.1, blue crosses mark the approximated resonances.

In order to compute condition numbers we need to set the values for n_{int} and n_{ext} in Example 4.1 and the left and right points of the cavity, x_l and x_r. For first computations we set $n_{\text{int}} = \sqrt{2}$, $n_{\text{ext}} = 1$, $x_l = -1$ and $x_r = 1$. By Equations (4.10) and (4.7) this leads to analytic solutions

$$\omega = \frac{k\pi}{2\sqrt{2}} - i\frac{1}{2\sqrt{2}} \ln\left(\frac{\sqrt{2}+1}{\sqrt{2}-1}\right).$$

These solutions are marked with red squares in Figure 5.2.

To obtain an numerical solution, we split $\mathbb{R}$ into $\Omega_{\text{int}} = [-2, 2]$ which contains the cavity and some surrounding air and $\Omega_{\text{ext}} = \mathbb{R}\backslash\Omega_{\text{int}}$. The interior Ω_{int} was discretized with first order finite elements with an equidistant mesh with a mesh width $h = \frac{1}{45} \approx 0.022$. The cavity that stretches from -1 to 1 was thus discretized with 90 degrees of freedom. Since as noted in Remark 4.1, we have integer multiples of half wavelengths inside the cavity, $l = k\frac{\text{wavelengths}}{2}$ for $k \in \mathbb{N}$, that means that for $k = 1$ we have 180 degrees

of freedom per wavelength which is a very good resolution. However, the resolution is fixed, so for larger values of k, the number of degrees of freedom per wavelength reduces, so the larger k, the worse the approximation of the solution in the interior. This is also reflected by the fact that in Figure 5.2 the computed resonances with a real part over $\Re(\omega) \sim 10$ are not perfectly aligned with the analytic resonances any more. For the exterior we used the pole condition with $L = 15$ terms of the series expansion in the exterior and a parameter value $s_0 = 0.4 - 1.0i$. The resonances computed by our algorithm are marked with blue crosses in Figure 5.2.

We will now compute the weighted condition numbers for each computed resonance for this example. In order to visualize the distribution of the condition numbers, we partition the set of partition numbers by using the k-means clustering algorithm included in MATLAB [Seb04, Spä85]. The partitioning of the weighted condition numbers for Example 4.1 is plotted in Figure 5.3. We make the following observations:

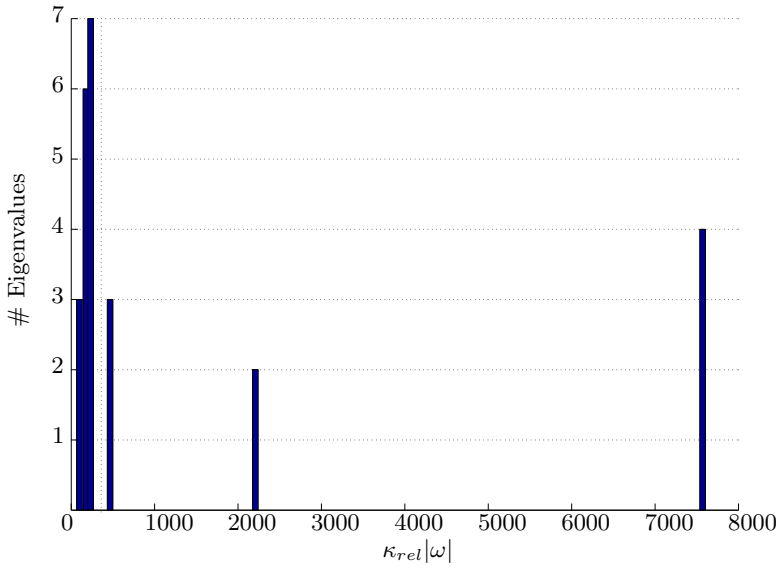

Figure 5.3: Distribution of the weighted condition numbers for Example 4.1. We can see several clusters of condition numbers. Most eigenvalues are contained in three clusters located below $\kappa_{rel}|\omega| \approx 300$ (the value is marked by a vertical dotted line) and separated from the rest by a gap in the weighted condition numbers.

1. The clusters for Example 4.1 are centered around $\kappa_{rel}|\omega| \approx 98.85$, $\kappa_{rel}|\omega| \approx 174.94$, $\kappa_{rel}|\omega| \approx 231.15$, $\kappa_{rel}|\omega| \approx 466.47$, $\kappa_{rel}|\omega| \approx 2208.43$ and $\kappa_{rel}|\omega| \approx 7569.03$.

2. The locations of these cluster centroids may vary slightly due to the

random element in the initial guess of the Arnoldi algorithm we used for the computation of the spectrum $\sigma(A, B)$, however this variation is not in a relevant order of magnitude to change the clusters.

3. The clusters above $\kappa_{rel}|\omega| \approx 300$ are well separated from the rest of the eigenvalues, indicating that they contain good candidates for spurious solutions.

4. Roughly two thirds of the computed eigenvalues have a weighted condition number below that threshold.

With these foundations we can now identify the spurious solutions within our spectrum. Figure 5.4 shows the computed eigenvalue spectrum of the discrete problem with blue crosses. The colored circles around the computed eigenvalue spectrum mark the cluster each eigenvalue belongs to. The cluster with the lowest weighted condition number is marked with green circles, the second lowest weighted condition number is marked with magenta circles, the third one is marked with cyan circles. The fourth cluster, the first one with a centroid above $\kappa_{rel}|\omega| \approx 300$ is marked with yellow circles, the cluster centered at $\kappa_{rel}|\omega| \approx 1694$ is marked with red circles and the cluster with the highest center is marked with black circles. In Figure 5.4 we can see that the eigenvalues corresponding to physical solutions are all within the first three clusters, thus all have weighted condition numbers well below 300. Thus the comparison of the weighted condition numbers gives a means of distinction between physical and spurious solutions. However, this detection is not always reliable as we will see in the next example.

Now we will turn to a second simple example in two space dimensions. Again, we will see, that the detection of spurious solutions via their condition number works well in this case. However, following this, will be two more complicated examples where this bold approach does not work so well any more. This necessitates a new approach which we will develop in Sections 5.3 and 5.4.

Example 5.1.

The geometry for this example is similar to that of Example 4.1 taken to two space dimensions. We wish to solve the Helmholtz resonance problem on a two-dimensional square cavity with side length d embedded in air. Its refractive index $n_{\text{int}}(x, y) = \text{const.}$ for $(x, y) \in \Omega_{\text{int}}$. Its geometry is sketched on the left hand side picture of Figure 5.5. For our example we use $n_{\text{int}} = 2.5$ and $d = 1$.

In order to obtain a reference solution for this problem, we use the commercial FEM software package JCMsuite [PBZS07, ZBKS06]. In order to ensure the quality of the calculations, we make use of its high order finite elements, adaptive refinement and sophisticated PML implementation [Zsc09]

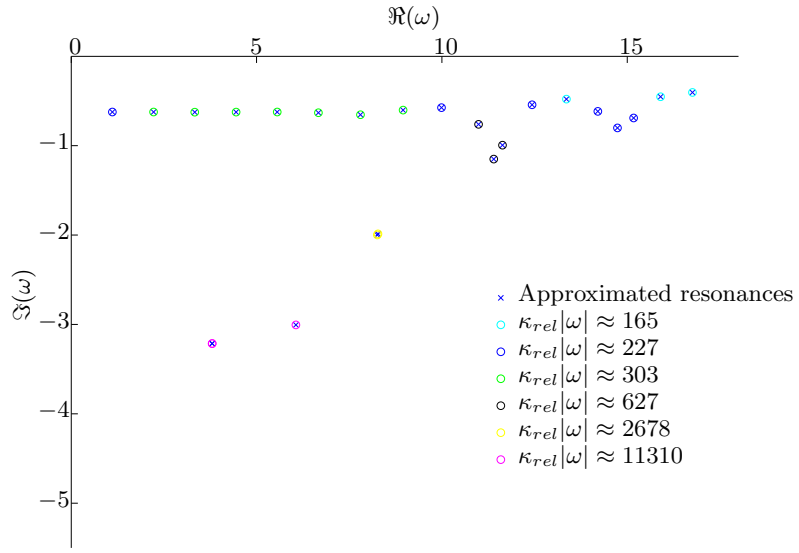

Figure 5.4: The spectrum $\sigma(A, B)$ of Example 4.1 where the eigenvalues are marked corresponding to their weighted condition numbers.

as transparent boundary condition. To obtain the reference solution for Example 5.1, we use second order finite elements and three adaptive refinement steps. Both solutions, the reference solution and our solution are computed on the same grid which is generated with JCMgeo, the triangulation tool included in JCMsuite. We use it since it creates high quality mixed meshes for two-dimensional geometries including the prismatoids we require for the implementation of the pole condition in the exterior. We chose not to attach the prismatiods for the exterior directly to the resonator but instead added a buffer layer of air to Ω_{int}. This increases the speed and precision of JCMsuite (see right-hand side image of Figure 5.5).

The comparison of the solution of our algorithm compared with the JCMsuite reference solution is shown in Figure 5.6. The fact that the agreement of both methods is not perfect stems from the fact that we computed the solutions with considerably higher accuracy for the reference solution using three adaptive refinement steps and a finite element order 3 as opposed to linear finite elements in the interior.

Next we will compute the condition numbers associated with the eigenvalues computed for Example 5.1. Figure 5.6 already suggests that the eigenvalues with imaginary parts $\Im(\omega) \approx -1$ are the physical resonances of the problem.

Figure 5.7 shows the distribution of the eigenvalues, computed as for Example 4.1. The centroids of the clusters computed by the kmeans algorithm are located at $\kappa_{rel}|\omega| \approx 987.67$, $\kappa_{rel}|\omega| \approx 1302.02$, $\kappa_{rel}|\omega| \approx 1765.07$,

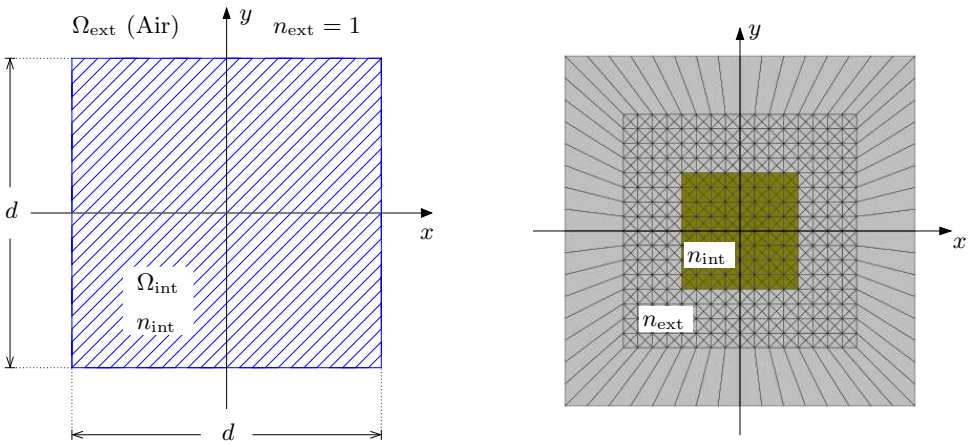

Figure 5.5: Left: Sketch of the structure for Example 5.1 Right: Mixed grid generated with JCMgeo that is used for both pole condition and JCMsuite reference calculations. The interior is discretized with triangles, the exterior with trapezes.

$\kappa_{rel}|\omega| \approx 2161.12$, $\kappa_{rel}|\omega| \approx 3062.56$ and $\kappa_{rel}|\omega| \approx 14656.09$. Again, it is possible to find a threshold that separates the physical from the spurious solutions of the problem. This threshold is even clearer for this example than for the previous example, since the last cluster is well apart from the rest. The threshold is marked with a dotted vertical line in Figure 5.7. In Figure 5.8 we again marked the eigenvalues in the spectrum computed for Example 5.1 by the cluster computed by kmeans they belong to. We see that this confirms our threshold and that the eigenvalues belonging to the first three clusters are the physical solutions of the problem while the eigenvalues with a higher condition number are spurious solutions.

The third example is again a one-dimensional example. Its layout is only slightly more complicated, its eigenmode structure however is far more involved due to leakage into the exterior cladding and the extra layers involved. It resembles an air-filled cavity surrounded by two materials with a higher refractive indices. It resembles the kind of cavity that can be found in photonic crystals.

Example 5.2.

Again we want to solve the Helmholtz resonance problem in one space dimension. The air-filled cavity stretches from $x = -1$ to $x = 1$. It is surrounded by a cladding with a material with refractive index $n = 3.5$ and a thickness of $d = 1$ on each side which again is embedded in an infinitely thick material with refractive index $n = 2.5$. The layout of this example is sketched in Figure 5.9.

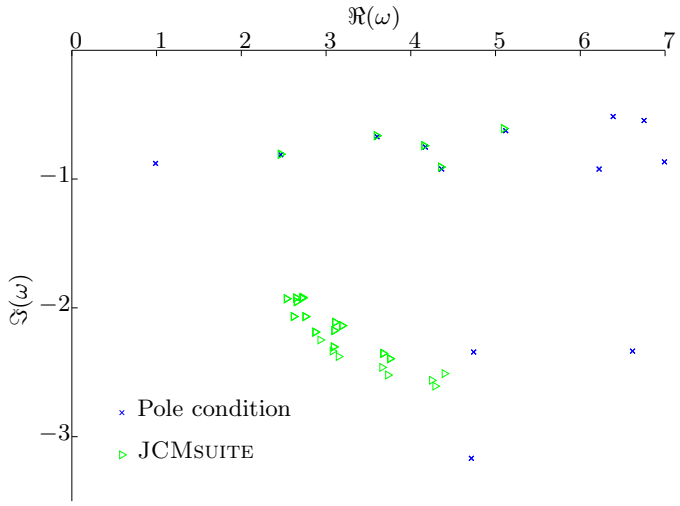

Figure 5.6: Eigenvalue spectrum computed for Example 5.1 with the pole condition (blue crosses) and reference solution obained with JCMsuite (green triangles).

As before we use JCMsuite to obtain a reference solution. For the reference solutions for Example 5.2 we use a finite element degree of 5 and two adaptive refinement steps. Figure 5.10 shows a comparison of the calculations using our implementation of the pole condition in the exterior and linear finite elements in the interior and the reference solutions obtained with JCMsuite. One can see that some of the eigenvalues computed with our implementation are in good agreement with the reference solution while others have no correspondence. By manual inspection it is possible to confirm that the solutions where both methods are in good agreement are the physical resonances of the problem. However this implies that even for this allegedly simple example, a commercial FEM package also outputs a significant number of spurious solutions that pollute the computed spectrum.

We will now compute the condition numbers for the modes computed with the pole condition. The distribution of the weighted condition numbers $\kappa_{rel}|\omega|$ is plotted in Figure 5.11. For the interior we use again first order finite elements on a grid with a step size of $h = 0.003$. In the exterior we use $L = 25$ Hardy modes and a parameter value $s_0 = 0.87 - 1.14i$

The eigenvalue clusters computed for this example with kmeans are centered around the values $\kappa_{rel}|\omega| \approx 0.0377 \cdot 10^5$, $\kappa_{rel}|\omega| \approx 0.1247 \cdot 10^5$, $\kappa_{rel}|\omega| \approx 0.2231 \cdot 10^5$, $\kappa_{rel}|\omega| \approx 0.5382 \cdot 10^5$, $\kappa_{rel}|\omega| \approx 0.9171 \cdot 10^5$ and $\kappa_{rel}|\omega| \approx 2.1416 \cdot 10^5$. Again we can see that roughly two thirds of the solutions have a weighted condition number that is below a certain threshold. This time, however, this contains the eigenvalues in the first cluster and,

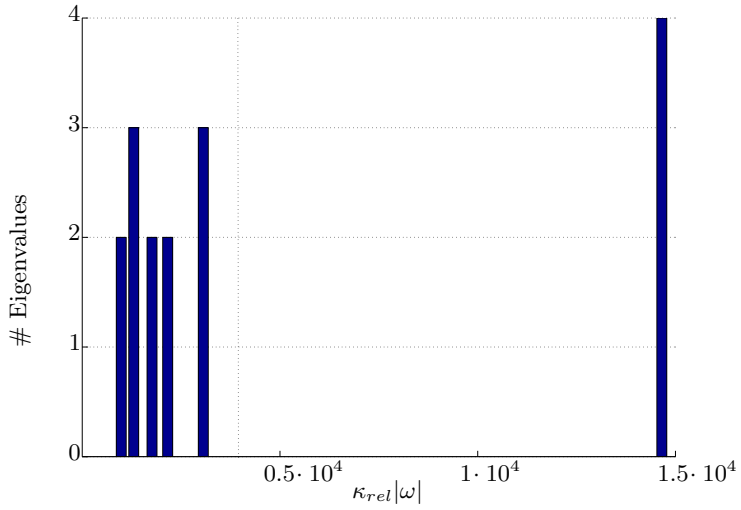

Figure 5.7: Distribution of the weighted condition numbers $\kappa_{rel}|\omega|$ for Example 5.1. The vertical dotted line marks the threshold separating the physical from the spurious solutions.

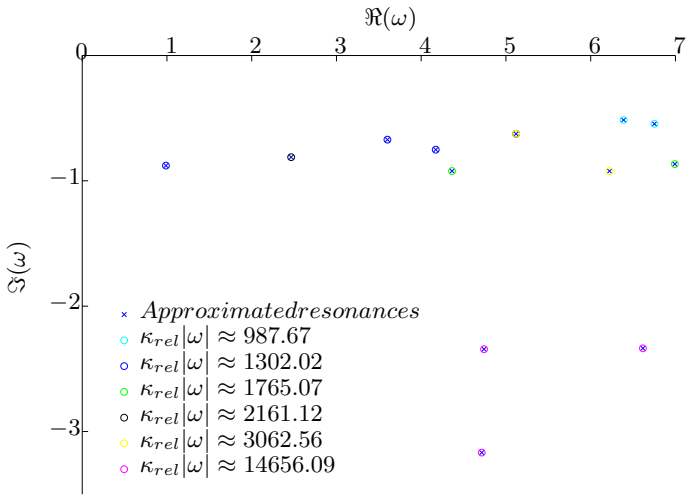

Figure 5.8: Eigenvalue spectrum $\sigma(A, B)$ for Example 5.1 with eigenvalues marked by membership to eigenvalue clusters.

