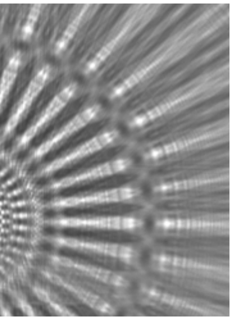
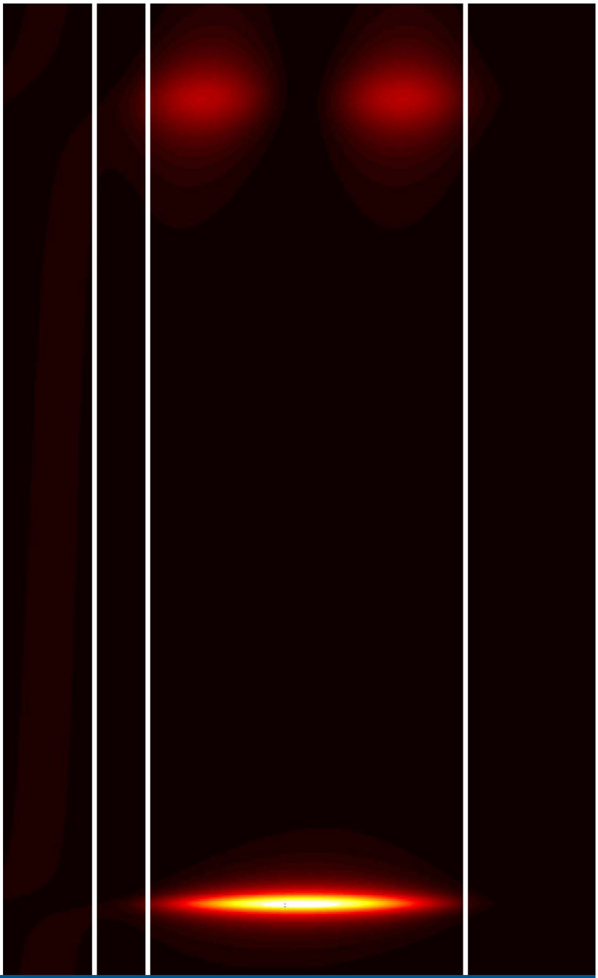
Göttingen Series in
X-ray Physics



Felix Schenk

Optimization of resonances for multilayer x-ray resonators



Universitätsverlag Göttingen

Felix Schenk

Optimization of resonances for multilayer x-ray resonators

This work is licensed under the
Creative Commons License 3.0 “by-nd”,
allowing you to download, distribute and print the
document in a few copies for private or educational
use, given that the document stays unchanged
and the creator is mentioned.
You are not allowed to sell copies of the free version.

Published in 2011 by Universitätsverlag Göttingen
as Volume 3 in the series „Göttingen series in x-ray physics“

Felix Schenk

Optimization of resonances for multilayer x-ray resonators

Göttingen series in x-ray physics
Volume 3

Universitätsverlag Göttingen
2011

Bibliographische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Supported by the Deutsche Forschungsgemeinschaft
via the Collaborative Research Center SFB 755

Address of the Author
Dr. Felix Schenk
Email: Felix.Schenk@gmx.de

Dissertation zur Erlangung des
mathematisch-naturwissenschaftlichen Doktorgrades
„Doctor rerum naturalium“
der Georg-August-Universität Göttingen
vorgelegt von Felix Schenk aus Lüneburg
Göttingen 2010

Referent: Prof. Dr. Thorsten Hohage
Koreferent: Prof. Dr. Tim Salditt
Tag der mündlichen Prüfung: 07.12.2010

This work is protected by German Intellectual Property Right Law.
It is also available as an Open Access version through the publisher's homepage and the Online Catalogue of the State and University Library of Goettingen (http://www.sub.uni-goettingen.de). Users of the free online version are invited to read, download and distribute it. Users may also print a small number for educational or private use. However they may not sell print versions of the online book.

Layout: Felix Schenk
Cover: Jutta Pabst
Cover image: Felix Schenk

© 2011 Universitätsverlag Göttingen
http://univerlag.uni-goettingen.de
ISBN: 978-3-86395-008-8
ISSN: 2191-9860

Preface of the series editor

The Göttingen series in X-ray physics is intended as a collection of research monographs in x-ray science, carried out at the Institute for X-ray Physics at the Georg-August-Universität in Göttingen, and in the framework of its related research networks and collaborations.
It covers topics ranging from X-ray microscopy, nano-focusing, wave propagation, image reconstruction, tomography, short X-ray pulses to applications of nanoscale x-ray imaging and biomolecular structure analysis.
In most but not all cases, the contributions are based on Ph.D. dissertations. The individual monographs should be enhanced by putting them in the context of related work, often based on a common long term research strategy, and funded by the same research networks. We hope that the series will also help to enhance the visibility of the research carried out here and help others in the field to advance similar projects.

Prof. Dr. Tim Salditt, Editor
Göttingen February 2011

Contents

1 **Introduction** **1**
  1.1 Motivation . . . 1
  1.2 Overview and Outline . . . 1
  1.3 Related work . . . 4

2 **Derivation of the mathematical problem** **7**
  2.1 Scattering problem . . . 7
  2.2 Optimization problem . . . 9
    2.2.1 Example of a multilayer x-ray resonator . . . 11
    2.2.2 First formulation of the optimization problem . . . 12
  2.3 Reformulation of the differential equation . . . 12
  2.4 Resonances . . . 15
  2.5 Optimization problem using resonances . . . 19

3 **Derivative of the total field and kinematic approximation** **23**
  3.1 Introduction to the kinematic approximation . . . 23
  3.2 General formalism . . . 24
    3.2.1 A Lippmann-Schwinger equation for steplike profiles . . . 24
    3.2.2 Fréchet derivatives and error estimates . . . 30
    3.2.3 Padé approximation . . . 32
  3.3 Application to step profiles/Kinematic approximation . . . 34
    3.3.1 Approximation formulas . . . 34
    3.3.2 Example . . . 37
  3.4 Conclusion . . . 40

4 **Analysis of the optimization problem** **41**
  4.1 Derivatives of eigenvalues . . . 41
    4.1.1 Introduction to derivatives of eigenvalues . . . 41
    4.1.2 Basic theory . . . 42
    4.1.3 Derivatives of the eigenvalues . . . 48
  4.2 Derivatives for generalized eigenvalue problem . . . 51
    4.2.1 Derivatives of the eigenvalues . . . 52
    4.2.2 Derivatives of the eigenvectors . . . 53
  4.3 Objective function and discretization . . . 55
    4.3.1 Approximation to the L^2-norm and asymptotic expansion of scattering solutions in the vicinity of resonances . 55
    4.3.2 Discretization . . . 58
    4.3.3 Example . . . 59
    4.3.4 Discrete objective function . . . 60
  4.4 Derivative of the discrete objective function . . . 62
    4.4.1 Existing work . . . 62
    4.4.2 Application of results from Section 4.2 . . . 65
  4.5 Derivative with respect to layer changes . . . 67

5 Numerical results **71**
5.1 Introductory examples 71
5.2 Matrix derivatives 75
5.2.1 Piecewise constant approximation 76
5.2.2 Approximation by splines of higher order 77
5.3 Optimization results 79
5.3.1 Optimization algorithms 79
5.3.2 Optimization of refractive indices for fixed layer thicknesses 80
5.3.3 Optimization of layer thicknesses for fixed refractive indices 85
5.3.4 Refractive indices and thicknesses simultaneously 90
5.3.5 Optimization with splines 94
5.3.6 Advanced aims and objective functions 97
5.3.7 Conclusions 104

6 Summary and Outlook **107**

Appendix **109**

Bibliography **111**

Nomenclature **117**

1 Introduction

1.1 Motivation

In the development of optical devices (for example in optical near-field microscopy) smaller and smaller wavelengths are necessary. Usually x-rays with spot sizes in the micrometer range are achieved by (coherent or incoherent) focussing, but the limits of these techniques have been reached. Since x-rays with smaller spots could open up a new range of applications in scattering, microscopy (for example high resolution tomography) and spectroscopy, especially in non-synchrotron experiments, one is interested in new approaches.

One approach is the use of multilayer systems for one-dimensional beam concentration. Such systems consist of several layers which can be varied in material, thickness and shape and support certain resonant states. These resonant states can be excited by x-ray beams under special grazing angles of incidence, corresponding to resonant frequencies of the system, leading to a very high field enhancement inside the system compared to the incident field (see *Pfeiffer et al.* [PSH$^+$00]). Field enhancement in x-ray waveguides can also be used to strengthen the signal of weakly-scattering biomolecular films (see *Pfeiffer, Mennicke and Salditt* [PMS02] and *Salditt et al.* [SPP$^+$03]) as well as for nuclear resonant x-ray scattering (see *Röhlsberger et al.* [RSKL05] and [RKS$^+$04]). The fabrication of the multilayer systems is typically done by thin film deposition techniques such as e-beam evaporation and/or sputtering. In real systems interface roughness is always a concern. The question to which extend the roughness parameters influence the optical properties is one motivation for the mathematical approach developed here.

It is the aim of this thesis to improve the existing systems, or more precisely the achievable field enhancement, using tools from optimization theory and the theory of resonances. We want to provide an optimization algorithm involving the field enhancement as objective function. Mathematically this means we have to optimize the field enhancement as a function of the refractive index. This leads to interesting optimization problems with objective functions involving resonances and resonance functions.

1.2 Overview and Outline

Resonances in open systems can be described by eigenvalue problems with a radiation condition at infinity and arise in various fields including acoustics, classical mechanics, quantum mechanics, and x-ray physics. In this thesis we focus on the optimization of resonances for multilayer x-ray resonators.

Let us make things a little more concrete. The propagation of polarized x-rays in layered media is described by the reduced wave equation

$$u'' + k_0^2 n^2 u = \nu u, \tag{1.1}$$

where the function n describes the refractive index and $\nu = k_0^2 \cos^2 \alpha$ parameterizes the angle of incidence α. We want to find a function n for which the field enhancement in the multilayer structure for a resonant angle of incidence is maximized subject to side constraints on n. The side constraints guarantee physical admissibility of the refractive profiles. In the objective function of our optimization problem we use an approximation of the solution u arising from an asymptotic expansion in the vicinity of resonances. This avoids taking a maximum over ν in the objective function and leads to an objective function involving complex resonances and corresponding resonant functions.

The problem is discretized via finite elements combined with Hardy space infinite elements in order to model radiation conditions in a way which preserves the linear eigenvalue structure of the problem (see *Hohage and Nannen* [HN09]). Most of the existing work on the optimization of eigenvalues for a matrix depending on parameters is restricted to the case of symmetric/Hermitian matrices or linear dependence on the parameters[1]. These assumptions are not satisfied for our problem. To compute the derivative of the objective function anyways, we derive analytic expressions for the derivatives of resonances and resonance functions with respect to n using perturbation theory of linear operators. We finish with numerical computation leading to improved multilayer x-ray resonators for several situations.

The outline of this thesis is as follows. In Chapter 2 we derive a complete formulation of our optimization problem and give the basic definitions needed for resonance problems. To this end, we formulate a scattering problem on all of $\mathbb{R}$ (differential equation and radiation conditions characterizing the behavior at ±infinity) describing the situation for a fixed multilayer system and all permitted angles of incidence. We also formulate it in weak form and provide an existence and uniqueness result. The scattering problem already leads to a first formulation of our optimization problem over a set of admissible n and ν. Since the optimization problem has a very complicated form, due to the maximum over all permitted ν, we use the concept of resonances to make it more accessible. Resonances are defined as eigenvalues to eigenfunctions which fulfill a radiation condition. Or to put it in another way, they are the singularities of the complex analytic continuation of the solution operator to the scattering problem with respect to ν. Resonances enable us to compute the resonant frequencies more directly as the real parts of the complex resonances approximate them (see Section 4.3.4). At the end of Chapter 2 we analyze the geometric multiplicities of the resonances. For the sake of differentiability we use in the objective function to our optimization problem the L^2-norm of the total field inside the system instead of the maximum norm. This replacement is very reasonable as the L^2-norm inside the system is proportional to the energy inside of it. Our numerical experiments suggest that the L^∞- and L^2-norm are almost proportional for the field at a resonant frequency.

As a first step to find the derivative of our objective function we analyze throughout Chapter 3 the dependence of the solution to the scattering problem on the refractive index n. Fréchet derivatives of a suitable solution operator are computed with the help of methods from integral equations. We provide a result which is of interest in its own, the mathematical derivation of the kinematic approximation. The kinematic approximation is an approximation to

---

[1] For details, we refer to the introduction to Section 4.1.

the reflectivity[2] in dependence of the angle of incidence α, which is widely used especially in the x-ray community. First, we rewrite the scattering problem as an integral equation using a Green's function (fundamental solution) for some admissible initial profile and analyze existence and uniqueness of solutions to the integral equation. The analyticity of the solution operator with respect to n at the initial profile is shown for α not equal to the critical angle, and formulas for the derivatives are given. From this we deduce approximation formulas not only on the reflectivity (in particular the kinematic approximation using the first derivative), but also on the complete solution to the scattering problem for small perturbations of the initial situation. We also get error estimates and higher order approximations for both using higher derivatives. In this sense we do not only justify the kinematic approximation, but also show how to improve it. Whenever a Green's function is known one can obtain approximation formulas more suited to the particular application and hope for even better approximation results. Close to the critical angle of incidence the approximation of the reflectivity is still not too good even if we use higher order Taylor expansions. This is due to the fact that the solution operator is not differentiable at the critical angle. Improvements in this region can be achieved by Padé approximations. We explain how they can be computed from the derivatives and show their success and the other approximation formulas for the reflectivity and the total field in a numerical example.

In Chapter 4 we analyze the differentiability of simple, isolated eigenvalues, and corresponding eigenvectors, of a general operator-valued function with respect to n. We apply the results to a generalized eigenvalue problem as it arises by formulating the scattering problem in operator form using the Hardy space formulation proposed in *Hohage and Nannen* [HN09]. For the eigenvalues we prove continuous differentiability with respect to n under the assumption that the underlying operator has this property. We derive expressions for their derivatives and for those of the corresponding eigenvectors. As the latter are non-unique, we show continuous differentiability assuming a special scaling for them. We derive a handy formula for an appropriate objective function, and its derivative can be deduced from our formulas for the derivatives of eigenvalues and eigenvectors.

In order to arrive at a fully discrete problem, n is assumed to be encoded in a vector $\mathbf{n}$, and the scattering problem is discretized by finite elements combined with Hardy space infinite elements. A discretization of n in the assumed form, where the resulting operator depends continuously differentiably on $\mathbf{n}$, can be obtained by a piecewise constant approximation or a spline of higher order. We explain why existing results on the optimization of eigenvalues of a parameter-dependent matrix cannot be applied to the completely discrete version of our objective function. Our results on the derivative of the objective function can still be applied as the discrete version is only a special case of the situation analyzed before. The required derivatives of the discretized operator are computed. At the end of Chapter 4 this is done with respect to the

---

[2] The reflectivity is defined as the squared absolute value of the reflection coefficient. For a definition of the reflection coefficient, we refer to Chapter 2, equation (2.6a).

positions of layer change for a piecewise constant refractive index and for other discretizations of n in Chapter 5.

Finally, in Chapter 5 we present numerical results showing what is achieved by our method and leading to improved multilayer systems. We discuss different discretizations of the refractive index and explain how we handle absorption effects. The absorption is not modelled as an independent variable as this always leads to the lowest admissible absorption. We start with the optimization of piecewise constant refractive profiles, optimizing refractive indices in the layers, layer thicknesses and both simultaneously. Especially the optimization of the layer thicknesses turns out to be very promising and leads to improvements up to 200% in the field enhancement compared to standard systems. More accurate approximations of the refractive index using splines of higher order suggest that one cannot achieve better L^2-norms using different shapes than potential wells with sharp layer changes. At the end we also involve the angular acceptance[3] into our objective function to improve the efficiency of the resonators in the practical application.

1.3 Related work

Eigenvalue optimization problems arise in many different areas like optimal design problems (see e.g. *Cox and Overton* [CO92] and *Cox and McLaughlin* [CM90]), problems of optimal control (see e.g. *Boyd et. al* [BGFB94] and *Burke, Lewis, and Overton* [BLO00], [BLO03]) and even in problems from the theory of graphs (see e.g. *Donath and Hoffman* [DH73] and *Overton and Womersley* [OW93]). We firstly checked if similar techniques apply to our problem. It turned out that the underlying assumptions are not fulfilled in our problem, and we are also interested in derivatives of eigenvectors which causes additional problems. In particular, there exists a lot of work by *Overton and Burke* on the optimization of eigenvalues of a matrix-valued function depending on several parameters, but most of this work is restricted to symmetric/Hermitian matrices which we do not have. Without this assumption severe problems occur in the variational analysis. Moreover, we will also need derivatives of specially scaled eigenvectors such that the existing results, that we are aware of, cannot be applied directly to our problem. We will comment on this in more detail in Chapter 4 and give a detailed introduction to the topic and the existing literature at the point where it becomes relevant and when we are more familiar with the special properties of our problem.

As announced before, the concept of resonances is a basic tool in this thesis. For a more detailed introduction than the one in Chapter 2, we refer to *Hohage and Nannen* [HN09], [Nan08] and the references therein. These works also form the basis for the numerics later since they provide the Hardy space infinite elements method.

The central issue examined in this thesis is a problem of optimal design/shape. We want to improve the field enhancement of multilayer systems by changing

---

[3] By the angular acceptance we measure how exact one must match a resonant angle to achieve at least half of the field enhancement at the resonant frequency itself.

their design. Closely related to our problem is the one studied in an article by *Heidler et al.* [HBKW08] from 2008. They aim for the improvement of micro- and nanoscale structures (in particular photonic crystals) to decrease energy loss. In the considered model the energy loss can be measured by the magnitude of the imaginary part of a (scattering) resonance. Gradient methods are applied to a quality factor which they define as the reciprocal of the absolute value of the imaginary part of the resonance there. In our case such an objective will turn out to be inappropriate since other quantities have also to be taken into account. A reason for this are absorption effects which in contrast to the article of *Heidler et al.* [HBKW08] cannot be neglected in our case. Formulas for the necessary derivatives in *Heidler et al.* [HBKW08] are derived in the continuous setting, but not in a completely rigorous way. Another difference to our aims is the restriction to piecewise constant functions in the numerics which simplifies the evaluation of the objective function (e.g. it is possible to use a linear system to solve the scattering problem). We want to consider more general bounded profiles in the optimization algorithm as well.

There are several other articles also using (generalized) gradient methods for different types of problems related to our setting like *Kao and Santosa* [KS07], *Lipton, Shipman and Venakides* [LSV03] and *Dobson and Santosa* [DS04]. Motivated by photonic band gap devices in *Kao and Santosa* [KS07] a quality factor (similar to the one in *Heidler et al.* [HBKW08]) is optimized using integral equation methods to design good resonators. Boundary integral methods are used in *Lipton, Shipman and Venakides* [LSV03] to optimize resonances in photonic crystal slabs where a quite different objective function is used. The transmission coefficient[4] and its variational gradient are central there. *Dobson and Santosa* [DS04] examine optimally localized eigenmodes of an inhomogeneous membrane which leads to a self-adjoint analogue to the problem considered by *Heidler et al.* [HBKW08]. It is different from the ones mentioned before because it does not deal with resonances but with Dirichlet eigenvalues on a bounded domain. A discretization leading to symmetric/Hermitian matrices can be assumed then which simplifies the situation for the eigenvalue optimization significantly as already mentioned above.

Somewhat further away from our studies, but absolutely worth mentioned, are the articles by *Burger, Osher and Yablonovitch* [BOY04] and *Felici and Engl* [FE01] where for different reasons (like non-existence of solutions and unstable dependence on the data) ill-posedness occurs in the optimization problems and hence the methods presented there involve regularization techniques. In *Burger, Osher and Yablonovitch* [BOY04] this is done in the context of the design of photonic crystals and in *Felici and Engl* [FE01] for the optimization of optical waveguides.

There is also the very interesting doctoral thesis by *Schneck* [Sch09] on the optimization of the reflection coefficient by constrained optimization in Hardy spaces. The starting point of this work is a mirror design problem, i.e. the design of a multilayer structure (by changing the refractive index describing

---

[4] For a definition see Chapter 2, equation (2.6b).

the structure) to meet a desired complex-valued reflection coefficient. The thesis does not aim for a new method. It examines the space of realizable reflection coefficients and looks at the arising optimization problem for mirror design from a more abstract point of view. This leads to the consideration of an optimization problem in Hardy spaces[5] subject to side constraints which is examined analytically (existence, uniqueness, extremal property) and numerically. Although many of the keywords sound familiar from our problem, it is completely different in some sense. One thing is our objective function involving the field enhancement which we want to maximize. Of course, one could also ask the question how to design a system to reach a certain field distribution, but the reflection coefficient has much nicer properties, especially if absorption can be neglected (as in *Schneck* [Sch09]), than the complete field distribution or respectively the achieved field enhancement. Apart from this, the work of *Schneck* provides existence and uniqueness results for the considered optimization problem only in a larger solution space. From this it gains information on the accuracy of solutions of the original optimization problem by rigorous bounds on its optimum. Such bounds would also be desirable in our problem, but it turns out that useful objective functions are considerably more complicated and depend on many parameters. Hence, rigorous and useful upper bounds on the achievable field enhancement seem out of reach.

The decision to use the L^2-norm instead of the infinity norm is very reasonable as explained above. One could also work directly with the infinity norm by rewriting the problem into one with additional side constraints for every discretization point. This approach is used e.g. in *Grund and Rösch* [GR01] in the context of a control problem and leads to a huge set of additional side constraints, especially for finer discretizations.

Another possibility to solve our optimization problem are genetic algorithms, to avoid the effort of computing the derivatives. One attempt to apply genetic algorithms in the optimization of x-ray waveguides can be found in *Karimov and Kurmaev* [KK03]. But only very small systems of three layers were examined for which the layer thicknesses and the materials (only very few materials allowed) in the cladding (outer layers) were optimized. In general, genetic algorithms have the advantage of global instead of local convergence. They do not require much information on the problem and can even be applied to discontinuous problems. But as our problem can be shown to be differentiable they will perform worse compared to gradient methods as genetic algorithms have the disadvantage that the computational cost grows extremely fast with the number of degrees of freedom. Especially in our problem, evaluations of the objective function are very expensive since they require solving a resonance problem.

---

[5] The occurrence of Hardy spaces might lead to the conjecture of an approach very closely related to ours, but they are used as solution spaces for the optimization there while we use them in the modelling of the radiation conditions.

2 Derivation of the mathematical problem

The problem of finding improved x-ray resonators described in the Introduction will be formulated mathematically in this chapter. We arrive at an optimization problem involving resonances.

The situation of a fixed multilayer system and a fixed angle of incidence gives rise to a one-dimensional scattering problem which we formulate in strong and weak form, and we provide an existence and uniqueness result. With the help of this we are able to give a first formulation of our optimization problem. Since the optimization problem has a complicated form, we use the concept of resonances to make it more accessible. Two equivalent formal definitions of resonances are given. The first one introduces them as eigenvalues whose corresponding eigenfunctions fulfill a radiation condition. We prove that all resonances can be excited by the incident field and have a corresponding one-dimensional (eigen-)space of resonance functions.

2.1 Scattering problem

Before we turn to the mathematical formulation of the optimization problem and to resonances, we give a short derivation of the underlying scattering problem.

We consider a multilayer system which is impinged on the surface by a x-ray beam under the angle of incidence α. The refractive index n is assumed to depend only on the cartesian variable z. We assume that the upper half-space $\{[x,y,z] \in \mathbb{R}^3 : z > 0\}$ is filled with air and that the half-space $\{[x,y,z] \in \mathbb{R}^3 : z < -a\}$ is filled with a substrate such that

$$n(z) = \begin{cases} 1, & z > 0 \\ n_{\text{sub}}, & z < -a. \end{cases} \tag{2.1a}$$

Thus, n is non-constant only within the strip $-a \leq z \leq 0$. A sketch of the situation can be found in Figure 2.1.

In the x-ray regime the refractive index is of the form $n = 1 - \delta + i\beta$ with $\delta, \beta \geq 0$ and $\delta, \beta \ll 1$ for all known materials[6], and β describes the absorption. For our analysis in addition to (2.1a) we only assume that

$$n \in L^\infty(\mathbb{R}) \quad \text{with} \quad \operatorname{Re}(n) > 0 \text{ and } \operatorname{Im}(n) \geq 0. \tag{2.1b}$$

The propagation of x-rays can be described[7] by the time-harmonic Maxwell equations

$$\operatorname{curl}\operatorname{curl}\vec{E} - k_0^2 n^2 \vec{E} = 0$$

for the electric field $\vec{E}$, with $k_0 := 2\pi/\lambda$ and the wavelength λ.

The incident[8] field $\vec{E}_\text{i}$ is given by

$$\vec{E}_\text{i} = \left[0, \operatorname{Re}\left(e^{ik_0(x\cos\alpha - z\sin\alpha)} e^{-i\omega t}\right), 0\right]^\top. \tag{2.2}$$

---

[6] Typically (the real part of) the refractive index for x-rays is some 10^{-6} below 1.

[7] An elaborate introduction to the physical background of multilayer x-ray resonators can be found in [Pfe02, Section 2.1 and 2.2].

[8] The propagation direction of waves will be discussed immediately.

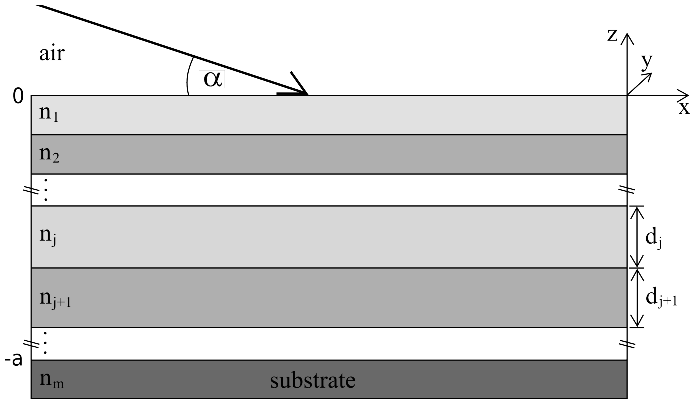


Figure 2.1: *example of a multilayer system*

If the electric field $\vec{E}$ is polarized in y-direction, the ansatz

$$\vec{E}(x,y,z) = [0, u(z)e^{i\sqrt{\nu}x}, 0]^\top \tag{2.3}$$

leads to the ordinary differential equation

$$-u''(z) - (k_0^2 n^2(z) - \nu)u(z) = 0 \tag{2.4}$$

on the whole real axis. The angle of incidence α is related to the parameter ν in (2.4) via

$$k_0 \cos\alpha = \sqrt{\nu}.$$

To have the chance of unique solvability for the differential equation (2.4) we need further conditions on the solution u.

Let us first consider the situation for $z > 0$. For $z > 0$ the differential equation (2.4) has two linearly independent solutions

$$u_{\mathrm{i}}(z) := e^{-i\sqrt{k_0^2-\nu}\,z} \quad \text{and} \quad e^{i\sqrt{k_0^2-\nu}\,z}, \tag{2.5}$$

remembering that $n \equiv 1$ for $z > 0$. If $\nu \in [0, k_0^2)$ $(0 < \alpha \leq \pi/2)$ the first solution corresponds to our incident field $\vec{E}_{\mathrm{i}}$ propagating in the direction $[\cos\alpha, 0, -\sin\alpha]^\top$, and the second solution corresponds to a reflected plane wave propagating in the direction $[\cos\alpha, 0, \sin\alpha]^\top$. This can be seen respectively from computing the time-averaged/complex Poynting vector which points in the direction of the energy flux or by directly examining the influence of the chosen time-dependence $e^{-i\omega t}$ on the solutions in (2.5).

Our discussion shows that we have to look for solutions u of the form

$$u(z) = u_{\mathrm{i}}(z) + \mathfrak{r}_\nu e^{i\sqrt{k_0^2-\nu}\,z} \qquad \text{for } z > 0 \tag{2.6a}$$

with a reflection coefficient $\mathfrak{r}_\nu \in \mathbb{C}$. The formulation of the problem is completed by requiring that there exists a transmission coefficient $\mathfrak{t}_\nu \in \mathbb{C}$ such that

$$u(z) = \mathfrak{t}_\nu e^{-i\sqrt{k_0^2 n_{\mathrm{sub}}^2-\nu}\,z} \qquad \text{for } z < -a. \tag{2.6b}$$

Note that all solutions to (2.4) in the interval $(-\infty, -a)$ are linear combinations of $e^{\pm i\sqrt{k^2 n_{\rm sub}^2 - \nu}\, z}$. If $\nu < \mathrm{Re}(k_0^2 n_{\rm sub}^2)$ we choose a wave propagating downward rather than upward, and if $\nu > \mathrm{Re}(k_0^2 n_{\rm sub}^2)$ we pick an exponentially decaying rather than an exponentially growing solution. Here and in the following, we always choose the branch of the square root function which has a cut at the negative imaginary axis

$$i\mathbb{R}_0^- := \{\nu \in \mathbb{C} : \mathrm{Re}(\nu) = 0 \text{ and } \mathrm{Im}(\nu) \leq 0\}\,. \tag{2.7}$$

We sum all this up in a definition and the complete formulation of the scattering problem.

Definition 2.1. *a) We call a solution v of (2.4) in $\{z \in \mathbb{R} : z > 0\}$ outgoing in the upper section if it is of the form $v(z) = \mathfrak{r}_\nu \cdot e^{i\sqrt{k_0^2 - \nu}\, z}$ for all $z > 0$ with a constant $\mathfrak{r}_\nu \in \mathbb{C}$.*
b) We call a solution v of (2.4) in $\{z \in \mathbb{R} : z < -a\}$ outgoing in the lower section if it is of the form $v(z) = \mathfrak{t}_\nu \cdot e^{-i\sqrt{k_0^2 n_{\rm sub}^2 - \nu}\, z}$ for all $z < -a$ with a constant $\mathfrak{t}_\nu \in \mathbb{C}$.

Scattering problem.

$$\begin{aligned}
&\text{Find } u \in H^2_{\rm loc}(\mathbb{R}), \text{ such that} \\
&-u''(z) - (k_0^2 n^2(z) - \nu)u(z) = 0 \quad \text{for } z \in \mathbb{R} && \text{(2.8a)} \\
&u = u_{\rm i} + u_{\rm s} && \text{(2.8b)} \\
&u_{\rm i}(z) = e^{-i\sqrt{k_0^2 - \nu} z} && \text{(2.8c)} \\
&u_{\rm s} \text{ is outgoing in the upper section} && \text{(2.8d)} \\
&u \text{ is outgoing in the lower section.} && \text{(2.8e)}
\end{aligned}$$

$H^2_{\rm loc}(\mathbb{R})$ denotes the space of functions whose restriction to $\mathfrak{K}$ belongs to the Sobolev space $H^2(\mathfrak{K})$ for every compact subset $\mathfrak{K} \subset \mathbb{R}$. The use of $H^2_{\rm loc}(\mathbb{R})$ is necessary as the expected solutions are not globally H^2-functions which becomes immediately clear from equation (2.6a). For a brief introduction to Sobolev spaces, we refer to [RR94, Section 6.4].

2.2 Optimization problem

As mentioned in the Introduction, for a suitable multilayer system there exist certain angles of incidence (we call these angles resonant angles, and in terms of the corresponding ν we will often speak of resonant frequencies) for which a very high field enhancement inside the system can be observed.
When the incident x-ray beam impinges on the surface of the multilayer system, a part of the incident field is reflected and the rest penetrates into the system. For a suitable combination of parameters a standing wavefield $\vec{E}$ with enhanced field inside the system occurs and propagates along the x-axis. Figure 2.2 sketches the described effect.
Let us motivate the effect a little more from a physical point of view and also discuss the interval of angles in which we expect resonant angles. As a first

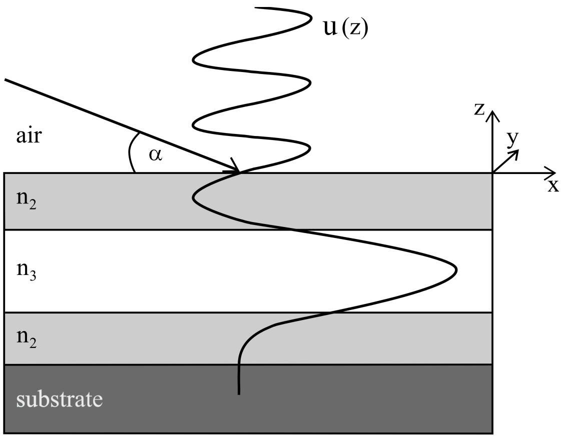


Figure 2.2: *sketch of a multilayer x-ray resonator with an impinging beam that excites a resonant state*

example, consider a system with piecewise constant refractive index consisting of three layers on a substrate with $\mathrm{Re}(n_2) < \mathrm{Re}(n_3) < 1$ as sketched in Figure 2.2. Neglecting absorption effects for a moment, for angles of incidence α above the critical angle[9] between air and n_2

$$\alpha_{\mathrm{cr},n_2} := \arccos\left(\mathrm{Re}(n_2)\right) \tag{2.9}$$

an (with α) increasing part of the incident field penetrates into the whole multilayer system. In contrast, for $\alpha < \alpha_{\mathrm{cr},n_3} := \arccos\left(\mathrm{Re}(n_3)\right)$ only a very small part of the incident field is transmitted into the system as a strongly decaying evanescent wave. Therefore, the interesting interval is

$$\alpha_{\mathrm{cr},n_3} \leq \alpha \leq \alpha_{\mathrm{cr},n_2}, \tag{2.10}$$

in which for a suitable combination of thicknesses and refractive indices there are certain angles of incidence for which a highly enhanced field can be observed inside the system. A part of the incident field tunnels through the top layer (with n_2) as an evanescent wave and only a very small part tunnels back. Especially almost everything is reflected on the thick bottom layer while the field propagates along the x-axis (for a more detailed description, we refer to [Pfe02, Section 2.2.2-2.2.5]).