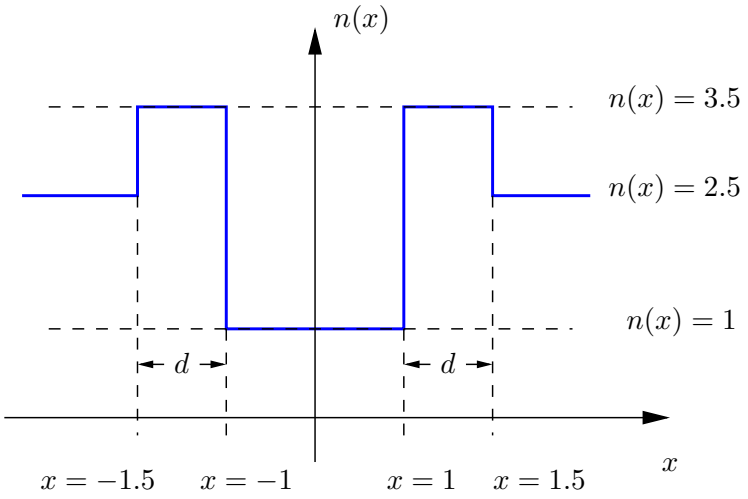

Figure 5.9: Sketch of the layout of the air-filled cavity used in Example 5.2.

as Figure 5.12 shows, this cluster already contains some spurious solutions and not all physical solutions. If we want to include all physical solutions, we will have to separate the spectrum at a weighted condition number of $\kappa_{rel}|\omega| \approx 0.3 \cdot 10^5$, which will also lead to almost all solutions, spurious or not, being below the threshold.

This means that Example 5.2 is an example that shows, that the bold approach of utilizing condition numbers is not satisfactory for the detection of spurious solutions. A fact that can be accounted to the ignorance of the condition numbers to the nature of the perturbation which does not reflect the fact that we made out the discretization in the exterior as the cause for spurious solutions. We will choose a related, yet alternative approach in the following section. This approach will take into account the perturbation we apply and then compute the perturbations of the solution directly without having to solve the entire problem twice or running into any identification issues when modes are close to each other.

Finally we will give another two-dimensional example. It has an exterior domain that is heterogeneous. This means that it contains multiple materials, making it more difficult than Example 5.1. It resembles a slot waveguide that is used for in many physical applications. We will model this air-filled resonator which is a slot cut into a highly conductive substrate. See the left hand side image of Figure 5.13 for a sketch of such a layout.

Example 5.3.

We will model a resonator in two space dimensions. The resonator itself given by an air-filled slot in a highly conductive substrate. It has a deptgh of $d = 0.8$ and a width of $w = 0.5$. For the computational domain, we add a padding of $p_x = 0.5$ in x-direction and a padding of $p_y = 0,7$ in

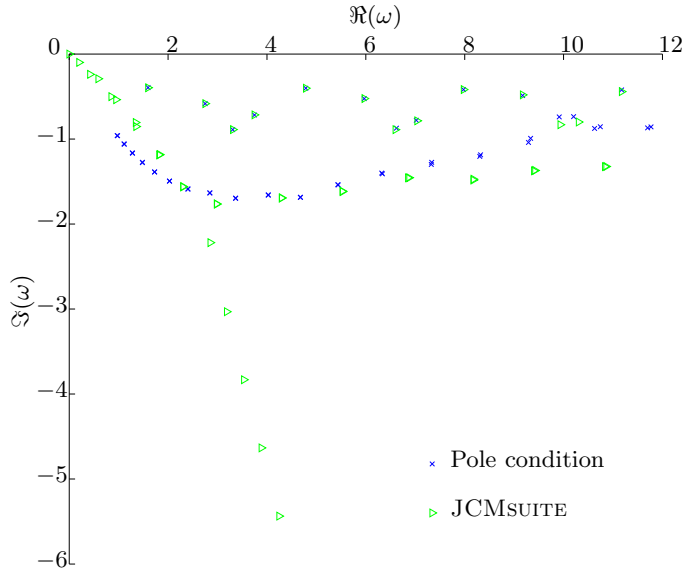

Figure 5.10: Eigenvalue spectrum computed for Example 5.2 with the pole condition (blue crosses) and reference solution obained with JCMsuite (green triangles).

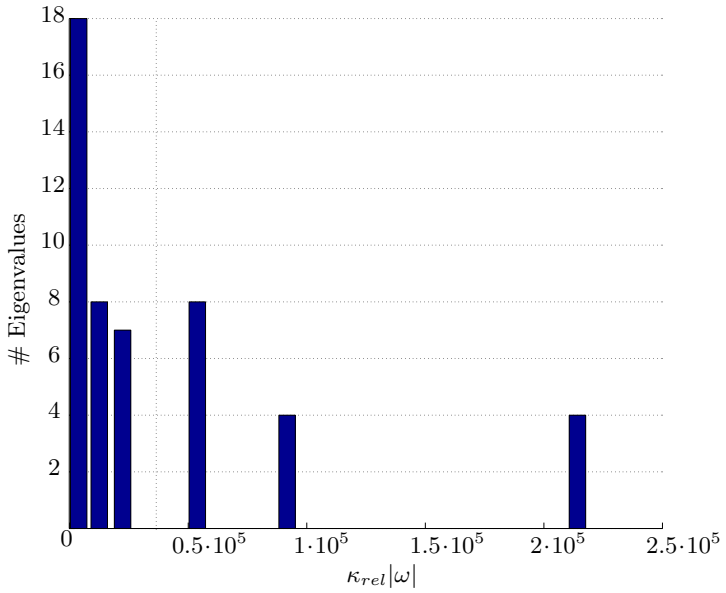

Figure 5.11: Distribution of the weighted condition numbers for Example 5.2. The dotted vertical line indicates the threshold for distinguishing between spurious and physical solutions.

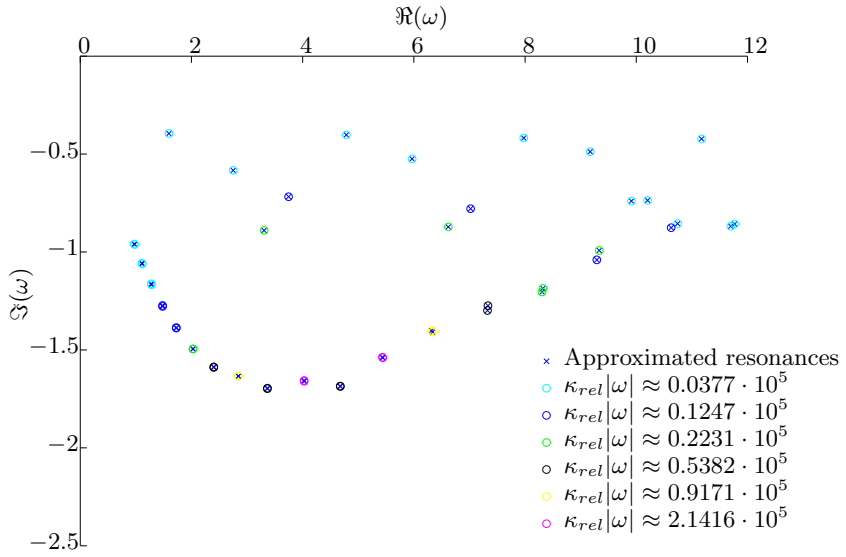

Figure 5.12: The spectrum $\sigma(A, B)$ of Example 5.2 where the eigenvalues are marked corresponding to their weighted condition numbers.

y-direction to each side of the waveguide. In order to make our material lossy, we use a complex refractive index of $n_{\text{sub}} = 0.01 + 3i$ for the substrate. See the left hand side image of Figure 5.13 for a schematic of this setup and the right hand side image for the finite element mesh used for the solution.

Figure 5.15 shows the comparison of the spectrum computed with our MATLAB-code compared with a reference solution obtained with JCMSUITE. Again we used second order finite elements and three adaptive refinement steps for the reference solution. The interior solution for the pole condition was again computed with linear finite elements. We chose $s_0 = 2.05 - 0.6i$ and $L = 10$ as pole condition parameters. Again, the eigenvalues that occur in both solutions and can be identified as physical solutions in the spectral region in question by manual inspection. The left-hand side image of Figure 5.14 shows the field distribution for such a physical solution while the right-hand side image of Figure 5.14 shows the field distribution of a solution we consider to be spurious since it does not reflect the basic properties of the underlying geometry such as the symmetry. For the field distribution shown here, it is straightforward to distinguish between physical and spurious solutions however, there are also field distributions where this distinction is not so easy, especially when it comes to more complex resonator geometries.

Now we will again compute the condition numbers for the eigenvalues

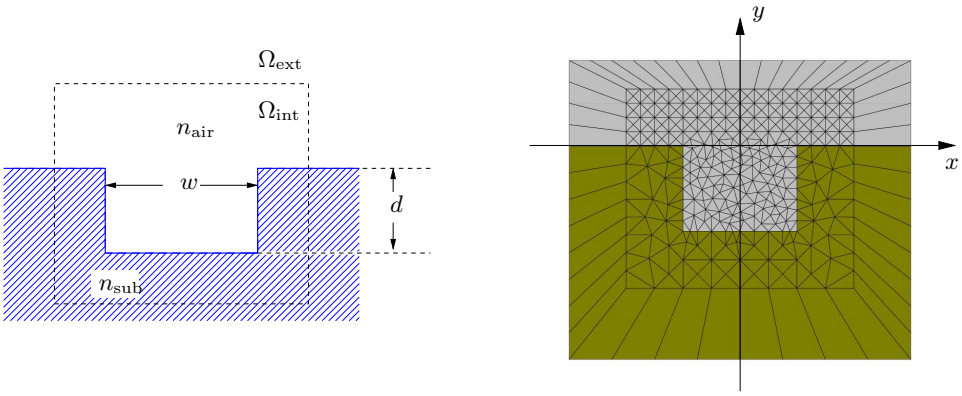

Figure 5.13: Left: Sketch of the structure for Example 5.3 Right: Mixed grid generated with JCMgeo that is used for both pole condition and JCMsuite reference calculations. The interior is discretized with triangles, the exterior with trapezes.

and group them in clusters, using the kmeans algorithm. This yields clusters centered around $\kappa_{rel}|\omega| \approx 379.00$, $\kappa_{rel}|\omega| \approx 475.34$, $\kappa_{rel}|\omega| \approx 895.48$, $\kappa_{rel}|\omega| \approx 1555.42$, $\kappa_{rel}|\omega| \approx 2787.85$ and $\kappa_{rel}|\omega| \approx 4533.88$. In Figure 5.16 we colored the eigenvalues corresponding to the cluster they belong to. It can be seen that similar to Example 5.2 there are clusters that contain both physical and spurious solutions making a distinction by condition number a very delicate, if not impossible issue for this example.

5.3 Exact Perturbation

We have already seen that the fact that the condition number is a purely algebraic feature that disregards all knowledge about the physics of the problem and about the nature of the perturbation, makes an identification of spurious solutions difficult. This suggests that there might be a better way to distinguish between physical and spurious solutions of a problem when including this knowledge in our considerations. We will now aim at deriving a condition that is void of the generality of the condition number but includes the special kind of perturbation we cause when changing the parameter s_0 in the pole condition used to discretize the exterior domain. Again, in order to differentiate between the eigenvalue that is computed of a concrete example and closely related to our problem statement, we will use ω for the eigenvalues of the Helmholtz equation and λ for eigenvalues that are not related to our equation.

Instead of utilizing condition numbers it might be a good idea to check the sensitivity of the eigenvalues with respect to changes of a parameter ρ. The problems connected with the use of condition numbers may be overcome

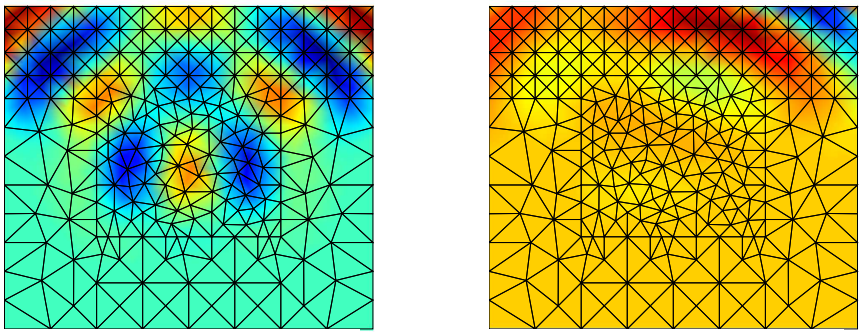

Figure 5.14: Left: Field distribution for the eigenvalue at $\omega \approx 7.02 - 0.97i$ which we consider a physical solution of the problem. Right: Field distribution for the eigenvalue at $\omega \approx 6.97 - 2.26i$ which we consider a spurious solution of the problem.

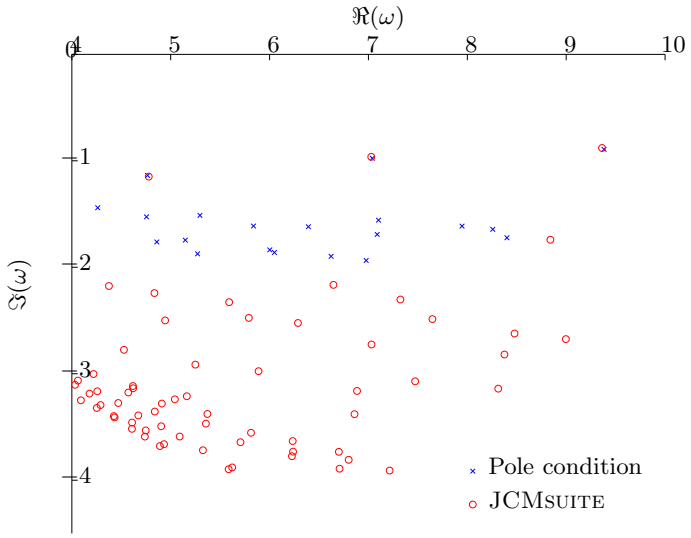

Figure 5.15: Comparison of JCMsuite reference solution with the solution computed with the pole condition and linear finite elements in the interior for Example 5.3.

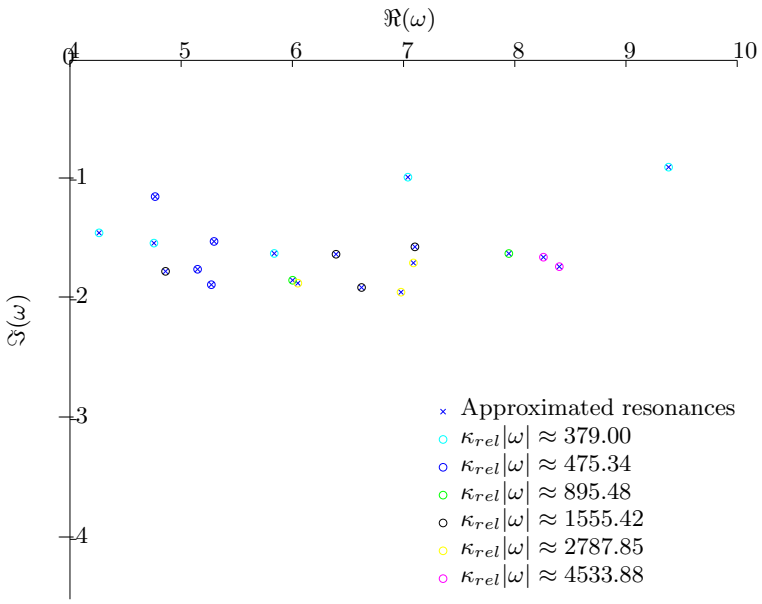

Figure 5.16: The spectrum $\sigma(A, B)$ of Example 5.3 where the eigenvalues are marked corresponding to their weighted condition numbers.

by directly computing the change of $\lambda(\widetilde{\rho})$ for perturbation $\widetilde{\rho} = \rho + \Delta\rho$. A way to obtain such a direct approximation stems from the backwards error analysis for the generalized eigenvalue problem [HH98].

Given the generalized eigenvalue problem from Definition 5.1, if we perturb the matrices A and B by ΔA and ΔB, this results in perturbed eigenvalue $\Delta\lambda$, right eigenvector $\Delta\mathbf{u}$ and left eigenvector $\Delta\mathbf{v}$. Since ΔA and ΔB arise from a variation of the pole condition parameter s_0, we know them explicitly, which will allow us to compute $\Delta\lambda$ directly.

Lemma 5.3.

Let $\mathbf{u}$ *and* $\mathbf{v}$ *be the left and right eigenvectors for the eigenvalue* λ *of the generalized eigenvalue problem* $(A - \lambda B)\mathbf{u} = 0$. *Let* ΔA *and* ΔB *be perturbations of* A *and* B. *This leads to perturbed eigenvalue* $\lambda + \Delta\lambda$ *and eigenvectors* $\mathbf{u} + \Delta\mathbf{u}$ *and* $\mathbf{v} + \Delta\mathbf{v}$.

Then in first order we can approximate $\Delta\lambda$ *by*

$$\Delta\lambda = \frac{\mathbf{v}^H \Delta A\mathbf{u} - \lambda\mathbf{v}^H \Delta B\mathbf{u}}{\mathbf{v}^H B\mathbf{u}} + \mathcal{O}(\varepsilon^2). \tag{5.8}$$

Proof.