For a general refractive index we can restrain the region where we expect resonant angles by

$$\arccos\left(\max_{z\in[-a,0]} \mathrm{Re}(n(z))\right) \leq \alpha \leq \arccos\left(\min_{z\in[-a,0]} \mathrm{Re}(n(z))\right) \tag{2.11}$$

---

[9] If the wavelength is small compared to the geometry, one may use geometrical optics as an approximation to the full physical optics model. The small absorption coefficient is neglected in the definition of the critical angle.

or respectively for the resonant frequencies by

$$k_0^2 \left(\min_{z\in[-a,0]} \mathrm{Re}(n(z)) \right)^2 \leq \nu \leq k_0^2 \left(\max_{z\in[-a,0]} \mathrm{Re}(n(z)) \right)^2 . \tag{2.12}$$

What we are interested in, is the improvement of the multilayer structures to get higher field enhancements or respectively better resonant states by modifying the refractive index. But before we formulate the optimization problem, let us look at an example to illustrate the situation.

2.2.1 Example of a multilayer x-ray resonator

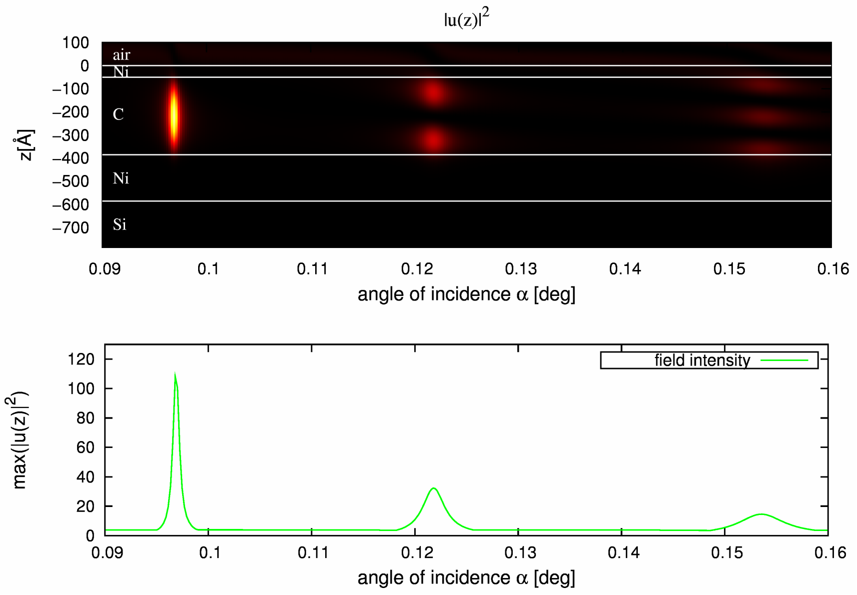


Figure 2.3: *upper panel: field intensity along the z-axis for different angles of incidence α, lower panel: maximum field intensity for different angles of incidence α; values: Table 2.1*

We consider again a simple multilayer system consisting of three layers and a substrate (a carbon layer between two nickel layers on silicon substrate) where the refractive index is assumed to be piecewise constant[10]. In this case we are able to compute the solution to the scattering problem (2.8) analytically from a linear system (known as matrix or Parratt algorithm, see e.g. the classical paper of *Parratt* [Par54] or [BW97] and [Lek87]). The used values can be found in Table 2.1 and the results in Figure 2.3. For the layer thicknesses we use the unit Ångström (1Å=0.1nm).
In the upper panel of Figure 2.3 the field intensity $|u_s(z)|^2$ along the z-axis is plotted for different angles of incidence and in the lower panel the maximum

---

[10] The presented example is taken from [PSH+00].

field intensity. The peaks in the graphs clearly show the resonant frequencies or respectively the resonant angles of incidence we are looking for.

layer	thickness (in Å)	refractive index
Ni	$d_1 = 50$	$1 - 4.45 \cdot 10^{-6} + 1.37 \cdot 10^{-7} i$
C	$d_2 = 335$	$1 - 1.15 \cdot 10^{-6} + 2.21 \cdot 10^{-10} i$
Ni	$d_3 = 200$	$1 - 4.45 \cdot 10^{-6} + 1.37 \cdot 10^{-7} i$
Si	infinite	$1 - 1.20 \cdot 10^{-6} + 4.56 \cdot 10^{-9} i$

Table 2.1: *values used in the computation for Figure 2.3 at* $\lambda = 0.62 \mathring{A}$

2.2.2 First formulation of the optimization problem

We are now able to formulate the verbally described optimization problem mathematically. By $u[\nu, n]$ we denote[11] the solution u of (2.8) restricted to $[-a, 0]$ (inside the system) for the refractive index n and the angle-dependent parameter ν.

OPTIMIZATION PROBLEM 2.2.

$$\max_{n} \max_{\nu} \; \|u[\nu, n]\|_\infty \text{ under (2.8) as side condition and side conditions on } n \text{ and } \nu.$$

Optimization Problem 2.2 is rather complicated because it is not only a parameter optimization problem in n with a differential equation as side condition, but also a nested problem due to the maximum over ν. Moreover, the interval $[0, k_0^2)$ for ν is quite large since the peaks in the intensity curve are very narrow as one can see in Figure 2.3. As we will later see (Section 4.3.2), it even holds: the higher a peak in the intensity is, the narrower it is. Therefore, we need a very fine resolution for ν because otherwise the best peaks are likely to be missed. If we want to reduce the set of permitted values of ν, we have to do it in dependence of n, as seen in equation (2.12), but side conditions depending on each other would make the problem even more complicated.

To make the problem more accessible, we use the concept of resonances. Resonances are complex numbers in our case, and their real parts approximate the resonant frequencies as we will see in Section 4.3.1. We introduce the basic definitions and basic theory in the two following sections.

2.3 Reformulation of the differential equation

For the following analysis and the numerical solution we have to replace the differential equation (2.8a) posed on $\mathbb{R}$ by a differential equation posed on the bounded interval $[-a, 0]$. Moreover, it will be essential to study problem (2.8) also for complex values of ν. The set of all admissible ν is given by

$$\mathfrak{Z} := \left\{ \nu \in \mathbb{C} : \operatorname{Re}(\nu) \in [0, k_0^2) \wedge (k_0^2 n_{\text{sub}}^2 - \nu) \notin i\mathbb{R}_0^- \right\}. \tag{2.13}$$

---

[11] Square brackets are used to denote the dependence on ν and n since u is already a function (of z) itself. We will often omit the square brackets, in particular in the differential equations, whenever it is clear from the context which n and ν are used.

Let us introduce the Dirichlet-to-Neumann numbers

$$\mathrm{DtN}_\nu^+ := i\sqrt{k_0^2-\nu} \quad \text{and} \quad \mathrm{DtN}_\nu^- := -i\sqrt{k_0^2 n_{\mathrm{sub}}^2-\nu}.$$

Looking at the general solution to (2.8a) in $(0,\infty)$ it is easy to see that u is of the form (2.6a) if and only if $u_{\mathrm{s}} = u - u_{\mathrm{i}}$ satisfies $u_{\mathrm{s}}'(0) = \mathrm{DtN}_\nu^+ u_{\mathrm{s}}(0)$. Similarly, (2.6b) is equivalent to $u'(-a) = \mathrm{DtN}_\nu^- u(-a)$. Using these relations we obtain the following equivalent formulation of problem (2.8).

Lemma 2.3. *Let* $\nu \in \mathfrak{Z}$. *If* $u \in H^2_{\mathrm{loc}}(\mathbb{R})$ *is a solution to* (2.8), *then* $u_{\mathrm{s}} := (u-u_{\mathrm{i}})|_{[-a,0]}$ *satisfies*

$$-u_{\mathrm{s}}''(z) - (k_0^2 n^2(z) - \nu)u_{\mathrm{s}}(z) = f_\nu(z), \qquad z \in [-a,0] \tag{2.14a}$$

$$u_{\mathrm{s}}'(0) = \mathrm{DtN}_\nu^+ u_{\mathrm{s}}(0) \tag{2.14b}$$

$$u_{\mathrm{s}}'(-a) = \mathrm{DtN}_\nu^- u_{\mathrm{s}}(-a) + \mathrm{DtN}_\nu^- u_{\mathrm{i}}(-a) - u_{\mathrm{i}}'(-a) \tag{2.14c}$$

with $f_\nu(z) := k_0^2(n^2(z)-1)e^{-i\sqrt{k_0^2-\nu}\,z}$.
Vice versa, if $u_{\mathrm{s}} \in H^2([-a,0])$ *is a solution to* (2.14), *then* $u(z) := e^{-i\sqrt{k_0^2-\nu}\,z} + u_{\mathrm{s}}(z)$ *is the restriction of a solution to* (2.8).

Remark 2.4. *The boundary conditions (2.14b) and (2.14c) are also known as transparent boundary conditions. In virtue of Lemma 2.3 and as it will be clear from the context, we do not distinguish in our notation between* u_{s} *on all of* $\mathbb{R}$ *and its restricted version to* $[-a,0]$.

Definition 2.5. *We call a function* $u_{\mathrm{s}} \in H^2([-a,0])$ *a strong solution if it fulfills (2.14).*

A weak formulation of (2.14) can be obtained in a standard way. We write it in operator form as

$$(I + T(\nu))u_{\mathrm{s}} = G(\nu), \tag{2.15}$$

where the bounded linear operators $I,\ T(\nu) \in \mathfrak{L}(H^1([-a,0]))$ and the function $G(\nu) \in H^1([-a,0])$ are implicitly given by the relations[12]

$$\langle Iu, v\rangle_{H^1} = \int_{-a}^0 u'\overline{v}' + u\overline{v}\,\mathrm{d}z, \tag{2.16a}$$

$$\begin{aligned}\langle T(\nu)u, v\rangle_{H^1} = {} & -k_0^2\int_{-a}^0 n^2 u\overline{v}\,\mathrm{d}z + (\nu-1)\int_{-a}^0 u\overline{v}\,\mathrm{d}z \\ & -\mathrm{DtN}_\nu^+ u(0)\overline{v}(0) + \mathrm{DtN}_\nu^- u(-a)\overline{v}(-a),\end{aligned} \tag{2.16b}$$

$$\langle G(\nu), v\rangle_{H^1} = \int_{-a}^0 f_\nu \overline{v}\,\mathrm{d}z - \Big(\mathrm{DtN}_\nu^- u_{\mathrm{i}}(-a) - u_{\mathrm{i}}'(-a)\Big)\,\overline{v}(-a), \tag{2.16c}$$

which hold for all $u, v \in H^1([-a,0])$.

Definition 2.6. *We call a function* $u_{\mathrm{s}} \in H^1([-a,0])$ *a weak solution if it fulfills (2.15).*

---

[12] Here and in the following we always use the convention that scalar products are antilinear in the second argument, and by $\mathfrak{L}(X)$ we denote the set of bounded linear operators on a normed vector space X.

So far, we have only stated the problem in different formulation. Existence and uniqueness of solutions are provided by the following proposition.

Proposition 2.7.

1. *$u_\mathrm{s} \in H^1([-a,0])$ is a solution to* (2.15) *if and only if it is a solution to* (2.14).
2. *The equivalent problems* (2.8), (2.14) *and* (2.15) *have a unique solution for all $\nu \in \mathfrak{Z}$ with* $\mathrm{Im}(\nu) \leq 0$.

Proof. 1. This follows by partial integration and elliptic regularity results, for details see e.g. [RR94, Thm. 8.53].
2. We show that $(I + T(\nu))$ has a bounded inverse using Riesz theory.
$T(\nu) \in \mathfrak{L}(H^1([-a,0]))$ is a compact operator which follows from the compactness of the embedding $H^1([-a,0]) \hookrightarrow L^2([-a,0])$ (see e.g. [RR94, Thm. 6.98]) and the fact that the DtN numbers define rank 1 operators.
To prove injectivity of $(I + T(\nu))$, assume that $(I + T(\nu))u_\mathrm{s} = 0$. Then

$$\begin{aligned} 0 = \mathrm{Im}\left(\langle (I + T(\nu))u_\mathrm{s}, u_\mathrm{s}\rangle_{H^1}\right) = &- k_0^2 \int_{-a}^{0} \mathrm{Im}(n^2)|u_\mathrm{s}|^2 \,\mathrm{d}z + \mathrm{Im}(\nu)\|u_\mathrm{s}\|_{L^2}^2 \\ &- \mathrm{Im}(\mathrm{DtN}_\nu^+)|u_\mathrm{s}(0)|^2 + \mathrm{Im}(\mathrm{DtN}_\nu^-)|u_\mathrm{s}(-a)|^2. \end{aligned}$$

Since by our assumptions all terms on the right hand side are nonpositive, all of them must vanish. As

$$\mathrm{Im}(\mathrm{DtN}_\nu^+) = \mathrm{Re}\left(\sqrt{k_0^2 - \nu}\right) > 0, \tag{2.17}$$

we obtain that $u_\mathrm{s}(0) = 0$. Using the equivalence of (2.15) and (2.14), it follows that $u_\mathrm{s}'(0) = 0$ and uniqueness results for ordinary differential equations (see Appendix) imply that $u_\mathrm{s} \equiv 0$.
Now, Riesz theory implies that $(I + T(\nu))$ has a bounded inverse (see e.g. [Kre89, Thm. 4.17]), and hence (2.15) has a unique solution. □

In view of the following section and subsequent considerations, we cite another equivalent weak formulation of (2.8). It is called Hardy space formulation, and in contrast to (2.15) it preserves the linear structure in ν.

Definition 2.8. *The Hardy space $H^+(S^1)$ is the set of all functions $f \in L^2(S^1)$, which are L^2-boundary values of a in the unit disk $\{s \in \mathbb{C} \,|\, |s| < 1\}$ holomorphic function g, in the sense that $\lim_{r \nearrow 1} \int_0^{2\pi} \left|g\left(re^{i\varphi}\right) - f\left(e^{i\varphi}\right)\right| \,\mathrm{d}\varphi = 0$, and for which the integrals $\int_0^{2\pi} |g(re^{i\varphi})|^2 \,\mathrm{d}\varphi$ are uniformly bounded for $r \in [0,1)$. Equipped with the L^2-scalar product $H^+(S^1)$ becomes a Hilbert space.*

Lemma 2.9. *Let $X^\mathrm{H} := H^+(S^1) \times H^1([-a,0]) \times H^+(S^1)$, $\nu \in \mathfrak{Z}$ and $u_s \in H^2_\mathrm{loc}(\mathbb{R})$ a solution to*

$$-u_\mathrm{s}''(z) - (k_0^2 n^2(z) - \nu)u_\mathrm{s}(z) = f_\nu(z) \quad \text{for } z \in \mathbb{R} \tag{2.18a}$$

$$u_\mathrm{s} \text{ is outgoing in the upper section} \tag{2.18b}$$

$$u_\mathrm{s} + u_\mathrm{i} \text{ is outgoing in the lower section} \tag{2.18c}$$

with f_ν *as in Lemma 2.3. Then we can define bilinear forms*[13] $\mathcal{B}[n], \mathcal{M} : X^{\mathrm{H}} \to \mathbb{C}$ *and* $\boldsymbol{f}(\nu) \in X^{\mathrm{H}}$ *for problem (2.18) and functions* $u^{\ominus}, u_{\mathrm{s}}^{\oplus} \in H^{+}(S^1)$ *such that* $\boldsymbol{u}_{\mathrm{s}} := \left(u^{\ominus}, u_{\mathrm{s}}|_{[-a,0]}, u_{\mathrm{s}}^{\oplus}\right) \in X^{\mathrm{H}}$ *fulfills*

$$\mathcal{B}[n](\boldsymbol{u}_{\mathrm{s}}, \boldsymbol{v}) - \nu\mathcal{M}(\boldsymbol{u}_{\mathrm{s}}, \boldsymbol{v}) = \mathcal{M}(\boldsymbol{f}(\nu), \boldsymbol{v}) \tag{2.19}$$

for all $\boldsymbol{v} \in X^{\mathrm{H}}$*. Vice versa if* $\boldsymbol{u}_{\mathrm{s}} = (u^{\ominus}, u_{\mathrm{s}}, u_{\mathrm{s}}^{\oplus}) \in X^{\mathrm{H}}$ *is a solution to (2.19), then* u_{s} *belongs to* $H^2([-a,0])$ *and is the restriction to* $[-a,0]$ *of a solution to (2.18).*
Equation (2.19) corresponds to an equivalent operator equation of the form

$$(B(n) - \nu M)\boldsymbol{u}_{\mathrm{s}} = M\boldsymbol{f}(\nu), \tag{2.20}$$

with a boundedly invertible operator $M \in \mathfrak{L}(X^{\mathrm{H}})$ *and* $(B(n) - \nu M) \in \mathfrak{L}(X^{\mathrm{H}})$ *is a Fredholm operator of index* 0*. If additionally* $\mathrm{Im}(\nu) \leq 0$*,* $(B(n) - \nu M)$ *is boundedly invertible.*

For details, we refer to [HN09, Theorem 2.4]. They are not explained here, as this leads us too far away from our main purpose. We remark the following: In contrast to the situation in [HN09, Theorem 2.4] we have radiation conditions on both sides and an inhomogeneous differential equation. Because of the two radiation conditions $\boldsymbol{u}_{\mathrm{s}}$ consists of three components. Considering the set $\mathfrak{Z}$ of admissible ν, one can observe that we allow an exponential decay in the lower section. This is also covered by the Hardy space formulation, we refer in particular to [HN09, Remark 2.8]. Since $\mathrm{supp}(f_\nu) \subset [-a,0]$, the inhomogeneous differential equation does not cause any problems in the exterior of $[-a,0]$, where we incorporate the radiation conditions into the weak formulation via the Hardy spaces. In particular, $\boldsymbol{f}(\nu) := [0, \widetilde{f}_\nu, 0] \in X^{\mathrm{H}}$ and $\mathcal{M}(\boldsymbol{f}(\nu), \boldsymbol{v})$ have no parts in the Hardy spaces and $\mathcal{M}(\boldsymbol{f}(\nu), \boldsymbol{v})$ is an analog to (2.16c). The part $\widetilde{f}_\nu \in H^1([-a,0])$ includes terms that allow for the fact that in the lower section the total field $u = u_{\mathrm{s}} + u_{\mathrm{i}}$ is outgoing and not the u_{s}, but it stays holomorphic in ν. However, the proof in [HN09, Theorem 2.4] carries over to our lemma. Note that (2.18) is an equivalent formulation of the scattering problem (2.8).
For the operator equation (2.20) observe that there exists an isometric bijection between the space of bounded linear operators on a Hilbert space and the space of bounded sesquilinear forms on this space (see [Con90], Theorem 2.2). That $(B(n) - \nu M)$ is a Fredholm operator of index 0 is shown in [HN09, Theorem 2.5 and Corollary 2.6]. The injectivity of this operator for $\mathrm{Im}(\nu) \leq 0$ follows from Proposition 2.7, using the equivalence of the formulations. By Riesz theory this implies the bounded invertibility.
Lemma 2.9 is the basis for the Hardy space method which we later use to discretize our problem.

2.4 Resonances

We give two equivalent definitions of resonances. The first one introduces them as eigenvalues with outgoing eigenfunctions while the second definition

---

[13] The notation $\mathcal{B}[n]$ indicates that this bilinear form includes the refractive index n.

is based on the analytic Riesz-Fredholm theory[14] and introduces them as poles of the resolvent. If we write simply “outgoing” for a function, we mean that it is outgoing in both sections.

THEOREM AND DEFINITION 2.10. *A complex number $\nu_\star \in \mathfrak{Z}$ is called a resonance if it satisfies one of the following equivalent conditions:*

1. *There exist nontrivial solutions $u_{s\star} \in H^2_{\mathrm{loc}}(\mathbb{R})$ (resonance functions) of the eigenvalue problem*

$$\left(-\frac{d^2}{dz^2} - k_0^2 n^2(z)\right) u_{s\star}(z) = -\nu_\star u_{s\star}(z) \quad \textit{for } z \in \mathbb{R} \tag{2.21a}$$

$$u_{s\star} \textit{ is outgoing} \tag{2.21b}$$

2. *$\nu_\star \in \mathfrak{Z}$ is a pole of the resolvent*

$$R : \mathfrak{Z} \to \mathfrak{L}(H^1([-a,0])), \quad R(\nu) := (I + T(\nu))^{-1}. \tag{2.22}$$

PROOF. Let $\nu_\star$ be a resonance after the first definition. Then we have a nontrivial solution $u_{s\star} \in H^2([-a,0])$ to the following homogeneous form of (2.14):

$$\begin{aligned} -u''_{s\star}(z) - (k_0^2 n^2(z) - \nu_\star) u_{s\star}(z) &= 0 \quad \text{for all } z \in [-a,0] && (2.23\text{a}) \\ u'_{s\star}(0) &= \mathrm{DtN}^+_{\nu_\star} u_{s\star}(0) && (2.23\text{b}) \\ u'_{s\star}(-a) &= \mathrm{DtN}^-_{\nu_\star} u_{s\star}(-a). && (2.23\text{c}) \end{aligned}$$

This $u_{s\star}$ solves the homogeneous form of the weak formulation:

$$(I + T(\nu_\star)) u_{s\star} = 0. \tag{2.24}$$

Therefore, the operator $(I + T(\nu_\star))$ is not invertible at $\nu_\star$ and $\nu_\star$ is a pole of $R(\nu)$ since the resolvent is a meromorphic function. The proof of this property can be found below in Theorem 2.12.
Starting now with a resonance $\nu_\star$ after the second definition which means $\nu_\star$ is a pole of the resolvent and $(I + T(\nu_\star))$ is therefore not invertible. Then there exists a $u_{s\star} \neq 0$ in $H^1([-a,0])$ with $(I + T(\nu_\star)) u_{s\star} = 0$ because for our operator injectivity and surjectivity are equivalent (cf. proof of Proposition 2.7). $u_{s\star}$ is then also a solution to (2.23) in $H^2([-a,0])$ and hence the restriction to $[-a,0]$ of a nontrivial solution to (2.21). □

We will show now an important result on the existence, number and location of resonances for our problem. As a preparation, we cite a famous theorem of *Steinberg*.

PROPOSITION 2.11. *For a Banach space X, a domain $\mathfrak{D} \subseteq \mathbb{C}$ and an operator-valued analytic function $T : \mathfrak{D} \to \mathfrak{L}(X)$, with $T(\nu)$ compact for all $\nu \in \mathfrak{D}$, either (i) or (ii) holds:*

(i) *$(I - T(\nu))$ is not invertible for any $\nu \in \mathfrak{D}$.*

(ii) *$(I - T(\nu))$ is invertible except for at most a discrete subset of $\mathfrak{D}$ and $(I - T(\nu))^{-1}$ is a meromorphic function in $\mathfrak{D}$.*

---

[14] For details on the analytic Riesz-Fredholm theory see e.g. [Kat95].

PROOF. See [Tay96, Ch.9, Proposition 7.4]. □

THEOREM 2.12. *1. The resolvent R is a meromorphic function on $\mathfrak{Z}$. In particular, we have at most a discrete set of resonances.*

2. There cannot exist any resonances $\nu_\star \in \mathfrak{Z}$ with $\operatorname{Im}\nu_\star \leq 0$.

PROOF. 1. The compactness of $T(\nu)$, which we have already discussed in the proof of Proposition 2.7, extends to all $\nu \in \mathfrak{Z}$. Moreover, the mapping $\nu \mapsto T(\nu)$ on $\mathfrak{Z}$ with values in $\mathfrak{L}(H^1([-a,0]))$ is an analytic function. For this, it suffices to show (see e.g. [Kat95, III.§3.1, Thm. 3.12]) that the right hand side of (2.16b) depends holomorphically on ν. This can be seen quite easily. Now we can apply Proposition 2.11.
2. In Proposition 2.7 we have seen that the resolvent is invertible for all $\nu \in \mathfrak{Z}$ with $\operatorname{Im}\nu \leq 0$. □

The question arises whether the resonances, if they exist, can be excited by the incident field. Mathematically this means to check if a pole $\nu_\star$ of R is also a pole of the mapping $\nu \mapsto R(\nu)G(\nu)$. We begin with a preparing Lemma.

LEMMA 2.13. *Let $\nu_\star$ be a resonance and $u_{s\star}$ a corresponding resonance function, i.e. a function $u_{s\star} \neq 0$ with $(I+T(\nu_\star))u_{s\star} = 0$. Then*
a) $(I+T(\nu_\star)^*)\overline{u_{s\star}} = 0$ and
b) $\langle G(\nu_\star), \overline{u_{s\star}}\rangle_{H^1} = -2i\sqrt{k_0^2-\nu_\star}\,u_{s\star}(0) \neq 0$.

PROOF. a) For arbitrary $v \in H^1([-a,0])$ we have

$$\begin{aligned}
&\langle \overline{v}, (I+T(\nu_\star)^*)\overline{u_{s\star}}\rangle_{H^1} = \langle (I+T(\nu_\star))\overline{v}, \overline{u_{s\star}}\rangle_{H^1}\\
&= \int_{-a}^{0} \overline{v}' u_{s\star}' - k_0^2 \int_{-a}^{0} n^2 \overline{v} u_{s\star}\,\mathrm{d}z + \nu_\star \int_{-a}^{0} \overline{v} u_{s\star}\,\mathrm{d}z\\
&\quad - \mathrm{DtN}^{+}_{\nu_\star}\overline{v}(0)u_{s\star}(0) + \mathrm{DtN}^{-}_{\nu_\star}\overline{v}(-a)u_{s\star}(-a)\\
&= \langle (I+T(\nu_\star))u_{s\star}, v\rangle_{H^1} = 0.
\end{aligned}$$

Because $v \in H^1([-a,0])$ was arbitrary, we obtain $(I+T(\nu_\star)^*)\overline{u_{s\star}} = 0$.
b) By definition we have $f_\nu = k_0^2(n^2-1)u_{i,\nu} = u_{i,\nu}'' + (k_0^2 n^2 - \nu)u_{i,\nu}$ with $u_{i,\nu}(z) = e^{-i\sqrt{k_0^2-\nu}z}$ and therefore

$$\begin{aligned}
\langle G(\nu_\star), \overline{u_{s\star}}\rangle_{H^1} = &\int_{-a}^{0} \left[u_{i,\nu_\star}'' + (k_0^2 n^2 - \nu_\star)u_{i,\nu_\star}\right] u_{s\star}\,\mathrm{d}z\\
&- \left[\mathrm{DtN}^{-}_{\nu_\star} u_{i,\nu_\star}(-a) - u_{i,\nu_\star}'(-a)\right] u_{s\star}(-a).
\end{aligned}$$

We perform twice a partial integration to get

$$\begin{aligned}
\langle G(\nu_\star), \overline{u_{s\star}}\rangle_{H^1} = &\int_{-a}^{0} u_{i,\nu_\star}\left[u_{s\star}'' + (k_0^2 n^2 - \nu_\star)u_{s\star}\right]\mathrm{d}z\\
&+ u_{s\star}(0)u_{i,\nu_\star}'(0) - u_{s\star}(-a)u_{i,\nu_\star}'(-a) - u_{s\star}'(0)u_{i,\nu_\star}(0) + u_{s\star}'(-a)u_{i,\nu_\star}(-a)\\
&- \left[\mathrm{DtN}^{-}_{\nu_\star} u_{i,\nu_\star}(-a) - u_{i,\nu_\star}'(-a)\right] u_{s\star}(-a). \qquad (2.25)
\end{aligned}$$

Since $u_{\mathrm{s}\star}$ is a resonance function to $\nu_\star$, we know that $u_{\mathrm{s}\star}$ is a solution to (2.23). In particular, $u''_{\mathrm{s}\star} + (k_0^2 n^2 - \nu_\star)u_{\mathrm{s}\star} = 0$ and together with the DtN conditions (2.23b) and (2.23c), equation (2.25) reduces to

$$\langle G(\nu_\star), \overline{u_{\mathrm{s}\star}}\rangle_{H^1} = -2i\sqrt{k_0^2 - \nu_\star}\, u_{\mathrm{s}\star}(0). \tag{2.26}$$

It remains to show that (2.26) is not equal to zero. Let us assume the contrary. But then we have $u_{\mathrm{s}\star}(0) = 0$ and moreover

$$u'_{\mathrm{s}\star}(0) = \mathrm{DtN}^+_{\nu_\star} u_{\mathrm{s}\star}(0) = 0, \tag{2.27}$$

which implies $u_{\mathrm{s}\star}(z) = 0$ for all $z \in [-a, 0]$, using uniqueness results from the theory of ordinary differential equations (see Appendix). This is a contradiction to our assumption that $u_{\mathrm{s}\star}$ is a resonance function. □

Theorem 2.14. *The two following statements are equivalent:*

1. $\nu_\star$ *is a pole of* $\nu \mapsto R(\nu)$.
2. $\nu_\star$ *is a pole of* $\nu \mapsto R(\nu)G(\nu)$.

Proof. Consider the mapping $G : \mathfrak{Z} \to H^1([-a, 0])$ with $G(\nu)$ defined by

$$\langle G(\nu), v\rangle_{H^1} = \int_{-a}^{0} f_\nu \overline{v}\, \mathrm{d}z - \Big(\mathrm{DtN}^-_\nu u_{i,\nu}(-a) - u'_{i,\nu}(-a)\Big)\, \overline{v}(-a). \tag{2.28}$$

One can proof the weak holomorphy of G, which means that $l(G)$ is holomorphic for every l in the dual space of $H^1([-a, 0])$, and weak holomorphy implies holomorphy[15]. With the help of the Riesz representation theorem it thus suffices to show that the right hand side of (2.28) depends holomorphically on ν. This can be deduced from Lebesgue's dominated convergence theorem. Therefore, G is a holomorphic function in ν, and it is obvious that every pole of the mapping $\nu \mapsto R(\nu)G(\nu)$ is a pole of $\nu \mapsto R(\nu)$.
Let $v_\star$ now be a pole of R and let us assume that it is not a pole of the mapping $\nu \mapsto R(\nu)G(\nu)$. Then the function $\varphi(\nu) := R(\nu)G(\nu)$ is analytic in the vicinity of $\nu_\star$ with

$$\varphi(\nu_\star) := \lim_{\nu \to \nu_\star} R(\nu)G(\nu). \tag{2.29}$$

Moreover, $\varphi(\nu)$ fulfills the equation

$$(I + T(\nu))\, \varphi(\nu) = G(\nu) \tag{2.30}$$

for all ν in the vicinity of $\nu_\star$ and particularly for $\nu_\star$. With a resonance function $u_{\mathrm{s}\star}$ to the resonance $\nu_\star$ by a) of Lemma 2.13 we can write

$$0 = \langle \varphi(\nu_\star), (I + T(\nu_\star)^*)\overline{u_{\mathrm{s}\star}}\rangle_{H^1} = \langle (I + T(\nu_\star))\varphi(\nu_\star), \overline{u_{\mathrm{s}\star}}\rangle_{H^1} = \langle G(\nu_\star), \overline{u_{\mathrm{s}\star}}\rangle_{H_1}.$$

But the last scalar product is not equal to zero by b) of Lemma 2.13 and the proof is finished. □

---

[15] A proof of the fact that from weak holomorphy follows holomorphy can be found in [Kat95, III.§1.6, Thm.1.37].

2.5 Optimization problem using resonances

As announced before, we will formulate our optimization problem in a different way using the concept of resonances.
So far, we have not commented on the connection between the resonant frequencies and the complex resonances. We will do this later (see Section 4.3.1) in more detail. It will turn out that the real parts of the complex resonances approximate the resonant frequencies.
We pick the best resonance for some initial system with admissible (cf.(2.1)) refractive index $\check{n}$, which supports at least one resonant state, i.e. we pick the resonance whose corresponding resonant frequency produces the highest field enhancement for the system with $n = \check{n}$. By Theorem 2.12 we know that this resonance is isolated. Hence, it can be separated from the other resonances by a closed curve enclosing this resonance but no other resonances. Let us assume that the picked resonance is simple (order of the pole of the resolvent is one), then there exists a neighborhood of $\check{n}$ in which the resonance changes continuously with n and stays simple (for a complete discussion of this fact, we refer to Chapter 4). We choose the biggest possible neighborhood for which this is true and can define a function $\nu_\diamond(n)$ inside, arising from the best resonance of the system with refractive index $\check{n}$. Note carefully that the resonance does not coalesce with other resonances while changing n under the given assumptions. This leads to the following optimization problem.

Optimization problem 2.15.

$$\max_{n} \ \|u[\mathrm{Re}(\nu_\diamond[n]), n]\|_\infty \quad \text{under side conditions on } n.$$

Since the infinity norm is not differentiable we replace it by the L^2-norm of the solution. It is justified by the numerical experience that both are almost proportional in our application and may also be motivated by the fact that $\int_{-a}^{0} |u(z)|^2 \,\mathrm{d}z$ is proportional to the energy inside the system.

Optimization problem 2.16.

$$\max_{n} \ \|u[\mathrm{Re}(\nu_\diamond(n)), n]\|_{L^2} \quad \text{under side conditions on } n.$$

Remark 2.17. *Note that we got rid of the maximum over the ν and do not have a nested optimization problem anymore. Although we have to compute resonances now, we have simplified the optimization problem significantly. In particular, we have also simplified the side conditions as we have seen that they get dependent on each other if we try to reduce the interval for ν in Optimization Problem 2.2.*

We close the section with an important result on the geometric multiplicity of the resonances. Although we treat our equation like a partial differential equation, we will take advantage of results for ordinary differential equations in the proof of the following theorem.

THEOREM 2.18. *All the resonances have geometric multiplicity* 1, *i.e. their eigenspaces of resonance functions are all one-dimensional.*

PROOF. Take any admissible refractive index n and let $\nu_\star$ a corresponding resonance after the first definition which means there exist non-trivial solutions to (2.21).
Consider for $\nu \in \mathfrak{Z}$ the initial value problem

$$-v''(z) - (k_0^2 n^2 - \nu)v(z) = 0 \quad \text{for } z \in \mathbb{R} \tag{2.34a}$$

$$v(0) = \gamma_1 \tag{2.34b}$$

$$v'(0) = \gamma_2 \tag{2.34c}$$

which has for all pairs $\gamma_1, \gamma_2 \in \mathbb{C}$ of initial values a unique solution in $H^2_{\text{loc}}(\mathbb{R})$ (see Appendix). Thus, there is a one-to-one correspondence between $\mathbb{C}^2$ and the space of solutions of the homogeneous ordinary differential equation (2.34a).
For all $\nu \in \mathfrak{Z}$ a fundamental system for the differential equation is given by the solutions to the two initial values problems with

$$\begin{aligned} &\gamma_1 = 1,\ \gamma_2 = i\sqrt{k_0^2 - \nu} \\ \text{and} \quad &\gamma_1 = 1,\ \gamma_2 = -i\sqrt{k_0^2 - \nu}. \end{aligned}$$

We denote these two solutions by $v^+[\nu]$ and $v^-[\nu]$, and clearly we have

$$v^\pm[\nu](z) = e^{\pm i\sqrt{k_0^2-\nu}z}, \quad \text{for } z \geq 0. \tag{2.35}$$

$v^+[\nu]$ and $v^-[\nu]$ are linearly independent for all $\nu \in \mathfrak{Z}$ since their initial values are linearly independent which implies the linear independence of the functions. Hence, any solution $v[\nu]$ to the differential equation (2.34a) can be written as

$$v[\nu] = c^+ v^+[\nu] + c^- v^-[\nu] \tag{2.36}$$

with some constants $c^\pm \in \mathbb{C}$ adapted to $v[\nu]$. For a resonance function $u_{\mathrm{s}\star}$ to the eigenvalue $\nu = \nu_\star$, which solves (2.34a) for $\nu_\star$, we must have $c^- = 0$ since the eigenfunctions $u_{\mathrm{s}\star}$ are required to be outgoing in the upper section, i.e. of the form $\mathfrak{r}_{\nu_\star} e^{i\sqrt{k_0^2-\nu_\star}z}$ for $z > 0$, which only fulfills $v^+[\nu_\star]$. Thus, the eigenspace has dimension 1. □

REMARK 2.19. *To give a little more insight to the look of resonance functions, we remark that there is another possibility for the definition of resonances, based on the considerations of the previous proof. As above, the initial value problem*

$$-\widetilde{v}''(z) - (k_0^2 n^2 - \nu)\widetilde{v}(z) = 0 \quad \textit{for } z \in \mathbb{R} \tag{2.37a}$$

$$\widetilde{v}(-a) = \gamma_3 \tag{2.37b}$$

$$\widetilde{v}'(-a) = \gamma_4 \tag{2.37c}$$

with initial values at $z = -a$ *has a unique solution for all initial values* $\gamma_3, \gamma_4 \in \mathbb{C}$ *and all* $\nu \in \mathfrak{Z}$*. In analogy to the argument in the proof we can find for the respective initial values linearly independent solutions* $\widetilde{v}^{\pm}[\nu]$ *with*

$$\widetilde{v}^{\pm}[\nu](z) = e^{\pm i\sqrt{k_0^2 n_{\mathrm{sub}}^2 - \nu}z}, \quad \text{for } z \leq -a. \tag{2.38}$$

They form a fundamental system for the differential equation (2.34a). Thus, we can express $v^+[\nu]$ *by*

$$v^+[\nu] = \widetilde{c}^+(\nu)\widetilde{v}^+[\nu] + \widetilde{c}^-(\nu)\widetilde{v}^-[\nu] \tag{2.39}$$

for all $\nu \in \mathfrak{Z}$ *with suitable constants* $\widetilde{c}^{\pm}(\nu)$*, depending on* ν *only. For a resonance* $\nu_\star$ *we must have* $\widetilde{c}^+(\nu_\star) = 0$*. This means that* $\nu \in \mathfrak{Z}$ *is a resonance if and only if* $v^+[\nu]$ *and* $\widetilde{v}^-[\nu]$ *are linearly dependent. Recall that* $v^+[\nu]$ *and* $\widetilde{v}^-[\nu]$ *are the unique (outgoing in the upper or respectively outgoing in the lower section) solutions to the two initial value problems (2.34) and (2.37) with* $\gamma_1 = 1, \gamma_2 = i\sqrt{k_0^2 - \nu}$ *and* $\gamma_3 = 1, \gamma_4 = -i\sqrt{k_0^2 n_{\mathrm{sub}}^2 - \nu}$*.*

Note that the obtained condition for ν *to be a resonance is also not easily checked since for general refractive indices* n *there are no simple analytic formulas for* $\widetilde{c}^{\pm}(\nu)$ *or respectively for* $\mathfrak{r}_\nu$ *and* $\mathfrak{t}_\nu$*. The latter is a problem of its own interest we will come back to in Chapter 3 (in particular, note the remarks in Section 3.1).*

3 Derivative of the total field and kinematic approximation

For a general refractive index n we are not able to compute an analytic solution of the scattering problem (2.8) and especially we have no chance to find the resonances analytically. This is why we aim to develop a numerical optimization algorithm to find better multilayer systems. The optimization methods we favor will all need the derivative of the objective function.
It seems reasonable to examine the dependence of the solution to the scattering problem on n in more detail first. Doing so, we will also provide a result which is of interest on its own. We present a mathematical derivation of the kinematic approximation, which is widely used, especially in the x-ray community. Actually, our derivation even yields more general approximations and error estimates. The scattering problem is written as an equivalent Lippmann-Schwinger integral equation, making use of a Green's function for some initial profile. Analyticity of the corresponding solution operator with respect to n at the initial profile is shown and approximation formulas are deduced from this. Padé approximations promise improved accuracy in the region of the critical angle where standard approximations usually fail. We finish this chapter with the application of the general results to step profiles and verify the achieved formulas numerically in an example. The approximation results are very good and especially the Padé approximations work very well.
Before we start our examinations, we give a brief introduction to what is meant by the kinematic approximation and a short overview of preliminary work on this topic.

3.1 Introduction to the kinematic approximation

Recall the definitions of reflection coefficient $\mathfrak{r}_\nu$ and transmission coefficient $\mathfrak{t}_\nu$ given in Chapter 2 (equations (2.6a) and (2.6b)). Often one is interested in the reflectivity and transmittance

$$\mathcal{R}_\nu := |\mathfrak{r}_\nu|^2 \quad \text{and} \quad \mathcal{T}_\nu := |\mathfrak{t}_\nu|^2, \tag{3.1}$$

which means the reflected or respectively transmitted intensity. For a general refractive index n it is often a lot of effort to solve many scattering problems for different values of ν to analyze the dependence of $\mathfrak{r}_\nu$ and $\mathfrak{t}_\nu$ on ν. Moreover, this does not deliver a functional connection between the reflectivity/transmittance and ν. Hence, one is interested in approximation formulas for reflectivity and transmittance (or respectively reflection and transmission coefficient) as functions of ν. The kinematic approximation is one of such approximations to the reflectivity and has broad applications, for example in the phase retrieval problem, an inverse problem, examined this way by *Hohage, Gieweckemeyer and Salditt* in [HGS08]. Especially the functional connection between ν and the reflectivity is important there.
First considerations on the mathematical derivation of approximation formulas for reflection and transmission coefficient in scattering from interfaces were

done in 1995 by *Caticha* [Cat95] and in the context of phase retrieval problems by *Klibanov, Sacks and Tikhonravov* [KST95]. In [Cat95] the differential equation is expressed as an integral equation with the help of a Green's function for a sharp interface between two layers where the transition point of the layers is left variable. Based on this, the approximation formulas are motivated and derived by approximating the functions under the integral. A drawback of this method is the arising ambiguity. Dependent on where self-consistency is imposed, one obtains different formulas for the transmission coefficient, but the formulas for the reflection coefficients are the same. We also rewrite our problem as an integral equation in our derivation, but we allow more general initial profiles adapted to the application and illustrate how the first order approximation formulas can be understood as a linearization of a suitable solution operator. In particular, we do not have any problems with ambiguity as it occurs in the self-consistency approach in [Cat95].
A more recent article by *Feranchuk et al.* [FFK$^+$03] from 2003 presents another ansatz for the computation of reflection and transmission coefficients which uses a different approximation for an integral equation similar to the one in [Cat95]. It has the advantage that it does not produce ambiguity and can be used for the computation of successive approximations by iterating the integral equation to improve accuracy. But in contrast to our results they are not able to prove convergence of the successive approximations for their approach mathematically ([FFK$^+$03, p.6]). In Section 3.2 we derive an error estimate on the accuracy of our approximations.

3.2 General formalism

In the scattering problem (2.8) studied in Chapter 2, we want to examine the dependence of the solution (total field) u on n^2 and its sensitivity. Hence, we are especially interested in the behavior for infinitely small perturbations in n^2 which leads to the Fréchet derivative of a suitable operator.

3.2.1 A Lippmann-Schwinger equation for steplike profiles

Before we start our analysis, we rewrite (2.8) such that both of the radiation conditions are imposed directly on the unknown function. We define

$$w := u - \theta u_i, \quad \text{with } \theta(z) := \begin{cases} 1 & \text{for } z > 0 \\ 0 & \text{for } z \leq 0 \end{cases}$$

and w shall then fulfill the following problem:

$$-w'' - (k_0^2 n^2 - \nu) w = 2\delta_0 u_\mathrm{i}' + \delta_0' u_\mathrm{i} \quad \text{for } z \in \mathbb{R} \tag{3.2a}$$

$$(w + \theta u_\mathrm{i}) \in H^2_\mathrm{loc}(\mathbb{R}) \text{ with } u_\mathrm{i}(z) = e^{-i\sqrt{k_0 - \nu} z} \tag{3.2b}$$

$$w \text{ is outgoing.} \tag{3.2c}$$

δ_0 denotes the delta distribution centered at 0 and (3.2a) has to be understood in the sense of distributions, i.e. by (3.2a) we mean

$$-\int_{-\infty}^{\infty} w v'' \,\mathrm{d}z - \int_{-\infty}^{\infty} (k_0^2 n^2 - \nu) w v \,\mathrm{d}z = 2u_\mathrm{i}'(0) v(0) - (u_\mathrm{i} v)'(0) \tag{3.3}$$

for all test functions $v \in C_0^\infty(\mathbb{R})$.
Equation (3.3) for w can be verified by the following computation. Let u a solution to (2.8), then twice partially integrating yields

$$\begin{aligned}
&- \int_{-\infty}^{\infty} wv''\,\mathrm{d}z - \int_{-\infty}^{\infty} (k_0^2 n^2 - \nu) wv\,\mathrm{d}z \\
&= - \int_{-\infty}^{\infty} (u - \theta u_\mathrm{i}) v''\,\mathrm{d}z - \int_{-\infty}^{\infty} (k_0^2 n^2 - \nu)(u - \theta u_\mathrm{i}) v\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} -uv''\,\mathrm{d}z - \int_{-\infty}^{\infty} (k_0^2 n^2 - \nu) uv\,\mathrm{d}z + \int_0^\infty u_\mathrm{i} v''\,\mathrm{d}z + \int_0^\infty (k_0^2 - \nu) u_\mathrm{i} v\,\mathrm{d}z \\
&= \int_{-\infty}^{\infty} (-u'' - (k_0^2 n^2 - \nu) u) v\,\mathrm{d}z + \int_0^\infty (u_\mathrm{i}'' + (k_0^2 - \nu) u_\mathrm{i}) v\,\mathrm{d}z \\
&\qquad\qquad\qquad\qquad\qquad\qquad - u_\mathrm{i}(0) v'(0) + u_\mathrm{i}'(0) v(0) \\
&= 2u_\mathrm{i}'(0) v(0) - (u_\mathrm{i} v)'(0) \qquad (3.4)
\end{aligned}$$

for all test functions $v \in C_0^\infty(\mathbb{R})$. Here we have used that $n \equiv 1$ for $z > 0$ and that $u_\mathrm{i}'' + (k_0^2 - \nu) u_\mathrm{i} = 0$ and $u'' + (k_0^2 n^2 - \nu) u = 0$ for all $z \in \mathbb{R}$.

Remark 3.1. *The following analysis does not rely on the special form of the term $(k_0^2 n^2 - \nu)$. It could be replaced by any function $m \in L^\infty(\mathbb{R})$. But since n can be easily substituted in a way to reach such a form, we do not change our notation.*

Let us fix $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \leq 0$ for the next considerations[16]. The solution to (3.2) is not known analytically for most n. Thus, we start with a well-studied admissible initial situation that we will perturb in the following to reach a representation of our problem as an integral equation. The admissible[17] initial refractive index is denoted by n_I and the corresponding solution to (3.2) by w_I.

Lemma and Definition 3.2. *Let $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \leq 0$. There exists a function*

$$\mathcal{G}_\nu \in L^1_\mathrm{loc}(\mathbb{R}^2 \setminus \{(x,x) : x \in \mathbb{R}\})$$

with the following properties:

1. $\left(-\frac{\partial^2}{\partial z^2} - (k_0^2 n_\mathrm{I}^2(z) - \nu)\right) \mathcal{G}_\nu(z,y) = \delta(z-y)$, *in the sense that*

   $$\int_{-\infty}^{\infty} -\mathcal{G}_\nu(z,y) v''(z) - (k_0^2 n_\mathrm{I}^2(z) - \nu) \mathcal{G}_\nu(z,y) v(z)\,\mathrm{d}z = v(y)$$

   for all test functions $v \in C_0^\infty(\mathbb{R})$ and all $y \in \mathbb{R}$.

2. $\mathcal{G}_\nu(\cdot, y)$ *is outgoing for all $y \in \mathbb{R}$.*

$\mathcal{G}_\nu$ *is called a Green's function for the equation* $-w_\mathrm{I}'' - (k_0^2 n_\mathrm{I}^2 - \nu) w_\mathrm{I} = 0$.
$\mathcal{G}_\nu(\cdot, y)$ *is locally twice weakly differentiable on $(-\infty, y)$ and (y, ∞) for all $y \in \mathbb{R}$. On all of $\mathbb{R}$ it is only locally once weakly differentiable. Moreover, $\mathcal{G}_\nu|_{[-a,0]^2}$ lies in $L^2([-a,0]^2)$.*

---

[16] $\mathfrak{Z} := \left\{\nu \in \mathbb{C} : \mathrm{Re}(\nu) \in [0, k_0^2) \wedge (k_0^2 n_\mathrm{sub}^2 - \nu) \notin i\mathbb{R}_0^-\right\}$ as defined in Section 2.3

[17] The assumptions on an admissible refractive index were given in the equations (2.1).

Remark 3.3. *In the definition above for* $y \notin [-a, 0]$ *"outgoing" has to be understood in the following sense:*
$y > 0$*:* $G(\cdot, y)$ *is outgoing in the lower section for* $z < -a$ *and outgoing in the upper section for* $z > y$.
$y < -a$*:* $G(\cdot, y)$ *is outgoing in the lower section for* $z < y$ *and outgoing in the upper section for* $z > 0$.

Proof (of Lemma 3.2). To show the existence of a Green's function, split

$$\mathcal{G}_\nu := \mathcal{G}_\nu^{\mathrm{S}} + \widetilde{\mathcal{G}_\nu} \tag{3.7}$$

with $\mathcal{G}_\nu^{\mathrm{S}}$ being a Green's function for the equation $-w'' - (k_0^2 n_{\mathrm{S}}^2 - \nu)w = 0$ with the single step

$$n_{\mathrm{S}}(z) := \begin{cases} 1 & \text{for } z > 0 \\ n_{\mathrm{sub}} & \text{for } z \leq 0. \end{cases} \tag{3.8}$$

We define

$$\kappa_1 := \sqrt{k_0^2 - \nu} \text{ and } \kappa_2 := \sqrt{k_0^2 n_{\mathrm{sub}}^2 - \nu}. \tag{3.9}$$

For fixed $\nu \in \mathfrak{Z}$ elementary computations show that a Green's function for n_{S} is given by[18]

$$\mathcal{G}_\nu^{\mathrm{S}}(z, y) = \begin{cases} \dfrac{i\,(\kappa_2 - \kappa_1)}{2\kappa_2\,(\kappa_1 + \kappa_2)} e^{-i\kappa_2 y} e^{-i\kappa_2 z} + \dfrac{i}{2\kappa_2} e^{i\kappa_2 |z-y|} & \text{for } y < 0,\ z < 0 \\ \dfrac{i}{\kappa_1 + \kappa_2} e^{-i\kappa_2 y} e^{i\kappa_1 z} & \text{for } y \leq 0,\ z \geq 0 \\ \dfrac{i}{\kappa_1 + \kappa_2} e^{i\kappa_1 y} e^{-i\kappa_2 z} & \text{for } y \geq 0,\ z < 0 \\ \dfrac{i\,(\kappa_1 - \kappa_2)}{2\kappa_1\,(\kappa_1 + \kappa_2)} e^{i\kappa_1 y} e^{i\kappa_1 z} + \dfrac{i}{2\kappa_1} e^{i\kappa_1 |z-y|} & \text{for } y > 0,\ z > 0. \end{cases}$$

$\widetilde{\mathcal{G}_\nu}$ must then fulfill the equation

$$\left(-\frac{\partial^2}{\partial z^2} - (k_0^2 n_{\mathrm{I}}^2(z) - \nu)\right) \widetilde{\mathcal{G}_\nu}(z, y) = k_0^2 (n_{\mathrm{I}}^2(z) - n_{\mathrm{S}}^2(z)) \mathcal{G}_\nu^{\mathrm{S}}(z, y), \ z \in \mathbb{R} \tag{3.10}$$

and has to be outgoing. $\widetilde{\mathcal{G}_\nu}(\cdot, y) \in H^2_{\mathrm{loc}}(\mathbb{R})$ since it is an outgoing solution of (3.10) which is an elliptic partial differential equation with a right hand side in $L^2(\mathbb{R})$ with support in $[-a, 0]$. This can be seen for all such right hand sides like in Proposition 2.7. Note here that $\operatorname{supp}(n_{\mathrm{I}}^2 - n_{\mathrm{S}}^2) \subset [-a, 0]$ as n_{I} and n_{S} are both admissible refractive indices, and they are therefore only different from each other in $[-a, 0]$ (cf. (2.1a)). Now we can argue completely analogously to Section 2.3 concerning existence, uniqueness and regularity. The assertion on the regularity of $\widetilde{\mathcal{G}_\nu}(\cdot, y)$ then follows from the continuity and differentiability properties of $\mathcal{G}_\nu^{\mathrm{S}}$. Moreover, as in Proposition 2.7, from existence and

---

[18] Note that the ν with $k_0^2 n_{\mathrm{sub}}^2 - \nu = 0$ is excluded (see (2.13) for the set of admissible ν).

uniqueness we have in particular by Riesz-Fredholm theory a continuous solution operator, which maps the right hand side on the corresponding solution. We obtain the estimate

$$\left\|\widetilde{\mathcal{G}_\nu}(\cdot,y)\right\|_{L^2([-a,0])} \le \left\|\widetilde{\mathcal{G}_\nu}(\cdot,y)\right\|_{H^1([-a,0])} \le c\left\|k_0^2(n_{\mathrm{I}}^2-n_{\mathrm{S}}^2)\mathcal{G}_\nu^{\mathrm{S}}(\cdot,y)\right\|_{L^2([-a,0])} \tag{3.11}$$

for all $y \in \mathbb{R}$ with some constant $c > 0$ which is independent of y as the differential operator is. Since $(n_{\mathrm{I}} - n_{\mathrm{S}})$ is bounded and $\mathcal{G}_\nu^{\mathrm{S}}$ continuous, equation (3.11) implies $\mathcal{G}_\nu, \frac{\partial \mathcal{G}_\nu}{\partial z} \in L^2([-a,0]^2)$ because

$$\sqrt{\int_{-a}^{0}\int_{-a}^{0} |\mathcal{G}_\nu(z,y)|^2 \,\mathrm{d}z\,\mathrm{d}y} = \sqrt{\int_{-a}^{0} \|\mathcal{G}_\nu(\cdot,y)\|^2_{L^2([-a,0])}\,\mathrm{d}y} < \infty \tag{3.12}$$

and analogously for $\frac{\partial \mathcal{G}_\nu}{\partial z}$. □

Remark 3.4. *Green's function can also be shown to be symmetric, but since we do not need this in the following we do not prove this property here.*

A central property of Green's function is given in the following lemma.

Lemma 3.5. *Let $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \le 0$ and $f \in L^2(\mathbb{R})$ with $\mathrm{supp}(f) \subset [-a,0]$. Then the function*

$$w_f(z) := \int_{-\infty}^{\infty} \mathcal{G}_\nu(z,y)f(y)\,\mathrm{d}y \tag{3.13}$$

is the unique solution to the problem

$$\begin{aligned} &\text{Find } w \in H^2_{\mathrm{loc}}(\mathbb{R}), \text{ such that} \\ &-w''(z) - (k_0^2 n_{\mathrm{I}}^2(z) - \nu)w(z) = f(z) \quad \text{for } z \in \mathbb{R} && \text{(3.14a)} \\ &w \text{ is outgoing.} && \text{(3.14b)} \end{aligned}$$

Proof. Uniqueness of solutions to problem (3.14) follows again as in Proposition 2.7. To show that w_f is a solution to (3.14), we compute for a test function $v \in C_0^\infty(\mathbb{R})$:

$$\begin{aligned} &-\int_{-\infty}^{\infty} w_f(z)v''(z) - (k_0^2 n_{\mathrm{I}}^2(z) - \nu)w_f(z)v(z)\,\mathrm{d}z \\ =&-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{G}_\nu(z,y)f(y)\,\mathrm{d}y\; v''(z)\,\mathrm{d}z \\ &\qquad -\int_{-\infty}^{\infty}(k_0^2 n_{\mathrm{I}}^2(z)-\nu)\int_{-\infty}^{\infty}\mathcal{G}_\nu(z,y)f(y)\,\mathrm{d}y\; v(z)\,\mathrm{d}z \\ =&-\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathcal{G}_\nu(z,y)v''(z)f(y)\,\mathrm{d}y\,\mathrm{d}z \qquad\qquad (3.15) \\ &\qquad -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(k_0^2 n_{\mathrm{I}}^2(z)-\nu)\mathcal{G}_\nu(z,y)v(z)f(y)\,\mathrm{d}y\,\mathrm{d}z \\ =&\int_{-\infty}^{\infty}\left(\int_{-\infty}^{\infty} -\mathcal{G}_\nu(z,y)v''(z) - (k_0^2 n_{\mathrm{I}}^2(z)-\nu)\mathcal{G}_\nu(z,y)v(z)\,\mathrm{d}z\right) f(y)\,\mathrm{d}y \\ =&\int_{-\infty}^{\infty} v(y)f(y)\,\mathrm{d}y. \end{aligned}$$

To find this we have used the fact that all integrals are finite because of the compact support of n_P and v, Fubini's theorem and the first property of $\mathcal{G}_\nu$. By the second property of $\mathcal{G}_\nu$ we deduce that w_f is in addition outgoing. □

By $\mathfrak{P}$ we denote the set of all admissible perturbations $n_\mathrm{P} \in L^\infty(\mathbb{R})$ with support in $[-a, 0]$ such that $\sqrt{n_\mathrm{I}^2 + n_\mathrm{P}}$ defines an admissible (see conditions (2.1)) refractive index. Define the operator

$$F_\nu : \mathfrak{P} \to H^2([-a,0]) \quad \text{with} \quad n_\mathrm{P} \mapsto w\left[\nu, \sqrt{n_\mathrm{I}^2 + n_\mathrm{P}}\right] \tag{3.16}$$

which maps admissible perturbations to the corresponding solution of (3.2) restricted to $[-a, 0]$. F_ν is well-defined for all $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \le 0$ by the unique solvability of the scattering problem, shown in Proposition 2.7. By Lemma 2.3 it suffices to know the values of a solution w to (3.2) at the points $-a$ and 0 to also know the behavior of the solution outside of $[-a, 0]$.
Our definition of F_ν is not very explicit so far, and it is not easy to see what its Fréchet derivative is. We will make use of the formulation of problem (3.2) as a Lippmann-Schwinger integral equation to express F_ν in a more accessible form. We start with a preparing Lemma.

LEMMA 3.6. *Let* $\nu \in \mathfrak{Z}$ *with* $\mathrm{Im}(\nu) \le 0$. *The integral operator*

$$S_\nu : L^2([-a,0]) \to H^2([-a,0]) \quad \textit{with} \quad (S_\nu f)(z) := \int_{-a}^{0} \mathcal{G}_\nu(z,y) f(y)\,\mathrm{d}y$$

is bounded.

PROOF. Let $f \in L^2([-a, 0])$. By Lemma 3.5 the operator S_ν maps f to a solution w_f of (3.14). There exists a unique solution to problem (3.14) for all $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \le 0$ (cf. Lemma 3.5). Again as in Proposition 2.7, the solution operator to problem (3.14) is bounded from $L^2([-a, 0])$ to $H^2([-a, 0])$. Thus, we deduce

$$\|S_\nu f\|_{H^2([-a,0])} = \|w_f\|_{H^2([-a,0])} \le c\,\|f\|_{L^2([-a,0])} \tag{3.18}$$

with some constant $c > 0$. □

Recall that w_I denotes the solution to (3.2) for the initial profile n_I, i.e. it is an outgoing solution to

$$-\,w_\mathrm{I}'' - (k_0^2 n_\mathrm{I}^2 - \nu) w_\mathrm{I} = 2\delta_0 u_\mathrm{i}' + \delta_0' u_\mathrm{i} \quad \text{for } z \in \mathbb{R} \tag{3.19}$$

in the sense of distributions (cf. equation (3.3)) with $(w_\mathrm{I} + \theta u_\mathrm{i}) \in H^2_\mathrm{loc}(\mathbb{R})$.

THEOREM 3.7. *Let* $n_\mathrm{P} \in \mathfrak{P}$ *and* $\nu \in \mathfrak{Z}$ *with* $\mathrm{Im}(\nu) \le 0$.

1. *Every solution* $w \in L^2([-a, 0])$ *to the Lippmann-Schwinger equation*

$$w(z) - \int_{-a}^{0} \mathcal{G}_\nu(z,y) k_0^2 n_\mathrm{P}(y) w(y)\,\mathrm{d}y = w_\mathrm{I}(z), \quad z \in [-a, 0] \tag{3.20}$$

can be continued by

$$w(z) = \int_{-a}^{0} \mathcal{G}_\nu(z,y) k_0^2 n_\mathrm{P}(y) w(y)\,\mathrm{d}y + w_\mathrm{I}(z), \quad z \in \mathbb{R}\backslash[-a,0] \tag{3.21}$$

to a solution of (3.2) and thus in particular $(w + \theta u_\mathrm{i}) \in H^2_\mathrm{loc}(\mathbb{R})$ *holds.*

2. *Vice versa, if* w *is a solution to (3.2), then the restriction* $w|_{[-a,0]}$ *is a solution to (3.20).*

Proof. 1. Let $w \in L^2([-a,0])$ be a solution to (3.20). By Lemma 3.6 we have $w \in H^2([-a,0])$ (note $n_\mathrm{P} \in L^\infty([-a,0]) \subset L^2([-a,0])$) and can continue it by (3.21) to a function on all of $\mathbb{R}$. The resulting

$$w(z) = \int_{-a}^{0} \mathcal{G}_\nu(z,y) k_0^2 n_\mathrm{P}(y) w(y)\,\mathrm{d}y + w_\mathrm{I}(z), \quad z \in \mathbb{R} \tag{3.22}$$

is outgoing in both sections which follows from the second property of $\mathcal{G}_\nu$ and the definition of w_I. Hence, we have $(w+\theta u_\mathrm{i}) \in H^2_\mathrm{loc}(\mathbb{R})$ by Lemma 3.5 together with $(w_\mathrm{I} + \theta u_\mathrm{i}) \in H^2_\mathrm{loc}(\mathbb{R})$. Again by Lemma 3.5

$$\widetilde{w} := \int_{-a}^{0} \mathcal{G}_\nu(z,y) k_0^2 n_\mathrm{P}(y) w(y)\,\mathrm{d}y \tag{3.23}$$

fulfills the equation $-\widetilde{w} - (k_0^2 n_\mathrm{I}^2 - \nu)\widetilde{w} = k_0^2 n_\mathrm{P} w$. Adding (3.19) yields (3.2a).
2. Let now w be a solution to (3.2) and w_2 be the right hand side of (3.21). Then, as seen above, w_2 can be continued to a solution of (3.2). Thus, $w - w_2$ solves (3.2) with zero right hand side and using Proposition 2.7 we deduce $w = w_2$. □

We define the multiplication operator

$$N(n_\mathrm{P}) : L^2([-a,0]) \to L^2([-a,0]) \quad \text{with} \quad f \mapsto k_0^2 n_\mathrm{P} f. \tag{3.24}$$

By E we denote the embedding operator from $H^2([-a,0])$ to $L^2([-a,0])$ and can express (3.20) in operator form by

$$(I - ES_\nu N(n_\mathrm{P}))w = w_\mathrm{I} \tag{3.25}$$

with $(I - ES_\nu N(n_\mathrm{P})) : L^2([-a,0]) \to L^2([-a,0])$. For an explicit representation of F_ν we have to invert the operator $(I - ES_\nu N(n_\mathrm{P}))$.

Lemma 3.8. *Let* $\nu \in \mathfrak{Z}$ *with* $\mathrm{Im}(\nu) \leq 0$. *For every* $n_\mathrm{P} \in \mathfrak{P}$ *the integral equation (3.25) has a unique solution.*

Proof. The operator $N(n_\mathrm{P})$ is continuous in $L^2([-a,0])$ as n_P is bounded. By Lemma 3.6 S_ν is also bounded and by the compactness of the embedding E (see e.g. [RR94, Thm. 6.98]), the operator $ES_\nu N(n_\mathrm{P})$ is therefore compact. Hence, $(I - ES_\nu N(n_\mathrm{P}))$ is a Fredholm operator of index 0 in $L^2([-a,0])$, and for the solvability of (3.25) (or respectively (3.20)) by Riesz theory it suffices to show that it has trivial nullspace (see e.g. [Kre89, Thm. 4.17]). Solving the integral equation (3.20) for $w_\mathrm{I} = 0$ is equivalent (in the sense of Lemma 3.7) to solving problem (3.2) for $u_\mathrm{i} = 0$, and we already know that we have a unique solution to this by Proposition 2.7. □

We conclude from Lemma 3.8 that $(I - ES_\nu N(n_{\mathrm{P}}))$ is boundedly invertible and can write[19]

$$F_\nu[n_{\mathrm{P}}] = (I - ES_\nu N(n_{\mathrm{P}}))^{-1} w_{\mathrm{I}}, \tag{3.26}$$

with the bounded operator $(I - ES_\nu N(\cdot))^{-1} : L^2([-a,0]) \to L^2([-a,0])$. Note that F_ν maps n_{P} on the solution to problem (3.2) restricted to $[-a,0]$ which can be expressed by $(I - ES_\nu N(n_{\mathrm{P}}))^{-1} w_{\mathrm{I}}$. Respresentation (3.26) also helps us to compute the Fréchet derivative of F_ν.

3.2.2 Fréchet derivatives and error estimates

Definition 3.9. *Let Y, Z be normed spaces, and let $\mathfrak{U}$ be an open subset of Y. A mapping $F : \mathfrak{U} \to Z$ is called Fréchet differentiable at $\phi \in \mathfrak{U}$ if there exists a bounded linear operator $F'[\phi] : Y \to Z$ such that*

$$\lim_{h\to 0} \frac{1}{\|h\|_Y} \|F[\phi + h] - F[\phi] - F'[\phi]h\|_Z = 0.$$

$F'[\phi]$ is called the Fréchet derivative of F at ϕ. If additionally $F' : \mathfrak{U} \to \mathfrak{L}(Y, Z)$ is continuous at ϕ, we say that F is continuously differentiable at ϕ.

Theorem 3.10. *For every $\nu \in \mathfrak{Z}$ with $\operatorname{Im}(\nu) \leq 0$ the operator F_ν is analytic at $n_{\mathrm{P}} = 0$, as it is given by*

$$F_\nu[n_{\mathrm{P}}] = \sum_{j=0}^{\infty} (ES_\nu N(n_{\mathrm{P}}))^j w_{\mathrm{I}}, \quad z \in [-a, 0] \tag{3.28}$$

in a neighborhood of $n_{\mathrm{P}} = 0$. Taking only the linear term we obtain for the first derivative of F at $n_{\mathrm{P}} = 0$

$$\left(F'_\nu[0]h\right)(z) = \int_{-a}^{0} k_0^2 \mathcal{G}_\nu(z, y) h(y) w_{\mathrm{I}}(y)\, \mathrm{d}y, \quad z \in [-a, 0]. \tag{3.29}$$

Proof. By Lemma 3.6 and since $\|Ev\|_{L^2} \leq \|v\|_{H^2}$ for all $v \in H^2([-a,0])$, we find

$$\begin{aligned} \|ES_\nu N(n_{\mathrm{P}}) f\|_{L^2} &\leq \|E\|_{H^2\to L^2} \|S_\nu\|_{L^2\to H^2} \|N(n_{\mathrm{P}})\|_{L^2\to L^2} \|f\|_{L^2} \\ &\leq c_S k_0^2 \|n_{\mathrm{P}}\|_\infty \|f\|_{L^2} \end{aligned} \tag{3.30}$$

for all $f \in L^2([-a,0])$, with some constant $c_S > 0$. Thus, if we choose the perturbation n_{P} small enough such that $\|ES_\nu N(n_{\mathrm{P}})\|_{L^2\to L^2} < 1$, the operator $(I - ES_\nu N(n_{\mathrm{P}}))$ is invertible with

$$(I - ES_\nu N(n_{\mathrm{P}}))^{-1} = \sum_{j=0}^{\infty} (ES_\nu N(n_{\mathrm{P}}))^j \tag{3.31}$$

by the Neumann series (for details see e.g. [Kre89, Section 2.3]). In this sense, the Taylor series of F_ν around 0 is given by

$$F_\nu[n_{\mathrm{P}}] = \sum_{j=0}^{\infty} (ES_\nu N(n_{\mathrm{P}}))^j w_{\mathrm{I}}, \quad z \in [-a, 0]. \tag{3.32}$$

□

---

[19] Again we use square brackets here because $F_\nu[n_{\mathrm{P}}]$ is itself a function for every n_{P}.