Using the perturbed left and right eigenvectors arising from a perturbation of A and B and the perturbed eigenvalue, we rewrite the entire perturbed problem as

$$(A + \Delta A)(\mathbf{u} + \Delta\mathbf{u}) = (\lambda + \Delta\lambda)(B + \Delta B)(\mathbf{u} + \Delta\mathbf{u}). \tag{5.9}$$

Next, we expand equation (5.9) and premultiply with $\mathbf{v}^H$, the left eigenvector for λ. Since $\mathbf{v}^H A = \lambda \mathbf{v}^H B$, we can cancel some of the resulting terms and get

$$\begin{aligned} \mathbf{v}^H \Delta A \mathbf{u} + \mathbf{v}^H \Delta A \Delta\mathbf{u} = \lambda \mathbf{v}^H \Delta B \mathbf{u} + \lambda \mathbf{v}^H \Delta B \Delta\mathbf{u} \quad (5.10)\\ + \Delta\lambda \mathbf{v}^H B\mathbf{u} + \Delta\lambda \mathbf{v}^H B \Delta\mathbf{u} + \Delta\lambda \mathbf{v}^H \Delta B \mathbf{u} + \Delta\lambda \mathbf{v}^H \Delta B \Delta\mathbf{u}. \end{aligned}$$

Isolating $\Delta\lambda$ in (5.10), we have

$$\Delta\lambda = \frac{\mathbf{v}^H \Delta A \mathbf{u} - \lambda \mathbf{v}^H \Delta B \mathbf{u} + \mathbf{v}^H \Delta A \Delta\mathbf{u} - \lambda \mathbf{v}^H \Delta B \Delta\mathbf{u}}{\mathbf{v}^H B\mathbf{u} + \mathbf{v}^H B \Delta\mathbf{u} + \mathbf{v}^H \Delta B \mathbf{u} + \mathbf{v}^H \Delta B \Delta\mathbf{u}}.$$

Neglecting the higher order terms we arrive at the first order approximation of $\Delta\lambda$:

$$\Delta\lambda = \frac{\mathbf{v}^H \Delta A \mathbf{u} - \lambda \mathbf{v}^H \Delta B \mathbf{u}}{\mathbf{v}^H B\mathbf{u}} + \mathcal{O}(\varepsilon^2). \tag{5.11}$$

□

Remark 5.3.

If we would not know ΔA and ΔB explicitly, we could bound them by the tolerance matrices E and F and derive an absolute norm-wise condition number:

Lemma 5.4.

Let λ be a simple, finite, nonzero eigenvalue of the pair (A, B) with left eigenvector $\mathbf{v}$ and right eigenvector $\mathbf{u}$. If $\Delta x \to 0$ for $\varepsilon \to 0$, then the absolute condition number of λ, $\kappa_{abs}(\lambda)$ is

$$\begin{aligned} \kappa_{abs}(\lambda) = \lim_{\varepsilon \to 0} \sup \quad & \left\{ \frac{|\Delta\lambda|}{\varepsilon|\lambda|} : (\widetilde{A} - \widetilde{\lambda}\widetilde{B})(\widetilde{\mathbf{u}}) = 0, \right. \\ & \widetilde{A} = A + \Delta A, \widetilde{B} = B + \Delta B, \\ & \widetilde{\lambda} = \lambda + \Delta\lambda, \widetilde{\mathbf{u}} = \mathbf{u} + \Delta\mathbf{u}, \\ & \left. \|\Delta A\|_2 \leq \varepsilon \|E\|_2, \|\Delta B\|_2 \leq \varepsilon \|F\|_2 \right\} \\ = \quad & \frac{\|\mathbf{u}\|_2 \|\mathbf{v}\|_2 (\|E\|_2 + |\lambda| \|F\|_2)}{|\lambda| |\mathbf{v}^H B \mathbf{u}|}. \end{aligned}$$

Proof.

For a proof, see [HH98].

□

We would again, however, sacrifice the specific knowledge of ΔA and ΔB in order to obtain this more general result.

Using the formula for $\Delta\lambda$, we can directly compute the effect of a perturbation of the pole condition parameter on an eigenvalue $\lambda \in \sigma(A, B)$. We know that the matrices A and B are dependent on s_0, the pole condition parameter. Our task is to find out the perturbed eigenvalue $\Delta\lambda$ for a perturbation $s_0 \to s_0 + \Delta s_0$. That means, we need to compute the change in the matrices A and B, ΔA and ΔB. Since $A(s_0 + \Delta s_0) := \widetilde{A} = A + \Delta A =: A(s_0) + \Delta A$ and $B(s_0 + \Delta s_0) := \widetilde{B} = B + \Delta B =: B(s_0) + \Delta B$, we can compute $\Delta A = A(s_0 + \Delta s_0) - A(s_0)$ and $\Delta B = B(s_0 + \Delta s_0) - B(s_0)$.

Since only the entries of the exterior, A_{ext} and B_{ext}, depend on s_0, all other entries cancel, if we compute ΔA and ΔB, hence the perturbations ΔA and ΔB of A and B will have the same structure as A_{ext} and B_{ext}, the entries of A and B for the exterior degrees of freedom (cf. Sections 3.4 and 3.5).

We recall that for the one-dimensional case

$$\begin{aligned} A_{\text{ext}} &= 2s_0 P_l^\top (\mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)}) P_l + 2s_0 P_r^\top (\mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)}) P_r \text{ and} \\ B_{\text{ext}} &= \frac{2n_l^2}{s_0} P_l^\top (\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)}) P_l + \frac{2n_r^2}{s_0} P_r^\top (\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)}) P_r. \end{aligned}$$

(cf. Equations (3.46) and (3.47)) and for the two-dimensional case

$$\begin{aligned} A_{\text{ext}} &= \sum_{T_i} P_i^\top A_{\text{loc},i}^{\text{ext}} P_i \text{ and} \\ B_{\text{ext}} &= \sum_{T_i} P_i^\top B_{\text{loc},i}^{\text{ext}} P_i. \end{aligned}$$

(cf. Equations (3.68) and (3.69)) with the local element matrices described in Equations (3.66) and (3.67):

$$\begin{aligned} A_{\text{loc},i}^{\text{ext}} &= s_0 A_{\text{loc},i}^{\text{ext},(1)} + \frac{1}{s_0} A_{\text{loc},i}^{\text{ext},(-1)} + A_{\text{loc},i}^{\text{ext},(0)} \\ B_{\text{loc},i}^{\text{ext}} &= n_i^2 \frac{1}{s_0} B_{\text{loc},i}^{\text{ext},(-1)} + n_i^2 \frac{1}{s_0^2} B_{\text{loc},i}^{\text{ext},(-2)}. \end{aligned}$$

So for the one-dimensional case, we can directly compute

$$\begin{aligned}
\Delta A &= A(s_0+\Delta s_0) - A(s_0) \qquad (5.12)\\
&= 2(s_0+\Delta s_0)\left(P_l^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_l + P_r^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_r\right)\\
&\quad -2(s_0)\left(P_l^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_l + P_r^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_r\right)\\
&= 2\Delta s_0\left(P_l^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_l + P_r^\top (\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)})P_r\right) \text{ and}\\
\Delta B &= B(s_0+\Delta s_0) - B(s_0) \qquad (5.13)\\
&= \frac{2}{s_0+\Delta s_0}\left(n_l^2 P_l^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_l + n_r^2 P_r^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_r\right)\\
&\quad -\frac{2}{s_0}\left(n_l^2 P_l^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_l + n_r^2 P_r^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_r\right)\\
&= -\frac{2\Delta s_0}{s_0(s_0+\Delta s_0)}\left(n_l^2 P_l^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_l + n_r^2 P_r^\top (\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)})P_r\right).
\end{aligned}$$

For the two-dimensional case, we have by the same computation

$$\begin{aligned}
\Delta A &= A(s_0+\Delta s_0) - A(s_0) \qquad (5.14)\\
&= \Delta s_0 \sum_{T_i} P_i^\top A_{\mathrm{loc},i}^{\mathrm{ext},(1)} P_i\\
&\quad -\frac{\Delta s_0}{s_0(s_0+\Delta s_0)}\sum_{T_i} P_i^\top A_{\mathrm{loc},i}^{\mathrm{ext},(2)} P_i \text{ and}\\
\Delta B &= B(s_0+\Delta s_0) - B(s_0) \qquad (5.15)\\
&= -\frac{\Delta s_0}{s_0(s_0+\Delta s_0)} n_i^2 \sum_{T_i} P_i^\top B_{\mathrm{loc},i}^{\mathrm{ext},(-1)} P_i\\
&\quad -\frac{\Delta s_0\,(2s_0+\Delta s_0)}{(s_0+\Delta s_0)^2\, {s_0}^2} n_i^2 \sum_{T_i} P_i^\top B_{\mathrm{loc},i}^{\mathrm{ext},(-2)} P_i.
\end{aligned}$$

Since $\frac{\Delta s_0}{s_0(s_0+\Delta s_0)} \to 0$ for $\Delta s_0 \to 0$ and $\frac{\Delta s_0(2s_0+\Delta s_0)}{(s_0+\Delta s_0)^2 s_0^2} \to 0$ for $\Delta s_0 \to 0$, a small perturbation of s_0 causes small perturbations ΔA and ΔB, as one would expect.

Again, we will have to introduce a scaling of $\Delta\lambda$, the quantity we will use to identify the spurious solutions. However, the scaling we will use differs from the scaling we introduced in Section 5.2 for the relative condition number κ_{rel}, since it takes into account the problem we are solving and the type of our perturbation. We established in Section 3.3, that the approximation of the Laplace transform of the exterior solution $\mathcal{L}\{u_{\mathrm{ext}}\}\circ\mathcal{M}_{s_0}$ by an power series expansion will be best near s_0. Thus, it is reasonable to expect the resonances to be more sensitive to perturbations of s_0 with increasing distance to the parameter. Thus, we will take the distance $|s_0-\omega|$ as scaling for the perturbation $\Delta\omega$.

We will now apply these findings to the examples from the previous section and revert to the notion of ω which is part of the problem statement. First we will revisit the simple one-dimensional cavity layout from Example 4.1 and assure ourselves that the perturbations $\Delta\omega$ really hold. For this we will solve the problem twice. As in the previous section, we will choose a value of $s_0 = 0.4 - 0.8i$ and then we will perturb s_0 to $\widetilde{s}_0 = 0.4 + 5 \cdot 10^{-3} - (0.8 + 5 \cdot 10^{-3})i$, that is $\Delta s_0 \approx 7 \cdot 10^{-3}$. We will compute the eigenvalue spectrum of the ω^2 eigenvalue problem for each value s_0 and then compute $\Delta_{exact}\omega = \omega(s_0) - \omega(\widetilde{s}_0)$ and evaluate the formula given in Lemma (5.3) to compute the first order approximation $\Delta_{approx}\omega$ from ω, $A, B, \Delta A$ and ΔB. Comparing the values for $\Delta\omega$ of both calculations, as expected, yields

$$\max_{\omega \in \sigma(A,B)} (\Delta_{exact}\omega - \Delta_{approx}\omega) \approx -1.5907 \cdot 10^{-6} - 9.3504 \cdot 10^{-7} i \lesssim (\Delta s_0)^2.$$

Figure 5.17 shows the distribution of the weighted perturbations $\frac{1}{|s_0 - \omega|}|\Delta\omega|$.

Again we use the k-means algorithm to compute clusters of weighted perturbations. The centroids of the clusters computed for this example are positioned at $1/|s_0 - \omega||\Delta\omega| \approx 0.0001$, $1/|s_0 - \omega||\Delta\omega| \approx 0.0006$, $1/|s_0 - \omega||\Delta\omega| \approx 0.0012$, $1/|s_0 - \omega||\Delta\omega| \approx 0.0019$, $1/|s_0 - \omega||\Delta\omega| \approx 0.0032$ and $1/|s_0 - \omega||\Delta\omega| \approx 0.0080$. We can see that the clusters at $1/|s_0 - \omega| \leq 0.0012$ that correspond to a relative perturbation of the order of Δs_0 contain most resonances. Hence, we have divided the spectrum $\sigma(A, B)$ into two sets of eigenvalues. Those reacting to perturbations of the exterior with a reaction that is at most in the order of the perturbation and those that react stronger than the perturbation. Since we expect the physical solutions to be well-converged in the interior and in the exterior, it is plausible to expect them to respond less strongly to perturbations of the exterior. Hence, we expect the first set, that is the resonances with $\Delta\omega = \mathcal{O}(\Delta s_0)$, to contain the physical solutions and the second set, that responds more strongly, to contain the spurious solutions.

In Figure 5.18 we marked eigenvalues that have a response of the order of magnitude of the perturbation with red circles. We can see that they correspond to the physical solutions of the problem.

The same mechanism works well for Example 5.2. Again we can use the order of magnitude of the perturbation Δs_0 as measure for the perturbation of the eigenvalues that we allow. On the left hand side plot of Figure 5.20 we see the distribution of the computed perturbations, on the right hand side plot again those eigenvalues whose perturbation is in the same order of magnitude as Δs_0 are marked with red circles. We can see that for wide parts of the spectrum they correspond to the physical solutions that can be seen in Figure 5.10. However as $\Re(\omega)$ increases, as expected the identification is less precise and some spurious solutions fullfill our criterion.

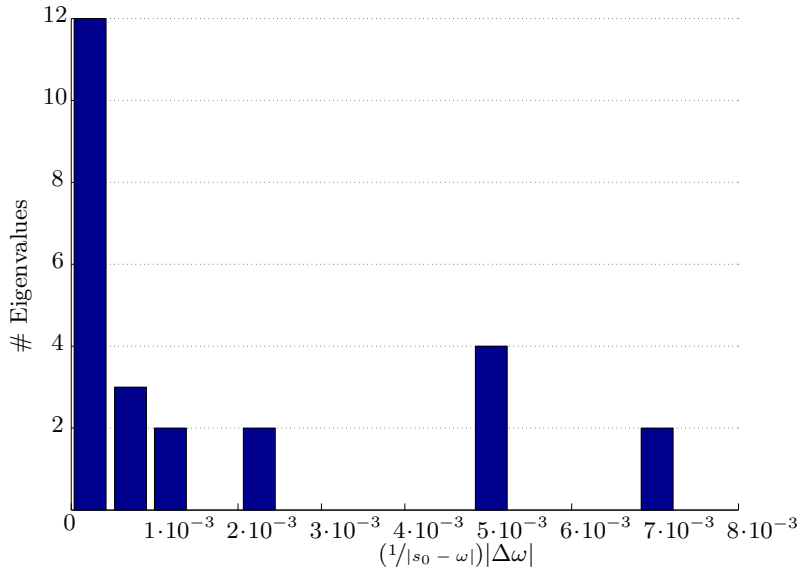

Figure 5.17: Distribution of the perturbations $(1/|s_0 - \omega|)|\Delta\omega|$ for Example 4.1.

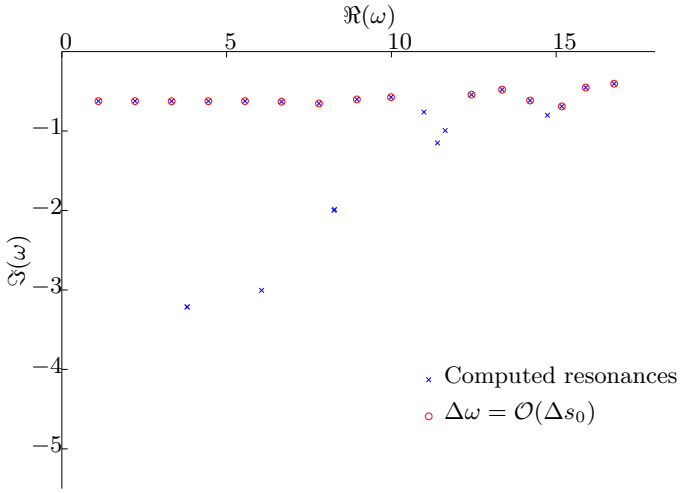

Figure 5.18: The spectrum of Example 4.1 (blue crosses) $\sigma(A, B)$ where the eigenvalues with $\Delta\omega = \mathcal{O}(\Delta s_0)$ or more precisely $\Delta\omega \leq 0.001$ are marked with red circles.

In the next section we will deal with the problem of identifying the region of the complex plane where we can rely on the perturbations we compute this way.

We will now turn to the two-dimensional examples of the previous section. For the simple square cavity of Example 5.1, the computed perturbations are given in Table 5.1. We can see that again, the modes we identified as physical solutions respond to a perturbation of the exterior with a relatively low perturbation while the response of the spurious solutions is significantly larger than the magnitude of the perturbation of s_0. Figure 5.21 visualizes this, having the modes whose perturbation is at most in the order of magnitude of Δs_0 marked with red circles.

For Example 5.3 we have the same situation. Here, we perturb the pole condition parameter s_0 with $\Delta s_0 = 0.05 - 0.05i$, hence we expect the physical resonances to respond with a perturbation in the order of $|\Delta s_0|$. Figure 5.22 shows the eigenvalue spectrum for this problem where the eigenvalues whose perturbation is $\mathcal{O}(\Delta s_0)$ are marked with red circles. We can see that these are exactly the eigenvalues that the spectrum computed with our method had in common with the spectrum of the reference solution computed with JCMsolve.

5.4 A Convergence Monitor for Resonances

The methods we derived in the previous sections for the detection of spurious solutions all suffer from the major drawback, that it is not possible to distinguish between solutions that react strongly to perturbations (i.e. are ill-conditioned) because they are spurious solutions and modes, that react strongly to perturbations because their approximation is not good enough in the exterior domain Ω_{ext}. In order to overcome this problem, we will complement the methods from the previous sections with a convergence monitor. This will give us a region in the complex plane in which the eigenvalues are well-converged for our choice of s_0 and the number of degrees of freedom L used in the computation.