For small n_{P} we may hope that the linearization

$$F_\nu[n_{\mathrm{P}}] \approx F_\nu[0] + F_\nu'[0] n_{\mathrm{P}} \tag{3.33}$$

is already a good approximation. In other words, we have found a closed formula which gives us an approximation to the total field u for small perturbations of the initial situation, which means perturbations n_{P} with $\|n_{\mathrm{P}}\|_\infty$ small.

As explained in the introduction to this section, an approximation to the reflection coefficient $\mathfrak{r}_\nu$ is of particular interest. So pick a small $n_{\mathrm{P}} \in \mathfrak{P}$. To approximate $\mathfrak{r}_\nu$, by the definition of w and condition (3.2c) we only need an approximation to $w(0) - u_{\mathrm{i}}(0) = w(0) - 1$ which can be directly deduced from (3.33). This leads to an approximation to the reflectivity $\mathcal{R}_\nu$. Analogously one gets an approximation to the transmission coefficient and transmittance from the approximation to $w(-a)$. For a more detailed discussion of the procedure, we refer to Section 3.3 where we apply the results to step profiles.

Clearly, it also possible to use higher derivatives of F_ν. The recursion scheme

$$\phi_0 := w_{\mathrm{I}}, \quad \phi_{l+1} := w_{\mathrm{I}} + S_\nu N(n_{\mathrm{P}}) \phi_l \tag{3.34}$$

or explicitly

$$\phi_l = \sum_{j=0}^{l} (S_\nu N(n_{\mathrm{P}}))^j w_{\mathrm{I}}, \quad l = 0, 1, 2, \ldots \tag{3.35}$$

leads to higher order approximations to $F[n_{\mathrm{P}}]$. This scheme is known as successive approximations ([Kre89, Section 10.5]). $\phi_l(0) - 1$ always gives an approximation to $\mathfrak{r}_\nu$ and $\phi_l(-a)$ to $\mathfrak{t}_\nu$. The scheme (3.34) is easy to implement because we only need to discretize the integral operator in (3.29) via numerical integration.

Remark 3.11. *The derivation of formula (3.31) does not depend on the particular norm. It is also possible to consider the integral equation in other spaces. The results with the Neumann series hold true whenever we can show* $\|S_\nu N(n_{\mathrm{P}})\| < 1$ *in an appropriate operator norm. If we use for example the maximum norm in* $[-a, 0]$*, we get*

$$\|S_\nu N(n_{\mathrm{P}})\|_\infty \leq a k_0^2 \|n_{\mathrm{P}}\|_\infty \sup_{z \in [-a,0]} \sup_{y \in [-a,0]} |\mathcal{G}_\nu(z, y)| \,. \tag{3.36}$$

We can also deduce an error estimate from the Neumann series:

$$\left\| F_\nu[n_{\mathrm{P}}] - \sum_{j=0}^{l} (S_\nu N(n_{\mathrm{P}}))^j w_{\mathrm{I}} \right\| \leq \frac{\|S_\nu N(n_{\mathrm{P}})\|^{l+1}}{1 - \|S_\nu N(n_{\mathrm{P}})\|} \|w_{\mathrm{I}}\| \,. \tag{3.37}$$

In concrete applications this estimate can be made more explicit as we will see in Section 3.3.

3.2.3 Padé approximation

Since F_ν is not differentiable with respect to ν at $\kappa_2 = 0$ (if we neglect absorption, this is exactly the critical angle[20]), the approximation close to the critical angle will get worse. F_ν was only defined for $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \leq 0$ where $\kappa_2 = 0$ is excluded. The crossover[21] from ν-values with $\mathrm{Re}(k_0^2 n_{\mathrm{sub}}^2 - \nu) > 0$ to $\mathrm{Re}(k_0^2 n_{\mathrm{sub}}^2 - \nu) < 0$ is only continuous but not differentiable. Therefore, in the region of the critical angle we have to use many terms of the Taylor series of F_ν, to find good approximations to the field distribution and the reflectivity in this region. Thus, we are interested in possible improvements which might also improve the approximation everywhere. A Padé approximation is an approximation of a function by a rational function of given numerator and denominator degree. It often performs better than truncating the Taylor series and can still converge in regions where the Taylor series does not.
In our definition we follow *Baker* (see [BGM96, p.21]). Let $g : \mathbb{R} \to \mathbb{C}$ be a function that is analytic at $x = 0$, i.e. g can be respresented by a power series

$$g(x) = \sum_{j=0}^{\infty} a_j x^j \tag{3.38}$$

in a neighboorhood of $x = 0$ with coefficients $a_j \in \mathbb{C}$.

Definition 3.12. *If there exist polynomials p and q of respectively degree* K *and* L *such that*

$$\frac{p(x)}{q(x)} = g(x) + \mathcal{O}(x^{\mathrm{K+L+1}}) \tag{3.39}$$

and

$$q(0) = 1, \tag{3.40}$$

then we call

$$\mathcal{P}_{\mathrm{K,L}} = \frac{p(x)}{q(x)} \tag{3.41}$$

a Padé approximation of g.

An equivalent definition is given if we replace (3.39) by

$$p(x) - g(x)q(x) = \mathcal{O}(x^{\mathrm{K+L+1}}), \tag{3.42}$$

provided that (3.40) is retained. A linear system for the coefficients $p_k, q_l \in \mathbb{C}$ of the polynomials

$$p(x) := \sum_{k=0}^{\mathrm{K}} p_k x^k \quad \text{and} \quad q(x) := \sum_{l=0}^{\mathrm{L}} q_l x^l \tag{3.43}$$

---

[20] Recall that the critical angle was given by $\alpha_{\mathrm{cr,sub}} = \arccos(\mathrm{Re}(n_{\mathrm{sub}}))$.

[21] Note here in particular that for $\kappa_2 = 0$ we have the Laplace equation in the substrate and obviously for $\kappa_2 \to 0$ the Green's function $\mathcal{G}_\nu$ does not converge to the one needed at $\kappa_2 = 0$.

can be deduced from (3.42) by comparison of coefficients for the powers of x up to order $\mathrm{K}+\mathrm{L}$. The equations for $\mathrm{K}+1,\dots,\mathrm{K}+\mathrm{L}$ do not depend on p_k such that we are led to a system for the q_l which must be solved first. The coefficients p_k follow then from evaluating the first K equations. The Padé approximation $\mathcal{P}_{\mathrm{K},0}$ is obviously the truncated Taylor series of g up to order K. Definition 3.12 already implies that a Padé approximation need not exist in general. For given g, K, L it can happen that the linear system for the coefficients has no solution, but when it exists, it is unique. Because we do not want to go into too much detail here, we refer to the book of *Baker and Graves-Morris* [BGM96] for a complete introduction to Padé approximations including theory and numerical methods for their computation. We only state the formulas for the computation of the coefficients. Set $q_0 = 1$ and if $j < 0$ define $a_j = 0$. Then the other coefficients for q can be computed from

$$\begin{pmatrix} a_{\mathrm{K-L+1}} & a_{\mathrm{K-L+2}} & \cdots & a_{\mathrm{K}} \\ a_{\mathrm{K-L+2}} & a_{\mathrm{K-L+3}} & \cdots & a_{\mathrm{K+1}} \\ \cdots & \cdots & \cdots & \cdots \\ a_{\mathrm{K}} & a_{\mathrm{K+1}} & \cdots & a_{\mathrm{K+L-1}} \end{pmatrix} \begin{pmatrix} q_{\mathrm{L}} \\ q_{\mathrm{L-1}} \\ \cdots \\ q_1 \end{pmatrix} = - \begin{pmatrix} a_{\mathrm{K+1}} \\ a_{\mathrm{K+2}} \\ \cdots \\ a_{\mathrm{K+L}} \end{pmatrix}. \tag{3.44}$$

The coefficients for p can be computed afterwards from the equations

$$p_0 = a_0 \tag{3.45a}$$
$$p_1 = a_1 + a_0 q_1 \tag{3.45b}$$
$$p_2 = a_2 + a_1 q_1 + a_0 q_2 \tag{3.45c}$$
$$\cdots \tag{3.45d}$$
$$p_{\mathrm{K}} = a_{\mathrm{K}} + \sum_{l=1}^{\min(\mathrm{K},\mathrm{L})} a_{\mathrm{K}-l} q_l. \tag{3.45e}$$

Exemplarily we give the linear systems for the Padé approximations $\mathcal{P}_{2,1}$ and $\mathcal{P}_{2,2}$. We obtain for $\mathcal{P}_{2,1}$:

$$q_0 = 1 \tag{3.46a}$$
$$q_1 = -\frac{a_3}{a_2} \tag{3.46b}$$
$$p_0 = a_0 \tag{3.46c}$$
$$p_1 = a_1 + a_0 q_1 \tag{3.46d}$$
$$p_2 = a_2 + a_1 q_1, \tag{3.46e}$$

and for $\mathcal{P}_{2,2}$:

$$q_0 = 1 \tag{3.47a}$$
$$\begin{pmatrix} a_1 & a_2 \\ a_2 & a_3 \end{pmatrix} \begin{pmatrix} q_2 \\ q_1 \end{pmatrix} = - \begin{pmatrix} a_3 \\ a_4 \end{pmatrix} \tag{3.47b}$$
$$p_0 = a_0 \tag{3.47c}$$
$$p_1 = a_1 + a_0 q_1 \tag{3.47d}$$
$$p_2 = a_2 + a_1 q_1 + a_0 q_2. \tag{3.47e}$$

Padé approximations can be used for our problem as follows: In Theorem 3.10 we have shown that F_ν is analytic at $n_\mathrm{P} = 0$, and in particular for every $z \in [-a, 0]$ we have

$$(F_\nu[n_\mathrm{P}])(z) = \left(\sum_{j=0}^{\infty}(S_\nu N(n_\mathrm{P}))^j w_\mathrm{I}\right)(z) \tag{3.48}$$

for $\|n_\mathrm{P}\|$ small enough (recall here especially Remark 3.11). For an $h > 0$ small enough, $\widetilde{n_\mathrm{P}} := \frac{n_\mathrm{P}}{h}$ and a fixed $z \in [-a, 0]$, the mapping

$$[-h, h] \to \mathbb{C}, \qquad t \mapsto (F_\nu[t\widetilde{n_\mathrm{P}}])(z) \tag{3.49}$$

is analytic at $t = 0$, and we can compute Padé approximations of this function around zero and evaluate them at $t = h$. In particular, for $z = 0$ we get approximations of the reflection coefficient $\mathfrak{r}_\nu$ by evaluating Padé approximations of the function[22]

$$[-h, h] \to \mathbb{C}, \qquad t \mapsto \sum_{j=0}^{\infty} t^j \left((S_\nu N(\widetilde{n_\mathrm{P}}))^j w_\mathrm{I}\right)(0) - 1 \tag{3.50}$$

at $t = h$.

The numerical results in Section 3.3.2 show the success of Padé approximations in our application.

3.3 Application to step profiles/Kinematic approximation

Let us apply the general formalism from Section 3.2 to step profiles, by which we mean piecewise constant refractive profiles. We achieve an explicit error estimate for this case, derive the kinematic approximation and show the obtained results in a numerical example, in particular the Padé approximations.

3.3.1 Approximation formulas

Using the notation from Section 3.2 we start with a single step as initial refractive profile

$$n_\mathrm{I}(z) := \begin{cases} 1 & \text{for } z > 0 \\ n_\mathrm{sub} & \text{for } z \leq 0. \end{cases} \tag{3.51}$$

Recall the Green's function for this situation:

$$\mathcal{G}_\nu^\mathrm{S}(z, y) = \begin{cases} \dfrac{i(\kappa_2 - \kappa_1)}{2\kappa_2(\kappa_1 + \kappa_2)} e^{-i\kappa_2 y} e^{-i\kappa_2 z} + \dfrac{i}{2\kappa_2} e^{i\kappa_2 |z-y|} & \text{for } y < 0,\ z < 0 \\ \dfrac{i}{\kappa_1 + \kappa_2} e^{-i\kappa_2 y} e^{i\kappa_1 z} & \text{for } y \leq 0,\ z \geq 0 \\ \dfrac{i}{\kappa_1 + \kappa_2} e^{i\kappa_1 y} e^{-i\kappa_2 z} & \text{for } y \geq 0,\ z < 0 \\ \dfrac{i(\kappa_1 - \kappa_2)}{2\kappa_1(\kappa_1 + \kappa_2)} e^{i\kappa_1 y} e^{i\kappa_1 z} + \dfrac{i}{2\kappa_1} e^{i\kappa_1 |z-y|} & \text{for } y > 0,\ z > 0, \end{cases}$$

---

[22] Note here again that $w[n_\mathrm{P}] = F[n_\mathrm{P}]$ represents the total field for $z < 0$ and $u_\mathrm{i}(0) = 1$.

with $\kappa_1 = \sqrt{k_0^2 - \nu}$ and $\kappa_2 = \sqrt{k_0^2 n_{\text{sub}}^2 - \nu}$. Moreover, we may compute the solution

$$w_{\text{I}}(z) := \begin{cases} \dfrac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2} e^{i\kappa_1 z} & \text{for } z > 0 \\ \dfrac{2\kappa_1}{\kappa_1 + \kappa_2} e^{-i\kappa_2 z} & \text{for } z \leq 0. \end{cases}$$

to (3.2) for the initial profile, from which follow the reflectivity and transmittance for the single step which are often called Fresnel reflectivity and Fresnel transmittance (of the substrate):

$$\mathcal{R}_{\text{F}} := \left| \frac{\kappa_1 - \kappa_2}{\kappa_1 + \kappa_2} \right|^2 \quad \text{and} \quad \mathcal{T}_{\text{F}} := \left| \frac{2\kappa_1}{\kappa_1 + \kappa_2} \right|^2. \tag{3.52}$$

We prove the following corollary of Theorem 3.10 for the case of step profiles:

Corollary 3.13. *Let $\nu \in \mathfrak{Z}$ and $\operatorname{Im}(\nu) \leq 0$. For an admissible refractive profile $n = \sqrt{n_{\text{I}}^2 + n_{\text{P}}}$ with n_{I} as defined in (3.51) and $n_{\text{P}} \in \mathfrak{P}$ we find the approximation*

$$\widetilde{u[\nu, n]}(z) = \frac{2\kappa_1}{\kappa_1 + \kappa_2} \left(e^{-i\kappa_2 z} + k_0^2 \int_{-a}^{0} \mathcal{G}_\nu^{\text{S}}(z, y) n_{\text{P}}(y) e^{-i\kappa_2 y} \, \mathrm{d}y \right) \tag{3.53}$$

of the total field distribution $u[\nu, n]$ to problem (2.8) restricted to $z \in [-a, 0]$, and if n_{P} is differentiable the approximation

$$\widetilde{\mathcal{R}_\nu} = \mathcal{R}_{\text{F}} \left| 1 + \frac{\kappa_1}{\kappa_2 (1 - n_{\text{sub}}^2)} \int_{-a}^{0} (n_{\text{P}})'(y) e^{-2i\kappa_2 y} \, \mathrm{d}y \right|^2 \tag{3.54}$$

to the reflectivity $\mathcal{R}_\nu$.
With the additional approximations

$$\kappa_1 \approx \kappa_2 \quad \text{and} \quad e^{-2i\kappa_2 y} \approx e^{-2i\kappa_1 y}, \tag{3.55}$$

we get the kinematic approximation

$$\widetilde{\widetilde{\mathcal{R}_\nu}} = \mathcal{R}_{\text{F}} \left| \frac{1}{\rho_\infty} \int_{-\infty}^{\infty} (n^2)'(y) e^{-2i\kappa_1 y} \, \mathrm{d}y \right|^2, \tag{3.56}$$

where $\rho_\infty := 1 - n_{\text{sub}}^2$ is the contrast.
If $\tau := k_0^2 a \eta \left\| n_{\text{P}} \right\|_\infty < 1$ with

$$\eta := \sup_{z \in [-a,0]} \sup_{y \in [-a,0]} \left(\left| \frac{\kappa_2 - \kappa_1}{2\kappa_2(\kappa_1 + \kappa_2)} e^{\operatorname{Im}(\kappa_2)(z+y)} \right| + \left| \frac{1}{2\kappa_2} e^{-\operatorname{Im}(\kappa_2)|z-y|} \right| \right),$$

the following error estimate for approximation (3.53) holds:

$$\left\| u[\nu, n] - \widetilde{u[\nu, n]} \right\|_\infty \leq \frac{\tau^2}{1 - \tau} \sqrt{\mathcal{T}_{\text{F}}}. \tag{3.57}$$

Proof. We apply the results from Section 3.2, in particular the equations (3.29) and (3.33), to find formula (3.53). Evaluating this formula at $z = 0$ and using the explicit formula for $\mathcal{G}_\nu^{\mathrm{S}}$ we deduce for the ν-dependent reflectivity the approximation

$$\begin{aligned}\widetilde{\mathcal{R}_\nu} &= \left|\frac{2\kappa_1}{\kappa_1+\kappa_2} + k_0^2\int_{-a}^{0}\frac{2i\kappa_1}{(\kappa_1+\kappa_2)^2} n_{\mathrm{P}}(y)e^{-2i\kappa_2 y}\,\mathrm{d}y - u_{\mathrm{i}}(0)\right|^2\\ &= \left|\frac{\kappa_1-\kappa_2}{\kappa_1+\kappa_2} - k_0^2\int_{-a}^{0}\frac{\kappa_1}{-\kappa_2(\kappa_1+\kappa_2)^2}(n_{\mathrm{P}})'(y)e^{-2i\kappa_2 y}\,\mathrm{d}y\right|^2\\ &= \left|\frac{\kappa_1-\kappa_2}{\kappa_1+\kappa_2}\right|^2\cdot\left|1+\frac{k_0^2\kappa_1}{\kappa_2(\kappa_1+\kappa_2)(\kappa_1-\kappa_2)}\int_{-a}^{0}(n_{\mathrm{P}})'(y)e^{-2i\kappa_2 y}\,\mathrm{d}y\right|^2\\ &= \left|\frac{\kappa_1-\kappa_2}{\kappa_1+\kappa_2}\right|^2\cdot\left|1+\frac{\kappa_1}{\kappa_2(1-n_{\mathrm{sub}}^2)}\int_{-a}^{0}(n_{\mathrm{P}})'(y)e^{-2i\kappa_2 y}\,\mathrm{d}y\right|^2\end{aligned}\tag{3.58}$$

by partial integration assuming n_{P} to be differentiable, the fact that $\mathrm{supp}(n_{\mathrm{P}}) \subset [-a, 0]$ compact and since

$$\frac{k_0^2\kappa_1}{\kappa_2(\kappa_1^2-\kappa_2^2)} = \frac{k_0^2\kappa_1}{\kappa_2(k_0^2-\nu-k_0^2n_{\mathrm{sub}}^2+\nu)} = \frac{\kappa_1}{\kappa_2(1-n_{\mathrm{sub}}^2)}.$$

With the approximations (3.55), again the fact that n_{P} has compact support in $[-a, 0]$ and observing

$$\frac{1}{1-n_{\mathrm{sub}}^2}\int_{-\infty}^{\infty}(n_{\mathrm{I}}^2)'(y)e^{-2i\kappa_2 y}\,\mathrm{d}y = 1, \tag{3.60}$$

equation (3.58) implies the kinematic approximation as $(n^2)' = (n_{\mathrm{I}}^2)' + n_{\mathrm{P}}'$. By $(n_{\mathrm{I}}^2)'$ we mean formally $(n_{\mathrm{I}}^2)' = (1 - n_{\mathrm{sub}})\delta_0$.

From (3.37), (3.36) and the fact that $\mathrm{Im}(\kappa_2) \geq 0$ (note $\nu \in \mathfrak{Z}$ with $\mathrm{Im}(\nu) \leq 0$, $\mathrm{Re}(n_{\mathrm{sub}}) > 0$ and $\mathrm{Im}(n_{\mathrm{sub}}) \geq 0$), we get in the maximum norm the following error estimate

$$\begin{aligned}&\|F_\nu[n_{\mathrm{P}}] - F_\nu[0] - F_\nu'[0]n_{\mathrm{P}}\|_\infty \leq \frac{\tau^2}{1-\tau}\|w_{\mathrm{I}}\|_\infty \quad \text{if } \tau := k_0^2 a\eta\,\|n_{\mathrm{P}}\|_\infty < 1,\\ &\quad\text{where } \eta = \sup_{z\in[-a,0]}\sup_{y\in[-a,0]}\left(\left|\frac{\kappa_2-\kappa_1}{2\kappa_2(\kappa_1+\kappa_2)}e^{\mathrm{Im}(\kappa_2)(z+y)}\right| + \left|\frac{1}{2\kappa_2}e^{-\mathrm{Im}(\kappa_2)|z-y|}\right|\right)\\ &\quad\text{and } \|w_{\mathrm{I}}\|_\infty = \left|\frac{2\kappa_1}{\kappa_1+\kappa_2}\sup_{z\in[-a,0]} e^{\mathrm{Im}(\kappa_2)z}\right| = \sqrt{\mathcal{T}_{\mathrm{F}}}.\end{aligned}\tag{3.61}$$

which implies (3.57). □

Remark 3.14. *Let us make a few remarks to the presented result.*

1. *The approximations (3.55) are valid for small contrasts in the materials or big angles of incidence. Note that by (3.54) we have proven an approximation to the reflectivity where these assumptions are not necessary.*

2. *Sometimes one uses the parameter* $q := \frac{4\pi}{\lambda}\sin\alpha$ *instead of* ν. *There is the following relation between the two:*

$$q = \frac{4\pi}{\lambda}\sin\alpha = 2k_0\sqrt{1-\cos^2\alpha} = 2\sqrt{k_0^2 - k_0^2\cos^2\alpha} = 2\sqrt{k_0^2-\nu} = 2\kappa_1.$$

3. *Equation (3.57) is an estimate on the maximal error over all points of* $[-a,0]$. *In particular, we can estimate the error in the approximation to the reflection coefficient by*

$$\left|\widetilde{\mathcal{R}_\nu} - \mathcal{R}_\nu\right| \le \left(2\sqrt{\widetilde{\mathcal{R}_\nu}} + \frac{\tau^2}{1-\tau}\|w_{\mathrm{I}}\|_\infty\right)\frac{\tau^2}{1-\tau}\|w_{\mathrm{I}}\|_\infty . \tag{3.63}$$

4. *Higher order approximations of the total field and the reflectivity can be obtained by iterating the integral operator*

$$(S_\nu N(n_{\mathrm{P}}))w = k_0^2\int_{-a}^{0} \mathcal{G}_\nu^{\mathrm{S}}(\cdot,y)\,n_{\mathrm{P}}(y)\,w(y)\,\mathrm{d}y \tag{3.64}$$

   as explained in Section 3.2.2.

3.3.2 Example

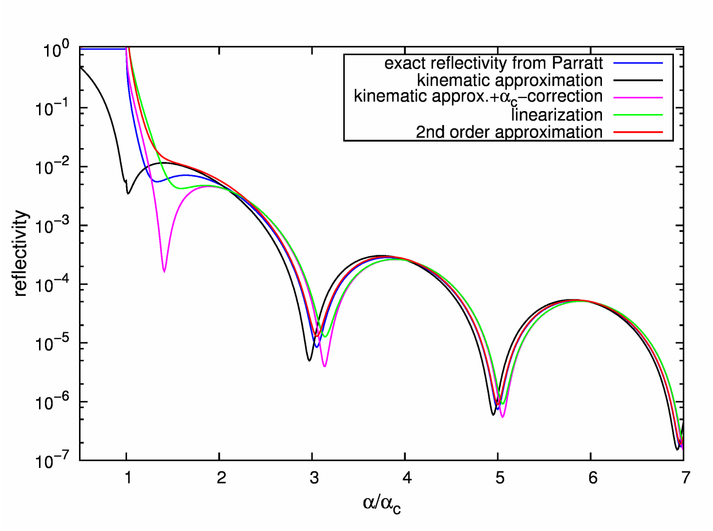


Figure 3.1: *reflectivity and several approximations to the reflectivity for a polymer film* ($\delta_{\mathrm{ps}} = 3.5\cdot10^{-6}$) *of* 100Å *on silicon substrate* ($\delta_{\mathrm{Si}} = 7.56\cdot10^{-6}$), *plotted against the angles of incidence as multiples of the critical angle* α_{c}; $\lambda = 1.54$ Å; *absorption neglected*

As a test example for the formulas from Corollary 3.13, higher order and Padé approximations, we use an example from [Tol99, p.76]. The test system is a polymer film ($\delta_{\mathrm{ps}} = 3.5\cdot10^{-6}$) of 100 Å on silicon substrate ($\delta_{\mathrm{Si}} = 7.56\cdot10^{-6}$)

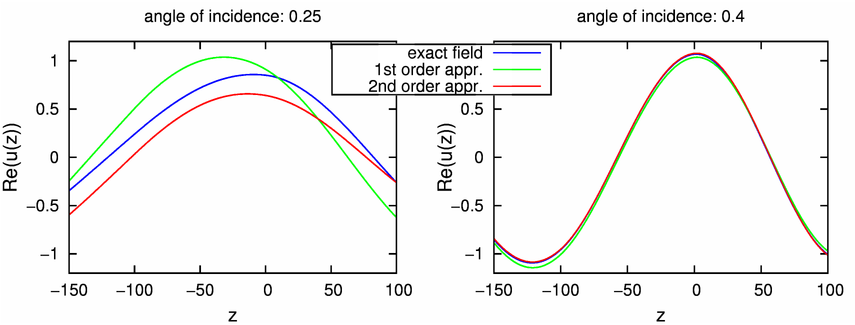


Figure 3.2: *approximation of total field distribution for a polymer film on silicon substrate, comparision of exact solution from Paratt algorithm to 1st and 2nd order Taylor approximation close to the critical angle* $\alpha_c \approx 0.22°$

at a wavelength of $\lambda = 1.54$ Å, sharp interfaces are assumed and absorption effects are neglected, i.e. an example of a three layer system with piecewise constant refractive index.

Figure 3.1 shows the exact reflectivity (computed by the Parratt algorithm) for different angles of incidence α compared to the kinematic approximation (with and without critical angle correction) and our approximation (3.54). The critical angle correction is a modification of the standard kinematic approximation to slightly improve the approximation close to the critical angle[23]. The scale for the angles of incidence indicates multiples of the critical angle α_c. In our test example the critical angle between air and substrate is about 0.22°. Below the critical angle the reflectivity is known to be 1 since we neglected absorption effects, but for angles slightly above it the approximations are bad. One way to improve the approximation globally is the iteration scheme given in Section 3.2. The black line in Figure 3.1 indicates the second order approximation computed from iterating the integral operator (3.64) twice.

We also want to compare our approximation of the total field distribution to the exact fields in the same example. The exact solutions are again computed by the Parratt algorithm. Figure 3.2 shows the exact real part of the total field compared with its first and second order Taylor approximation at angles of incidence of 0.25° and 0.4°.

Both figures illustrate that the second order approximation fits the exact solution already a lot better than first order approximations in this example, but in the region of the critical angle it is still quite bad. For this reason and to test their quality in general, we also computed Padé approximations of the reflection coefficient in the test example and plotted the resulting curves. Figures 3.3 and 3.4 show different details of the reflectivity curve shown above. Note that the dip in the fifth order approximation in Figure 3.4 arises from the

---

[23] For the critical angle correction in equation (3.56) the term $-2i\kappa_1 (= -2ik_0 \sin\alpha)$ in the exponential function under the integral is replaced by $-2ik_0 \sin\sqrt{\alpha^2 - \alpha_c^2}$, see [Tol99, p.76/77].

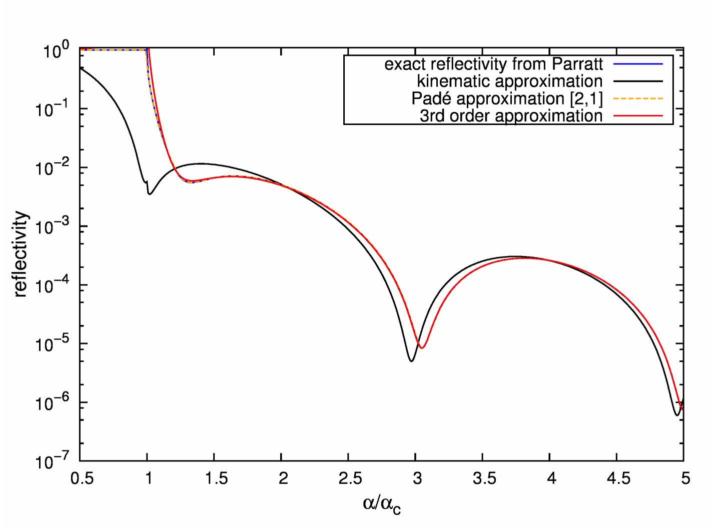


Figure 3.3: *kinematic and 3rd order approximation compared to Padé $\mathcal{P}_{2,1}$ for the example from Figure 3.1*

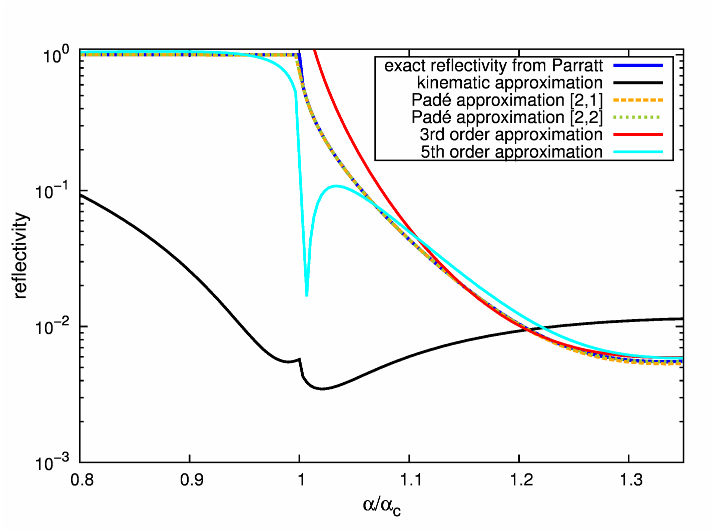


Figure 3.4: *kinematic and higher order approximations compared to Padé $\mathcal{P}_{2,1}$ and $\mathcal{P}_{2,2}$ close to the critical angle, a detail of Figure 3.3*

fact that F_ν is not differentiable when α reaches the critical angle (recall that absorption effects are neglected in this example). This is also the reason for the jags at the critical angle in the other curves and figures. One can clearly see that the exact reflectivity curve is only continuous, but not differentiable with respect to α (or respectively ν) at the critical angle. The approximation to the reflectivity resulting from Padé approximations is an significant improvement. In particular close to the critical angle the truncated Taylor series converges quite slowly while already the Padé approximation $\mathcal{P}_{2,1}$ fits the exact reflectivity almost perfect. Even the fifth order Taylor approximation is still relatively far away from the exact solution in this region. In contrast, the Padé approximations $\mathcal{P}_{2,1}$ and $\mathcal{P}_{2,2}$ are hard to distinguish from the exact reflectivity curve.

3.4 Conclusion

In Section 3.2 we have derived a general approach to linearize the solution operator for problems of the form (2.4) whenever a Green's function for some initial situation is known. The approach is therefore not limited to step profiles, shown exemplarily in Section 3.3. One can also start from other initial profiles. Moreover, we have seen how one can come to higher order approximations to the field distribution and in particular to the reflection coefficient/reflectivity in an easy way by iterating a suitable integral operator. Especially in x-ray physics, the kinematic approximation is a commonly used approximation to the reflectivity. We have presented a rigorous derivation, applying our general approach to step profiles in Section 3.3 and using two further approximations (3.55). But even in the case of step profiles, compared to the kinematic approximation, our derived first order approximation formula (3.54) has the advantage that it is neither restricted to small contrasts in the initial materials nor do we have to neglect absorption effects. Higher order approximations improve the approximation of the field distribution and the reflectivity significantly, but very close to the critical angle they are in general still quite bad. We have shown that Padé approximations can be used to improve the approximation in this region considerably. Note here in particular that Padé approximations can be directly computed from the derivatives which we need anyways in the truncated Taylor series.

Although we have deduced a formula for the derivative of the solution operator, it turns out to be very inconvenient to compute the derivative of the solution operator to the scattering problem (2.8) by formula (3.29). To evaluate it at an arbitrary n we need the Green's function corresponding to this n.

4 Analysis of the optimization problem

Recall the Optimization Problem 2.16 formulated in Chapter 2 using resonances and in particular the used objective function which is the L^2-norm of the solution at the resonance frequency in dependence on n. As already mentioned we aim for a method using derivatives. So far, we have only commented on the derivative of the solution operator, but because of the formulation in Optimization Problem 2.16 we also need the derivatives of the resonances. Hence, in this chapter we analyze the differentiability of simple, isolated eigenvalues and corresponding eigenvectors of a general operator-valued function with respect to a parameter n in a Banach space and apply the results to a generalized eigenvalue problem as it arises by rewriting the scattering problem in operator form using the Hardy space formulation (cf. Lemma 2.9). The differentiability of the eigenvalues does not hold in general even if the operator depends continuously differentiably on n. But for isolated and simple eigenvalues we can show that they depend differentiably on n when the operator does. We derive expressions for their derivatives and for the corresponding eigenvectors. Since the latter are non-unique, we will show the differentiability assuming a special scaling.

In particular, the results apply to the discrete version of our problem which we obtain by a finite elements discretization combined with Hardy space infinite elements assuming that n can be encoded in a vector using some discretization for it. We explain why existing results on the optimization of eigenvalues of a parameter-dependent matrix cannot be applied and state formulas for the derivative of our objective function.

This chapter is closed by the computation of the derivative of the discrete operator with respect to the positions of layer changes if a piecewise constant refractive index is assumed.

4.1 Derivatives of eigenvalues

4.1.1 Introduction to derivatives of eigenvalues

Although our analysis is not restricted to the finite-dimensional case we start this section with some comments on such problems. All the main problems for isolated eigenvalues already arise in this setting. As already pointed out, we aim for the derivative of eigenvalues and eigenvectors of an operator(matrix)-valued function. Consider the matrix-valued function $r \mapsto \mathbf{A}(r)$ depending on the single parameter $r \in \mathbb{R}$ in a subdomain of $\mathbb{R}$. The eigenvalues of $\mathbf{A}(r)$ are the roots of the characteristic equation

$$\det(\mathbf{A}(r) - \zeta) = 0. \tag{4.1}$$

Let each entry of $\mathbf{A}(r)$ and hence $\mathbf{A}(r)$ depend differentiably on r. We would like to know whether the roots of the characteristic equation also depend differentiably on r. This question is far from being trivial because the dependence of the roots of a polynomial on its coefficients is highly nonlinear, and it is well-known that for a polynomial degree of greater than five there are no explicit formulas for the roots.

Another important issue is the definition of functions of r representing the eigenvalues of $\mathbf{A}(r)$ since the number of eigenvalues may change with r in a very complicated manner when $\mathbf{A}(r)$ is assumed to depend only differentiably (not analytically) on r. Consider the following two small examples[24].

$$\mathbf{A}_{\mathrm{I}}(r) = \begin{pmatrix} 1 & r \\ 0 & 0 \end{pmatrix}, \quad \mathbf{A}_{\mathrm{II}}(r) = \begin{pmatrix} r & 1 \\ 0 & 0 \end{pmatrix}. \tag{4.2}$$

The matrix $\mathbf{A}_{\mathrm{I}}$ has the eigenvalues 0 and 1 for all $r \in \mathbb{R}$. We can easily distinguish between the two eigenvalues and define functions $\mu_1(r) = 0$ and $\mu_2(r) = 1$, and they are obviously both differentiable. In contrast, the matrix $\mathbf{A}_{\mathrm{II}}$ has the eigenvalues 0 and r for all $r \in \mathbb{R}$, and for $r = 0$ we have the double eigenvalue 0. The eigenvalues are continuous everywhere independent of how we connect the branches at $r = 0$, but the differentiability at $r = 0$ might be destroyed if we sort them in ascending order for example. Everywhere else they are obviously differentiable. These examples already give some indication in which way the assumption of simple eigenvalues can help us.

For the case of a matrix-valued function there exists a large literature on derivatives of eigenvalues and optimality conditions for certain objective functions involving eigenvalues. We will give a short overview of the existing work in Section 4.4 and explain in detail why the results do not apply to our problem. Most of them are restricted to the case of symmetric/Hermitian matrices or linear dependence on the parameters. These assumptions are not satisfied for our problem. Moreover, we also need derivatives of eigenvectors, which is not possible in a general way as they are not unique.

Hence, we do our analysis from a more general point of view and apply the results to the discrete problem as a special case of the general theory. If we use the Hardy space formulation for the continuous problem, the results of the following sections also apply to it. We will comment on this later.

4.1.2 Basic theory

We consider an operator-valued function

$$W : \mathfrak{N} \subset Y \to \mathfrak{L}(X), \quad n \mapsto W(n), \tag{4.3}$$

on an open subset $\mathfrak{N}$ of a Banach space Y with values in the space of bounded, linear operators $\mathfrak{L}(X)$, where X is a Hilbert space. We want to find formulas for the derivatives with respect to n of simple, isolated eigenvalues and corresponding eigenvectors of the eigenvalue problem: Find $\nu_\star$ such that there exists a nonzero $v_\star \in X$ with

$$W(n)v_\star = \nu_\star v_\star. \tag{4.4}$$

$\nu_\star$ is called an eigenvalue if $\ker(W(n) - \nu_\star)$ is non-trivial. The dimension of this null space is called the geometric multiplicity of the eigenvalue $\nu_\star$ which we assume to be 1 (simple eigenvalue). We suppose that the reader is familiar with

---

[24] These examples can also be found in [Kat95, p.64].

eigenvalues and spectra of operators in Banach spaces and therefore repeat only a few definitions and results needed here. For a complete introduction, we refer to [Kat95, III.§6] and standard literature on functional analysis and spectral theory.

Our following analysis relies on results in the monograph [Kat95] by *Kato*. We will often refer to results stated there but we want to present the results more suited to our assumptions and application. Moreover, such a general result for Fréchet derivatives is not explicitly stated there, and in particular we will need more information on specially scaled eigenvectors.

Remark 4.1. *It suffices to examine the situation for a general operator-valued function W around 0 to deduce results for all admissible n from this result since we can always transform the problem by introducing a shifted operator: Consider the point $n = \check{n}$, where we want to compute the derivative, as an initial refractive index. Perturb this situation by an admissible $\hat{n}$ ($\hat{n}$ such that $\check{n} + \hat{n}$ defines an admissible refractive index) and examine the operator-valued function $W_{\check{n}}$ with*

$$W_{\check{n}}(\hat{n}) := W(\check{n} + \hat{n}) \tag{4.5}$$

around 0. $W_{\check{n}}(0) = W(\check{n})$ represents the operator for the initial refractive index. We will later assume that W is continuously differentiable for all considered n and therefore in particular $W_{\check{n}}$ around $\hat{n} = 0$ for each admissible initial $\check{n}$.

For $W(n)$ define the resolvent[25]

$$\widetilde{R}(\nu, n) := (W(n) - \nu)^{-1} \in \mathfrak{L}(X) \tag{4.6}$$

for all $\nu \in \mathbb{C}$ for which $(W(n) - \nu)$ is boundedly invertible. The set of these ν is called the resolvent set $\Theta\left(W(n)\right)$ and by $\Sigma\left(W(n)\right)$ we denote the spectrum which is the complementary set of $\Theta\left(W(n)\right)$. Since $W(n) \in \mathfrak{L}(X)$ for all n, neither $\Theta(W(n))$ nor $\Sigma(W(n))$ is empty for any n (see [Kat95, III.§6.2]), but in contrast to finite-dimensional operators it may well happen that the spectrum is an uncountable set. Observe that $\widetilde{R}(\nu, n)$, $\Theta\left(W(n)\right)$ and $\Sigma\left(W(n)\right)$ define again functions of n.

Motivated by Theorem 2.12 we are interested in the sensitivity analysis of isolated eigenvalues of $W(n)$. An isolated eigenvalue is an eigenvalue in the spectrum which can be separated from the rest of the spectrum by a rectifiable, simple closed curve enclosing this point and no other points of the spectrum. We have not yet explained how to define functions of eigenvalues. The following lemma and the sketched proof clarify the situation.

Lemma 4.2. *Let $W : \mathfrak{N} \to \mathfrak{L}(X)$ with $n \mapsto W(n)$ be continuous in n and let $\nu_\star(0)$ be an isolated and simple eigenvalue of $W(0)$. Then, in a neighboorhood of $n = 0$ we can define a function $n \mapsto \nu_\star(n)$ of simple, isolated eigenvalues of $W(n)$ corresponding to the eigenprojection*

$$P_\star(n) := -\frac{1}{2\pi i} \int_{\Gamma_\star} \widetilde{R}(\nu, n)\, \mathrm{d}\nu, \tag{4.7}$$

---

[25] Note that the resolvent $\widetilde{R}(\nu, n)$ is slightly different from the one we used in Chapter 2 where the resolvent corresponded to a nonlinear eigenvalue problem.

with a rectifiable, simple closed curve $\Gamma_\star$ *in the resolvent set of* $W(0)$ *enclosing* $\nu_\star(0)$*, but no other values in the spectrum of* $W(0)$*.*
Moreover, the resolvent $\widetilde{R}(\nu, n)$ *has the following Laurent expansion around* $\nu = \nu_\star(n)$ *then:*

$$\widetilde{R}(\nu, n) = -(\nu - \nu_\star(n))^{-1} P_\star(n) + \sum_{m=0}^{\infty} (\nu - \nu_\star(n))^m S_\star(n)^{m+1}, \tag{4.8}$$

where the reduced resolvent $S_\star(n)$ *is given by*

$$S_\star(n) := \frac{1}{2\pi i} \int_{\Gamma_\star} (\nu - \nu_\star(n))^{-1} \widetilde{R}(\nu, n) \, \mathrm{d}\nu. \tag{4.9}$$

Furthermore, $S_\star(n), P_\star(n) \in \mathfrak{L}(X)$*, the relations*

$$W(n) P_\star(n) = \nu_\star(n) P_\star(n), \tag{4.10a}$$

$$S_\star(n) P_\star(n) = P_\star(n) S_\star(n) = 0 \tag{4.10b}$$

and

$$(W(n) - \nu_\star(n)) S_\star(n) = I - P_\star(n) \tag{4.10c}$$

hold true, and the mapping $n \mapsto P_\star(n)$ *is continuous.*

Remark 4.3. *Throughout this chapter we will use integrals of continuous vector- and operator-valued functions. They can be defined as for numerical functions. We will make use of them as well as of formulas for integration and differentiation without further comments, provided that they hold true in this general context. Similarly we proceed with results from complex analysis which we will need in the following. For a more detailed introduction to this topic, see [Kat95, I.§1.7 and p.152].*

Proof (of Lemma 4.2). We only briefly sketch the proof. The details can be found in [Kat95, III.§6.4/§6.5, IV.§3.4].
Let $\Gamma_\star$ a rectifiable, simple closed curve enclosing $\nu_\star(0)$ but no other values in the spectrum of $W(0)$. The spectrum of the operator $W(0)$ is separated into $\nu_\star(0)$ and $\Sigma(W(0)) \setminus \{\nu_\star(0)\}$ by $\Gamma_\star$. For sufficiently small n, i.e. with sufficiently small $\|n\|$, from $\Gamma_\star \subset \Theta\left(W(0)\right)$ it follows $\Gamma_\star \subset \Theta(W(n))$. This is due to continuity reasons (see [Kat95, IV.§3.1, Theorem 3.1]). For all fixed sufficiently small and admissible n the spectrum of the operator $W(n)$ is therefore separated into two parts by $\Gamma_\star$.
Let us restrict to these values of n in the following as we are interested in a function of eigenvalues in a neighborhood of $n = 0$. A calculation shows that for $P_\star(n)$ defined by (4.7) it holds that $P_\star(n) \in \mathfrak{L}(X)$ and

$$(P_\star(n))^2 = P_\star(n), \tag{4.11}$$

i.e. $P_\star(n)$ is a projection. In particular, if we denote its range by $\mathfrak{B}_\star(n)$ and by $\mathfrak{B}^c_\star(n)$ the range of $(I - P_\star(n))$, then $X = \mathfrak{B}_\star(n) \oplus \mathfrak{B}^c_\star(n)$, and one can show

that both subspaces are invariant under $W(n)$ (see [Kat95, III.§6.4, Theorem 6.17]) or equivalently that $W(n)$ and $P_\star(n)$ commute (see [Kat95, III.§5.6]).
Moreover, it can be shown that for all sufficiently small n the spectrum of $W(n)|_{\mathfrak{B}_\star(n)}$ consists of exactly one simple eigenvalue which we denote by $\nu_\star(n)$. This follows from the continuity of the eigenprojection $P_\star$ in n, which we discuss at the end of the proof, as it implies that the ranges of the eigenprojection are isomorphic for all sufficiently small n (cf. [Kat95, IV.§3.4,I.§4.6]). Hence, we can define a function $n \mapsto \nu_\star(n)$ of simple, isolated eigenvalues of $W(n)$ in a neighborhood of $n = 0$, corresponding to $P_\star$. For any fixed sufficiently small and admissible n the spectrum of the operator $W(n)$ is separated into $\nu_\star(n)$ and $\Sigma(W(n)) \setminus \{\nu_\star(n)\}$ by $\Gamma_\star$. $W(n)|_{\mathfrak{B}^c_\star(n)}$ has the spectrum $\Sigma(W(n)) \setminus \{\nu_\star(n)\}$ and its resolvent is holomorphic at $\nu_\star(n)$. Thus, the complete resolvent for $W(n)$ can then be splitted in the form

$$\widetilde{R}(\nu, n) = \widetilde{R}(\nu, n)P_\star(n) + \widetilde{R}(\nu, n)\left(1 - P_\star(n)\right) \tag{4.12}$$

and shown to have the expansion

$$\widetilde{R}(\nu, n) = -(\nu - \nu_\star(n))^{-1}P_\star(n) + \sum_{m=0}^{\infty}(\nu - \nu_\star(n))^m S_\star(n)^{m+1}. \tag{4.13}$$

Equation (4.13) is the Laurent expansion of $\widetilde{R}(\nu, n)$ at the isolated singularity $\nu_\star(n)$.
The relations (4.10b) and (4.10c) for the reduced resolvent can be found in [Kat95, III.§6.5, (6.34)]. Exemplarily we compute

$$\begin{aligned}(W(n) - \nu_\star(n))S_\star(n) &= \frac{1}{2\pi i}\int_{\Gamma_\star}(\nu - \nu_\star(n))^{-1}(W(n) - \nu_\star(n))\widetilde{R}(\nu, n)\,\mathrm{d}\nu \\ &= \frac{1}{2\pi i}\int_{\Gamma_\star}(\nu - \nu_\star(n))^{-1}I + \widetilde{R}(\nu, n)\,\mathrm{d}\nu = I - P_\star(n)\end{aligned} \tag{4.14}$$

by adding a zero and noting that $\Gamma_\star$ winds around $\nu_\star(n)$ once and therefore the integral over $(\nu - \nu_\star(n))^{-1}$ yields the factor $2\pi i$.
It remains to show the continuity of the mapping $n \mapsto P_\star(n)$. This can be deduced from the continuity of the resolvent in ν and n which follows directly from the continuity result shown in [Kat95, IV.§3.3, Theorem 3.15]. Since $\Gamma_\star$ is compact, for any sufficiently small $n_0 \in \mathfrak{N}$ it holds the following: for any $\epsilon > 0$ there exists a $\delta(n_0)$, independent of ν, such that

$$\left\|\widetilde{R}(\nu, n) - \widetilde{R}(\nu, n_0)\right\| < \epsilon \quad \text{if } \|n - n_0\| < \delta(n_0). \tag{4.15}$$

Regarding (4.7), this implies the continuity of $n \mapsto P_\star(n)$. □

As mentioned in the lemma above, we call $P_\star(n)$ eigenprojection for $\nu_\star(n)$. The dimension of its range is called the algebraic multiplicity of the eigenvalue. Thus, in our examinations $P_\star(n)$ is a degenerated operator with one-dimensional range for all admissible $n \in \mathfrak{N}$ with sufficiently small $\|n\|$ as we want to assume that $\nu_\star(0)$ is an isolated and simple eigenvalue of $W(0)$ (or respectively with some initial $\check{n} \in \mathfrak{N}$, note Remark 4.1). Throughout this and the following section we consider only such n and assume that W depends at least continuously on n such that a function $n \mapsto \nu_\star(n)$ of simple, isolated eigenvalues is always well-defined in the sense of the lemma above. Before we turn to the derivatives, we need some theory for degenerated operators and especially how one may define the trace of such operators.

Let X still be a Hilbert space. We call an operator $A \in \mathfrak{L}(X)$ degenerate if $\operatorname{rank}(A) := \dim(\mathfrak{R}(A)) < \infty$, where $\mathfrak{R}(A)$ denotes the range of A. Therefore, A is in particular compact and by singular value decomposition (see e.g. [Kat95, V.§2.3]) there exist orthonormal sequences $v_1, \dots, v_m$ and $g_1, \dots, g_m$ in X and values $\mu_1, \dots, \mu_m > 0$, where $m = \operatorname{rank}(A)$, such that for all $v \in X$

$$Av = \sum_{j=1}^{m} \mu_j \langle v, g_j \rangle v_j, \tag{4.16}$$

$g_j = \frac{1}{\mu_j} A^* v_j$ and $v_j = \frac{1}{\mu_j} A g_j$. The adjoint operator A^* is given by

$$A^* g = \sum_{j=1}^{m} \mu_j \langle g, v_j \rangle g_j \tag{4.17}$$

for all $g \in X$, and A^* is degenerate as well with $\operatorname{rank}(A) = \operatorname{rank}(A^*)$.

In accordance with the finite-dimensional case the trace of a degenerated operator A can be defined[26] as

$$\operatorname{tr}(A) := \sum_{j=1}^{m} \langle A v_j, v_j \rangle \tag{4.18}$$

with some orthonormal basis $v_1, \dots, v_m$ of $\mathfrak{R}(A)$ and $m = \operatorname{rank}(A)$. The trace is linear and its value is independent of the choice of the orthonormal basis[27]. Moreover,

$$\operatorname{tr}(A_1 A_2) = \operatorname{tr}(A_2 A_1) \tag{4.19}$$

for all degenerate operators A_1, A_2. For a detailed discussion of the properties of the trace, we refer to [Kat95, III.§4.3 and X.§1.4]).

LEMMA 4.4. *Let X be a Hilbert space. For any degenerate operator $A : X \to X$ the inequality*

$$|\operatorname{tr}(A)| \leq \operatorname{rank}(A) \, \|A\| \tag{4.20}$$

holds with the canonical operator norm on $\mathfrak{L}(X)$.

---

[26] Actually, the trace can be defined for a much wider class of operators (for operators in Hilbert spaces see e.g. [Kat95, X.§1.3/4]), but for our purpose the degenerate operators are sufficient.

[27] In fact, $\operatorname{tr}(\cdot^* \cdot)$ defines an inner product on the space of Hilbert-Schmidt operators, and by the polarization principle the definition of the trace is therefore independent of the chosen basis (see [Kat95, V.§2.4]).

Proof. Let $m = \operatorname{rank}(A)$ and $v_1, \ldots, v_m \in X$ be an orthonormal basis of $\mathfrak{R}(A)$. Then

$$|\operatorname{tr}(A)| \leq \sum_{j=1}^{m} |\langle Av_j, v_j \rangle| \leq \sum_{j=1}^{m} \|Av_j\| \, \|v_j\| \leq \sum_{j=1}^{m} \|A\| \, \|v_j\|^2 = m \, \|A\| \, . \tag{4.21}$$

□

Corollary 4.5. *Let $Q : \mathfrak{N} \subset Y \to \mathfrak{L}(X)$ be an operator-valued mapping with $Q(n)$ degenerate and with $\operatorname{rank}(Q(n)) \leq m_1$ for all n. Further, let Q be differentiable at $n = n_0 \in \mathfrak{N}$ with derivative $Q'(n_0)$ and let the derivative also be a degenerate operator with $\operatorname{rank}(Q'(n_0)h) \leq m_2$ for all $h \in Y$. Then the function $\operatorname{tr}(Q)$ is differentiable at $n = n_0$ with*

$$(\operatorname{tr}(Q))'(n_0)h = \operatorname{tr}(Q'(n_0)h). \tag{4.22}$$

If Q is even continuously differentiable at n_0 and $\operatorname{rank}(Q'(n)h) \leq m_2$ for all $h \in Y$ and all n in a neighborhood of n_0, then also $\operatorname{tr}(Q)$ is continuously differentiable at n_0.

Proof. With the help of Lemma 4.4 we find

$$\begin{aligned}
&\lim_{h\to 0} \frac{1}{\|h\|} \left|\operatorname{tr}\left(Q(n_0+h)\right) - \operatorname{tr}\left(Q(n_0)\right) - \operatorname{tr}\left(Q'(n_0)h\right)\right| \\
&= \lim_{h\to 0} \left|\operatorname{tr}\left(\frac{1}{\|h\|}\left(Q(n_0+h) - Q(n_0) - Q'(n_0)h\right)\right)\right| \\
&\leq \lim_{h\to 0} (2m_1+m_2) \frac{1}{\|h\|} \left\|Q(n_0+h) - Q(n_0) - Q'(n_0)h\right\| = 0
\end{aligned} \tag{4.23}$$

by the differentiability of Q at $n = n_0$.
If Q is even continuously differentiable at n_0, the continuous differentiability of $\operatorname{tr}(Q)$ follows from

$$\begin{aligned}
|\operatorname{tr}(Q'(n)h - Q'(n_0)h)| &\leq 2m_2 \left\|Q'(n)h - Q'(n_0)h\right\| \\
&\leq 2m_2 \left\|Q'(n) - Q'(n_0)\right\| \|h\|
\end{aligned} \tag{4.24}$$

for all $h \in Y$. □

For the eigenprojection $P_\star(n)$ we find the following:

Lemma 4.6. *Let $n \in \mathfrak{N}$ and $v_\star[n]$ be an eigenvector to the simple, isolated eigenvalue $\nu_\star(n)$ of $W(n)$. Then there exists an eigenvector $b_\star[n]$ of $W^*(n)$ to the eigenvalue $\overline{\nu_\star(n)}$. If additionally $b_\star[n]$ is normalized by*

$$\langle v_\star[n], b_\star[n] \rangle = 1, \tag{4.25}$$

then:

1. *The eigenprojection $P_\star(n)$ to the eigenvalue $\nu_\star(n)$ of $W(n)$ can be written as*

   $$P_\star(n)v = \langle v, b_\star[n] \rangle \, v_\star[n] \tag{4.26}$$

   for all $v \in X$.

2. *For any operator* $A \in \mathfrak{L}(X)$ *it holds*

$$\operatorname{tr}(AP_\star(n)) = \langle Av_\star[n], b_\star[n] \rangle \quad \textit{and} \quad \operatorname{tr}(P_\star(n)) = 1. \tag{4.27}$$

Proof. For any fixed n we have: $P_\star(n)$ is a degenerated operator with one-dimensional range since the eigenvalue $\nu_\star(n)$ is assumed to be simple. Thus, by our discussion of degenerate operators (see equation (4.16)) to any eigenvector $v_\star[n]$ there exists some vector $b_\star[n]$ to write $P_\star(n)$ in the form (4.26). It is left to prove that $b_\star[n]$ must be an eigenvector of $W^*[n]$ to the eigenvalue $\overline{\nu_\star(n)}$, normalized by (4.25). Since $P_\star^*(n)$ is given by

$$P_\star^*(n)g = \langle g, v_\star[n] \rangle \, b_\star[n], \tag{4.28}$$

for all $g \in X$, and is the eigenprojection to the simple and isolated eigenvalue $\overline{\nu_\star(n)}$ of $W^*(n)$ ([Kat95, III.§6.6, Thm. 6.22, (6.54)]), $b_\star[n]$ is an eigenvector to $\overline{\nu_\star(n)}$. The normalization condition follows from the fact that the eigenprojection maps on the eigenspace and $P_\star(n)^2 = P_\star(n)$ (cf. proof of Lemma 4.2), as

$$v_\star[n] = P_\star(n)v_\star[n] = \langle v_\star[n], b_\star[n] \rangle \, v_\star[n]. \tag{4.29}$$

For any $A \in \mathfrak{L}(X)$ the operator $AP_\star(n)$ is degenerate, of rank ≤ 1 and can be expressed by

$$AP_\star(n)v = \langle v, b_\star[n] \rangle \, Av_\star[n] \tag{4.30}$$

for all $v \in X$. We obtain for the trace of $AP_\star(n)$

$$\begin{aligned}
\operatorname{tr}(AP_\star(n)) &= \frac{1}{\|Av_\star[n]\|^2} \langle AP_\star(n)Av_\star[n], Av_\star[n] \rangle \\
&= \frac{1}{\|Av_\star[n]\|^2} \langle A \langle Av_\star[n], b_\star[n] \rangle \, v_\star[n], Av_\star[n] \rangle \\
&= \langle Av_\star[n], b_\star[n] \rangle \, .
\end{aligned}$$

□

4.1.3 Derivatives of the eigenvalues

We now prove formulas for the derivatives of simple isolated eigenvalues and corresponding eigenprojections.

Lemma 4.7. *Let* $W : \mathfrak{N} \to \mathfrak{L}(X)$ *with* $n \mapsto W(n)$ *be continuous. A simple and isolated eigenvalue* $\nu_\star(n)$ *of* $W(n)$ *can be expressed as*

$$\nu_\star(n) = \operatorname{tr}\left(W(n)P_\star(n)\right) \tag{4.31}$$

and the mapping $n \mapsto \nu_\star(n)$ *defined in Lemma 4.2 is continuous.*

Proof. If $\nu_\star(n)$ is a simple and isolated eigenvalue of $W(n)$, the operator $W(n)P_\star(n)$ is degenerate and of rank 1. Let $v_\star[n]$ be an eigenvector of $W(n)$ to $\nu_\star(n)$ with $\|v[n]\| = 1$. Then $\mathfrak{R}(W(n)P_\star(n))$ is spanned by this eigenvector and

$$\operatorname{tr}\left(W(n)P_\star(n)\right) = \langle W(n)v_\star[n], v_\star[n] \rangle = \nu_\star(n). \tag{4.32}$$

Since W is assumed to be continuous in n and $n \mapsto P_\star(n)$ is continuous by Lemma 4.2, with Lemma 4.4, we obtain for any $n_0 \in \mathfrak{N}$:

$$\begin{aligned} |\nu_\star(n_0) - \nu_\star(n)| &= |\mathrm{tr}\,(W(n_0)P_\star(n_0)) - \mathrm{tr}\,(W(n)P_\star(n))| \\ &\le 2\,\|W(n_0)P_\star(n_0) - W(n)P_\star(n)\| \to 0, \end{aligned} \tag{4.33}$$

if $\|n_0 - n\| \to 0$. This shows the continuity of the mapping $n \mapsto \nu_\star(n)$. □

THEOREM 4.8. *Let* $W : \mathfrak{N} \to \mathfrak{L}(X)$ *with* $n \mapsto W(n)$ *be continuously differentiable at* $n = 0$ *with the derivative* $W' : Y \to \mathfrak{L}(X)$ *and let* $n \mapsto \nu_\star(n)$ *be the function of simple, isolated eigenvalues from Lemma 4.2. Then the eigenprojection, i.e. the mapping* $n \mapsto P_\star(n)$*, and the eigenvalue, i.e. the mapping* $n \mapsto \nu_\star(n)$*, are both continuously differentiable at* $n = 0$ *with*

$$P'_\star(0)h = -P_\star(0)\,(W'(0)h)\,S_\star(0) - S_\star(0)\,(W'(0)h)\,P_\star(0) \tag{4.34}$$

and

$$\nu'_\star(0)h = \mathrm{tr}\,((W'(0)h)\,P_\star(0))\,, \tag{4.35}$$

for all $h \in Y$*.* $S_\star(n)$ *and* $P_\star(n)$ *are defined as in Lemma 4.2.*

PROOF. The operator inversion is differentiable for bounded operators $A \in \mathfrak{L}(X)$ whenever $A^{-1} \in \mathfrak{L}(X)$ exists (This can be seen via the Neumann series, for details see e.g. [Zei86, 4.7, p.154].). Thus, for a differentiable mapping $Q : \mathfrak{N} \to \mathfrak{L}(X)$ the mapping $Q^{-1} : \mathfrak{N} \to \mathfrak{L}(X)$, $n \mapsto Q(n)^{-1}$ is differentiable as a composition of differentiable mappings (see e.g. [Zei86, Proposition 4.10]) whenever $Q(n)^{-1}$ exists in $\mathfrak{L}(X)$ and $Q'(n)$ exists. It holds the formula

$$(Q^{-1})'(n)h = -Q(n)^{-1}\,(Q'(n)h)\,Q(n)^{-1} \tag{4.36}$$

for all $h \in Y$. If Q is even continuously differentiable, so is Q^{-1}.
Hence, the continuous differentiability of $n \mapsto W(n)$ at $n = 0$ implies the continuous differentiability of the resolvent $\widetilde{R}(\nu, n)$ at $n = 0$ with

$$\left[\frac{\partial}{\partial n}\widetilde{R}(\nu, n)\right]_{n=n_0} h = -\widetilde{R}(\nu, n_0)\,(W'(n_0)h)\,\widetilde{R}(\nu, n_0) \tag{4.37}$$

for all n_0 in a small neighborhood of $n = 0$, $h \in Y$.
In equation (4.15) we have already seen that $\|\widetilde{R}(\nu, n) - \widetilde{R}(\nu, n_0)\|$ is small uniformly (in the sense stated there) for $\nu \in \Gamma_\star$, if $\|n - n_0\|$ is sufficiently small. In this sense also the derivative (4.37) exists uniformly in ν. If we recall the expression (4.7) for $P_\star(n)$, equation (4.37) then implies the continuous differentiability of the mapping $n \mapsto P_\star(n)$ at $n = 0$ with

$$\begin{aligned} P'_\star(n_0)h &= -\frac{1}{2\pi i}\int_{\Gamma_\star} \left[\frac{\partial}{\partial n}\widetilde{R}(\nu, n)\right]_{n=n_0} h\,\mathrm{d}\nu \\ &= \frac{1}{2\pi i}\int_{\Gamma_\star} \widetilde{R}(\nu, n_0)\,(W'(n_0)h)\,\widetilde{R}(\nu, n_0)\,\mathrm{d}\nu. \end{aligned} \tag{4.38}$$

for all n_0 in a small neighborhood of $n = 0$. For each n_0 the resolvent $\widetilde{R}(\nu, n_0)$ has by Lemma 4.2 the Laurent expansion

$$\widetilde{R}(\nu, n_0) = -(\nu - \nu_\star(n_0))^{-1} P_\star(n_0) + \sum_{m=0}^{\infty} (\nu - \nu_\star(n_0))^m S_\star^{m+1}(n_0). \tag{4.39}$$

We plug (4.39) into (4.38) and note that only terms with $(\nu - \nu_\star(n_0))^{-1}$ contribute to the integral because all the other terms possess antiderivatives. For all n_0 in a small neighborhood of $n = 0$ we find

$$P_\star'(n_0)h = -P_\star(n_0)\,(W'(n_0)h)\,S_\star(n_0) - S_\star(n_0)\,(W'(n_0)h)\,P_\star(n_0), \tag{4.40}$$

where we have also used that $\Gamma_\star$ winds around $\nu_\star(n_0)$ once and therefore the integral of $(\nu - \nu_\star(n_0))^{-1}$ yields the factor $2\pi i$.
Applying Lemma 4.7 and (2.) of Lemma 4.6 we can write

$$\begin{aligned} \nu_\star(n) &= \operatorname{tr}(W(n)P_\star(n)) = \operatorname{tr}(W(n)P_\star(n) - \nu_\star(n_0)P_\star(n)) + \nu_\star(n_0)\operatorname{tr}(P_\star(n)) \\ &= \nu_\star(n_0) + \operatorname{tr}((W(n) - \nu_\star(n_0))\,P_\star(n)) \end{aligned} \tag{4.41}$$

for all n_0 with sufficiently small $\|n_0\|$ (such that $\nu_\star(n_0)$ is simple and isolated). We have seen above that the mapping $n \mapsto P_\star(n)$ is continuously differentiable at $n = 0$, and since W is continuously differentiable by assumption, also $n \mapsto (W(n) - \nu_\star(n_0))\,P_\star(n)$ is continuously differentiable at $n = 0$. Since $(W(n) - \nu_\star(n_0))\,P_\star(n)$ and the derivative (4.40) are degenerate and of rank 1 respectively, the continuous differentiability of $n \mapsto \nu_\star(n)$ at $n = 0$ follows from Corollary 4.5. Using the product rule ([Zei86, Prop.4.11]) we find

$$\begin{aligned} \frac{d}{dn}\left((W(n) - \nu_\star(0))\,P_\star(n)\right)\Big|_{n=0} h &= (W'(0)h)\,P_\star(0) + (W(0) - \nu_\star(0))\,P_\star'(0)h \\ &= P_\star(0)\,(W'(0)h)\,P_\star(0). \end{aligned} \tag{4.42}$$

The last equality can be deduced from (4.40) as

$$\begin{aligned} (W(0) - \nu_\star(0))\,P_\star'(0)h &= 0 - (W(0) - \nu_\star(0))\,S_\star(0)\,(W'(0)h)\,P_\star(0) \\ &= -\,(W'(0)h)\,P_\star(0) + P_\star(0)\,(W'(0)h)\,P_\star(0), \end{aligned} \tag{4.43}$$

where we also used the two identities $(W(0) - \nu_\star(0))\,S_\star(0) = I - P_\star(0)$ and $(W(0) - \nu_\star(0))\,P_\star(0) = 0$ (cf. equation (4.10)).
We obtain

$$\nu_\star'(0)h = \operatorname{tr}(P_\star(0)\,(W'(0)h)\,P_\star(0)) \tag{4.44}$$

and conclude

$$\nu_\star'(0)h = \operatorname{tr}((W'(0)h)\,P_\star(0)), \tag{4.45}$$

because the trace does not change if we commute the operators (see (4.19)).□

We want to avoid the computation of $P_\star(n)$ as an integral. With our results for the eigenprojections and degenerate operators, we can rewrite the statement for the eigenvalues.