First we will analyze the situation in the one-dimensional case. For this analysis we discard all other information and just look at one part of the exterior domain. Discarding the matrices P_l and P_r that map the local degrees of freedom to global degrees of freedom, the matrices in each part of Ω_{ext}, according to Equations (3.46) and (3.47) are

$$\begin{aligned} A_{\text{ext}} &= 2s_0 \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \\ B_{\text{ext}} &= \frac{2n_{\text{ext}}^2}{s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)}, \end{aligned}$$

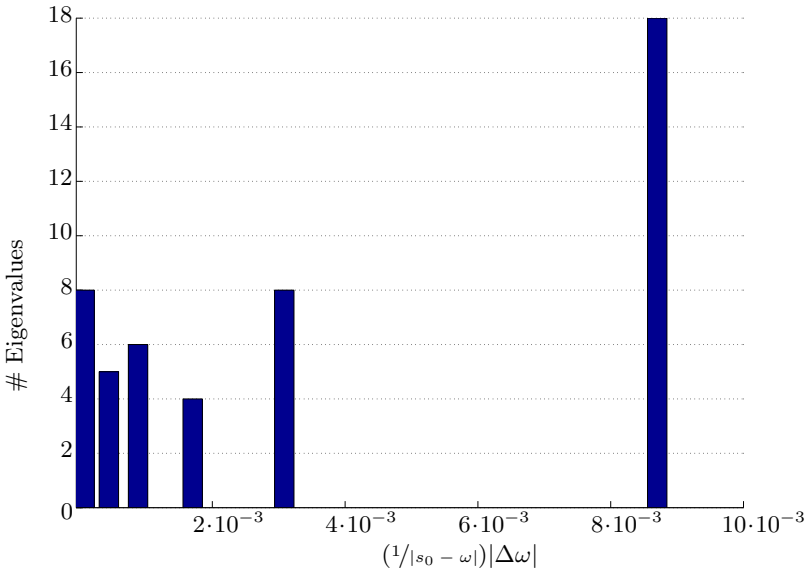

Figure 5.19: Distribution of the perturbations computed for Example 5.2. The perturbation of the exterior is $\Delta s_0 \approx 0.007$.

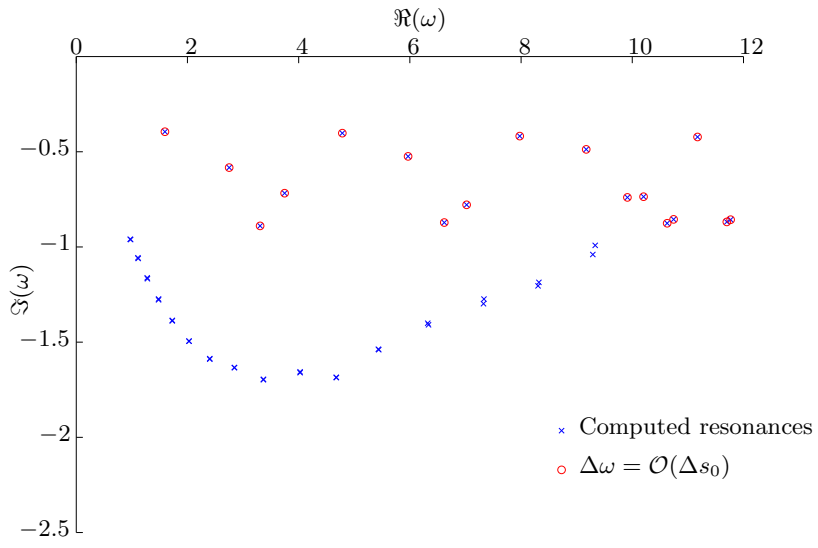

Figure 5.20: Spectrum for Example 5.2 (blue crosses) with eigenvalues whose perturbation is below a threshold marked (red circles).

ω	$\Delta\omega$	$(\lvert s_0-\omega\rvert)^{-1}\lvert\Delta\omega\rvert$
$0.9879-0.8783i$	$-0.1293+0.2056i4$	0.0865
$2.4653-0.8112i$	$-0.3676-0.1974i$	0.0595
$3.6017-0.6714i$	$-0.0330-0.4275i$	0.0321
$4.1717-0.7518i$	$-0.0375-0.6301i$	0.0353
$4.3646-0.9219i$	$0.5716-0.2886i$	0.0322
$5.1199-0.6245i$	$0.2352-0.1327i$	0.0102
$4.7418-2.3434i$	$-6.2911+9.9605i$	0.4134
$4.7142-3.1680i$	$2.3877+5.3463i$	0.1772
$6.2217-0.9224i$	$0.7460+0.7706i$	0.0273
$6.3881-0.5147i$	$0.0291+0.1300i$	0.0033
$6.7526-0.5452i$	$0.5148-0.0535i$	0.0114
$6.6171-2.3360i$	$-4.3000+12.1518i$	0.2605
$6.9928-0.8666i$	$-0.1614+0.3781i$	0.0083

Table 5.1: Computed perturbations of the eigenvalues of Example 5.1 for a perturbation in the exterior with $|\Delta s_0| \approx 0.07$. Eigenvalues whose perturbation is larger than Δs_0 (that is: spurious solutions) are marked red.

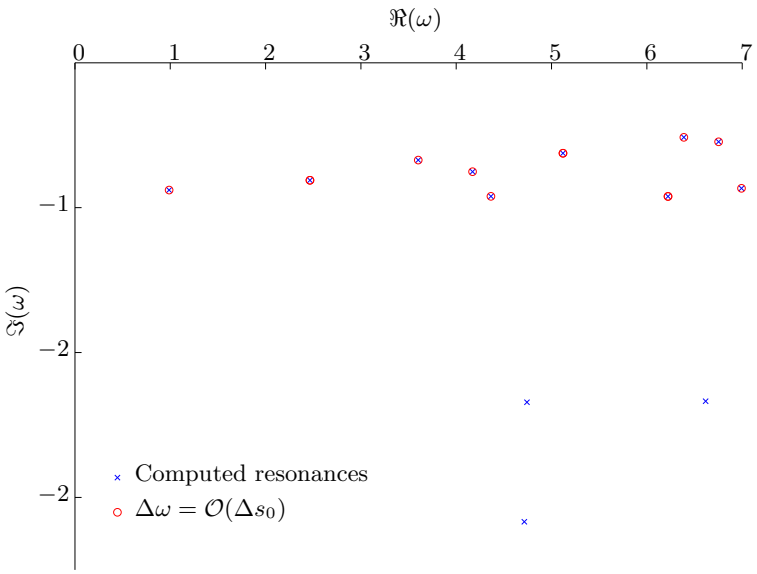

Figure 5.21: Spectrum for Example 5.1 (blue crosses) with eigenvalues whose perturbation is below a threshold marked with red circles.

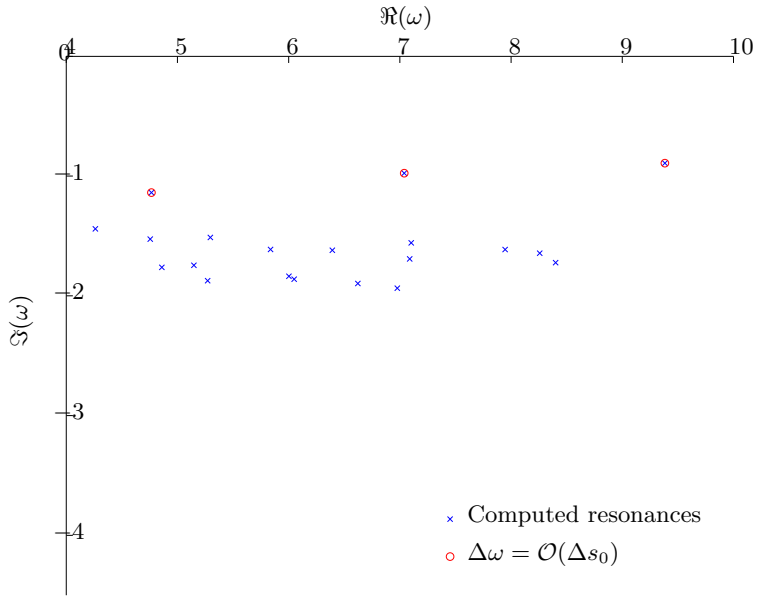

Figure 5.22: Spectrum for Example 5.3 (blue crosses) with eigenvalues whose perturbation is below a threshold marked with red circles.

thus in the exterior it holds that

$$2s_0 \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \mathbf{u}_{\text{ext}} - \omega^2 \frac{2n_{\text{ext}}^2}{s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \mathbf{u}_{\text{ext}} = 0, \tag{5.16}$$

where $\mathbf{u}_{\text{ext}}$ are the degrees of freedom for the respective part of the exterior domain and n_{ext} is the refractive index therein. Inserting the definitions of $\mathcal{T}_L^{(+)}$ and $\mathcal{T}_L^{(-)}$ from Equation (3.45), and naming the entries of $\mathbf{u}_{\text{ext}} = (z_0, z_1, \dots z_L)^\top$, the matrix form of Equation (5.16) reads:

$$s_0 \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & 1 & \frac{1}{2} & \\ & \frac{1}{2} & 1 & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \end{pmatrix} - \omega^2 \frac{n_{\text{ext}}^2}{s_0} \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & & \\ -\frac{1}{2} & 1 & -\frac{1}{2} & \\ & -\frac{1}{2} & 1 & \\ & & & \ddots \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ z_2 \\ \vdots \end{pmatrix} = 0.$$

For $l \geq 3$, the matrix equation corresponds to a linear second order recurrence with coefficients depending on ω^2. Using general coefficients, the $l+2$nd line of this second order recurrence reads

$$\widetilde{a}_2(\omega^2) z_{l-2} + \widetilde{a}_1(\omega^2) z_{l-1} + \widetilde{a}_0(\omega^2) z_l = 0. \tag{5.17}$$

We will now establish a lemma about linear recurrence relations. Its proof can be easily found by direct calculations.

Lemma 5.5.

Given a second order linear recurrence relation $a_0 x_k = a_1 x_{k-1} + a_2 x_{k-2}$ with the characteristic polynomial

$$\chi(t) = a_0 t^2 - a_1 t - a_2.$$

Let t_1 and t_2 be the roots of $\chi(t)$. Then t_1^k and t_2^k each solve the recurrence relation. The convergence rate of the solution corresponding to t_i^k is given by $|t_i^{k+1}| = \kappa |t_i^k|$, hence $\kappa = \frac{|t_i^{k+1}|}{|t_i^k|} = |t_i|$.

The stability of the solution of a linear second order recurrence relation therefore depends on the roots of its characteristic polynomial and the convergence rate on their modulus. We therefore have a closer look at the characteristic polynomial of the second order linear recurrence relation (5.17) arising in our implementation. Its characteristic polynomial is

$$\chi_{\omega^2}(z) = \widetilde{a}_0(\omega^2) z^2 - \widetilde{a}_1(\omega^2) z - \widetilde{a}_2(\omega^2). \tag{5.18}$$

The subscript ω^2 indicates that we see ω^2 as parameter in this setting. Inserting the known coefficients, Equation (5.18) reads

$$\chi_{\omega^2}(z) = \left(\frac{s_0}{2} + \omega^2 \frac{n_{\text{ext}}^2}{2 s_0} \right) z^2 + \left(s_0 - \omega^2 \frac{n_{\text{ext}}^2}{s_0} \right) z + \left(\frac{s_0}{2} + \omega^2 \frac{n_{\text{ext}}^2}{2 s_0} \right). \tag{5.19}$$

The roots of Equation (5.19) are

$$z_1 = \frac{n_{\text{ext}}\omega + i s_0}{n_{\text{ext}}\omega - i s_0} \quad \text{and} \quad z_2 = \frac{n_{\text{ext}}\omega - i s_0}{n_{\text{ext}}\omega + i s_0}.$$

Clearly it holds that $z_1 z_2 = 1$, hence, if the solution corresponding to z_1 is asymptotically stable, the solution corresponding to z_2 diverges and vice versa. Similar to the argumentation used for Equation (3.18), we can identify one of the roots, z_1 with an outgoing solution and the other root, z_2 with an incoming solution.

We can therefore restrict ourselves to investigating the convergence behavior of the solution connected with z_2. In order that the solution converges with a convergence rate $0 < \kappa < 1$, we require that $|z_2| = \kappa$, hence

$$|n_{\text{ext}}\omega - i s_0| = \kappa |n_{\text{ext}}\omega + i s_0| \tag{5.20}$$

and

$$|n_{\text{ext}}\omega - i s_0|^2 = \kappa^2 |n_{\text{ext}}\omega + i s_0|^2.$$

This can be reformulated

$$\begin{aligned}(n_{\text{ext}}\omega - is_0)\overline{(n_{\text{ext}}\omega - is_0)} &= \kappa^2 (n_{\text{ext}}\omega + is_0)\overline{(n_{\text{ext}}\omega + is_0)} \\ \left(\frac{n_{\text{ext}}\omega}{is_0} - 1\right)\overline{\left(\frac{n_{\text{ext}}\omega}{is_0} - 1\right)} &= \kappa^2 \left(\frac{n_{\text{ext}}\omega}{is_0} + 1\right)\overline{\left(\frac{n_{\text{ext}}\omega}{is_0} + 1\right)}\end{aligned}$$

Splitting by real and imaginary part of $n_{\text{ext}}\omega/is_0$, we get:

$$\left(\Re\left(\frac{n_{\text{ext}}\omega}{is_0}\right) - 1\right)^2 + \Im\left(\frac{n_{\text{ext}}\omega}{is_0}\right)^2 = \kappa^2 \left[\left(\Re\left(\frac{n_{\text{ext}}\omega}{is_0}\right) + 1\right)^2 + \Im\left(\frac{n_{\text{ext}}\omega}{is_0}\right)^2\right].$$

By expanding the quadratic terms, dividing by $1-\kappa^2$ and completing the square, we can isolate $\Re(n_{\text{ext}}\omega/is_0)$ and $\Im(n_{\text{ext}}\omega/is_0)$. We then obtain

$$\left[\Re\left(\frac{n_{\text{ext}}\omega}{is_0}\right) - \frac{1+\kappa^2}{1-\kappa^2}\right]^2 + \Im\left(\frac{n_{\text{ext}}\omega}{is_0}\right)^2 = \left(\frac{1+\kappa^2}{1-\kappa^2}\right)^2 - 1. \tag{5.21}$$

Since $(\Re(a)-b)^2+\Im(a)^2 = (\Re(a)-b)^2-(i\Im(a))^2 = (\Re(a)+i\Im(a)-b)(\Re(a)-i\Im(a)-b)$, we can isolate ω in Equation (5.21) and derive the domain C where the series converges with a convergence rate of κ to be

$$C = \left\{\omega \in \mathbb{C} : \left|\omega n_{\text{ext}} - is_0\frac{1+\kappa^2}{1-\kappa^2}\right| \le |is_0|\sqrt{\left(\frac{1+\kappa^2}{1-\kappa^2}\right)^2 - 1}\right\}. \tag{5.22}$$

Now we will revisit Example 4.1 and apply the results of the previous section before deriving a higher-dimensional version. In Figure 5.23 the spectrum of the cavity is plotted again with blue crosses. The red box marks the value of s_0. The grey circles, where the outermost, darkest circle corresponds to a rate of convergence of $\kappa = 0.9$ and the radius of the circles decreases for decreasing values of κ up to the innermost (lightest) circle which corresponds to a rate of convergence of $\kappa = 0.3$.

We can see that for a rate of convergence of $\kappa = 0.6$, the circle contains only physical and no spurious solutions. Since, as before, we chose $L = 15$ terms in the power series that approximates $\mathcal{M}_{s_0}\mathcal{L}\{u_{\text{ext}}\}$, this gives a value of approximately $4 \cdot 10^{-4}$ times the boundary value, which for our setting is in the order of one. This means it is reasonable to take these solutions to be converged.

For Example 5.2 the domains of convergence are displayed in Figure 5.24 for our choice of s_0. The circles correspond to rates of convergence of $\kappa = 0.5$ to $\kappa = 0.8$. So by setting a rate of convergence that seems reasonable, e.g. for $\kappa = 0.7$, $L = 25$ degrees of freedom give about $1.45 \cdot 10^{-4}$ times the boundary value, which we consider well-converged. As expected, we can see that there are only solutions that we identified as physical solutions within the circle corresponding to $\kappa = 0.7$.

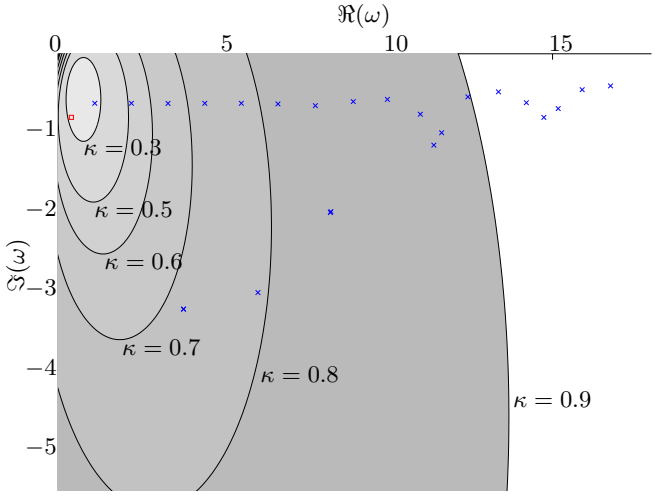

Figure 5.23: The spectrum (blue crosses) $\sigma(A, B)$ of Example 4.1. The shaded gray circles are the regions of convergence of the linear second order recurrence relation in the exterior domain for different rates of convergence $\kappa = 0.3$ (lightest, innermost circle) to $\kappa = 0.9$ (darkest, outermost circle). The circles are cropped to the area of interest.

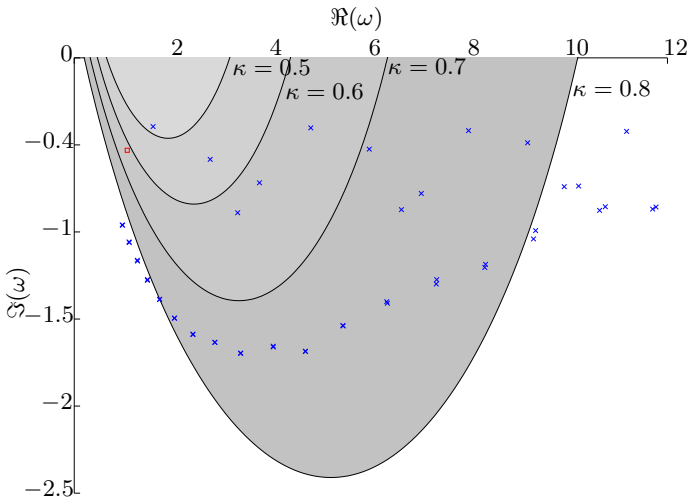

Figure 5.24: The spectrum (blue crosses) $\sigma(A, B)$ of Example 5.2. The gray circles are the regions of convergence of the linear second order recurrence relation in the exterior domain for different rates of convergence $\kappa = 0.5$ (lightest, innermost circle) to $\kappa = 0.8$ (darkest, outermost circle). The circles are cropped to the area of interest.