Corollary 4.9. *Let* $W : \mathfrak{N} \to \mathfrak{L}(X)$ *be continuously differentiable for all admissible* n *and have the simple, isolated eigenvalue* $\nu_\star(\check{n})$ *for some* $\check{n} \in \mathfrak{N}$. *Then there exists a continuous function* $n \mapsto \nu_\star(n)$ *of simple isolated eigenvalues in a neighborhood of* $\check{n}$.
Let further $v_\star(n)$ *be an eigenvector of* $W(n)$ *corresponding to the eigenvalue* $\nu_\star(n)$, *then there exists an eigenvector* $b_\star(n)$ *of* $W(n)^*$ *corresponding to the eigenvalue* $\overline{\nu_\star(n)}$. *If additionally* $b_\star(n)$ *is normalized by* $\langle v_\star(n), b_\star(n)\rangle = 1$, *the function* $n \mapsto \nu_\star(n)$ *is continuously differentiable and has at* $n = \check{n}$ *the derivative*

$$\nu_\star'(\check{n})h = \langle (W'(\check{n})h)\, v_\star(\check{n}), b_\star(\check{n})\rangle \,. \tag{4.46}$$

Proof. In virtue of Remark 4.1 the result follows directly from Theorem 4.8 and Lemma 4.6. □

4.2 Derivatives of eigenvalues and eigenvectors for a generalized eigenvalue problem

Let us consider now the generalized eigenvalue problem

$$B(n)v_\star = \nu_\star M v_\star, \tag{4.47}$$

with the differentiable operator-valued function $B : \mathfrak{N} \mapsto \mathfrak{L}(X)$, where $\mathfrak{N}$ is an open subset of a Banach space Y and X a Hilbert space. We want to assume that $M \in \mathfrak{L}(X)$ does not depend on n and is boundedly invertible. Then $(\nu_\star, v_\star)$ is a classic eigenpair of the operator $M^{-1}B(n)$ and we can define functions of eigenvalues $n \mapsto \nu_\star(n)$ as in the previous section in the sense of Lemma 4.2. In particular, we choose the largest possible subset of admissible n around some initial $\check{n} \in \mathfrak{N}$ in which we can define a function of eigenvalues. Moreover, we assume that $B(n)$ and M are conjugation-symmetric ($\mathcal{C}$-symmetric):

Definition 4.10. *Let* X *be a Hilbert space. We call a mapping*

$$\mathcal{C} : X \to X \tag{4.48}$$

a conjugation if it has the following properties:

1. $\mathcal{C}^2 = I$
2. $\mathcal{C}(cv) = \overline{c}\mathcal{C}(v)$ *for all* $v \in X$ *and all* $c \in \mathbb{C}$
3. $\|\mathcal{C}(v)\| = \|v\|$ *for all* $v \in X$.

An operator $A \in \mathfrak{L}(X)$ *is called conjugation-symmetric* ($\mathcal{C}$*-symmetric) if it fulfills*

$$\langle Av, \mathcal{C}(w)\rangle = \langle v, \mathcal{C}(Aw)\rangle \tag{4.49}$$

for all $v, w \in X$.

If X is a Hilbert space of complex-valued functions, there is a canonical conjugation given by complex conjugation: $\mathcal{C}(f) := \overline{f}$. However, since we work in an abstract Hilbert space setting so far, the existence of a conjugation has to be assumed in the following.

4.2.1 Derivatives of the eigenvalues

Lemma 4.11. *Assume that X is equipped with a conjugation $\mathcal{C}$ and that M and $B(n)$ are $\mathcal{C}$-symmetric for all $n \in \mathfrak{N}$. Let $B : \mathfrak{N} \to \mathfrak{L}(X)$ depend continuously differentiably on n, M be boundedly invertible and let $v_\star(n)$ be a simple and isolated eigenvalue to (4.47) in a neighborhood of some $\check{n} \in \mathfrak{N}$. Further, let $v_\star[n] \in X$ be a corresponding eigenvector, scaled*[28] *such that*

$$\langle v_\star[n], \mathcal{C}(Mv_\star[n])\rangle = 1. \tag{4.50}$$

Then $n \mapsto \nu_\star(n)$ is continuously differentiable and the Fréchet derivative is given by

$$\nu_\star'(n)h = \langle (B'(n)h)\, v_\star[n], \mathcal{C}(v_\star[n])\rangle . \tag{4.51}$$

Proof. $(\nu_\star(n), v_\star[n])$ is a classic eigenpair of the operator $M^{-1}B(n)$. Thus, by Corollary 4.9 we find

$$\nu_\star'(n)h = \left\langle M^{-1}\left(B'(n)h\right) v_\star[n], b_\star[n]\right\rangle , \tag{4.52}$$

with $b_\star[n]$ being an eigenvector of $(M^{-1}B(n))^*$ to the eigenvalue $\overline{\nu_\star(n)}$ and the normalization condition

$$\langle v_\star[n], b_\star[n]\rangle = 1. \tag{4.53}$$

Since $B(n)$ and M are $\mathcal{C}$-symmetric it holds

$$\begin{aligned}\left\langle v, \left(M^{-1}B(n)\right)^* \mathcal{C}(Mv_\star[n])\right\rangle &= \left\langle \left(M^{-1}B(n)\right) v, \mathcal{C}(Mv_\star[n])\right\rangle \\ &= \langle B(n)v, \mathcal{C}(v_\star[n])\rangle = \langle v, \mathcal{C}(B(n)v_\star[n])\rangle \\ &= \left\langle v, \overline{\nu_\star[n]}\mathcal{C}(Mv_\star[n])\right\rangle \end{aligned} \tag{4.54}$$

for all $v \in X$. Therefore, $\mathcal{C}(Mv_\star[n])$ is an eigenvector of $(M^{-1}B[n])^*$ to the eigenvalue $\overline{\nu_\star[n]}$, and by (4.50) it fulfills the normalization condition (4.53). Hence,

$$\nu_\star'(n)h = \left\langle M^{-1}\left(B'(n)h\right) v_\star[n], \mathcal{C}(Mv_\star[n])\right\rangle = \langle (B'(n)h)\, v_\star[n], \mathcal{C}(v_\star[n])\rangle . \tag{4.55}$$

□

Remark 4.12. *Under the assumptions of Lemma 4.11 the eigenprojection $P_\star(n)$ (corresponding to $M^{-1}B(n)$) is given by*

$$P_\star(n)v = \langle v, \mathcal{C}(Mv_\star[n])\rangle\, v_\star[n] \tag{4.56}$$

for all $v \in X$. This follows from Lemma 4.6 and the fact that $\mathcal{C}(Mv_\star[n])$ is an eigenvector of $(M^{-1}B(n))^*$ to the eigenvalue $\overline{\nu_\star(n)}$ (see equation (4.54)).

---

[28] The scaling is possible since the eigenspace is one-dimensional, M is invertible and $\langle cv, \mathcal{C}(Mcv)\rangle = c^2 \langle v, \mathcal{C}(Mv)\rangle$ for all $v \in X$ and $c \in \mathbb{C}$.

4.2.2 Derivatives of the eigenvectors

As our objective will also involve eigenvectors, we also comment on derivatives of generalized eigenvectors with respect to n. So far, we have only worked with the closely related eigenprojections. In contrast to eigenprojections, the eigenvectors are not uniquely determined. But as we assume simple, isolated eigenvalues, with the normalization

$$\langle v_\star[n], \mathcal{C}(Mv_\star[n])\rangle = 1 \tag{4.57}$$

the eigenvector $v_\star[n]$ is uniquely determined up to its sign since

$$\frac{1}{\sqrt{\langle cv_\star[n], \mathcal{C}(Mcv_\star[n])\rangle}} cv_\star[n] = \frac{c}{\sqrt{c^2}} v_\star[n] = \pm v_\star[n] \tag{4.58}$$

for all $c \in \mathbb{C}$. We will later see that the eigenvector only appears in quadratic terms in the objective function such that we can choose one sign without changing the values of the objective function or its derivative.

Lemma 4.13. *Let the assumptions of Lemma 4.11 be fulfilled. Then there exists a continuous function $n \mapsto v_\star[n]$ of eigenvectors corresponding to $\nu_\star(n)$ and fulfilling the normalization condition (4.57). It is even continuously differentiable, and its derivative is given by*

$$\begin{aligned} v'_\star[n]h = -S^{\mathrm{M}}_\star(n)\left(B'(n)h\right)v_\star[n] + \frac{1}{2}v_\star[n]\Big(\Big\langle S^{\mathrm{M}}_\star(n)\left(B'(n)h\right)v_\star[n], \mathcal{C}\left(Mv_\star[n]\right)\Big\rangle \\ + \Big\langle Mv_\star[n], \mathcal{C}\left(S^{\mathrm{M}}_\star(n)\left(B'(n)h\right)v_\star[n]\right)\Big\rangle\Big) \end{aligned} \tag{4.59}$$

with

$$S^{\mathrm{M}}_\star(n) := \frac{1}{2\pi i}\int_{\Gamma_\star} (\nu - \nu_\star(n))^{-1}\left(B(n) - \nu M\right)^{-1} \mathrm{d}\nu, \tag{4.60}$$

where $\Gamma_\star$ is a rectifiable, simple closed curve in the resolvent set enclosing $\nu_\star(\check{n})$ but no other values in the spectrum of $M^{-1}B(\check{n})$.

Proof. As in Section 4.1, it suffices to analyze the situation for a general operator around $n = 0$. Let $v_\star[0]$ be a generalized eigenvector of (4.47) (and thus an usual eigenvector of $M^{-1}B(n)$) to the simple and isolated eigenvalue $\nu_\star(n)$, normalized such that $\langle P_\star(0)v_\star[0], \mathcal{C}\left(MP_\star(0)v_\star[0]\right)\rangle = 1$. Note that we mean here always the eigenprojection $P_\star(n)$ to $W(n) = M^{-1}B(n)$, and the same is true for the reduced resolvent $S_\star(n)$ used later.

By Theorem 4.8 for simple and isolated eigenvalues $\nu_\star(n)$ the mapping $n \mapsto P_\star(n)$ is differentiable and therefore in particular continuous. Thus, we have $P_\star(n)v_\star[0] \neq 0$ for all n with sufficiently small $\|n\|$ because

$$P_\star(0)v_\star[0] = v_\star[0] \neq 0. \tag{4.61}$$

From the assumption that the eigenvalue $\nu_\star(n)$ is simple we know that $P_\star(n)$ is one-dimensional which ensures that $P_\star(n)v_\star[0]$ can only map on an eigenvector

$v[n]$ to $\nu_\star(n)$ or be equal to 0. This means $P_\star(n)v_\star[0]$ defines a continuous function of eigenvectors in a neighborhood of $n = 0$.
Therefore, we have the following appropriate representation of the scaled eigenvector

$$v_\star[n] = \frac{P_\star(n)v_\star[0]}{\sqrt{\langle (P_\star(n)v_\star[0]), \mathcal{C}(MP_\star(n)v_\star[0])\rangle}}, \tag{4.62}$$

for all n with sufficiently small $\|n\|$. This is exactly what need for the derivative at $n = 0$. Moreover, we can now define a continuous function $n \mapsto v_\star(n)$ for all admissible n since the scaled eigenvector is unique up to its sign. We have already shown the continuous differentiability of the mapping $n \mapsto P_\star(n)$ at $n = 0$ in Theorem 4.8 and therefore $n \mapsto v_\star[n]$ is continuously differentiable at $n = 0$. Note that we can differentiate into scalar product and conjugation since both are continuous with respect to the norm on X.
We compute

$$\begin{aligned} v_\star'[0]h = (P_\star'(0)h)\, v_\star[0] - \frac{1}{2} v_\star[0] \left(\langle (P_\star'(0)h)\, v_\star[0], \mathcal{C}\left(Mv_\star[0]\right)\rangle \right. \\ \left. + \langle v_\star[0], \mathcal{C}\left(M\left(P_\star'(0)h\right) v_\star[0]\right)\rangle \right), \end{aligned} \tag{4.63}$$

using the product rule and $\langle P_\star(0)v_\star[0], \mathcal{C}\left(MP_\star(0)v_\star[0]\right)\rangle = 1$.
Again by Theorem 4.8 we have

$$P_\star'(0)h = -P_\star(0)\left(M^{-1}B'(0)h\right) S_\star(0) - S_\star(0)\left(M^{-1}B'(0)h\right) P_\star(0) \tag{4.64}$$

with

$$\begin{aligned} S_\star(n) &= \frac{1}{2\pi i}\int_{\Gamma_\star} (\nu - \nu_\star(n))^{-1} \left(M^{-1}B(n) - \nu\right)^{-1} \mathrm{d}\nu \\ &= \left(\frac{1}{2\pi i}\int_{\Gamma_\star} (\nu - \nu_\star(n))^{-1} \left(B(n) - \nu M\right)^{-1} \mathrm{d}\nu\right) M. \end{aligned}$$

Since $S_\star(0)P_\star(0) = 0$ (see Lemma 4.2, (4.10b)), the expression $(P_\star'(0)h)\, v_\star[0]$ simplifies to

$$(P_\star'(0)h)\, v_\star[0] = -S_\star^{\mathrm{M}}(0)\left(B'(0)h\right) v_\star[0]. \qquad \square$$

To evaluate the derivative (4.59) of the eigenvector, we need the operator $S_\star^{\mathrm{M}}(n)$, defined in (4.60). Like the eigenprojection we do not want to compute it as an integral. The following lemma gives us insight.

Lemma 4.14. *Let the assumptions of Lemma 4.13 be fulfilled. For the modified reduced resolvent, defined in (4.60), it holds*

$$S_\star^{\mathrm{M}}(n) = \left(B(n) - \nu_\star(n)M + MP_\star(n)\right)^{-1} - P_\star(n)M^{-1}. \tag{4.65}$$

Proof. Let us consider some fixed n. By the relations (4.10) we find

$$\begin{aligned} \left[M^{-1}B(n) - \nu_\star(n) + P_\star(n)\right]\left[S_\star(n) + P_\star(n)\right] = (M^{-1}B(n) - \nu_\star(n))S_\star(n) \\ + P_\star(n)S_\star(n) + (M^{-1}B(n) - \nu_\star(n))P_\star(n) + (P_\star(n))^2 = I, \end{aligned} \tag{4.66}$$

with $P_\star(n)$ and $S_\star(n)$ (for $\nu_\star(n)$) corresponding to $M^{-1}B(n)$. We conclude that the operator $(M^{-1}B(n) - \nu_\star(n) + P_\star(n))$ is invertible and

$$S_\star(n) = \left(M^{-1}B(n) - \nu_\star(n) + P_\star(n)\right)^{-1} - P_\star(n). \tag{4.67}$$

By definition we have $S_\star^{\mathrm{M}}(n) = S_\star(n)M^{-1}$ and therefore

$$\begin{aligned} S_\star^{\mathrm{M}}(n) &= \left[\left(M^{-1}B(n) - \nu_\star(n) + P_\star(n)\right)^{-1} - P_\star(n)\right] M^{-1} \\ &= \left[M\left(M^{-1}B(n) - \nu_\star(n) + P_\star(n)\right)\right]^{-1} - P_\star(n)M^{-1} \\ &= (B(n) - \nu_\star(n)M + MP_\star(n))^{-1} - P_\star(n)M^{-1}. \end{aligned} \tag{4.68}$$

□

Remark 4.15. *Observe that under the given assumptions for all admissible n the operator $P_\star(n)M^{-1}$ is given by*

$$P_\star(n)M^{-1}v = \langle v, \mathcal{C}(v_\star[n])\rangle\, v_\star[n] \tag{4.69}$$

for all $v \in X$, where we used equation (4.56).

In the next section we discuss how the shown results apply to our problem. We have shown in Chapter 2 that the resolvent has only isolated poles (the resonances) and assumed that the examined best physical resonance is simple. If we consider the resonance problem on the whole real axis, we have to deal with $H^2_{\mathrm{loc}}(\mathbb{R})$ which is not a Hilbert space. In the weak formulation (2.15), which is a way out, the linear structure in ν of the problem is lost due to the DtN numbers. This inconvenience can be overcome using the Hardy space formulation (see Lemma 2.9), and our results can be applied.

4.3 Objective function and discretization

As explained in Chapter 2, we want to optimize the field enhancement as a function of the refractive index n. We have already pointed out in Section 2.5 that we replace the non-differentiable infinity-norm in our objective function by the L^2-norm of the field. But the dependence of the L^2-norm at the best resonant frequency on the refractive index is not very explicit so far. We are interested in an explicit formula of a suitable differentiable objective function.

4.3.1 Approximation to the L^2-norm and asymptotic expansion of scattering solutions in the vicinity of resonances

Let us briefly discuss how the results from Sections 4.1 and 4.2 apply to our problem formulated in Chapter 2. For simplicity let us assume we have rewritten the scattering problem (2.8) as an operator equation of the form

$$(B(n) - \nu M)\mathcal{u}_{\mathrm{s}} = Mf(\nu), \tag{4.70}$$

with a differentiable operator-valued function $B : \mathfrak{N} \subset Y \mapsto \mathfrak{L}(X)$, a boundedly invertible operator $M \in \mathfrak{L}(X)$ and some right hand side $Mf(\nu) \in X$, where X is a Hilbert space. $\mathfrak{N} \subset L^\infty(\mathbb{R})$ is in our problem the set of admissible refractive indices as defined in (2.1) and $M\boldsymbol{f}(\nu)$ is an analog to the $G(\nu)$ in the weak formulation (2.15). Furthermore, we assume that X is equipped with a conjugation $\mathcal{C}$ and that the operators $B(n)$ and M are $\mathcal{C}$-symmetric for all $n \in \mathfrak{N}$ and that $(B(n) - \nu M)$ is boundedly invertible for all $n \in \mathfrak{N}$ and all ν, except for a discrete set of isolated poles. Such a formulation can be achieved by the Hardy space formulation (cf. Lemma 2.9). In particular, we have $X = X^{\mathrm{H}}$, and $\boldsymbol{u}_{\mathrm{s}}$ and $\boldsymbol{f}(\nu)$ are defined as in Lemma 2.9. A conjugation on $X = X^{\mathrm{H}}$, for which $B(n)$ and M are $\mathcal{C}$-symmetric, is given by

$$\mathcal{C} : X^{\mathrm{H}} \to X^{\mathrm{H}} \quad \text{with} \quad \mathcal{C}\left(u^{\ominus}, u_{\mathrm{s}}, u_{\mathrm{s}}^{\oplus}\right) := \left(\overline{u^{\ominus}(\overline{\cdot})}, \overline{u_{\mathrm{s}}}, \overline{u_{\mathrm{s}}^{\oplus}(\overline{\cdot})}\right), \tag{4.71}$$

where the bar is the standard complex conjugation (for details, see [HN09, p.978, eq.(2.20)]). By Lemma 2.9 the operator $(B(n) - \nu M)$ is a Fredholm operator of index 0 for all $n \in \mathfrak{N}$ and all $\nu \in \mathfrak{Z}$. It is injective for all $\nu \in \mathfrak{Z}$ except for a discrete set, which follows from Theorem 2.12 together with the equivalence of Hardy space formulation and weak formulation (cf. remarks below Lemma 2.9).

Choose some admissible refractive index $\check{n}$ supporting at least one resonant state, and choose a resonance $\nu_\star(\check{n})$ (generalized eigenvalue of (4.70)) to it, which we assume to be simple. Then by Lemma 4.2 in a neighborhood of $\check{n}$ the resolvent $\widetilde{R}(\nu, n) = (M^{-1}B(n) - \nu)^{-1}$ has the expansion

$$\widetilde{R}(\nu, n) = -(\nu - \nu_\star(n))^{-1} P_\star(n) + \sum_{m=0}^{\infty} (\nu - \nu_\star(n))^m S_\star(n)^{m+1}. \tag{4.72}$$

Hence, when ν tends to $\nu_\star(n)$, the expression (4.72) is dominated by the first term (the second term is a convergent geometric series for $|\nu - \nu_\star(n)| < \|S_\star(n)\|$) and

$$\widetilde{R}(\nu, n)\boldsymbol{f}(\nu) \approx -(\nu - \nu_\star(n))^{-1} P_\star(n)\boldsymbol{f}(\nu_\star(n)) \tag{4.73}$$

is a first order approximation of the solution $\boldsymbol{u}_{\mathrm{s}}[\nu, n]$ to (4.70), provided that $P_\star(n)\boldsymbol{f}(\nu_\star(n)) \neq 0$ and $P_\star(n)\boldsymbol{f}(\nu) \approx P_\star(n)\boldsymbol{f}(\nu_\star(n))$. We discuss in the equations (4.78) how one can motivate that these conditions are fulfilled in our problem. Before, recall that if $v_\star[n]$ is an eigenvector corresponding to $\nu_\star(n)$ and scaled such that

$$\langle v_\star[n], \mathcal{C}(Mv_\star[n])\rangle_X = 1, \tag{4.74}$$

the eigenprojection $P_\star(n)$ is given by

$$P_\star(n)v = \langle v, \mathcal{C}(Mv_\star[n])\rangle_X \, v_\star[n] \tag{4.75}$$

for all $v \in X$ (see Remark 4.12).

We explain now the application to the Hardy space formulation. Elements $\boldsymbol{u}_{\mathrm{s}} \in X^{\mathrm{H}}$ consist of three components. The one in the middle is in $H^1([-a, 0])$ and is the restriction to $[-a, 0]$ of a solution u_{s} to problem (2.18) (cf. Lemma 2.9). Therefore, by the middle component of (4.73) we get an approximation to the solution of problem (2.18), and by adding u_{i} we get an approximation to the solution u of problem (2.8) in $[-a,0]$. We also remind that $Mf(\nu)$ in Lemma 2.9 has no parts in the Hardy spaces (see again the remarks below Lemma 2.9), and find

$$P_\star(n)f(\nu) = \left\langle Mf(\nu), \mathcal{C}(v_\star[n])\right\rangle_X v_\star[n] = \left\langle \underline{Mf(\nu)}, \overline{\underline{v_\star[n]}}\right\rangle_{H^1} v_\star[n], \tag{4.76}$$

where the underlined quantities indicate that we only use the part corresponding to the H^1-part of X^{H} there. In Lemma 2.13 we have shown the relation

$$\left\langle G(\nu_\star(n)), \overline{\underline{v_\star[n]}}\right\rangle_{H^1} = -2i\sqrt{k_0^2-\nu_\star(n)}\,\underline{v_\star[n]}(0) \neq 0. \tag{4.77}$$

Since the Hardy space formulation of problem (2.18) is equivalent to the weak formulation (2.15) (see Lemma 2.9), we conclude

$$P_\star(n)f(\nu_\star(n)) \neq 0. \tag{4.78a}$$

Simultaneously, $G(\nu)$ depends holomorphically on ν as already discussed in the proof of Theorem 2.14. This implies

$$P_\star(n)f(\nu) \approx P_\star(n)f(\nu_\star(n)) \tag{4.78b}$$

if $|\nu-\nu_\star(n)|$ is sufficiently small.

We are interested in the L^2-norm of u_{s} in $[-a,0]$. Hence, we again want to use only the part of (4.76), which corresponds to the H^1-component. With the approximation (4.73) we obtain the following approximation to the L^2-norm of u_{s} in $[-a,0]$:

$$\begin{aligned}
\|u_{\mathrm{s}}[\nu,n]\|_{L^2}^2 &\approx \frac{1}{|\nu-\nu_\star(n)|^2}\left\langle \underline{P_\star(n)f(\nu_\star(n))}, \underline{P_\star(n)f(\nu_\star(n))}\right\rangle_{L^2} \\
&= \frac{1}{|\nu-\nu_\star(n)|^2}\left|\left\langle \underline{Mf(\nu_\star(n))}, \overline{\underline{v_\star[n]}}\right\rangle_{H^1}\right|^2 \left\|\underline{v_\star[n]}\right\|_{L^2}^2 \\
&\approx \frac{4\left|k_0^2-\nu_\star(n)\right|}{|\nu-\nu_\star(n)|^2}\left|\underline{v_\star[n]}(0)\right|^2 \left\|\underline{v_\star[n]}\right\|_{L^2}^2,
\end{aligned} \tag{4.79}$$

where the last approximation is again motivated by equation (4.77).

Note that approximation formula (4.79) is differentiable by the results from the previous sections, as long as we are away from zero[29]. We have provided all necessary derivatives to obtain a closed formula for its derivative. As we cannot compute the resonances analytically, we discretize our problem now and derive a discrete analog of (4.79) as the objective function for our optimization process. Its derivative can be computed by applying the results of the previous sections.

Observe that approximation formula (4.79) also indicates that the real parts of the resonances approximate the (real-valued) resonant frequencies as the expression (4.79) is maximized by $\nu = \mathrm{Re}(\nu_\star(n))$ for ν in the real numbers. Also observe that the width of the peak in the intensity decreases if $\mathrm{Im}(\nu_\star(n))$ does (cf. comments in Section 2.2.2).

---

[29] This is important for the absolute values to be differentiable and will be discussed in the discrete setting, see Section 4.4.2.

4.3.2 Discretization

The discretization of the scattering problem is done by finite elements combined with Hardy space infinite elements. The Hardy space method is a Galerkin method with special ansatz functions based on the Hardy space formulation. Finite elements basis functions in the bounded domain $(-a, 0)$ are coupled with basis functions in the Hardy space to incorporate the radiation conditions. A favorable feature of this method is the preservation of the linear eigenvalue structure, which is surprising if one looks at the form of the DtN numbers which are nonlinear in ν. Since the assembly of the finite elements matrices is rather standard and a complete explanation of the Hardy space method leads us too far away from our purpose, we refer again to [HN09, Section 2.4] and [Nan08].

REMARK 4.16. *For the finite elements discretization one can use Lagrange elements or other standard finite elements. In our numerical experiments we use basis functions based on integrated Legendre polynomials (see e.g. [Sch98]) which stay stable for high polynomial orders.*
Also the grid for the finite elements can be chosen quite arbitrarily but sometimes it will turn out to be helpful to choose a certain discretization to make the computation of the derivatives easier. For example in the case of a piecewise constant approximation of the refractive index we often choose our discretization in a way such that the index of refraction does not change inside finite elements, but on intersection points between two elements. Nevertheless, our analysis does not rely on a certain kind of finite elements discretization.

We find the following discrete equation:

$$\mathbf{B}(n)\mathbf{u}_{\mathrm{s}}[\nu, n] - \nu\mathbf{M}\mathbf{u}_{\mathrm{s}}[\nu, n] = \mathbf{M}\mathbf{f}[\nu, n] \tag{4.80}$$

with vectors $\mathbf{f}[\nu, n], \mathbf{u}_{\mathrm{s}}[\nu, n] \in \mathbb{C}^J$ and matrices $\mathbf{B}(n), \mathbf{M} \in \mathbb{C}^{J\times J}$ which are complex(-conjugation)-symmetric (non-Hermitian) for all admissible n, i.e.

$$\mathbf{B}^\top(n) = \mathbf{B}(n) \quad \text{and} \quad \mathbf{M}^\top = \mathbf{M} \quad \text{for all admissible } n. \tag{4.81}$$

The matrix $\mathbf{M}$ is invertible and the matrix $\mathbf{B}(n)$ depends on n, but also contains parts (the stiffness matrix) which do not depend on n. Equation (4.80) is a discrete approximation to the weak formulation (2.15) of the scattering problem and the vector $\mathbf{u}_{\mathrm{s}}[\nu, n]$ approximates the solution to problem (2.14) (see [HN09, Section 2.4]).
Hence, (4.80) is a special case of equation (4.70) and the resonances may be computed as generalized eigenvalues of

$$\mathbf{B}(n)\mathbf{u}_{\mathrm{s}} = \nu\mathbf{M}\mathbf{u}_{\mathrm{s}}. \tag{4.82}$$

Let $\nu_j(n)$ be a simple eigenvalue of (4.82) and $\mathbf{v}_j[n]$ a corresponding eigenvector, scaled such that

$$(\mathbf{v}_j[n])^\top\mathbf{M}\mathbf{v}_j[n] = 1. \tag{4.83}$$

Clearly, all eigenvalues of a matrix are isolated and applying the results from Section 4.3.1, we can represent $\mathbf{u_s}[\nu, n]$ by the following sum[30]:

$$\mathbf{u_s}[\nu, n] = \frac{-1}{\nu - \nu_j(n)} \mathbf{P}_j(n)\mathbf{f}[\nu_j(n), n] + \sum_{m=0}^{\infty} (\nu - \nu_j(n))^m \mathbf{S}_j(n)^{m+1}\mathbf{f}[\nu_j(n), n], \tag{4.84}$$

with the reduced resolvent[31]

$$\mathbf{S}_j(n) := \frac{1}{2\pi i} \int_{\Gamma_j(n)} (\nu - \nu_j(n))^{-1} \left(\mathbf{M}^{-1}\mathbf{B}(n) - \nu\mathbf{I}\right)^{-1} \mathrm{d}\nu \tag{4.85}$$

and the eigenprojection

$$\mathbf{P}_j(n) := -\frac{1}{2\pi i} \int_{\Gamma_j(n)} \left(\mathbf{M}^{-1}\mathbf{B}(n) - \nu\mathbf{I}\right)^{-1} \mathrm{d}\nu. \tag{4.86}$$

As explained above (see equation (4.75)), the eigenprojection is given by

$$\mathbf{P}_j(n) = \mathbf{v}_j[n](\mathbf{M}\mathbf{v}_j[n])^\top, \tag{4.87}$$

and as a first order approximation to $\mathbf{u_s}[\nu, n]$ in the vicinity of the resonance $\nu_j(n)$, we use the formula

$$\mathbf{u_s}[\nu, n] \approx \frac{-1}{\nu - \nu_j(n)} \mathbf{v}_j[n]\mathbf{v}_j[n]^\top \mathbf{M}\mathbf{f}[\nu_j(n), n]. \tag{4.88}$$

(4.88) is a discrete analog to (4.73). Before we proceed, we want to emphasize by an example what we have derived here.

4.3.3 Example

Let us return to our motivating example from Section 2.2.1. Figure 4.1 shows the same plot as Figure 2.3, but here we have also plotted the numerically computed[32] resonances (blue crosses in the lower panel) and an approximation of the maximal field intensity in the vicinity of the different resonances (red lines in the lower panel) which was computed from approximation (4.88) of the scattered field.

The results show how nice the resonant frequencies meet the points of maximal field enhancement, and also the approximation formula fits the observed curves around the resonant frequencies very well, in particular the values at the resonant frequencies themselves.

---

[30] The reader should keep in mind that this holds for all n in a neighborhood of some initial $\check{n}$.

[31] $\Gamma_j(n)$ is again a rectifiable, simple closed curve enclosing $\nu_j(n)$ but no other generalized eigenvalues of (4.82).

[32] The computations were done with an averaged length of 50Å per finite element, a polynomial degree of 10 and 50 degrees of freedom in each of the Hardy spaces for the modelling of the two radiation conditions.

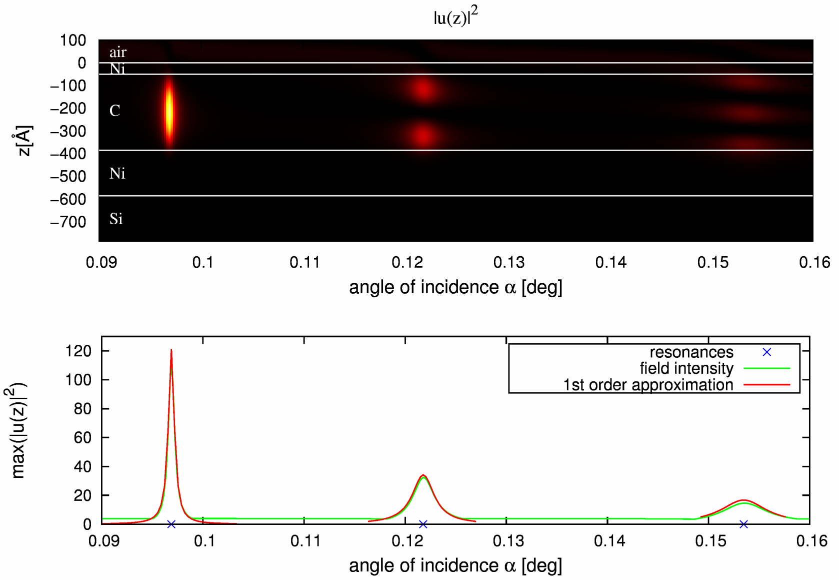


Figure 4.1: *upper panel: field intensity along the z-axis for different angles of incidence α, lower panel: maximum field intensity for different angles of incidence α, resonances and approximation of field intensity using formula (4.88); values: Table 2.1*

4.3.4 Discrete objective function

The approximation formula (4.88) leads as in Section 4.3.1 to an approximation of the L^2-norm of the solution in $[-a, 0]$ in dependence of ν in the vicinity of the resonance $\nu_j(n)$:

$$\begin{aligned}\|u_{\mathrm{s}}[\nu,n]\|_{L^2}^2 &\approx \underline{\mathbf{u}}_{\mathbf{s}}[\nu,n]^*\underline{\mathbf{M}}\;\underline{\mathbf{u}}_{\mathbf{s}}[\nu,n] = \overline{\underline{\mathbf{u}}_{\mathbf{s}}[\nu,n]}^\top\underline{\mathbf{M}}\;\underline{\mathbf{u}}_{\mathbf{s}}[\nu,n] \\ &\approx \frac{1}{|\nu-\nu_j(n)|^2}\left|\underline{\mathbf{v}_j(n)}^\top\underline{\mathbf{M}\,\mathbf{f}[\nu_j(n),n]}\right|^2\overline{\underline{\mathbf{v}}_j[n]}^\top\underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[n]. \qquad (4.89)\end{aligned}$$

The underlined vectors shall indicate that we use there only terms from the interior and no entries corresponding to the Hardy space parts. Motivated by equation (4.79) we rewrite this to[33]:

$$\|u_{\mathrm{s}}[\nu,n]\|_{L^2}^2 \approx \frac{4\,|k_0^2-\nu_j(n)|}{|\nu-\nu_j(n)|^2}\left|\underline{\mathbf{v}}_j[n](0)\right|^2\underline{\mathbf{v}}_j[n]^*\underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[n]. \qquad (4.90)$$

By this formula we can make our optimization problem more accessible. Choosing j such that $\nu_j(\check{n})$ is the best resonance for some initial system with $n = \check{n}$, we will use the following formulation in the discrete setting:

---

[33] The expression $\underline{\mathbf{v}}_j[n](0)$ has to be interpreted as the entry of the j-th (discrete) eigenvector $\underline{\mathbf{v}}_j$ which corresponds to the position $z = 0$. This degree of freedom exists independently of the finite elements discretization because $z = 0$ is a boundary point of our interior domain.