We will now try to generalize the reasoning that gave the regions of convergence in the one-dimensional case to higher space dimensions. As we have already established in Equations (3.63) and (3.64), in the two-dimensional case the matrices in the exterior are

$$\begin{aligned} A^{\text{ext}}_{\text{loc},i} &:= T^{\text{ext}}_{\text{loc},i,2} \otimes \left[\frac{-2h_\xi}{s_0} \mathcal{T}_L^{(-)\top} \left(h_\eta \,\mathrm{id} + \frac{a+b}{2s_0} \mathcal{D}_L \right)^{-1} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,3} \otimes \left[(-2) \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,4} \otimes \left[(-2) \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,5} \otimes \left[\frac{-2s_0}{h_\xi} \mathcal{T}_L^{(+)\top} \left(h_\eta \,\mathrm{id} + \frac{a+b}{2s_0} \mathcal{D}_L \right) \mathcal{T}_L^{(+)} \right] \text{ and} \end{aligned} \tag{5.23}$$

$$B^{\text{ext}}_{\text{loc},i} := n_i^2 T^{\text{ext}}_{\text{loc},i,1} \otimes \left[\frac{-2h_\xi}{s_0} \mathcal{T}_L^{(-)\top} \left(h_\eta \,\mathrm{id} + \frac{a+b}{s_0} \mathcal{D}_L \right) \mathcal{T}_L^{(-)} \right]. \tag{5.24}$$

In order to obtain a formulation that we are able to deal with, we will investigate the simpler case where the infinite edges of the prismatoid are parallel and perpendicular to the boundary of Ω_{int}, that is $a = b = 0$.

This will give us the simpler local stiffness and mass matrices

$$\begin{aligned} A^{\text{ext}}_{\text{loc},i} &= T^{\text{ext}}_{\text{loc},i,2} \otimes \left[\frac{-2h_\xi}{s_0} \mathcal{T}_L^{(-)\top} \left(h_\eta \,\mathrm{id} \right)^{-1} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,3} \otimes \left[(-2) \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(+)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,4} \otimes \left[(-2) \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,5} \otimes \left[\frac{-2s_0}{h_\xi} \mathcal{T}_L^{(+)\top} \left(h_\eta \,\mathrm{id} \right) \mathcal{T}_L^{(+)} \right] \\ &= T^{\text{ext}}_{\text{loc},i,2} \otimes \left[\frac{-2h_\xi}{h_\eta s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,3} \otimes \left[(-2) \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(+)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,4} \otimes \left[(-2) \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(-)} \right] \\ &\quad + T^{\text{ext}}_{\text{loc},i,5} \otimes \left[\frac{-2h_\eta s_0}{h_\xi} \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] \text{ and} \end{aligned} \tag{5.25}$$

$$\begin{aligned} B^{\text{ext}}_{\text{loc},i} &= n_i^2 T^{\text{ext}}_{\text{loc},i,1} \otimes \left[\frac{-2h_\xi}{s_0} \mathcal{T}_L^{(-)\top} \left(h_\eta \,\mathrm{id} \right) \mathcal{T}_L^{(-)} \right] \\ &= n_i^2 T^{\text{ext}}_{\text{loc},i,1} \otimes \left[\frac{-2h_\xi h_\eta}{s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right]. \end{aligned} \tag{5.26}$$

Due to the absence of the tridiagonal $\mathcal{D}_L$ which caused $A^{\text{ext}}_{\text{loc},i}$ and $B^{\text{ext}}_{\text{loc},i}$ to have more nonzero diagonals, both matrices now have a tridiagonal structure. Moreover, for our simplified setting the finite integrals $T^{\text{ext}}_{\text{loc},i,3}$ and $T^{\text{ext}}_{\text{loc},i,4}$ (defined in Equations (3.59c) and (3.59d)) are zero. Also in order for our discretization with prismatiods to be valid for the case of parallel infinite sides, we require, that h_ξ is equal for each prismatoid, without restriction of generality we can therefore set $h_\xi = 1$, leaving h_η, which on the ith prismatoid we will index as $h_{\eta,i}$, to determine the coupling of the two infinite sides of each prismatoid. Putting it all together, we can give discrete ξ-directional part of the local infinite element stiffness matrix $A^{\text{ext}}_{\text{loc},i}$ as

$$\begin{aligned} A^{\text{ext}}_{\text{loc},i} &= T^{\text{ext}}_{\text{loc},i,2} \otimes \left[\frac{(-2)}{h_{\eta,i} s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] + \\ & \quad T^{\text{ext}}_{\text{loc},i,5} \otimes \left[(-2) h_{\eta,i} s_0 \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right]. \end{aligned}$$

and the local infinite element mass matrix $B^{\text{ext}}_{\text{loc},i}$ as

$$B^{\text{ext}}_{\text{loc},i} = n_i^2 T^{\text{ext}}_{\text{loc},i,1} \otimes \left[\frac{(-2) h_{\eta,i}}{s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right].$$

If we choose standard finite elements to discretize the interior Ω_{int}, then the traces of these elements on the boundary Γ between Ω_{int} and Ω_{ext} are standard one-dimensional finite elements and $T^{\text{ext}}_{\text{loc},i,1}$, $T^{\text{ext}}_{\text{loc},i.2}$ and $T^{\text{ext}}_{\text{loc},i,5}$ are the standard one-dimensional finite element matrices. Using linear finite edge elements in the interior and inserting the well-known one-dimensional matrices for the boundary integrals, the local infinite element stiffness matrix $A^{\text{ext}}_{\text{loc},i}$ is

$$\begin{aligned} A^{\text{ext}}_{\text{loc},i} &= \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes \left[\frac{(-2)}{h_{\eta,i} s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] + \\ & \quad \frac{1}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \otimes \left[(-2) h_{\eta,i} s_0 \mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] \\ &= \frac{1}{h_{\eta,i} s_0} \begin{pmatrix} (-2) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] & 2 \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \\ 2 \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] & (-2) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \end{pmatrix} + \\ & \quad \frac{h_{\eta,i} s_0}{6} \begin{pmatrix} (-4) \left[\mathcal{T}_L^{(+)\top} \mathcal{T}^{(+)} \right] & (-2) \left[\mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] \\ (-2) \left[\mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] & (-4) \left[\mathcal{T}_L^{(+)\top} \mathcal{T}_L^{(+)} \right] \end{pmatrix}. \end{aligned} \tag{5.27}$$

The local infinite element mass matrix $B^{\text{ext}}_{\text{loc},i}$ is

$$\begin{aligned} B^{\text{ext}}_{\text{loc},i} &= \frac{n_i^2}{6} \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \otimes \left[\frac{(-2) h_{\eta,i}}{s_0} \mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \\ &= \frac{n_i^2 h_{\eta,i}}{6 s_0} \begin{pmatrix} (-4) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}^{(-)} \right] & (-2) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \\ (-2) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] & (-4) \left[\mathcal{T}_L^{(-)\top} \mathcal{T}_L^{(-)} \right] \end{pmatrix}. \end{aligned} \tag{5.28}$$

In order to carry out a convergence analysis that is similar to the one-dimensional approach, we will now reformulate the two-dimensional problem as linear second order matrix recurrence relation. As in the one-dimensional case, we will first discard the discretization in the interior Ω_{int} and assume that it is discretized using N degrees of freedom. Furthermore, we will discard the first Hardy modes that couple to the interior and only look at the modes k for $k \in \{2, ..., L\}$ on each infinite trapezoid of the exterior domain Ω_{ext}. We will assume that there are k such trapezoids and reference them using the index j.

We will now explore the way these exterior degrees of freedom enter the global stiffness and mass matrices. We have established before in Equations (5.27) and (5.28) that they both consist of two coupled tri-diagonal block matrices for each trapezoid, dividing the corresponding vector of unknowns into two blocks. Each of these blocks corresponds to one infinite side of the trapezoid. Since each infinite ray is the boundary of two prismatoids, each of these tri-diagonal block matrices couples with two other blocks, giving the global layout that is sketched in Figure 5.25 in the right-hand side plot. In order to obtain a second order linear recurrence, we will assume the interior degrees of freedom to be well converged and only take into account the exterior degrees of freedom. In the global vector of unknowns $\mathbf{u}$ they are located at the indices $N, ..., N + kL$. The unknowns $\mathbf{u}_{N+iL}, ...\mathbf{u}_{N+(i+1)L-1}, i \in \{0, ..., k\}$ correspond to the Hardy modes on one infinite ray.

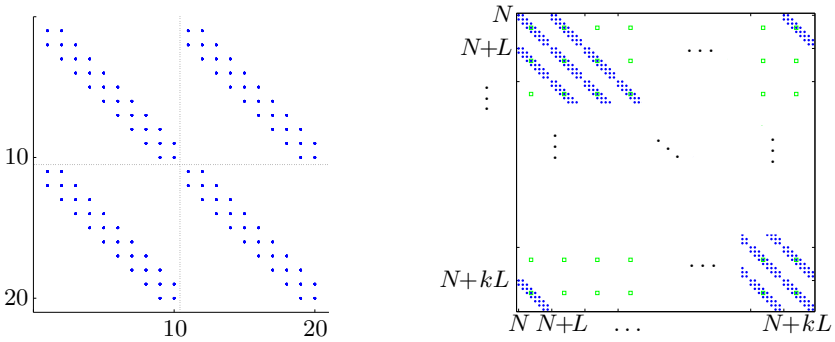

Figure 5.25: Left: The structure of the local infinite element stiffness matrix for one infinite trapezoid in the simplified case. We can see two coupled second order recurrence relations, one for each infinite side of the trapezoid. The discretization is done with $L = 10$ degrees of freedom. Right: coupling of the local element matrices in the global stiffness matrix, zoom into the structure for the exterior degrees of freedom.

In order to obtain a formulation that allows for a similar treatment as in the one-dimensional situation and allows for computation of a domain of

convergence, we will now collect the degrees of freedom that correspond to the same Hardy mode on each ray. That is, for $m \in \{0, ..., L-1\}$ we will create a vector that collects the unknowns corresponding to the mth Hardy modes on each ray:

$$\mathbf{u}_{\text{ext}}^{(m)} = (\mathbf{u}_{N+m}, \mathbf{u}_{N+L+m}, \mathbf{u}_{N+2L+m}, ..., \mathbf{u}_{N+kL+m})^\top . \tag{5.29}$$

Discarding the interior discretization and coupling with it which only affects the first degree of freedom, in each of the tri-diagonal blocks, the unknown $\mathbf{u}_{N+iL+m}$ is related with two neighboring unknowns $\mathbf{u}_{N+iL+m-1}$ and $\mathbf{u}_{N+iL+m+1}$ where $i \in \{0, ..., k\}$ and $m \in \{0, ..., L-1\}$. Naming the entries of the global stiffness matrix $\alpha_{i,j}$ and of the global mass matrix $\beta_{i,j}$, this relation is given by

$$\begin{aligned} &(\alpha_{N+iL+m,N+iL+m-1} - n_i^2\omega^2\beta_{N+iL+m,N+iL+m-1}) && \mathbf{u}_{N+iL+m-1} \\ &+(\alpha_{N+iL+m,N+iL+m} - n_i^2\omega^2\beta_{N+iL+m,N+iL+m}) && \mathbf{u}_{N+iL+m} \\ &+(\alpha_{N+iL+m,N+iL+m+1} - n_i^2\omega^2\beta_{N+iL+m,N+iL+m+1}) && \mathbf{u}_{N+iL+m+1} \quad = 0. \end{aligned}$$

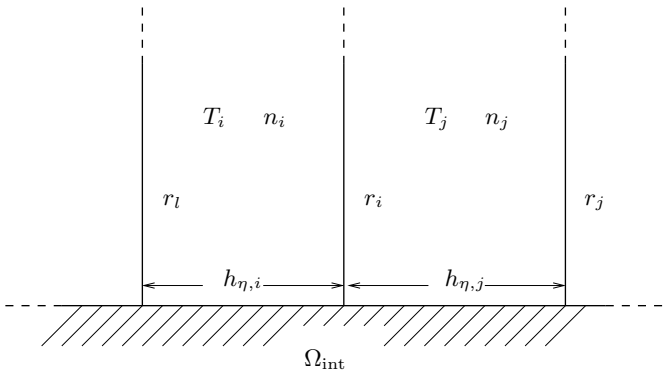

Figure 5.26: The situation in the exterior domain for our simplified problem. The ith and jth infinite trapezoids T_i and T_j with the refractive indices n_i and n_j share one infinite side r_i. The infinite rays to the right and to the left of r_i are labeled r_j and r_l.

However, since each infinite ray r_i couples with two other rays, r_j and r_l as depicted in Figure 5.26, we have such a relation for three different values of i in each row. Using the vectors $\mathbf{u}_{\text{ext}}^{(m)}$ defined in Equation (5.29), we have a linear second order matrix recurrence relation

$$M_\omega^{(0)}\mathbf{u}_{\text{ext}}^{(m-1)} + M_\omega^{(1)}\mathbf{u}_{\text{ext}}^{(m)} + M_\omega^{(2)}\mathbf{u}_{\text{ext}}^{(m+1)} = 0. \tag{5.30}$$

The coefficient matrices $M_\omega^{(0)}, M_\omega^{(1)}$ and $M_\omega^{(2)}$ are complex $k \times k$ matrices that depend on ω^2 and contain three nonzero entries composed of $\alpha_{i,j}$ and $\beta_{i,j}$ per row. In the following paragraph we will determine $M_\omega^{(0)}, M_\omega^{(1)}$ and

$M_\omega^{(2)}$. For this we will insert the known forms of $A_{\mathrm{loc},i}^{\mathrm{ext}}$ and $B_{\mathrm{loc},i}^{\mathrm{ext}}$ from Equations (5.27) and (5.28) and recall that

$$2\mathcal{T}_L^{(+)\top}\mathcal{T}_L^{(+)} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & 1 & \frac{1}{2} & \\ & \ddots & \ddots & \\ & & \frac{1}{2} & 1 \end{pmatrix} \text{ and}$$

$$2\mathcal{T}_L^{(-)\top}\mathcal{T}_L^{(-)} = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} & & \\ -\frac{1}{2} & 1 & -\frac{1}{2} & \\ & \ddots & \ddots & \\ & & -\frac{1}{2} & 1 \end{pmatrix}.$$

We will now compute the nonzero entries for $M_\omega^{(0)}$, $M_\omega^{(1)}$ and $M_\omega^{(2)}$. First, we will deal with $M_\omega^{(1)}$ which is the entry arising due to the main diagonals of the tri-diagonal blocks in the local element matrices. One of the entries for $M_\omega^{(1)}$ is caused by the main diagonal the local element matrices for each trapezoid the ray belongs to. We index them T_i and T_j, hence $A_{\mathrm{loc},i}^{\mathrm{ext}}$ and $A_{\mathrm{loc},j}^{\mathrm{ext}}$ both contribute to this entry in $M_\omega^{(1)}$ which therefore reads

$$-\left(\frac{1}{3}h_{\eta,i} + \frac{1}{3}h_{\eta,j}\right)s_0 - \left(\frac{1}{h_{\eta,i}} + \frac{1}{h_{\eta,j}}\right)s_0^{-1}. \tag{5.31}$$

In the same way, the main diagonals of $B_{\mathrm{loc},i}^{\mathrm{ext}}$ and $B_{\mathrm{loc},j}^{\mathrm{ext}}$ add to the same entry of $M_\omega^{(1)}$:

$$-\left(\frac{1}{3}h_{\eta,i}n_i^2 + \frac{1}{3}h_{\eta,j}n_j^2\right)s_0^{-1}. \tag{5.32}$$

The upper right and lower left blocks of $A_{\mathrm{loc},i}^{\mathrm{ext}}$, $A_{\mathrm{loc},j}^{\mathrm{ext}}$, $B_{\mathrm{loc},i}^{\mathrm{ext}}$ and $B_{\mathrm{loc},j}^{\mathrm{ext}}$ contribute two entry to each row of $M_\omega^{(1)}$. Put together, the entries from the local element stiffness matrices read

$$\frac{1}{h_{\eta,i}s_0} - \frac{1}{6}h_{\eta,i}s_0 \quad \text{and} \quad \frac{1}{h_{\eta,j}s_0} - \frac{1}{6}h_{\eta,j}s_0, \tag{5.33}$$

and the entries from the local mass matrices read

$$-n_i^2\frac{h_{\eta,i}}{6s_0} \quad \text{and} \quad -n_j^2\frac{h_{\eta,j}}{6s_0}. \tag{5.34}$$

So the diagonal entries of $M_\omega^{(1)}$ that correspond to the infinite ray r_i are made up of the summands computed in Equations (5.31) and (5.32) and are

$$-\left(\frac{1}{3}h_{\eta,i} + \frac{1}{3}h_{\eta,j}\right)s_0 - \left(\frac{1}{h_{\eta,i}} + \frac{1}{h_{\eta,j}}\right)s_0^{-1} + \omega^2\left(\frac{1}{3}h_{\eta,i}n_i^2 + \frac{1}{3}h_{\eta,j}n_j^2\right)s_0^{-1}$$

while Equations (5.33) and (5.34) give the two off-diagonal entries in the columns corresponding to the infinite rays r_j and r_l of the trapezoids T_i and T_j, to the left and to the right of the ith ray. These entries are

$$\frac{1}{h_{\eta,i}s_0} - \frac{1}{6}h_{\eta,i}s_0 + \omega^2 n_i^2 \frac{h_{\eta,i}}{6s_0}$$

and

$$\frac{1}{h_{\eta,j}s_0} - \frac{1}{6}h_{\eta,i}s_0 + \omega^2 n_j^2 \frac{h_{\eta,j}}{6s_0}.$$

The entries in $M_\omega^{(0)}$ and $M_\omega^{(2)}$ can be computed in the same way and differ from the entries of $M_\omega^{(1)}$ by a factor $\pm 1/2$ and read

$$-\left(\frac{1}{6}h_{\eta,i} + \frac{1}{6}h_{\eta,j}\right)s_0 + \frac{1}{2}\left(\frac{1}{h_{\eta,i}} + \frac{1}{h_{\eta,j}}\right)s_0^{-1} - \omega^2\left(\frac{1}{6}h_{\eta,i} + \frac{1}{6}h_{\eta,j}\right)s_0^{-1}$$

on the diagonal and

$$-\frac{1}{2h_{\eta,i}s_0} - h_{\eta,i}s_0\frac{1}{12} - \omega^2 n_i^2 \frac{h_{\eta,i}}{12s_0}$$

and

$$-\frac{1}{2h_{\eta,j}s_0} - h_{\eta,j}s_0\frac{1}{12} - \omega^2 n_j^2 \frac{h_{\eta,j}}{12s_0}$$

for the off-diagonal entries. Since in Section 3.5 we assumed the boundary Γ between Ω_{int} and Ω_{ext} to be an arbitrary convex polygon, we can now reorder the unknowns in a way that the trapezoid to the left of T_i is T_{i-1} and the trapezoid to the right of T_i is T_{i+1} for $2 \leq i \leq k-1$. The $k \times k$ matrices $M_\omega^{(0)}$, $M_\omega^{(1)}$ and $M_\omega^{(2)}$ for the linear second order matrix recurrence relation then are tri-diagonal matrices. Since T_1 couples with T_k, they have one off-diagonal entry at $(1,k)$ and one at $(k,1)$.