Optimization problem 4.17.

$$\max_{n} \mathfrak{f}(n) \text{ under side conditions on } n,$$

with

$$\mathfrak{f}(n) := \frac{4\,|k_0^2 - \nu_j(n)|}{|\mathrm{Im}(\nu_j(n))|^2} \left|\underline{\mathbf{v}}_j[n](0)\right|^2 \underline{\mathbf{v}}_j[n]^* \underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[n]. \tag{4.92}$$

With (4.92) we have found a suitable objective function defined by a handy formula. By the results from the preceding sections, it is also differentiable when $\mathbf{B}(n)$ depends differentiably on n.

Remark 4.18. *To Optimization Problem 4.17 we remark the following:*

1. *It is not sufficient to minimize the absolute value of the imaginary part of the resonance in dependence of n since in (4.90) not only the eigenvalue $\nu_j(n)$ changes with n, but also the eigenvectors which influence the other terms. These terms should be as big as possible while the imaginary part of $\nu_j(n)$ gets small. In particular, the term $\underline{\mathbf{v}_j[n]}^\top \underline{\mathbf{Mf}[\nu_j(n), n]}$ in (4.89) would be zero if the excitation $\underline{\mathbf{Mf}[\nu_j(n), n]}$, corresponding to the incident field, was orthogonal to $\overline{\underline{\mathbf{v}_j[n]}}$. But this can be excluded as already discussed in Section 4.3.1 (see equation (4.78a)). Nevertheless, the term can get small when n changes and has to be taken into account.*

2. *One may further be interested in the width of the peak around $\mathrm{Re}(\nu_j(n))$, produced by formula (4.88). We ask for which value $\nu = \nu_{1/p} \in \mathbb{R}$ only $1/p$ of its maximum value at $\mathrm{Re}(\nu_j(n))$ is left. An elementary computation shows:*

$$\nu_{1/p} = \sqrt{p-1}\,|\,\mathrm{Im}(\nu_j(n))|. \tag{4.93}$$

   Thus, better resonances will in general produce more narrow peaks, as already mentioned. Note here that the cosine is strictly monotonically increasing in $[0, \pi/2]$ such that the peak is also more narrow if we plot it against α as in the previous examples.

4.4 Derivative of the discrete objective function

4.4.1 Existing work

There exists a large literature on derivatives of eigenvalues and optimality conditions for certain objective functions involving eigenvalues since eigenvalue optimization problems arise in many areas (e.g. optimal design problems and problems of optimal control) as already shortly described in the introduction to this thesis. Of course, it was our first idea to carry over these results to our problem and apply them to the discrete version of our problem. However, it turned out that most of the results are not applicable here as they are restricted to symmetric matrices, special dependencies on the parameters or other objective functions. Usually they also do not cover derivatives of eigenvectors. We give a brief overview which may be incomplete and subjective. For further details, we refer to the articles we mention and the references therein. Mainly motivated by applications from control theory there are plenty of articles (for example [Ove92] by *Overton* and [LO96] by *Lewis and Overton*) dealing with the problem of a matrix depending on parameters and optimizing some objective function involving its eigenvalues. For instance, it may be desirable to minimize the largest[34] eigenvalue of a symmetric matrix in dependence on some parameters to control stability of a certain system (see [Ove92]). Most of the literature is restricted to symmetric or Hermitian matrices like the article [Ove92]. As we have seen, this assumption is not fulfilled in our problem. Let us briefly explain the main difference. In the case of symmetric or Hermitian matrices the largest eigenvalue is a convex function of the matrix entries. More precisely, one can show that the maximum eigenvalue may be written as the pointwise maximum of a set of linear functions which always defines a convex function. The subdifferential of a convex function $f : \mathbb{R}^k \to \mathbb{R}$ at a point $x_0 \in \mathbb{R}^k$ is defined as

$$\partial f(x_0) = \left\{ g \in \mathbb{R}^k : f(x) - f(x_0) \geq g^\top (x - x_0) \text{ for all } x \in \mathbb{R}^k \right\}. \tag{4.94}$$

It turns out (see *Overton* [Ove92, Theorem 1 and 2]) that for a symmetric matrix $\mathbf{B}$ the subdifferential of the largest eigenvalue μ_1 with respect to the matrix entries is given by

$$\partial \mu_1(\mathbf{B}) = \operatorname{conv} \left\{ \mathbf{b}\mathbf{b}^\top : \mathbf{b} \text{ is a normalized eigenvector to } \mu_1(\mathbf{B}) \right\}, \tag{4.95}$$

where conv denotes the convex hull and identifying $\mathbb{R}^{k\times k}$ and $\mathbb{R}^{k^2}$ in the canonical way. This leads to formulas for generalized derivatives of the largest eigenvalue when the matrix is assumed to depend continuously differentiably on the parameters. One makes use of so-called generalized gradients introduced by *Clarke* [Cla83] and expressions for the generalized derivatives follow from a chain rule (see *Lewis and Overton* [LO96, Theorem 3]). Whether the dependence of the matrix on the parameters is linear or nonlinear is not so important here as long as it depends differentiably on them. Note that the calculation

---

[34] The eigenvalues may be ordered in that case as they are all real since the considered matrix-valued function is assumed to map on symmetric matrices.

of (4.95) at many different points is not altogether easy whenever the largest eigenvalue is not simple since it requires a complete set of orthonormal eigenvectors corresponding to the largest eigenvalue (see e.g. *Overton* [Ove92]).

But why are such results not applicable to our problem? First of all and most important, we do not have symmetric/Hermitian matrices. The results for standard eigenvalue problems with symmetric/hermitian matrices would directly carry over to generalized eigenvalue problems of the form $\mathbf{B}(n)\mathbf{u_s} = \nu\mathbf{M}\mathbf{u_s}$ if $\mathbf{B}$ and $\mathbf{M}$ are assumed to be symmetric/Hermitian and moreover $\mathbf{M}$ is assumed to be positive semidefinite (see e.g. [Ove92, Section 7]). But none of these assumptions is fulfilled in our case because of the entries in the matrices which come from the radiation conditions or rather the Hardy space infinite elements. In the case of non-symmetric matrices the situation for points where the eigenvalues coalesce is much worse. What is particularly lost in the non-symmetric/non-Hermitian case is the existence of a complete orthonormal system with respect to the standard scalar product or to the one induced by $\mathbf{M}$ in the generalized eigenvalue problem respectively. This goes along with the loss of the standard Rayleigh representation for the largest eigenvalue.

To clarify this point, we consider again a small example. The matrix-valued function

$$\mathbf{B}(q) = \begin{pmatrix} 0 & 1 \\ -1 & 2q \end{pmatrix} \tag{4.96}$$

depends analytically on q in every subdomain of $\mathbb{R}$ but $\mathbf{B}(q)$ is not symmetric. Its eigenvalues are given by $q \pm \sqrt{q^2-1}$ for all $q \in \mathbb{R}$ but one cannot define two independent functions $\mu_1(q), \mu_2(q)$ for the eigenvalues which are differentiable at $q = 1$. Even worse, the function

$$\operatorname{Im}(\mu_j(q)) = \begin{cases} 0, & q \geq 1 \\ \pm\sqrt{1-q^2}, & q < 1 \end{cases} \tag{4.97}$$

is not only non-differentiable at $q = 1$, it is even not a Lipschitz function anymore such that the theory of generalized gradients also fails. But such things only happen at points where eigenvalues coalesce. We assumed simple eigenvalues anyway. However, it is also possible to work out variational properties of certain objective functions at points where two eigenvalues coalesce but then we need knowledge on the Jordan form of the matrix at these points (see e.g. *Burke, Lewis and Overton* [BLO00] and *Burke and Overton* [BO01]) and the theory gets quite complicated.

In general, there are much less articles dealing with non-symmetric matrices. *Lewis and Overton* [LO96] show that many eigenvalue optimization problems can be rephrased into semidefinite programming (SDP) which they analyze from a Fenchel duality perspective and can be solved by primal-dual interior points algorithms. But again the analysis heavily relies on the assumption of symmetric/Hermitian matrices. The suggestions made in Section 16 of [LO96] to rephrase non-Hermitian problems into Hermitian ones have serious drawbacks like many extra variables or even ill-conditioning. Another very general article by *Burke and Overton* [BO94] works out variational properties of the spectral radius/spectral abscissa[35] for the general case that the blocks corresponding to the considered eigenvalues may have any Jordan structure, but only under the stronger assumption that the examined matrix depends analytically on the parameters. The article most closely related to our purpose is a quite old one by *Overton and Womersley* from 1988 [OW88], providing necessary and sufficient conditions for the spectral radius to have a first-order local minimum, for the case of a nonsymmetric real-valued matrix depending affinely on a real parameter and assuming that all eigenvalues achieving the minimum are semisimple at the point where it is achieved. They make use of the eigenvalue perturbation theory by *Rellich* [Rel69] and *Kato* [Kat95] to find directional derivatives. This theory is also the basis for our considerations. To overcome the assumption of an affine-linear dependence on the parameters they suggest to locally expand the matrix into a Taylor series and to truncate it as an approximation to the exact matrix. Since such an advance may produce additional errors and requires higher order derivatives of the matrix-valued function, we do not follow it.

Another crucial point, we have already touched upon, is the objective function. Our objective function not only involves eigenvalues as in most of the articles (largest or smallest eigenvalue: *Overton* [Ove92], spectral radius: *Burke and Overton* [BO94] and *Overton and Womersley* [OW88], spectral abscissa: *Burke and Overton* [BO94] and [BO93]) but also eigenvectors, and thus we also need their derivatives. Minimizing the spectral abscissa, which is the maximal real part of the eigenvalues of a matrix-valued function, seems at first sight to be closely related to a reasonable objective function for our problem, namely minimizing the minimal imaginary part of the eigenvalues. But to focus on the imaginary part of the best resonance in the optimization process is not sufficient as explained above and moreover we would have to deal with the problem that some artificial resonances coming from the discretization might have smaller imaginary parts and we would then optimize physically meaningless values. Or to put it in another way, we have to analyze the eigenvalues which are true resonances and not necessarily the ones with smallest imaginary part or biggest real part (corresponding to smallest angle of incidence) computed from our discretized problem.

A basis for algorithms with very general objective functions and dependencies on the parameters (like the non-Lipschitz function above) can be found for example in *Burke, Overton and Lewis* [BLO02] and *Vanbiervliet et al.* [VVMV08]. *Burke, Lewis and Overton* [BLO02] suggest the use of so-called “gradient sampling” for problems where the gradient or respectively the subdifferential is very hard to compute. The gradient at a point is approximated by surrounding points where the gradient is easy to calculate under the assumption that the considered function is differentiable almost everywhere. Such techniques require in general many function evaluations but are on the other hand often the only way to allow for practicable algorithms for very general

---

[35] The spectral abscissa of a matrix-valued function $q \mapsto \mathbf{B}(q)$ for q in a subdomain of $\mathbb{R}^k$ is defined as $\max\{\mathrm{Re}(\mu) : \mu \text{ is an eigenvalue of } \mathbf{B}(q)\}$ and the spectral radius is defined as $\max\{|\mu| : \mu \text{ is an eigenvalue of } \mathbf{B}(q)\}$.

problems. But since under suitable assumptions it is possible to compute the derivative of our objective function directly, we do not make use of the existing algorithms for efficiency reasons and more reliable results using the exact derivative.

We have only given a short overview of the work by *Overton et al.* There is a tremendous number of further nice publications concerning optimization problems with eigenvalues and related topics. A list of those can be found on Overton's homepage *http://cs.nyu.edu/overton/papers.html.*

4.4.2 Application of results from Section 4.2

To reach a fully discrete formulation of our optimization problem we also have to discretize n. We do this before we discretize the problem with finite elements to avoid inconveniences. Let us assume that we can encode n in a vector

$$\mathbf{n} = [\tilde{\mathrm{n}}_1, \tilde{\mathrm{n}}_2, \ldots, \tilde{\mathrm{n}}_L] \in \mathbb{R}^L, \tag{4.98}$$

which is for example the case when the refractive index n is approximated by a piecewise constant function or by a spline of higher order. In the case of a piecewise constant approximation the vector $\mathbf{n}$ can consist of refractive indices[36] and/or layer widths. The air that surrounds the system and the substrate are kept fix as well as the width of the whole system and are not included in $\mathbf{n}$. In the case of higher order approximations the vector $\mathbf{n}$ consists of the coefficients of higher order spline which approximates n in $[-a, 0]$ (see Section 5.2 for details).

To compute the derivatives of our objective function

$$\mathfrak{f}[\mathbf{n}] = \frac{4\,|k_0^2 - \nu_j(\mathbf{n})|}{|\mathrm{Im}(\nu_j(\mathbf{n}))|^2}\,|\mathbf{v}_j[\mathbf{n}](0)|^2\,\overline{\underline{\mathbf{v}}_j[\mathbf{n}]}^\top \underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[\mathbf{n}], \tag{4.99}$$

we need the derivatives of the generalized simple eigenvalue $\nu_j(\mathbf{n})$ and the corresponding scaled eigenvector of

$$\mathbf{B}(\mathbf{n})\mathbf{u_s} = \nu\mathbf{M}\mathbf{u_s}. \tag{4.100}$$

with the complex-symmetric (non-Hermitian) matrices $\mathbf{B}(\mathbf{n})$ and $\mathbf{M}$. This is a special case of the problem discussed in Section 4.2 if we assume that $\mathbf{n} \mapsto \mathbf{B}(\mathbf{n})$ depends continuously differentiably[37] on $\mathbf{n}$. To present the result clearly arranged, we define the functions

$$\begin{aligned} \mathfrak{q}(\mathbf{n}) &:= \left|\underline{\mathbf{v}}_j[\mathbf{n}](0)\right|^2, & \mathfrak{r}(\mathbf{n}) &:= \underline{\mathbf{v}}_j[\mathbf{n}]^*\underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[\mathbf{n}], \\ \mathfrak{s}(\mathbf{n}) &:= |\mathrm{Im}(\nu_j(\mathbf{n}))|^2, & \mathfrak{t}(\mathbf{n}) &:= 4\left|k_0^2 - \nu_j(\mathbf{n})\right| \end{aligned} \tag{4.101}$$

---

[36] We can put real and imaginary part as independent variables into the $\mathbf{n}$, but as we will later see, it makes more sense to couple the absorption to the real part of the refractive index (see Section 5.1).

[37] That $\mathbf{n} \mapsto \mathbf{B}(\mathbf{n})$ depends continuously differentiably on $\mathbf{n}$ will be discussed for different types of discretizations of the refractive index in Section 4.5 and Chapter 5.

and their l-th partial derivatives we denote by

$$\mathfrak{g}_l(\mathbf{n}) := \frac{\partial \mathfrak{g}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n}), \quad \mathfrak{g} = \mathfrak{q}, \mathfrak{r}, \mathfrak{s}, \mathfrak{t}. \tag{4.102}$$

In this notation $\mathfrak{f}$ is given by

$$\mathfrak{f}(\mathbf{n}) = \frac{\mathfrak{t}(\mathbf{n})\mathfrak{q}(\mathbf{n})\mathfrak{r}(\mathbf{n})}{\mathfrak{s}(\mathbf{n})} \tag{4.103}$$

and has the gradient

$$\nabla \mathfrak{f} = \left[\frac{1}{\mathfrak{s}^2}\Big(\big[(\mathfrak{t}_l\mathfrak{q} + \mathfrak{t}\mathfrak{q}_l)\,\mathfrak{r} + \mathfrak{t}\mathfrak{q}\mathfrak{r}_l\big]\mathfrak{s} - \mathfrak{t}\mathfrak{q}\mathfrak{r}\mathfrak{s}_l\Big)\right]_{l=1,\ldots,L}. \tag{4.104}$$

Computing the partial derivatives of $\mathfrak{q}, \mathfrak{r}, \mathfrak{s}$ and $\mathfrak{t}$ completes the derivative of the objective function. We obtain

$$\begin{aligned}
\mathfrak{q}_l(\mathbf{n}) &= 2\operatorname{Re}\left(\underline{\mathbf{v}}_j[\mathbf{n}](0)\right)\operatorname{Re}\left(\frac{\partial \mathbf{v}_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}](0)\right) + 2\operatorname{Im}\left(\underline{\mathbf{v}}_j[\mathbf{n}](0)\right)\operatorname{Im}\left(\frac{\partial \mathbf{v}_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}](0)\right),\\
\mathfrak{r}_l(\mathbf{n}) &= \underline{\left(\frac{\partial \mathbf{v}_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}]\right)}^{*}\underline{\mathbf{M}}\left(\underline{\mathbf{v}}_j[\mathbf{n}]\right) + \left(\underline{\mathbf{v}}_j[\mathbf{n}]\right)^{*}\underline{\mathbf{M}}\,\underline{\left(\frac{\partial \mathbf{v}_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}]\right)},\\
\mathfrak{s}_l(\mathbf{n}) &= 2\operatorname{Im}\left(\nu_j(\mathbf{n})\right)\operatorname{Im}\left(\frac{\partial \nu_j}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right) \quad \text{and}\\
\mathfrak{t}_l(\mathbf{n}) &= \frac{-2}{|k_0^2 - \nu_j(\mathbf{n})|}\left(2\operatorname{Re}\left(k_0^2 - \nu_j(\mathbf{n})\right)\operatorname{Re}\left(\frac{\partial \nu_j}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right) + \mathfrak{s}_l(\mathbf{n})\right).
\end{aligned} \tag{4.105}$$

Note that we have differentiated real and imaginary part of the functions separately as functions from $\mathbb{R}^L$ to $\mathbb{R}$ and that we are away from zero due to our assumptions[38] on n (or respectively $\mathbf{n}$) and ν such that the norms and absolute values are differentiable.

What is left to do, is to write down formulas for the derivatives of the eigenvalues and eigenvectors. Applying the results from Section 4.2 we find for the partial derivative of the eigenvalue $\nu_j(\mathbf{n})$

$$\frac{\partial \nu_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}] = \mathbf{v}_j[\mathbf{n}]^\top \left(\frac{\partial \mathbf{B}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right)\mathbf{v}_j[\mathbf{n}] \tag{4.106}$$

and for the partial derivatives of the eigenvector $\mathbf{v}_j[\mathbf{n}]$

$$\begin{aligned}
\frac{\partial \mathbf{v}_j}{\partial \tilde{\mathrm{n}}_l}[\mathbf{n}] = &-\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})\left(\frac{\partial \mathbf{B}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right)\mathbf{v}_j[\mathbf{n}] + \frac{1}{2}\mathbf{v}_j[\mathbf{n}]\left(\left[\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})\left(\frac{\partial \mathbf{B}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right)\mathbf{v}_j[\mathbf{n}]\right]^\top \mathbf{M}\mathbf{v}_j[\mathbf{n}]\right.\\
&\left. + \mathbf{v}_j[\mathbf{n}]^\top \mathbf{M}\left[\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})\left(\frac{\partial \mathbf{B}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n})\right)\mathbf{v}_j[\mathbf{n}]\right]\right),
\end{aligned} \tag{4.107}$$

---

[38] We assumed $\operatorname{Re}(\nu) \in [0, k_0^2)$ and by the second part of Theorem 2.12 the imaginary part of the resonances is positive. Also $\underline{\mathbf{v}}_j[\mathbf{n}](0) \neq 0$ as the value it approximates is not equal to zero by part b) of Lemma 2.13.

with

$$\mathbf{S}_j^{\mathbf{M}}(\mathbf{n}) = \Big(\mathbf{B}(\mathbf{n}) - \nu_j(\mathbf{n})\mathbf{M} + \mathbf{M}\mathbf{v}_j[\mathbf{n}]\mathbf{v}_j[\mathbf{n}]^\top\mathbf{M}\Big)^{-1} - \mathbf{v}_j[\mathbf{n}]\mathbf{v}_j[\mathbf{n}]^\top \tag{4.108}$$

and always using the normalization

$$\mathbf{v}_j[\mathbf{n}]^\top\mathbf{M}\mathbf{v}_j[\mathbf{n}] = 1. \tag{4.109}$$

For the representation (4.108) of $\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})$, recall Lemma (4.14) and the equations (4.69) and (4.56).

Remark 4.19. *Observe that we mainly need the the partial derivatives with respect to* $\mathbf{n}$ *of* $\mathbf{n} \to \mathbf{B}(\mathbf{n})$ *and the matrix* $\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})$ *to compute the gradient of the objective function (see formula (4.104)) or respectively the derivatives of the eigenvalues and eigenvectors. The best eigenvalue and its corresponding eigenvector have to be computed anyways, when we solve the resonance problem. But since we do not need any other eigenvalues or eigenvectors, we can use numerical methods which compute only this or a few eigenpairs. Moreover, only the application of the matrix* $\mathbf{S}_j^{\mathbf{M}}(\mathbf{n})$ *to vectors is required in the partial derivatives (4.107) of the scaled eigenvectors. By (4.108) we basically have to solve a linear system for this. If we also implement the partial derivatives of* $\mathbf{n} \to \mathbf{B}(\mathbf{n})$ *by their application to vectors, the evaluation of the derivative can be done in a very efficient way. This is in particular true for the derivative with respect to layer changes which we discuss now.*

4.5 Derivative with respect to layer changes

As already mentioned in a side note, we are not only interested in the optimal choice of the refractive indices for a fixed number of layers, but also in the optimal thicknesses for given materials (or respectively given refractive indices). If we model n by a spline of higher order, this is done implicitly, but let us consider here the idealized setting of a piecewise constant approximation of the refractive index n. We want to examine the dependence of the objective function on the positions of layer change in this setting while the index of refraction in the layers is kept fix. This means the vector $\mathbf{n}$ in the objective function (4.99) does not encode the indices of refraction, but the positions of layer change now. To avoid confusion in the notation we rather denote the vector for the positions of layer change by $\mathbf{s}$ instead of $\mathbf{n}$.

We keep the width of the whole system fixed, i.e. $-a$ and 0 are kept fix, and encode all other positions, where the refractive index changes, in the vector $\mathbf{s} := [\mathrm{s}_1, \dots, \mathrm{s}_L] \in \mathbb{R}^L$. L layer changes correspond to a system of $L+1$ layers. The set of all admissible $\mathbf{s}$ we denote by $\mathfrak{S}$, and it is determined by the conditions

$$-a < \mathrm{s}_{l+1} < \mathrm{s}_l < 0, \qquad \text{for } l = 1, \dots, L-1. \tag{4.110}$$

Define the mapping

$$\mathfrak{S} \subset \mathbb{R}^L \to \mathfrak{N} \subset L^\infty([-a, 0]), \qquad \mathbf{s} \mapsto n[\mathbf{s}], \tag{4.111}$$

which maps the position vector $\mathbf{s}$ to the corresponding admissible piecewise constant refractive index with inner layer changes at the positions $\mathbf{s}$. This means

$$n[\mathbf{s}](z) := \begin{cases} 1, & z > 0 \\ n_1, & \mathrm{s}_1 < z \leq 0 \\ n_2, & \mathrm{s}_2 < z \leq \mathrm{s}_1 \\ \dots & \\ n_l, & \mathrm{s}_l < z \leq \mathrm{s}_{l-1} \\ \dots & \\ n_{L+1}, & -a < z \leq \mathrm{s}_L \\ n_{\mathrm{sub}}, & z \leq -a, \end{cases} \tag{4.112}$$

with fixed $n_l \in \mathbb{C}$, $l = 1, \dots, L+1$.
As explained, our objective reads now

$$\mathfrak{f}(\mathbf{s}) := \frac{4\,|k_0^2 - \nu_j(\mathbf{s})|}{|\mathrm{Im}(\nu_j(\mathbf{s}))|^2}\, |\mathbf{v}_j[\mathbf{s}](0)|^2\, \underline{\mathbf{v}}_j[\mathbf{s}]^* \underline{\mathbf{M}}\, \underline{\mathbf{v}}_j[\mathbf{s}], \tag{4.113}$$

and the optimization problem in the fully discrete version reads

$$\max_{\mathbf{s}} \mathfrak{f}(\mathbf{s}) \text{ under the side conditions (4.110).}$$

The formulas from Section 4.4.2 can be applied if we can show continuous partial differentiability of the matrix-valued function $\mathbf{B} : \mathfrak{S} \mapsto \mathbb{C}^{J \times J}$ with

$$\mathbf{B}(\mathbf{s}) := \mathbf{B}(n[\mathbf{s}]) \tag{4.115}$$

for all $\mathbf{s} \in \mathfrak{S}$. Note that for a finite-dimensional matrix all norms are equivalent. Thus, from continuous partial differentiability of all entries follows continuous differentiability of every column and using the column-sum-norm the continuous differentiability of the whole matrix-valued function. But we have to be careful because the mapping in (4.111) is not differentiable with respect to $\mathbf{s}$ in the classical sense.
For the l-th partial derivative consider the point s_l which is the changing point between the l-th and the $(l+1)$-th layer. The part of $\mathbf{B}(\mathbf{s})$ corresponding to the stiffness matrix[39] does not change with s_l since the stiffness matrix does not depend on n. Also the Hardy space parts (which approximate the DtN operators) are not involved since we do not change the refractive index outside $(-a, 0)$ (In particular, we do not change starting and end point of the system where exterior domain and interior domain couple.). The part that changes, is the discrete version $\mathbf{K}(n)$ of the operator K given by the relation

$$\langle K(n)u, v\rangle_{H^1} \;=\; \int_{-a}^{0} -k_0^2 n^2 u\bar{v}\,\mathrm{d}z, \tag{4.116}$$

for all $u, v \in H^1([-a, 0])$.

---

[39] The stiffness matrix approximates the operator A given by the relation $\langle Au, v\rangle_{H^1} = \int_{-a}^{0} u'v'\,\mathrm{d}z$ for all $u, v \in H^1([-a, 0])$.

We are interested in the following limit

$$\lim_{h\to 0}\frac{1}{h}\left(\mathbf{B}(\mathbf{s}+h\mathbf{e}_l)-\mathbf{B}(\mathbf{s})\right)=\lim_{h\to 0}\frac{1}{h}\left(\mathbf{K}(\mathbf{s}+h\mathbf{e}_l)-\mathbf{K}(\mathbf{s})\right), \tag{4.117}$$

where $\mathbf{e}_l$ denotes the l-th standard basis vector of $\mathbb{R}^L$.
To avoid inconveniences in the notation we only discuss the case $h>0$, the case $h<0$ is completely analogous. In the matrix $\mathbf{K}(\mathbf{s})$ we have entries of the form

$$\mathbf{k}_{j,i}(\mathbf{s})=\langle K(n[\mathbf{s}])b_i,b_j\rangle_{H^1} \tag{4.118}$$

with finite element basis functions b_i,b_j with $i,j=1,\dots,\mathrm{DOF}_{\mathrm{FEM}}$, where $\mathrm{DOF}_{\mathrm{FEM}}$ denotes the number of degrees of freedom in the finite elements part. For every $h>0$ with $\mathrm{s}_{l+1}<\mathrm{s}_l+h<\mathrm{s}_{l-1}$ ($\mathrm{s}_0:=0$ and $\mathrm{s}_{L+1}:=-a$ if needed) it holds

$$(n[\mathbf{s}+h\mathbf{e}_l])^2-(n[\mathbf{s}])^2=c_l\chi_{[\mathrm{s}_l,\mathrm{s}_l+h]}, \tag{4.119}$$

where $c_l:=n_{l+1}^2-n_l^2\in\mathbb{C}$ denotes the jump between the l-th and the $(l+1)$-th layer and $\chi_{[\mathrm{s}_l,\mathrm{s}_l+h]}$ is the indicator function[40] of the interval $[\mathrm{s}_l,\mathrm{s}_l+h]$. We compute

$$\begin{aligned}\frac{\partial}{\partial \mathrm{s}_l}\mathbf{k}_{j,i}(\mathbf{s})&=\lim_{h\to 0}\frac{1}{h}\Big(\langle\mathbf{B}(\mathbf{s}+h\mathbf{e}_l)b_i,b_j\rangle_{H^1}-\langle\mathbf{B}(\mathbf{s})b_i,b_j\rangle_{H^1}\Big)\\&=\lim_{h\to 0}\frac{1}{h}\int_{-a}^{0}-k_0^2\left((n[\mathbf{s}+h\mathbf{e}_l])^2-(n[\mathbf{s}])^2\right)b_ib_j\,\mathrm{d}z\\&=\lim_{h\to 0}\frac{1}{h}\int_{-a}^{0}-k_0^2c_l\chi_{[\mathrm{s}_l,\mathrm{s}_l+h]}b_ib_j\,\mathrm{d}z\\&=-k_0^2c_l\lim_{h\to 0}\frac{1}{h}\int_{\mathrm{s}_l}^{\mathrm{s}_l+h}b_ib_j\,\mathrm{d}z=-k_0^2c_lb_i(\mathrm{s}_l)b_j(\mathrm{s}_l),\end{aligned} \tag{4.120}$$

for all $i,j=1,\dots,\mathrm{DOF}_{\mathrm{FEM}}$ and all $l=1,\dots,L$, using the mean value theorem for integration (applied to real and imaginary part separately) and (4.119). The partial derivatives are all continuous in $\mathbf{s}$ since standard finite element basis function are continuous.
Moreover, the computed partial derivatives of $\mathbf{B}$ have the following property:

$$\mathbf{v}^*\left(\frac{\partial\mathbf{B}}{\partial \mathrm{s}_l}(\mathbf{s})\right)\mathbf{u}=-k_0^2c_l\mathbf{u}(\mathrm{s}_l)\overline{\mathbf{v}}(\mathrm{s}_l) \tag{4.121}$$

for all $\mathbf{u},\mathbf{v}\in\mathbb{C}^{\mathrm{J}\times\mathrm{J}}$, where $\mathbf{u}(\mathrm{s}_l)$ and $\overline{\mathbf{v}}(\mathrm{s}_l)$ have to be interpreted as the approximation of the function they represent, evaluated at the point $z=\mathrm{s}_l$.

REMARK 4.20. *1. The computation (4.120) can be done in the same way for the continuous operator $K(n[\mathbf{s}])$. This leads to derivatives with respect to $\mathbf{s}$ of the operator $B(n[\mathbf{s}])$ in the Hardy space formulation (2.20) or some other similar weak formulation. Again only the part which corresponds to $K(n)$ changes with $\mathbf{s}$. The partial derivatives also fulfill an analog to (4.121) for all continuous functions u,v.*

---

[40] The indicator function $\chi_{[\mathrm{s}_l,\mathrm{s}_l+h]}(z)$ is 1 for z inside the interval $[\mathrm{s}_l,\mathrm{s}_l+h]$ and 0 outside.

2. *As already pointed out in Remark 4.16, we are free to choose our discretization for a piecewise constant approximation of the refractive index in a way such that the changing point between two layers is always a boundary point of two neighboring finite elements. The choice of the discretization adapted to the current* $\mathbf{s}$ *does not cause any difficulties in the derivatives as, apart from resolution and accuracy, the discrete approximations of resonances and resonance functions are independent of it. For the described discretization there is only one basis function which is not equal to zero at* $z = \mathrm{s}_l$*. This holds for Lagrange elements as well as for the finite elements we use, following [Sch98], because* $z = \mathrm{s}_l$ *is a nodal point. Hence, we can evaluate expression (4.121) by simply picking the values from the vectors* $\mathbf{u}$ *and* $\overline{\mathbf{v}}$ *which correspond to the position* s_l *and multiplying them by* $-k_0^2 c_l$*.*

5 Numerical results

In the preceding chapter we have provided the theory needed to set up a numerical optimization algorithm for the optimization problem described in Section 2. The derivatives of $\mathbf{s} \mapsto \mathbf{B}(\mathbf{s})$ with respect to the layer changes have already been computed in Section 4.5. What is left to compute, are the derivatives of $\mathbf{n} \mapsto \mathbf{B}(\mathbf{n})$ for other discretizations of n, encoded in $\mathbf{n}$. We discuss piecewise constant approximations of n and approximations by splines of higher order and explain how to deal with absorption effects in both cases. Afterwards we present our optimization results, first for piecewise constant approximations and then also for splines. At the end of this Chapter we modify our objective function to take a further aim (angular acceptance, see Section 5.3.6) into account.
The computations in the subsequent sections were all done with the following discretization parameters[41]:

averaged width per finite element: 25(Å)
polynomial degree for finite element basis: 10
degrees of freedom in the Hardy space for air: 20
degrees of freedom in the Hardy space for substrate: 70

Table 5.1: *discretization parameters*

The number of degrees of freedom in the Hardy space for the substrate has to be chosen higher to accurately cover two different cases. Dependent on the angle of incidence the resulting field may decrease exponentially in the substrate or keep on oscillating (cf. Section 2.1). The achieved accuracy with the discretization parameters from Table 5.1 is absolutely sufficient. On the one hand, for the considered (very small) angles of incidence the solution to the scattering problem (2.8) oscillates very slowly (at most two or three wavelengths) inside the multilayer system, and on the other hand in practice the angle of incidence can only be adjusted up to a precision of 0.0001° due to experimental limitations.

5.1 Introductory examples

Let us motivate our procedure and discuss some central features based on a few small examples. Recall the multilayer system with piecewise constant refractive index already considered in Sections 2.2.1 and 4.3.3 which consists of a carbon layer between two nickel layers on silicon substrate (cf. Table 2.1 for the exact values). The parameter, we want to vary first, is the location of the interface between the top layer (nickel) and the guiding layer (carbon) while all the other parameters are kept fix. This means an optimization of the top layer thickness while in particular the width of the complete system (585Å) is kept fixed. Figure 5.1 shows the effect on the three best resonances if we

---

[41] Moreover, in all our computations we used the wavelength $\lambda = 0.62$Å which corresponds to a photon energy of about 20keV.