Hence, we can compute the $k \times k$ coefficient matrices $M_\omega^{(0)}, M_\omega^{(1)}$ and $M_\omega^{(2)}$ of the linear second order matrix recurrence relation

$$M_\omega^{(0)}\mathbf{u}_{\text{ext}}^{(m-1)} + M_\omega^{(1)}\mathbf{u}_{\text{ext}}^{(m)} + M_\omega^{(2)}\mathbf{u}_{\text{ext}}^{(m+1)} = 0. \tag{5.35}$$

In order to be able to gain some insight about the stability of its solutions, we will transform it to a more convenient form by rephrasing it as linear first order matrix recurrence relation with a $2k \times 2k$ coefficient matrix. This is done by solving Equation (5.35) for $\mathbf{u}_{\text{ext}}^{(m+1)}$ and then concatenating the two

vectors $\mathbf{u}_{\text{ext}}^{(j+1)}$ and $\mathbf{u}_{\text{ext}}^{(j)}$ into a vector:

$$\begin{aligned} 0 =& M_\omega^{(0)}\mathbf{u}_{\text{ext}}^{(m-1)} + M_\omega^{(1)}\mathbf{u}_{\text{ext}}^{(m)} + M_\omega^{(2)}\mathbf{u}_{\text{ext}}^{(m+1)} \qquad (5.36)\\ \Leftrightarrow \quad \mathbf{u}_{\text{ext}}^{(m+1)} =& \left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(1)}\mathbf{u}_{\text{ext}}^{(m)} + \left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(0)}\mathbf{u}_{\text{ext}}^{(m-1)} \\ \Leftrightarrow \quad \begin{pmatrix}\mathbf{u}_{\text{ext}}^{(m+1)}\\ \mathbf{u}_{\text{ext}}^{(m)}\end{pmatrix} =& \underbrace{\begin{pmatrix}\left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(1)} & \left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(0)}\\ \text{id} & 0\end{pmatrix}}_{=:C} \begin{pmatrix}\mathbf{u}_{\text{ext}}^{(m)}\\ \mathbf{u}_{\text{ext}}^{(m-1)}\end{pmatrix} \end{aligned}$$

Remark 5.4.

The existence of $\left(M_\omega^{(2)}\right)^{-1}$ can be deducted from the structure of $M_\omega^{(2)}$. To see this, we partition $M_\omega^{(2)}$ as

$$M_\omega^{(2)} = \left(\begin{array}{c|c} M_1 & M_2 \\ \hline M_3 & M_4 \end{array}\right).$$

Then M_1 and M_4 are tri-diagonal matrices and thus invertible. That means that their Schur complements S_{M_1} and S_{M_4} exist. Using these Schur complements, we can give the inverse of $M_\omega^{(2)}$. Direct calculations [Ber09] then show that $\left(M_\omega^{(2)}\right)^{-1}$ reads

$$\left(M_\omega^{(2)}\right)^{-1} = \left(\begin{array}{c|c} (S_{M_4})^{-1} & -M_1^{-1}M_2\,(S_{M_1})^{-1} \\ \hline -M_4^{-1}M_3\,(S_{M_4})^{-1} & (S_{M_1})^{-1} \end{array}\right).$$

As in the one-dimensional case, it is an established fact, that the stability of the solutions of a linear vector iteration $x_{n+1} = Cx_n$ with $C \in \mathbb{C}^{n\times n}$ depends on the corresponding eigenvalues of the coefficient matrix C (see e.g. [DB02, Theorem 3.33]). However, due to the inversion of $M_\omega^{(2)}$ and the computation of the eigenvalues of C, it is not possible to find a closed formula for the computation of the domain of convergence in the two-dimensional case. Instead, we will have to do a sampling for different values of ω and compute the eigenvalues of C for each ω. That means for each value of ω, we compute the matrix C and its eigenvalues. This gives $2k$ eigenvalues for each ω. It can be seen however, that these $2k$ eigenvalues are clustered around two centers c_1 and c_2, one corresponding to an outgoing solution and one corresponding to an incoming solution. The two centers and the mean deviation from them can be computed using a kmeans algorithm. As in the one-dimensional case, the modulus of the eigenvalue corresponding to the outgoing solution determines the convergence rate of the solution. The steps for computing the clusters on a grid in the complex plane are given in Algorithm 1. Even though Algorithm 1 requires one matrix inversion and

1: **for** $0 \leq \Re(\omega) \leq \Re_{max}$ **do**
2: **for** $\Im_{min} \leq \Im(\omega) \leq 0$ **do**
3: compute $M_\omega^{(2)}$, $M_\omega^{(1)}$ and $M_\omega^{(0)}$
4: assemble $C = \begin{pmatrix} \left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(1)} & \left(-M_\omega^{(2)}\right)^{-1} M_\omega^{(0)} \\ \mathrm{id} & 0 \end{pmatrix}$
5: $\mathbf{e}$ = eigenvalues of C # *This gives* $2k$ *eigenvalues*
6: **if** partitioning of $\mathbf{e}$ with kmeans possible **then**
7: c_1, c_2 = centers of partitions
8: store modulus of center corresponding to outgoing solution in map M
9: **else**
10: **return** Error: Computation failed
11: **end if**
12: **end for**
13: **end for**
14: **return** M

Algorithm 1: Computing the convergence rate in the exterior on a grid in the complex plane.

the computation of the eigenvalues of a matrix for each value of ω, its costs are moderate since the size of the matrices involved, k, corresponds to the number of prismatoids used for the discretization of the exterior which is modest for typical applications.

Figure 5.27 shows the convergence rate computed for a homogeneous two-dimensional exterior domain with $k = 25$ prismatoids. The complex plane is discretized for $0 \leq \Re(\omega) \leq 10$ and $-10 \leq \Im(\omega) \leq 0$ with a rectangular mesh with step size $h = 0.1$. The 10201 calculations done for computing the convergence rate in this area were done on a standard PC in $t = 129s$. For typical applications the region that is scanned can be reduced and the step size increased, yielding much lower computation times.

We will now apply this to the two-dimensional examples introduced in Section 5.2. First we will cover Example 5.1. We use Algorithm 1 to compute the regions of the complex plane, where the exterior converges with certain rates of convergence. The result can be seen in Figure 5.28. We can see that for a rate of convergence of $\kappa = 0.6$ no spurious solutions are within the region where we can expect convergence, if we relax the admissible rate of convergence to $\kappa = 0.7$ then the first spurious solution enters the spectrum. The shape and position of the plateaus for the different values of κ can be adjusted by altering the choice of s_0. For the final Example 5.3, the result can be seen in Figure 5.29. For this example for a rate of convergence of $\kappa = 0.5$, no spurious solutions are considered to be converged. If we allow for a rate of convergence that is lower than $\kappa = 0.5$, spurious solutions are

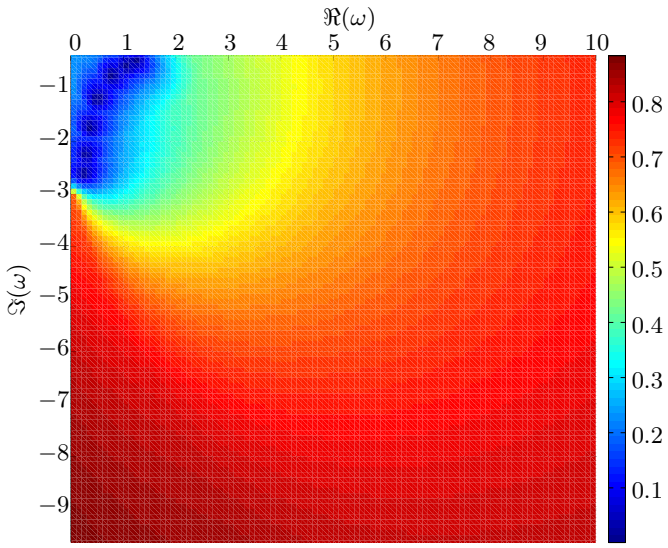

Figure 5.27: The computed convergence rate color coded for a region in the complex plane.

considered to be converged for this example.

We conclude that we have presented a method for the detection of spurious solutions. Contrary to condition numbers, the method of computing the perturbation of the eigenvalues directly is dependent on the precise perturbation we apply to the pole condition parameter, making it sensitive only to the transparent boundary condition and not to the solution of the inner problem. This makes for a good way of detecting spurious solutions that are caused by the transparent boundary condition. To obtain statements on the validity of these predictions, we have demonstrated how to complement it with a convergence monitor.

5.5 Putting All Together: An Algorithm for Removing Spurious Solutions

Now we have assembled all the tools required to formulate the central result of this thesis: an algorithm that solves an eigenvalue problem and uses the methods derived in this thesis to remove the spurious solutions from the computed eigenvalue spectrum. The steps this algorithm performs are as follows

1. **Input**: Read the problem geometry together with the discretization parameters such as finite element degree for the solution in Ω_{int}, pole condition parameter s_0, its perturbation Δs_0 and the number of Hardy

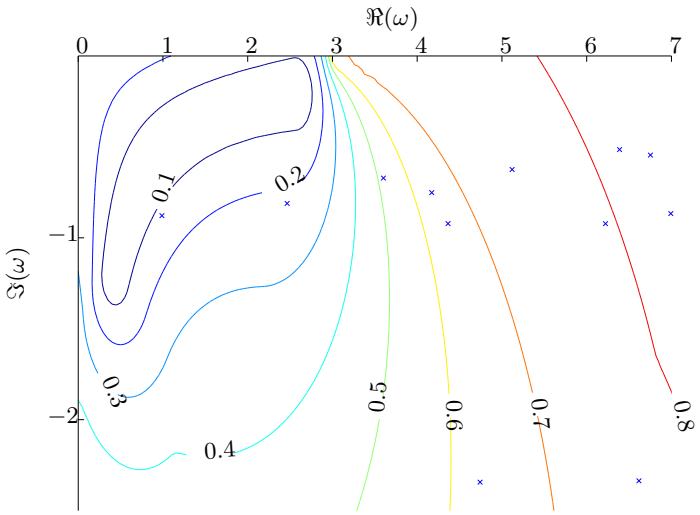

Figure 5.28: Spectrum of Example 5.1 with the regions of convergence. For a rate of convergence of $\kappa = 0.6$ we can see that no spurious solutions are converged, allowing for a lower rate of convergence of $\kappa = 0.7$, the first spurious solution is included in the spectrum.

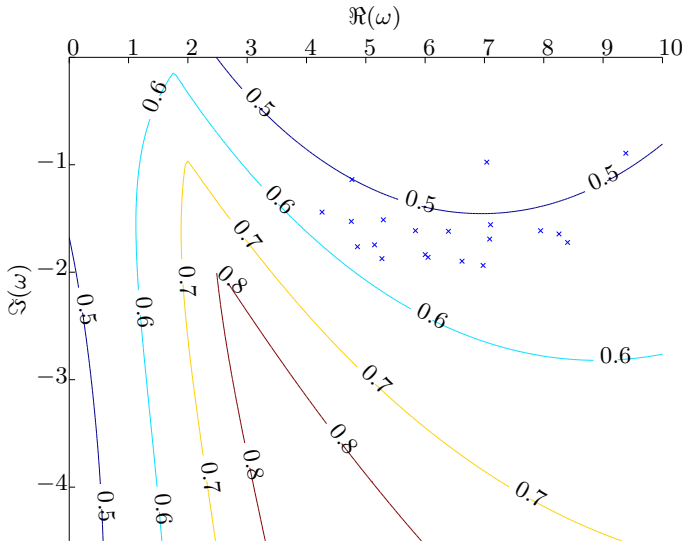

Figure 5.29: Spectrum of Example 5.3 with the regions of convergence. For a rate of convergence of $\kappa = 0.5$ we can see that no spurious solutions are converged, allowing for a lower rate of convergence of $\kappa = 0.6$, the first spurious solutions are included in the spectrum.

modes L used for the pole condition.

2. **Assemble**: Use standard finite element assembly for the degrees of freedom in Ω_{int} and the techniques described in Chapter 3 to assemble the problem matrices A and B for the pole condition parameter s_0, compute the perturbed matrices ΔA and ΔB according to Equations (5.12) and (5.13) in the one-dimensional case and (5.14) and (5.15) in the two-dimensional case.

3. **Solve**: Use an out of the box sparse Arnoldi algorithm to compute the spectrum $\sigma(A, B)$.

4. **Detect**: For each eigenvalue $\omega \in \sigma(A, B)$, compute the left and right eigenvectors. Use them to compute the perturbation $\Delta\omega$ by the formula given in Lemma 5.3. If $\Delta\omega = \mathcal{O}(\Delta s_0)$, find out if it is reasonable to assume the eigenvalue to be converged by evaluating the formula for the convergence monitor at ω. That is, in the one-dimensional case solve (5.20) for κ and in the two-dimensional case evaluate (5.36).

5. **Decide**: If the eigenvalue passes both tests in the previous step, it is considered a physical solution of the problem and included in the output, if it fails one of the tests, it is considered a spurious solution and removed from the spectrum.

The entire process is depicted in the flow-chart clearly arranged in Figure 5.30, where the out of the box algorithms are marked green, the computational steps for the detection are marked in dark red and the decisions to be taken for each eigenvalue are marked in magenta.

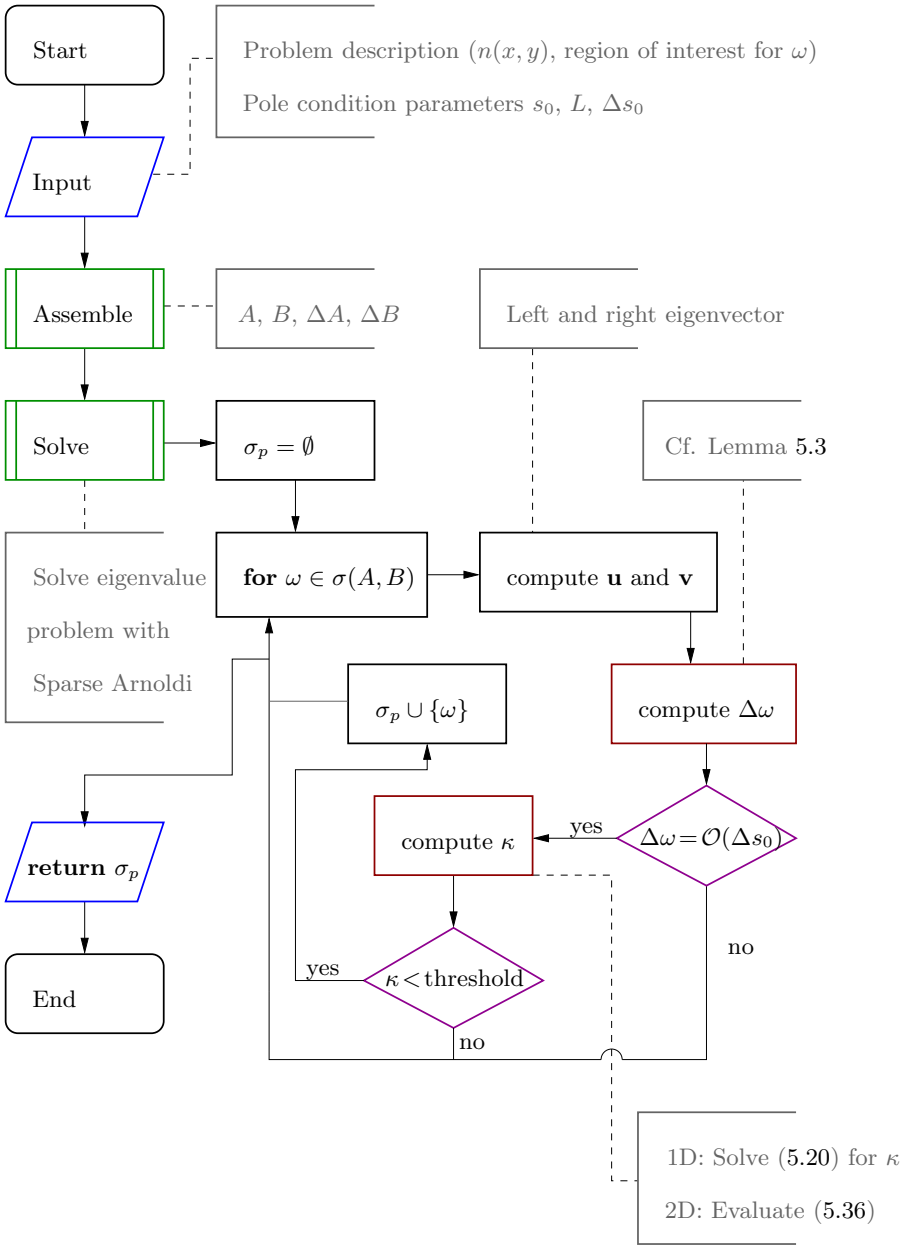

Figure 5.30: The algorithm for removing spurious solutions from the computed eigenvalue spectrum. Green color denotes out of the box algorithms, the central computations for the detection are marked in dark red and the decisions for determining spurious solutions are marked magenta.

Chapter 6

Summary and Outlook

In this thesis we have presented a robust algorithm for detecting spurious solutions within the computed eigenvalue spectrum of a resonance problem. The outline of the work performed here was as follows:

1. After an introduction, deriving the equations in question and briefly highlighting the physical and technical background of the problem, we derived the central tool for our considerations, the pole condition [Sch02] in Chapter 3. We followed the outline of Hohage, Nannen [NS11, Nan08, HN09], Schädle and Ruprecht [RSSZ08] to obtain a formulation that fits well into the finite element context. We gave an implementation in one and two space dimensions.

2. We reviewed the issue of spurious solutions and based on a one-dimensional example verified the hypothesis that there are spurious solutions that are caused by the implementation of transparent boundary conditions. We saw that these spurious solutions appear for both the perfectly matched layers and the pole condition. We concluded that the spurious solutions that are caused by the boundary condition are not an artifact of the particular method used but a systematic discretization error.

3. Based on these findings, we dedicated Chapter 5 to the detection of spurious solutions that are caused by the transparent boundary conditions. For this means, we first highlighted the perturbation theory for generalized eigenvalue problems.

4. We found, however, that the perturbation theory for generalized eigenvalue problems is too general to be a useful tool for the detection of spurious solutions. This is due to the fact, that this theory is a general purpose theory that can not take into account any special kind of perturbations. The perturbations we cause in order to detect spurious solutions are however very well defined, so a better detection means needed to be derived.

5. We then derived a formula for computing the exact effect a perturbation of the system matrices has on an eigenvalue of the generalized eigenvalue problem. Since our perturbations only affect the exterior parts of the system matrices, the formula can be easily evaluated in order to compute the perturbation of the eigenvalues.

6. In order to obtain statements concerning the validity of these perturbations, we complemented the formula for the perturbation with a convergence monitor for the resonances. Under some simplifications we were able to construct such a convergence monitor for the one- and the two-dimensional case. This convergence monitor together with the perturbation formula makes it possible for the first time to robustly detect spurious solutions in resonance spectra without a priori knowledge of the expected field distributions or spectral distributions of the eigenvalues.

We will now summarize the central results we obtained:

The Pole Condition - Hardy Space Infinite Elements

The pole condition detects outgoing solutions by the location of the poles of their Laplace transform in the complex plane. In the one-dimensional case the Helmholtz equation reads

$$\partial_{xx} u(x) n(x)^2 \omega^2 u(x) = 0 \quad \text{for } x \in \mathbb{R}$$

Dividing $\mathbb{R}$ into a bounded interior domain Ω_{int} and an unbounded exterior domain Ω_{ext} and applying the Laplace transform to the solution in Ω_{ext}, we found that the continuation of the Laplace transform in the exterior has poles and that the location of these poles is different for incoming and outgoing solutions. This enabled us to split the complex plane $\mathbb{C}$ into two sub-domains, $\mathbb{C}_{\text{in}}$ and $\mathbb{C}_{\text{out}}$ where the corresponding solutions are incoming or outgoing. Requiring of solutions that they are outward radiating then is equivalent to demanding that the continuation of the Laplace transform of the solution is analytic in $\mathbb{C}_{\text{in}}$

Using the Möbius transform $\mathcal{M}_{s_0}$, we could map the half-space $\mathbb{C}_{\text{in}}$ to the unit disc D and form a connection between their function spaces $H^-(\mathbb{C}_{\text{in}})$ and $H^+(D)$. The Möbius transform $\mathcal{M}_{s_0}$ depends on a parameter that we will use later on. An implementation of a series expansion in $H^+(D)$ could be given using the trigonometric monomials as ansatz functions. The implementation of the pole condition in the one-dimensional case was then reduced to the implementation of two bidiagonal $L \times L$ matrices $\mathcal{T}_L^{(\pm)}$.

In the two-dimensional case, a tensor product ansatz was chosen, using $\mathcal{T}_L^{(\pm)}$ in the radial direction away from the boundary $\partial\Omega$ and the traces of the finite elements in the interior alongside $\partial\Omega$. The implementation of the

radial parts then reduced to the known matrices $\mathcal{T}_L^{(\pm)}$ and an additional tridiagonal matrix $\mathcal{D}_L$.

Condition Numbers and Direct Perturbation of Eigenvalues

The discretization of Helmholtz resonance problems results in generalized eigenvalue problem $(A - \lambda B)\mathbf{u} = 0$ with large sparse complex matrices A and B. They can be rewritten as $(\beta A - \alpha B)\mathbf{u} = 0$. The solution $\mathbf{u}$ is called the (right) eigenvector, $\mathbf{v}$ is a left eigenvector, if $\mathbf{v}^H(A - \lambda B) = 0$. Given a left eigenvector $\mathbf{v}$ and a right eigenvector $\mathbf{u}$ for the eigenvalue $\lambda = \beta/\alpha$, the relative condition number $\kappa_{\text{rel}}(\lambda)$ is

$$\kappa_{\text{rel}}(\lambda) = \frac{\|\mathbf{u}\|_2\|\mathbf{v}\|_2}{\sqrt{|\alpha|^2 + |\beta|^2}}.$$

This number, however, is of limited use when detecting spurious solutions of a resonance problem. Due to the fact that it is a purely algebraic feature of the problem, it disregards all knowledge about the physics of the underlying problem and the nature of the perturbation in question. This suggests that there are better ways for detecting the spurious solutions within a given spectrum. Such a better way was found by computing the direct reaction of the eigenvalues to the well-defined perturbation caused by a change of the pole condition parameter.

This led to the following formula

$$\Delta\lambda \approx \frac{\mathbf{v}^H \Delta A\mathbf{u} - \lambda \mathbf{v}^H \Delta B\mathbf{u}}{\mathbf{v}^H B\mathbf{u}}.$$

The perturbations ΔA of A and ΔB of B are caused by perturbing the pole condition parameter s_0 with a perturbation Δs_0. Due to the explicit knowledge of Δs_0, the perturbations ΔA and ΔB can be directly computed. Since s_0 is only present in the matrix entries that are due to degrees of freedom in the exterior, the perturbations have zero entries for all interior degrees of freedom which means that they are indifferent to the discretization in the exterior and only take into account the sensitivity of the eigenvalues towards s_0. For detecting the spurious solutions we now only had to compute ΔA and ΔB and evaluate the formula for each λ. It was seen that the physical solutions reacted to a perturbation of s_0 with a perturbation that was $\mathcal{O}(\Delta s_0)$ while the spurious solutions reacted much stronger.

Convergence Monitor for the Pole Condition

It was however necessary to complement the detection method described above with a convergence monitor for it is possible for physical solutions to react strongly to perturbations of s_0 when they are not well-converged in

the exterior. For that end, we made the observation that the pole condition corresponds to a linear second order recurrence relation $\widetilde{a}_2(\omega^2)z_{l-2} + \widetilde{a}_1 z_{l-1} + \widetilde{a}_0 z_l = 0$. The stability conditions for such a relation depend on the roots of its characteristic polynomial and are well-known. We were able to apply them to the relation we obtained when applying the pole condition to the exterior domain of a one-dimensional problem. This enabled us to compute regions of the complex plane where we expect the discretization of the exterior domain to converge with a predefined convergence rate κ as

$$C(\kappa) = \left\{ \omega \in \mathbb{C} : \left| \omega n_{ext} - i s_0 \frac{1+\kappa^2}{1-\kappa^2} \right| \leq |i s_0| \sqrt{\left(\frac{1+\kappa^2}{1-\kappa^2} \right)^2 - 1} \right\}.$$

Outside these circles we cannot expect the solution to be converged and therefore not rely on the results of the perturbation described above. For an extension to two-dimensional problems, we derived under some simplifications a vector valued linear second order recurrence relation. Its rate of convergence could not be computed directly as in the one-dimensional case, however, it is possible to compute the regions of convergence numerically. This requires recasting the vector valued linear second order recurrence relation to a linear first order recurrence relation and computing the eigenvalues of the resulting coefficient matrix. An algorithm for obtaining the regions of convergence for such a problem was presented.

Outlook and Final Example

An interesting continuation of this work would be to investigate the possibility of including the conditions for the detection of spurious solutions directly into an iterative eigenvalue solver thus avoiding the computation of spurious solutions right at the step of solving the eigenvalue problem instead of filtering the results at a later stage. Furthermore, the combination of a convergence monitor and its extension to other equations for which implementations of the pole condition exist, would be a good starting point for thinking about an adaptive pole condition where the parameter s_0 and the number of degrees of freedom L are chosen automatically in a way such that the resulting eigenvalue spectrum contains as many of the eigenvalues of interest as possible.

After most of the examples in this thesis so far being purely academic or stemming from the area of optics, we will conclude by giving an example from the field of acoustics and model the resonances of the bell mentioned in the introductory poem. Figure 6.1 shows the computational mesh used for this example alongside a false color plot of the field distribution we would expect for a resonance. Our question is what frequencies make up for the special chime of such a bell.

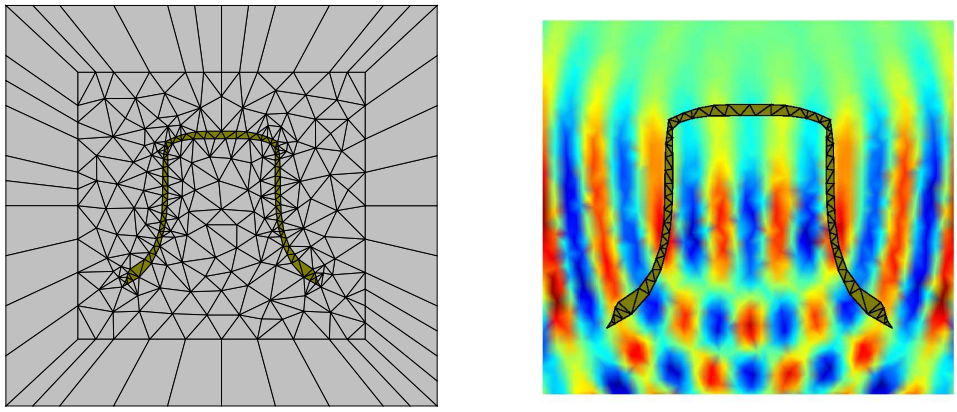

Figure 6.1: Left: Unstructured grid used to compute the resonances of a church bell. Right: Exemplary field distribution of a resonant state. The field distribution seems uneven due to linear interpolation in the plot-routines.

To answer this question we solve the resonance problem on the mesh depicted in Figure 6.1. The resulting resonance spectrum can be seen in the left-hand side image of Figure 6.2. Now we expect the occurrence of spurious solutions which pollute the spectrum of the bell. The right-hand side image of Figure 6.2 shows the full spectrum of the problem. As pole condition parameters we used $s_0 =$ and $L = 15$ and for the interior problem we used a finite element degree of $n = 3$. If we would add all the computed frequencies into the expected spectrum of the bell and mimic an analysis of the resulting sound with an oscilloscope, we would obtain the waveform shown in the left-hand side image of Figure 6.2. The sound we could expect from such a waveform void of regular patterns is just noise.

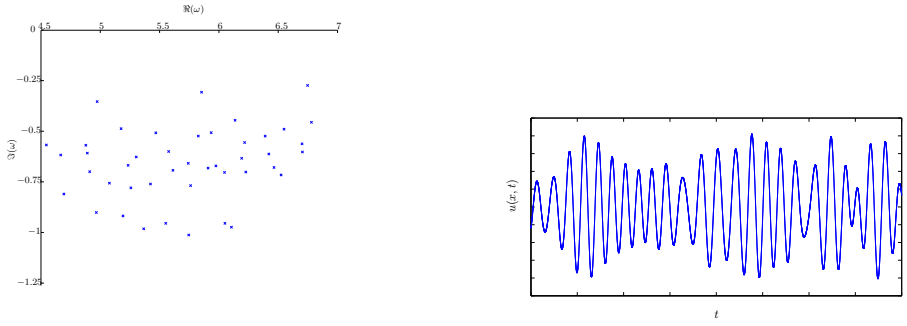

Figure 6.2: Left: Spectrum computed for the bell. Right: Oscilloscope plot of the spectrum resulting in an excitation of all computed modes simultaneously.

However, after applying the algorithms for the detection of the spurious

solutions that we presented in this thesis, we see that only the frequencies shown in Figure 6.3 qualify as physical solutions, which can again be verified by manual inspection. The resulting waveform has a clear recurring pattern and is shown in the left-hand side image of Figure 6.3. This corresponds to a chime with one fundamental frequency and two harmonics together making for the clear sound of the bell described in the introductory poem.

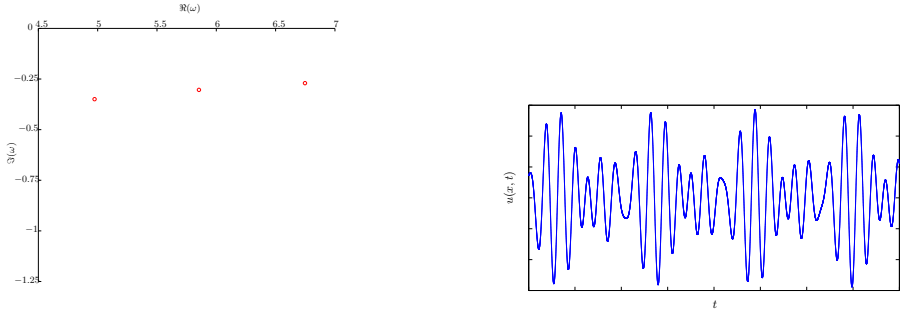

Figure 6.3: Left: Spectrum computed for the bell after removing spurious solutions. Right: Oscilloscope plot of the spectrum resulting in an excitation of the physical modes simultaneously.

Chapter 7

Zusammenfassung

Um partielle Differentialgleichungen auf unbeschränkten Gebieten numerisch zu lösen, wird das unbeschränkte Gebiet üblicherweise in einen beschränkten Innenraum und einen unbeschränkten Außenraum zerteilt. Die Gleichung wird dann nur auf dem beschränkten Innenraum gelöst und am Übergang zwischen Innenraum und Außenraum werden transparente Randbedingungen verwendet, die das Verhalten der Lösung im unbeschränkten Außenraum approximieren.

In der vorliegenden Arbeit wird die Helmholtz-Gleichung als Resonanzproblem auf unbeschränkten Gebieten gelöst. Dabei verursachen die transparenten Randbedingungen unphysikalische Lösungen, die das berechnete Frequenzspektrum verunreinigen. Diese Lösungen sind Artefakte, die durch die Diskretisierung mit transparenten Randbedingungen zurückzuführen sind. In der Praxis ist es oftmals schwierig, diese von den physikalischen Lösungen des Problems zu unterscheiden, wenn man kein a priori Wissen über das erwartete Eigenwertspektrum des untersuchten Objekts oder die Feldverteilung in seinem Inneren hat. Dabei gibt es zwei Klassen von unphysikalischen Lösungen: solche, die durch die Diskretisierung im Innenraum verursacht werden und für deren Vermeidung geeignete Strategien existieren und solche, die durch die transparenten Randbedingungen verursacht werden. Für die zweite Klasse von unphysikalischen Lösungen existiert bislang keine einheitliche Theorie und kein globaler Ansatz zu ihrer Vermeidung. In der vorliegenden Arbeit wurde ein Algorithmus entwickelt, der auf zuverlässige Art und Weise die zweite Art von unphysikalischen Lösungen im Frequenzspektrum erkennt und sie daraus entfernt.

Dieser Algorithmus verwendet als transparente Randbedingung die Polbedingung [Sch02], insbesondere ihre Implementierung als infinite Hardy-Raum Elemente [HN09, Nan08, NS11, RSSZ08]. Diese Methode hat den Vorteil, dass ein Parameter existiert, der in einem gewissen Rahmen frei gewählt werden kann. Da die zweite Art von unphysikalischen Lösungen von der Randbedingung verursacht werden, hängen sie auch stärker von der

Variation dieses Parameters ab als die physikalischen Lösungen des Problems. In der Arbeit wurde diese Abhängigkeit zunächst auf der algebraischen Seite mit Hilfe von Konditionszahlen für allgemeine Eigenwertprobleme untersucht. Die Konditionszahl erwies sich aber als zu allgemeines Werkzeug, um das Problem zuverlässig zu lösen, weshalb eine geschlossene Formel hergeleitet wurde, die aus der Variation des Polbedingungsparamters direkt die Reaktion der Eigenwerte berechnet. Diese Formel ist unabhängig von der Diskretisierung im Innenraum und deshalb besser geeignet, um die Problemstellung zu behandeln.

Diese Methode kann aber nur funktionieren, wenn die Lösung im Außenraum konvergiert ist, weshalb die Methode um einen neu entwickelten Konvergenz-Monitor ergänzt wurde, der es ermöglicht, zu jeder Resonanzfrequenz die Konvergenzrate der Polbedingung zu bestimmen. Die Kombination beider Methoden ermöglicht so eine zuverlässige und robuste Identifizierung der unphysikalischen Lösungen in den berechneten Spektralbereichen. Der Algorithmus, der beide Methoden vereint, wurde in der Arbeit auf eine Reihe von Beispielen aus der Nano-Optik und der Akustik angewendet.

Literaturverzeichnis

[AAB+07] X. Antoine, A. Arnold, C. Besse, M. Ehrhardt, and A. Schädle. A Review of Transparent and Artificial Boundary Conditions Techniques for Linear and Nonlinear Schrödinger Equations. Technical Report 07-34, ZIB, Takustr.7, 14195 Berlin, 2007.

[BBG00] D. Boffi, F. Brezzi, and L. Gastaldi. On the Problem of Spurious Eigenvalues in the Approximation of Linear Elliptic Problems in Mixed Form. *Mathematics of computation*, 69(229):121–140, 2000.

[Ber94] J. P. Berenger. A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *Journal of computational physics*, 114(2):185–200, 1994.

[Ber09] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

[BFGP99] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia. Computational Models of Electromagnetic Resonators: Analysis of Edge Element Approximation. *SIAM Journal on Numerical Analysis*, 36(4):1264–1290, 1999.

[BFLP99] D. Boffi, P. Fernandes, Gastaldi L, and I. Perugia. Computational Models of Electromagnetic Resonators: Analysis of Edge Element Approximation. *SIAM journal on numerical analysis*, pages 1264–1290, 1999.

[Bof01] D. Boffi. A Note on the Discrete Compactness Property and the de Rham Complex. *Applied Mathematics. Letters*, 14:33–38, 2001.

[Bos88] A. Bossavit. Whitney Forms: A Class of Finite Elements for Three-Dimensional Computations in Electromagnetism. *Physical Science, Measurement and Instrumentation, Management and Education-Reviews, IEE Proceedings A*, 135(8):493–500, 1988.

[Bos90] A. Bossavit. Solving Maxwell equations in a closed cavity, and the question of 'spurious modes'. *IEEE Transactions on Magnetics*, 26(2):702–705, 1990.

[BPSZ11] S. Burger, J. Pomplun, F. Schmidt, and L. Zschiedrich. Finite-Element Method Simulations of High-Q Nanocavities with 1D Photonic Bandgap. *Arxiv preprint arXiv:1102.4510*, 2011.

[Bra10] D. Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer Verlag, 2010.

[BSZ10] S. Burger, F. Schmidt, and L. Zschiedrich. Numerical Investigation of Photonic Crystal Microcavities in Silicon-on-Insulator Waveguides. *Arxiv preprint arXiv:1003.2314*, 2010.

[BZS10] S. Burger, L. Zschiedrich, and F. Schmidt. FEM Simulation of Plasmon Laser Resonances. In *AIP Conference Proceedings*, volume 1281, page 1613, 2010.

[CD72] D. Corr and J. Davies. Computer Analysis of the Fundamental and Higher Order Modes in Single and Coupled Microstrip. *Microwave Theory and Techniques, IEEE Transactions on*, 20(10):669–678, 1972.

[Cen91] Z. Cendes. Vector Finite Elements for Electromagnetic Field Computation. *Magnetics, IEEE Transactions on*, 27(5):3958–3966, 1991.

[CFR95] S. Caorsi, P. Fernandes, and M. Raffetto. Edge Elements and the Inclusion Condition [EM Eigenproblems]. *Microwave and Guided Wave Letters, IEEE*, 5(7):222–223, 1995.

[CFR96] S. Caorsi, P. Fernandes, and M. Raffetto. Towards a Good Characterization of Spectrally Correct Finite Element Methods in Electromagnetics. *COMPEL: Int J for Computation and Maths. in Electrical and Electronic Eng.*, 15(4):21–35, 1996.

[CFR97] S. Caorsi, P. Fernandes, and M. Raffetto. Do Covariant Projection Elements Really Satisfy the Inclusion Condition? *Microwave Theory and Techniques, IEEE Transactions on*, 45(9):1643–1644, 1997.

[CFR01a] S. Caorsi, P. Fernandes, and M. Raffetto. On the Convergence of Galerkin Finite Element Approximations of Electromagnetic Eigenproblems. *SIAM Journal on Numerical Analysis*, pages 580–607, 2001.

[CFR01b] S. Caorsi, P. Fernandes, and M. Raffetto. Spurious-Free Approximations of Electromagnetic Eigenproblems by Means of Nedelec-Type Elements. *ESAIM: Mathematical Modelling and Numerical Analysis*, 35(02):331–354, 2001.

[CHC99] J. T. Chen, C. X. Huang, and K. H. Chen. Determination of Spurious Eigenvalues and Multiplicities of True Eigenvalues Using the Real-Part Dual BEM. *Computational Mechanics*, 24:41–51, 1999.

[Che95] W. Chew. *Waves and Fields in Inhomogenous Media*. IEEE press, 1995.

[CK98] D. Colton and R. Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93. Springer Verlag, 1998.

[CRD03] W. Cecot, W. Rachowicz, and L. Demkowicz. An hp-adaptive Finite Element Method for Electromagnetics. Part 3: A three-dimensional Infinite Element for Maxwell's Equations. *International journal for numerical methods in engineering*, 57(7):899–921, 2003.

[CS70] Z. Csendes and P. Silvester. Numerical Solution of Dielectric Loaded Waveguides: I-Finite-Element Analysis. *Microwave Theory and Techniques, IEEE Transactions on*, 18(12):1124–1131, 1970.

[CSJ88] C. Crowley, P. Silvester, and H. Hurwitz Jr. Covariant Projection Elements for 3D Vector Field Problems. *Magnetics, IEEE Transactions on*, 24(1):397–400, 1988.

[DB02] P. Deuflhard and F. Bornemann. *Numerische Mathematik. II: Gewöhnliche Differentialgleichungen*. deGruyter Lehrbuch, 2 edition, 2002.

[DFP82] J. Davies, F. Fernandez, and G. Philippou. Finite Element Analysis of all Modes in Cavities with Circular Symmetry. *Microwave Theory and Techniques, IEEE Transactions on*, 30(11):1975–1980, 1982.

[DLW94] B. Dillon, P. Liu, and J. Webb. Spurious Modes in Quadrilateral and Triangular Edge Elements. *Compel-International Journal for Computation and Math in Electrical and Electronic Eng*, 13:311–316, 1994.

[Dur70] P. Duren. *Theory of H^p spaces*. Academic Press, 1970.

[FA76] A. Farrar and A. Adams. Computation of Propagation Constants for the Fundamental and Higher Order Modes in Microstrip (Short Papers). *Microwave Theory and Techniques, IEEE Transactions on*, 24(7):456–460, 1976.

[Flö11] D. Flöß. Numerische Analyse optischer Flüstergalerieresonatoren zur Biodetektion. Diploma thesis, 2011.

[FR02a] P. Fernandes and M. Raffetto. Characterization of Spurious-Free Finite Element Methods in Electromagnetics. *COMPEL: Int J for Computation and Maths. in Electrical and Electronic Eng.*, 21(1):147–164, 2002.

[FR02b] P. Fernandes and M. Raffetto. Counterexamples to the Currently Accepted Explanation for Spurious Modes and Necessary and Sufficient Conditions to Avoid Them. *IEEE Transactions on Magnetics*, 38(2):653 – 656, 2002.

[FS92] P. Fernandes and G. Sabbi. On the Spurious Modes in Electromagnetic Eigenproblems. In *Proceedings of the International Conference on Electromagnetic Field Problems and Applications, Hangzhou, China, International Academic Publishers, Beijing, China*, pages 89–92, 1992.

[GL96] G. H. Golub and C. F. Van Loan. *Matrix computations*. Johns Hopkins studies in the mathematical sciences. Johns Hopkins University Press, 1996.

[GS77] A. Ganguly and B. Spielman. Dispersion Characteristics for Arbitrarily Configured Transmission Media (Short Papers). *Microwave Theory and Techniques, IEEE Transactions on*, 25(12):1138–1141, 1977.

[GSH+10] T. Grossmann, S. Schleede, M. Hauser, M. B. Christiansen, C. Vannahme, C. Eschenbaum, S. Klinkhammer, T. Beck, J. Fuchs, G. U. Nienhaus, U. Lemmer, A. Kristensen, T. Mappes, and H. Kalt. Low-Threshold Conical Microcavity Dye Lasers. *Applied Physics Letters*, 97(6):063304, 2010.

[Hag99] T. Hagstrom. Radiation Boundary Conditions for the Numerical Simulation of Waves. *Acta numerica*, 8(1):47–106, 1999.

[Hag03] T. Hagstrom. New Results on Absorbing Layers and Radiation Boundary Conditions. *Topics in computational wave propagation*, pages 1–42, 2003.

[Ham07] M. Hammer. Hybrid Analytical/Numerical Coupled-Mode Modeling of Guided-Wave Devices. *Journal of Lightwave Technology*, 25(9):2287–2298, 2007.

[Har15] G. Hardy. The Mean Value of the Modulus of an Analytic Function. *Proceedings of the London Mathematical Society*, 2(1):269, 1915.

[HH98] D. J. Higham and N. J. Higham. Structured Backward Error and Condition of Generalized Eigenvalue Problems. *j-SIMAX*, 20(2):493–512, 1998.

[HN09] T. Hohage. and L. Nannen. Hardy Space Infinite Elements for Scattering and Resonance Problems. *SIAM Journal on Numerical Analysis*, 47(2):972–996, 2009.

[Hof62] K. Hoffman. *Banach Spaces of Analytic Functions*, volume 172. Prentice-Hall Englewood Cliffs, 1962.

[Hoy65] H. Hoyt. Numerical Studies of the Shapes of Drift Tubes and Linac Cavities. *Nuclear Science, IEEE Transactions on*, 12(3):153–155, 1965.

[HSR66] H. Hoyt, D. Simmonds, and W. Rich. Computer Designed 805 MHz Proton Linac Cavities. *Review of Scientific Instruments*, 37(6):755–762, 1966.

[HSZ02] T. Hohage, F. Schmidt, and L. Zschiedrich. A New Method for the Solution of Scattering Problems. Technical Report 02-01, Zuse Institute Berlin, 2002.

[HSZ03a] T. Hohage, F. Schmidt, and L. Zschiedrich. Solving Time-Harmonic Scattering Problems Based on the Pole Condition I: Theory. *SIAM Journal on Mathematical Analysis*, 35(1):183–210, 2003.

[HSZ03b] T. Hohage, F. Schmidt, and L. Zschiedrich. Solving Time-Harmonic Scattering Problems Based on the Pole Condition II: Convergence of the PML Method. *SIAM Journal on Mathematical Analysis*, 35:547, 2003.

[HWFK83] M. Hara, T. Wada, T. Fukasawa, and F. Kikuchi. A Three Dimensional Analysis of RF Electromagnetic Fields by the Finite Element Method. *Magnetics, IEEE Transactions on*, 19(6):2417–2420, 1983.

[Ihl98] F. Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132. Springer Verlag, 1998.

[Ket06] B. Kettner. Ein Algorithmus zur prismatoidalen Diskretisierung von unbeschränkten Außenräumen in 2D und 3D unter Einhaltung von Nebenbedingungen. Diploma thesis, 2006.

[KS08] B. Kettner and F. Schmidt. Meshing of heterogeneous unbounded domains. pages 26–30. Springer-Verlag, October 2008.

[Lei86] R. Leis. *Initial Boundary Value Problems in Mathematical Physics*. Teubner Stuttgart, 1986.

[LL87] L. Landau and E. Lifshitz. *Fluid Mechanics*. Pergamon, 1987.

[LP89] P. Lax and R. Phillips. *Scattering Theory*, volume 26. Academic Press, 1989.

[LSC91] J. F. Lee, D. Sun, and Z. Cendes. Tangential Vector Finite Elements for Electromagnetic Field Computation. *Magnetics, IEEE Transactions on*, 27(5):4032–4035, 1991.

[MD01] P. Monk and L. Demkowicz. Discrete Compactness and the Approximation of Maxwell's Equations in $\mathbb{R}^3$. *Mathematics of computation*, 70(234):507–524, 2001.

[Mon03] P. Monk. *Finite Element Methods for Maxwells Equations*. New York: Oxford University Press, 2003.

[MP98] O.J.F. Martin and N.B. Piller. Electromagnetic Scattering in Polarizable Backgrounds. *Physical Review E*, 58(3):3909, 1998.

[MS84] G. Menzala and T. Schonbek. Scattering Frequencies for the Wave Equation with a Potential Term. *Journal of functional analysis*, 55(3):297–322, 1984.

[Nan08] L. Nannen. *Hardy-Raum Methoden zur numerischen Lösung von Streu- und Resonanzproblemen auf unbeschränkten Gebieten*. Phd thesis, 2008.

[Néd80] J. Nédélec. Mixed Finite Elements in $\mathbb{R}^3$. *Numerische Mathematik*, 35(3):315–341, 1980.

[Néd01] J. Nédélec. *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, volume 144. Springer Verlag, 2001.

[NS11] L. Nannen and A. Schädle. Hardy Space Infinite Elements for Helmholtz-Type Problems with Unbounded Inhomogeneities. *Wave Motion*, 48(2):116–129, 2011.

[PBZS07] J. Pomplun, S. Burger, L. Zschiedrich, and F. Schmidt. Adaptive Finite Element Method for Simulation of Optical Nano Structures. *phys. stat. sol. (b)*, 244:3419 – 3434, 2007.

[PCS88] A. Pinchuk, C. Crowley, and P. Silvester. Spurious Solutions to Vector Diffusion and Wave Field Problems. *Magnetics, IEEE Transactions on*, 24(1):158–161, 1988.

[PL91] K. D. Paulsen and D. R. Lynch. Elimination of Vector Parasites in Finite Element Maxwell Solutions. *IEEE Transactions on Microwave Theory and Techniques*, 39(3):395–404, 1991.

[PM01] M. Paulus and O. Martin. Greens Tensor Technique for Scattering in Two-Dimensional Stratified Media. *Physical Review E*, 63(6):066615, 2001.

[Rec05] M. Rechberger. Numerical Methods for the Simulation of Acoustic Resonances. Diploma thesis, 2005.

[RSSZ08] D. Ruprecht, A. Schädle, F. Schmidt, and L. Zschiedrich. Transparent Boundary Conditions For Time Dependent Problems. *SIAM J. Sci. Comput.*, 30(5):2358–2385, 2008.

[Saa80] Y. Saad. Variations on Arnoldi's Method for Computing Eigenelements of Large Unsymmetric Matrices. *Linear Algebra and its Applications*, 34(0):269–295, 1980.

[SB84] E. Schweig and W. Bridges. Computer Analysis of Dielectric Waveguides: A Finite-Difference Method. *Microwave Theory and Techniques, IEEE Transactions on*, 32(5):531–541, 1984.

[Sch98] F. Schmidt. An Alternative Derivation of the Exact DtN-Map on a Circle. Technical report, Zuse Institute Berlin, 1998.

[Sch02] F. Schmidt. *Solution of Interior-Exterior Helmholtz-Type Problems Based on the Pole Condition Concept: Theory and Algorithms*. Habilitation thesis, Free University Berlin, 2002.

[Seb04] G.A.F. Seber. *Multivariate observations*. Wiley series in probability and statistics. Wiley-Interscience, 2004.

[Set92] V. Seth. *Three Chinese Poets: Translations of Poems by Wang Wei, Li Bai, and Du Fu*. HarperPerennial, 1992.

[SMYC95] D. Sun, J. Manges, X. Yuan, and Z. Cendes. Spurious Modes in Finite-Element Methods. *IEEE Antennas and Propagation Magazine*, 37(5):12–24, 1995.

[Som49] A. Sommerfeld. *Partial Differential Equations in Physics*, volume 6. Academic Press, 1949.

[Spä85] H. Späth. *Cluster Dissection and Analysis: Theory, FORTRAN programs, Examples. Translated by J. Goldschmidt.* New York: Halsted Press, 1985.

[SS90] G. W. Stewart and J. Sun. *Matrix Perturbation Theory.* Academic Press, 1990.

[SSA92] M. Swaminathan, T. Sarkar, and A. Adams. Computation of TM and TE Modes in Waveguides Based on a Surface Integral Formulation. *Microwave Theory and Techniques, IEEE Transactions on*, 40(2):285–297, 1992.

[Ste72] G. W. Stewart. On the Sensitivity of the Eigenvalue Problem $Ax = \lambda Bx$. *SIAM Journal on Numerical Analysis*, 9(4):669–686, 1972.

[Ste01] G. W. Stewart. *Matrix Algorithms Volume II: Eigensystems.* Matrix Algorithms. Society for Industrial and Applied Mathematics, 2001.

[Su85] C. Su. Origin of Spurious Modes in the Analysis of Optical Fibre Using the Finite-Element or Finite-Difference Technique. *Electronics Letters*, 21(19):858–860, 1985.

[TH00] T. Tischler and W. Heinrich. The Perfectly Matched Layer as Lateral Boundary in Finite-Difference Transmission-Line Analysis. *Microwave Theory and Techniques, IEEE Transactions on*, 48(12):2249–2253, 2000.

[Tis03] T. Tischler. *Die Perfectly-Matched-Layer-Randbedingung in der Finite-Differenzen-Methode im Frequenzbereich: Implementierung und Einsatzbereiche.* PhD thesis, 2003.

[Tsy98] S. Tsynkov. Numerical Solution of Problems on Unbounded Domains. A Review. *Applied Numerical Mathematics*, 27(4):465–532, 1998.

[TZ00] S. Tang and M. Zworski. Resonance Expansions of Scattered Waves. *Communications on Pure and Applied Mathematics*, 53(10):1305–1334, 2000.

[WC88] S. Wong and Z. Cendes. Combined Finite Element-Modal Solution of Three-Dimensional Eddy Current Problems. *Magnetics, IEEE Transactions on*, 24(6):2685–2687, 1988.

[WI91] J. Wang and N. Ida. Eigenvalue Analysis in Electromagnetic Cavities Using Divergence Free Finite Elements. *Magnetics, IEEE Transactions on*, 27(5):3978–3981, 1991.

[Wil88] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Monographs on numerical analysis. Clarendon Press, 1988.

[WMS88] M. Wei, G. Majda, and W. Strauss. Numerical Computation of the Scattering Frequencies for Acoustic Wave Equations. *Journal of Computational Physics*, 75(2):345–358, 1988.

[XZX89] C. Xu, L. Zhou, and A. Xu. An Improved Theory of Microwave Open Resonators. *International Journal of Infrared and Millimeter Waves*, 10(1):55–62, 1989.

[ZBKS06] L. Zschiedrich, S. Burger, B. Kettner, and F. Schmidt. Advanced Finite Element Method for Nano-Resonators. In M. Osinski, F. Henneberger, and Y. Arakawa, editors, *Physics and Simulation of Optoelectronic Devices XIV*, volume 6115, pages 164 – 174, 2006.

[Zsc09] L. Zschiedrich. *Transparent Boundary Conditions for Maxwell's Equations: Numerical concepts beyond the PML method*. PhD thesis, 2009.

[Zwo99] M. Zworski. Resonances in Physics and Geometry. *Notices of the AMS*, 46(3):319–328, 1999.