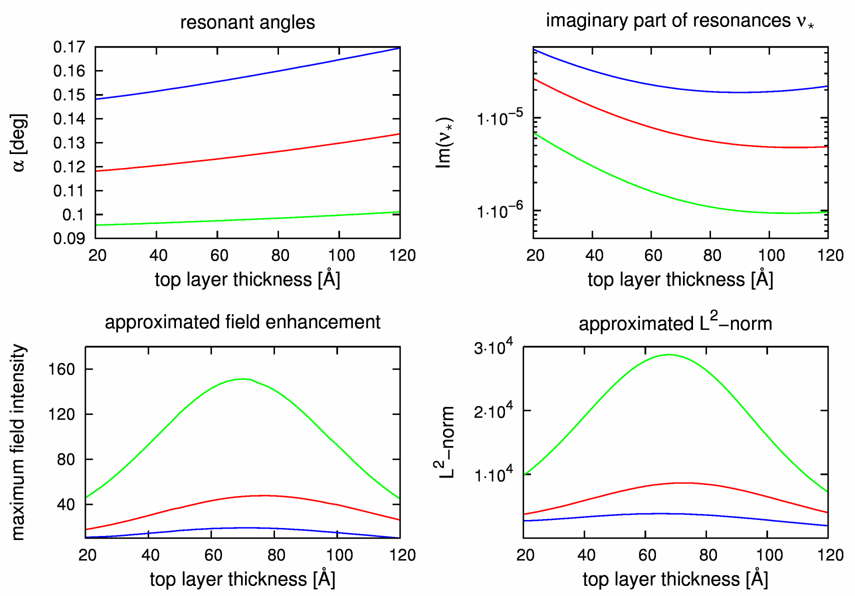


Figure 5.1: *top layer thickness varied from 20 to 120Å, absorption included; blue, red and green line show the results for the three best resonances of Ni-C-Ni system;* $\lambda = 0.62$*Å; refractive indices: Table 2.1*

vary the location of the interface between nickel and carbon from $z = 20$ to $z = 120$(Å). The upper left panel shows the dependence of the resonant angles[42] (computed as real parts of complex resonances) and the upper right panel the imaginary part of the resonances on a logarithmic scale. In the lower left panel the approximate field enhancement, given by formula (4.88), at the resonant frequencies $\mathrm{Re}(\nu_j)$ (or respectively the resonant angles) is plotted and in the lower right panel the approximation (4.90) to the L^2-norm at $\mathrm{Re}(\nu_j)$. The plots in the lower panels confirm our formulated conjecture that the L^∞-norm and the L^2-norm are approximately proportional and thus interchangeable in our problem. The use of the L^2-norm changes the values of the objective function, but the position of the optimum only slightly. In our objective function (4.113) we just use the best resonant frequency, i.e. the green line of the lower right panel, which has a readily identifiable maximum at about $z = 70$(Å) in this simple example.

A lot of comments and results stated in the previous chapters show up in this small example. In the upper panels of Figure 5.1 one may observe that the functions of eigenvalues do not coalesce and that it is not sufficient to minimize the imaginary part to achieve our aim to maximize the L^2-norm. Whereas the best resonance (green line) reaches the minimal imaginary part at about $z = 100$ and stays almost constant for larger z, the corresponding

---

[42] One should carefully note, if one compares the results to the discussion in [Pfe02, p.30], that not only the top layer thickness but also the guiding layer changes here.

field enhancement and the L^2-norm, shown in the lower panels, attain their maximum at about $z = 70$ and their values decrease significantly for larger top layer thicknesses. The decrease can be explained physically by absorption effects combined with a decreasing guiding layer thickness. Our observations also indicate that other terms than $(\mathrm{Im}(\nu_j))$ in (4.88) or respectively (4.90) vary for larger thicknesses and cannot be neglected.

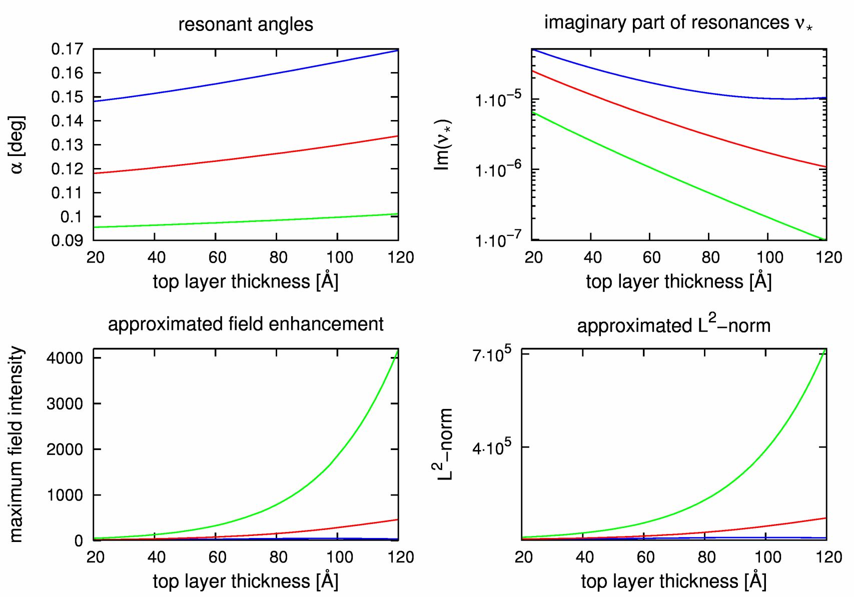
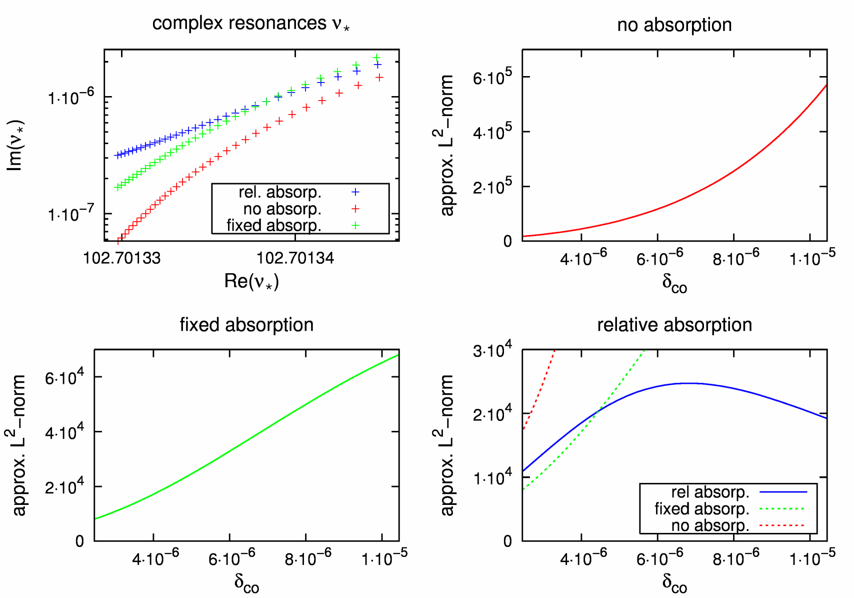


Figure 5.2: *top layer thickness varied from 20 to 120Å, absorption neglected; blue, red and green line show the results for the three best resonances of Ni-C-Ni system;* $\lambda = 0.62$*Å; refractive indices: Table 2.1*

One can also see that absorption effects cannot be neglected in general. This is illustrated in Figure 5.2, where the same quantities as in Figure 5.1 are shown, but absorption is neglected. The behavior is completely different. One may observe that the imaginary part of the best resonance (green line) now decreases more strongly and the two target values in the lower panels increase with the top layer thickness. Thus, for the L^2-norm of the solution at the best resonant frequency we have a maximum at the boundary of the considered interval of thicknesses. If we keep increasing the location of the interface between top and guiding layer to values larger than $z = 120$, field enhancement and L^2-norm will decrease from some point on. Note that this can already be observed for the weakest (blue line) of the three resonances in the considered interval. However, this is only due to the fact that we keep the width of the complete system fixed and therefore the guiding layer gets too small for a better resonance effect at some point. The increase of the top layer thickness with no decrease of the guiding layer was also examined under the given conditions. For this, we left the system width variable and found the following result. If we neglect absorption, the top layer thickness should go to infinity leading to an infinitely large field enhancement. But this result is useless in practice (in physical experiments), because absorption effects can indeed be reduced (using higher photon energies), but not be eliminated completely. In particular, for fixed energies we have to take absorption effects into account to get practically relevant results.


complex resonances ν_*

no absorption

fixed absorption

relative absorption

Figure 5.3: *dependence of resonances and resonant states on the refractive index in the coating for different treatments of absorption; red line: no absorption, green line: fixed absorption, blue line: relative absorption (5.1) with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$*; thicknesses:* 50/335/200Å*; guiding layer: air;* $\lambda = 0.62\mathring{A}$

In a second example, shown in Figure 5.3, the refractive index in the coating (the two nickel layers) was varied. The thicknesses (50/335/200Å) were kept fixed, and we replaced carbon by air in the guiding layer. Recall that the refractive index for x-rays can be written as $n = 1 - \delta + i\beta$ and $\delta_{\mathrm{air}} = 0$. The refractive index in the coating n_{co} was varied by the δ_{co} (from $2.5 \cdot 10^{-6}$ (nickel) to $10.5 \cdot 10^{-6}$), and the absorption term β_{co} was handled in three different ways:

1. no absorption: $\beta_{\mathrm{co}} = 0$

2. fixed absorption: ($\beta_{\mathrm{co}} = \beta_{\mathrm{Ni}} = 1.37 \cdot 10^{-7}$, independent of δ_{co})

3. relative absorption:

$$\beta_{\mathrm{co}} = c_{\beta/\delta}\delta_{\mathrm{co}}, \tag{5.1}$$

   with a constant $c_{\beta/\delta} \in \mathbb{R}$ describing the relation between β and δ.

The use of the relative absorption can be motivated by physical reasons: It is a commonly accepted approach in x-ray physics, e.g. in the selection of materials for x-ray optics, to use the ratio between absorption (β) and dispersion (δ) which dictates the performance. In the upper left panel of Figure 5.3 the change of the best resonance with the refractive index in the complex plane is shown for the three different absorption models. The other parts of the figure display our objective function (4.99) in dependence of δ_{co} for the three possibilities to handle absorption; for the relative absorption $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$ was chosen. One can observe that the real part of the resonances (i.e. the corresponding resonant frequency) is not changed by the different absorption models while the objective function changes dramatically. Note carefully the different scales in the subfigures. If we neglect absorption effects, the objective grows unbounded with δ_{co}, i.e. when the potential barrier is increased. Adding a fixed absorption in the coating changes the situation mainly quantitatively, which is clear as the layer thickness is kept fixed. The more realistic model of relative absorption changes the situation substantially. It is no longer optimal to choose δ_{co} as big as possible to maximize our objective function. The maximum is attained at about $\delta_{\mathrm{co}} \approx 7 \cdot 10^{-6}$. To compare the resulting objective function to the ones with other absorption models, we plotted them again as dashed lines in the lower right panel. The intersection between the blue and the green curve is of course δ_{Ni}. Note that the relative absorption has no effect on the guiding layer as $\delta_{\mathrm{air}} = 0$. Again it becomes obvious that the decrease of the imaginary parts of the resonances alone does not describe the behavior of the blue curve in an appropriate way.

To sum up, we have to take absorption effects into account, but we cannot model it as an independent variable since this would always lead to the smallest admissible absorption as these simple examples already show. A coupling of δ and β, in particular by the relative absorption introduced here, can help us to find the best compromise between the height of the potential barrier and low energy losses due to absorption.

5.2 Matrix derivatives

Having these facts in mind, we now want to compute derivatives of the discretized operator with respect to n since they are needed in the derivatives of the objective function we have obtained in Section 4.4. A piecewise constant approximation of the refractive index and B-splines of higher order will be used in the following. We always assume a fixed system width, i.e. 0 and $-a$ are the fixed starting and end point of the regarded systems.

So, let us assume that we are given an approximation to the refractive index which can be encoded in a vector $\mathbf{n} := [\tilde{\mathrm{n}}_1, \ldots, \tilde{\mathrm{n}}_L]$, i.e. a mapping

$$\mathfrak{N}_{\mathrm{dis}} \subset \mathbb{R}^L \to \mathfrak{N} \subset L^\infty([-a, 0]), \qquad \mathbf{n} \mapsto \tilde{n}[\mathbf{n}], \tag{5.2}$$

with $\mathfrak{N}$ the set of admissible refractive indices and $\mathfrak{N}_{\mathrm{dis}}$ the corresponding set of admissible values for $\mathbf{n}$. Let the mapping (5.2) further be continuously differentiable with respect to $\mathbf{n}$. Then as in Section 4.5, for the derivatives of the matrix-valued function $\mathbf{n} \mapsto \mathbf{B}(\mathbf{n}) \in \mathbb{C}^{J \times J}$ in (4.100), it suffices to consider the discrete version $\mathbf{K}(\mathbf{n})$ of the operator $K(\mathbf{n})$, given by the relation

$$\langle K(\mathbf{n})u, v\rangle_{H^1} \quad = \quad \int_{-a}^{0} -k_0^2 \left(\tilde{n}[\mathbf{n}]\right)^2 u\bar{v}\, \mathrm{d}z \tag{5.3}$$

for all $u, v \in H^1([-a, 0])$. The rest of $\mathbf{B}(\mathbf{n})$ does not change with $\mathbf{n}$. In the matrix $\mathbf{K}(\mathbf{n})$ we have again entries of the form

$$\mathbf{k}_{j,i}(\mathbf{n}) = \langle K(\mathbf{n})b_i, b_j\rangle_{H^1} \tag{5.4}$$

with finite element basis functions b_i, b_j with $i, j = 1, \ldots, \mathrm{DOF}_{\mathrm{FEM}}$, where $\mathrm{DOF}_{\mathrm{FEM}}$ denotes the number of degrees of freedom in the finite elements part of $\mathbf{B}(\mathbf{n})$. We deduce

$$\frac{\partial \mathbf{B}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n}) = \frac{\partial \mathbf{K}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n}), \quad l = 1, \ldots, L \tag{5.5a}$$

with

$$\frac{\partial k_{j,i}}{\partial \tilde{\mathrm{n}}_l}(\mathbf{n}) = \int_{-a}^{0} -2k_0^2 \tilde{n}[\mathbf{n}] \left(\frac{\partial \tilde{n}}{\partial \tilde{n}_l}[\mathbf{n}]\right) b_i b_j\, \mathrm{d}z, \quad i, j = 1, \ldots, \mathrm{DOF}_{\mathrm{FEM}}, \tag{5.5b}$$

and the partial derivatives are all continuous in $\mathbf{n}$. Hence, for the continuous differentiability of the mapping $\mathfrak{N}_{\mathrm{dis}} \to \mathbb{C}^{J \times J}$ with $\mathbf{n} \mapsto \mathbf{B}(\mathbf{n})$, the continuous differentiability of appropriate discretizations for n is left to be discussed.

5.2.1 Piecewise constant approximation

In the case of a piecewise constant approximation, the refractive index in $[-a, 0]$ is approximated by L layers in each of which the refractive index is constant. Starting from $z = 0$ we denote the (admissible) constant refractive values in the layers by $n_l = 1 - \delta_l + i\beta_l$, $l = 1, \ldots, L$ and encode them in a vector

$$\mathbf{n} = [\tilde{\mathrm{n}}_1, \tilde{\mathrm{n}}_2, \ldots, \tilde{\mathrm{n}}_{2L}] = [1 - \delta_1, \beta_1, 1 - \delta_2, \beta_2, \ldots, 1 - \delta_L, \beta_L] \in \mathbb{R}^{2L}. \tag{5.6}$$

For $z > 0$ we have by assumption $n \equiv 1$ and for $z < -a$ we have $n \equiv n_{\mathrm{sub}} = 1 - \delta_{\mathrm{sub}} + i\beta_{\mathrm{sub}}$. Both values are kept fixed as well as finite elements and Hardy space discretization.

As already pointed out in Section 5.1, we do not want to model the absorption as an independent variable since this would always lead to the minimal admissible absorption. Instead we model the absorption as a continuously differentiable function $\tilde{\beta} : \mathbb{R} \to \mathbb{R}$ of δ, i.e.

$$\beta_l = \tilde{\beta}(\delta_l), \quad l = 1, \ldots, L. \tag{5.7}$$

Therefore, we only have to analyze the dependence of

$$\mathbf{B}(\boldsymbol{\delta}) := \mathbf{B}\left(\left[1 - \delta_1, \tilde{\beta}(\delta_1), 1 - \delta_2, \tilde{\beta}(\delta_2), \ldots, 1 - \delta_L, \tilde{\beta}(\delta_L)\right]\right) \tag{5.8}$$

on $\boldsymbol{\delta} := [\delta_1, \ldots, \delta_L] \in \mathbb{R}^L$. The mapping

$$\tilde{n}[\boldsymbol{\delta}] = (\delta_1 + i\tilde{\beta}(\delta_1))\chi_{\mathrm{layer1}} + \ldots + (\delta_L + i\tilde{\beta}(\delta_L))\chi_{\mathrm{layer}L} \tag{5.9}$$

with the indicator functions χ_{layerl} for the layers depends continuously differentiable on $\boldsymbol{\delta}$. For a constant absorption in all layers (i.e. independent of δ_l) we obtain

$$\widetilde{\beta}(\delta_l) = \beta_{\text{all}} \in \mathbb{C} \quad \text{and} \quad \widetilde{\beta}'(\delta_l) = 0 \tag{5.10}$$

for all admissible δ_l. In Section 5.1 we have already discussed the possibility to choose the absorption proportional to δ. If we use the relative absorption

$$\widetilde{\beta}(\delta) = c_{\beta/\delta}\delta \tag{5.11}$$

with a constant $c_{\beta/\delta} \in \mathbb{C}$, we find $\widetilde{\beta}'(\delta) = c_{\beta/\delta}$.

Remark 5.1. *If we choose a discretization such that the refractive index only changes on intersection points (the layer thicknesses are assumed to be fixed in this examination) between two elements, the matrix* $\mathbf{B}(\mathbf{n})$ *in (4.100) is given by*

$$\begin{aligned} \mathbf{B}(\mathbf{n}) &= \mathbf{B}_0 + \mathrm{n}_1^2\mathbf{B}_1 + \mathrm{n}_2^2\mathbf{B}_2 + \ldots + \mathrm{n}_L^2\mathbf{B}_L \\ &= \mathbf{B}_0 + (\tilde{\mathrm{n}}_1 + i\tilde{\mathrm{n}}_2)^2\,\mathbf{B}_1 + (\tilde{\mathrm{n}}_3 + i\tilde{\mathrm{n}}_4)^2\,\mathbf{B}_2 + \ldots + (\tilde{\mathrm{n}}_{2L-1} + i\tilde{\mathrm{n}}_{2L})^2\,\mathbf{B}_L, \end{aligned} \tag{5.12}$$

where $\mathbf{B}_0$ *contains the Hardy space parts and all parts of our discretization that do not depend on* $\mathbf{n}$. *The matrix* $\mathbf{B}_l$ *contains the contribution of all finite elements in the* l*-th layer to the matrix* $\mathbf{B}(\mathbf{n})$, $l = 1, \ldots, L$. *For this discretization the partial derivatives of* $\mathbf{B}$ *with respect to* $\boldsymbol{\delta}$ *are simply*

$$\frac{\partial \mathbf{B}}{\partial \delta_l}(\boldsymbol{\delta}) = 2\left(1 - \delta_l + i\widetilde{\beta}(\delta_l)\right)\left(-1 + i\widetilde{\beta}'(\delta_l)\right)\mathbf{B}_l, \quad l = 1, \ldots, L. \tag{5.13}$$

Note that the derivative of $\mathbf{B}(\boldsymbol{\delta})$ *in the form (5.13) is rather easy to implement as it only involves matrices we have to compute anyways when we assemble* $\mathbf{B}(\boldsymbol{\delta})$.

5.2.2 Approximation by splines of higher order

For a more accurate approximation of the refractive index in $[-a, 0]$ we use B-splines (basis splines) of higher order. This allows us in particular to model continuous refractive indices more exactly. An introduction to B-splines can be found e.g. in [DR06] and [Kre98]. The B-splines Ψ_l are a special basis for spline spaces and have the advantage that one may compute the corresponding coefficients in a stable way. Furthermore, B-splines have the following properties:

1. positivity: $\Psi_l(z) \geq 0$ for all $z \in [-a, 0]$, $l = 1, \ldots, L$
2. partition of unity: $\sum_{l=1}^{L} \Psi_l(z) = 1$, $z \in [-a, 0]$
3. local support: Ψ_l is non-zero only in a subinterval of $[-a, 0]$, $l = 1, \ldots, L$,

where L is the dimension of a considered spline space (see below). From the local support it follows that local changes in the spline coefficients, change the resulting spline only locally. Especially in our application, changes of the refractive index in some subinterval shall not influence the refractive index globally. For a proof of the mentioned facts, see e.g. [dB78] and [dB90].

We want to approximate a refractive index $n = 1 - \delta + i\beta$, where δ and β are real-valued functions, by a spline. As in the case of a piecewise constant approximation, we write β as a function of δ, i.e. $\beta(z) = \widetilde{\beta}(\delta(z))$ with a continuously differentiable function $\widetilde{\beta} : \mathbb{R} \to \mathbb{R}$. We approximate the function δ by a spline of order m with respect to some fixed subdivision $-a = z_0 < z_1 < \ldots < z_r = 0,\ r > m$, i.e.

$$\delta(z) \approx \sum_{l=1}^{L} \delta_l \Psi_l(z), \quad z \in [-a, 0] \tag{5.14}$$

with $L := r + m$, the B-splines Ψ_l of order m and coefficients δ_l. Collecting the new variables in a vector $\boldsymbol{\delta} = [\delta_1, \delta_2, \ldots, \delta_L]$ we have discretized[43] δ. The corresponding refractive index is then approximated by

$$n(z) \approx \widetilde{n}[\boldsymbol{\delta}](z) := 1 - \sum_{l=1}^{L} \delta_l \Psi_l(z) + i\widetilde{\beta}\left(\sum_{l=1}^{L} \delta_l \Psi_l(z)\right), \tag{5.15}$$

and with respect to $\boldsymbol{\delta}$ we have the partial derivatives

$$\frac{\partial \widetilde{n}}{\partial \delta_l}[\boldsymbol{\delta}] = -\Psi_l(z) + i\widetilde{\beta}'\left(\sum_{l=1}^{L} \delta_l \Psi_l(z)\right) \Psi_l(z). \tag{5.16}$$

Observe that neither the B-splines nor the finite elements basis functions depend on $\boldsymbol{\delta}$. Thus, as above all the partial derivatives depend continuously on $\boldsymbol{\delta}$, and $\mathbf{B}$ is continuously differentiable with the derivative (5.5). In contrast to the piecewise constant approximation of n we have to assemble necessarily additional matrices for the partial derivatives of $\mathbf{B}$, but each of them has only very few entries since the B-splines have local support. Hence, for the l-th derivative it suffices to evaluate (5.5b) on elements which have a non-empty intersection with $\operatorname{supp}(\Psi_l)$.

As the B-splines are all positive and form a partition of unity, the box constraints

$$\delta_l^{\triangleright} \leq \delta_l \leq \delta_l^{\triangleleft}, \quad l = 1, \ldots, L \tag{5.17}$$

with the lower and upper bounds $\delta_l^{\triangleright}, \delta_l^{\triangleleft} \in \mathbb{R}$ guarantee that the spline (5.14) fulfills

$$\min_l \delta_l^{\triangleright} \leq \sum_{l=1}^{L} \delta_l \Psi_l(z) \leq \max_l \delta_l^{\triangleleft} \tag{5.18}$$

---

[43] Note here that one can also use other basis functions in (5.14) to approximate δ, provided that this leads to a differentiable dependence on the coefficients $\boldsymbol{\delta}$.

for all $z \in [-a, 0]$. Therefore, in an optimization algorithm we can prescribe realizable refractive indices using box constraints. In particular, we can control that the approximation of the refractive index is always below 1. Apart from physical admissibility, refractive indices above 1 would cause problems if we use for example relative absorption as the imaginary part of the refractive index gets negative then and does no longer act absorbing.

5.3 Optimization results

The question of an optimal system among all feasible systems is much too general. This is due to several reasons as already discussed partly in the introductory examples. There are only systems which are optimal in different situations.

Our optimization is done for a continuous range of refractive indices although only a discrete set of values is realizable practically because of the existing materials. Clearly, the optimization techniques we favor, are restricted to this assumption, but there is also a lot of practical motivation for this. First of all, the systems are very complicated to produce which leads to surface roughness. Therefore, a theoretically promising system cannot be reproduced perfectly either way. Moreover, one might think of the possibility to mix different materials to achieve a certain refractive index. And even if we cannot produce a certain optimal system, we get an impression of what is achievable under certain conditions.

Side conditions will always be necessary to avoid unrealistic refractive indices or infinite layers. As already shown in Section 5.1, higher potential barriers are often desirable and with an initial guess close to 1 the optimization algorithm might move to refractive indices bigger than 1 which is even more likely with the relative absorption model as the absorption effect is then inverted and enlarges the energy inside the system. But this is an unrealistic effect since there are no sources inside the system. Note in this context that the solution to the scattering problem (2.8) depends continuously on the refractive index n (see Chapter 3) and in none of our arguments we used that $\mathrm{Re}(n)$ is close to 1. The assumption $\mathrm{Re}(n) > 0$ was sufficient.

5.3.1 Optimization algorithms

With the objective function and its gradient we have provided all what is needed to apply the optimization algorithms (using derivatives of the objective function) we favor. It is not in the focus of this thesis to develop new optimization algorithms. Hence, we have rather implemented a routine which evaluates our objective function and its derivatives in the way explained above, involving resonances and resonance functions which are computed by the Hardy space method. This routine is handed over to the optimization toolbox of *MATLAB*[44] which we use for the optimization. We have mainly used a sequential quadratic programming (SQP) algorithm and an interior points algorithm. We only explain in a couple of words what is behind these algorithms. For details,

---

[44] A numerical computing environment developed by MathWorks, www.mathworks.de

we refer to the *MATLAB* documentation for the function *fmincon*[45] and the references we mention below.

- SQP: The algorithm seeks for a solution of the Karush–Kuhn–Tucker (KKT) conditions which are necessary conditions for optimality in a constrained optimization problem. In each major optimization step a constrained quadratic programming subproblem is solved. This subproblem results from a positive definite approximation of the Hessian of the Lagrangian function, using a quasi-Newton update formula, e.g. the Broyden–Fletcher–Goldfarb–Shanno (BFGS) update. The solution of the subproblem is used to form a search direction in an appropriate line search which is performed afterwards. For a detailed introduction to SQP, we refer to [Fle87], and for the basic version of the SQP-algorithm implemented in *fmincon*, we refer to [NW06, Chapter 18].
- interior points: The algorithm solves a family of minimization problems with equality constraints approximating the original problem. The solution of the approximating problems is either done by searching a solution to the KKT conditions of this problem using a linearization approach or if this is impossible, a conjugate gradient (CG) method is applied in a trust region. For a description of the complete algorithm, we refer to [BHN99].

5.3.2 Optimization of refractive indices for fixed layer thicknesses

We start with the presentation of optimization results for piecewise constant refractive indices. In this section we consider two examples where we only optimize the refractive index in a fixed number of layers, but not their thicknesses (cf. Section 5.2.1). The system is subdivided into L pieces with refractive indices $n_l = 1 - \delta_l + i\beta_l$. We use the box constraints

$$0 \leq \delta_l \leq 8 \cdot 10^{-6}, \quad \text{for } l = 1, \dots, L \tag{5.19}$$

and the relative absorption (5.11) with

$$c_{\beta/\delta} = \frac{\beta_{\text{Ni}}}{\delta_{\text{Ni}}} \approx 0.030786. \tag{5.20}$$

Figure 5.4 shows the results for a subdivision into 6 pieces. The exact values of the subdivision and the optimization results can be found in Table 5.2. Here and below, “true” in the tables has always to be understood as the immediate finite elements approximation, and by “approximated” we mean the values of the approximation formulas (4.88) and (4.89). A comparison of the values demonstrates the accuracy of the approximation formulas. As the initial value for the optimization process we used a profile where a vacuum layer is enclosed by silicon substrate ($n_{\text{sub}} = n_{\text{Si}} = 1 - 1.2 \cdot 10^{-6} + 4.56 \cdot 10^{-9} i$). Recall again that $n \equiv 1$ for $z > 0$ and $n \equiv n_{\text{sub}}$ for $z < -a$. In this example we chose $a = 550(\text{Å})$

---

[45] http://www.mathworks.com/help/toolbox/optim/ug/fmincon.html

to get an equidistant subdivision except for the first layer. Motivated by the standard Ni-C-Ni system we set the thickness of the first layer to 50. The initial profile and the field intensity of the best resonant frequency of the system are plotted in blue, and by the green line the same is displayed for the optimal profile. With the dashed red line we indicate the energy of the incident field needed to excite the best resonant frequency[46]. We use this color scheme in all of following optimization figures. The initial system in Figure 5.4 is a very flat potential well producing a field enhancement of about 20, whereas the optimal system achieves about 150, which is an improvement of approximately 30% compared to the standard system with nickel and carbon (field enhancement: 114). The calculations in this section were done using SQP (see Section 5.3.1), and the optimal solution mainly profits from higher potential barriers. But the constraint in the first layer is inactive, which is already a strong indication that variable layer thicknesses should be taken into account. This becomes even more evident if we refine the subdivision. Figure 5.5 presents the optimization result for a subdivision into 14 pieces. The optimal solution is not too different from the solution for 6 pieces concerning its shape the and achieved field enhancement. A more detailed analysis of the result suggests a thinner top layer. Another undesirable effect in the optimal solution of Figure 5.5 is the reduction of the system width by setting the first layer to air. Here, the optimization stuck in a local minimum, which we know, as increasing the guiding layer (in this case consisting of air) by moving the top layers to the right, leads to an improved L^2-norm ($4.0516 \cdot 10^4$) and field enhancement (161.7450). However, we decided to present the local optimum result to motivate how we advance in the following: We want to keep the system width fixed as thicker systems promise larger L^2-norms. Since the core of the system can still be narrowed by additional substrate layers as the lowermost and non-varied layer is modelled as an infinite substrate anyways, the described effect from Figure 5.5 is avoided in the following by requiring that the first layer is at least as high as the substrate, i.e. the side condition

$$\delta_1 \geq 1.2 \cdot 10^{-6}. \tag{5.21}$$

This condition is very reasonable since we cannot expect the described resonance effect (cf. Section 2.2) if we do not have at least a small potential barrier.

Generally, in the optimization of the refractive indices in a fixed number of layers, concerning the layer thicknesses we observed even for fine subdivisions a strong dependence of the optimization result on the initial value. This can be explained as follows. Since the thickness of the layers cannot be varied in this regard, small changes of a promising potential well (local optimum) debase

---

[46] energy of incident field for the resonant angle: When we neglect the comparatively small absorption term, the behavior of a wave under the fixed angle of incidence α in a medium with refractive index $n = 1-\delta$ is determined by $\kappa_\alpha(\delta) := k_0^2((1-\delta)^2 - \cos^2(\alpha))$. If $\kappa_\alpha(\delta) < 0$ we have an exponentially decaying wave ($\sqrt{\kappa_\alpha(\delta)} \in i\mathbb{R}$) and if $\kappa_\alpha(\delta) > 0$ the wave oscillates with wave number $\sqrt{\kappa_\alpha(\delta)} \in \mathbb{R}$. The crossover is located at $\delta = 1 - \cos(\alpha)$ and corresponds to the energy the incident field has. For a top layer with $\delta > 1 - \cos(\alpha)$ the incident field only tunnels into it as an evanescent wave.

the value of the objective function. Only very large changes increase the value of the objective function, but as our optimization procedure is local we cannot leave the local optimum. By the optimization of the layer thicknesses we overcome this problem.

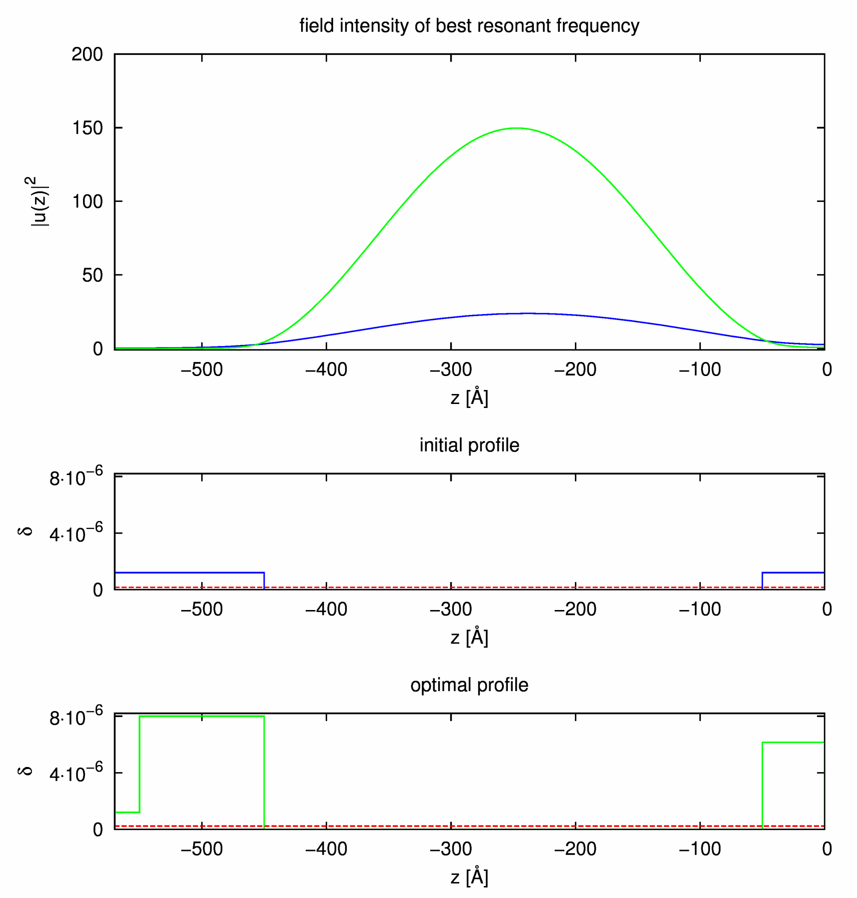


Figure 5.4: *optimization of refractive index in 6 layers subject to the box constraints (5.19), fixed layer widths, relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$*, silicon substrate starts at* $z = -550$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.2*

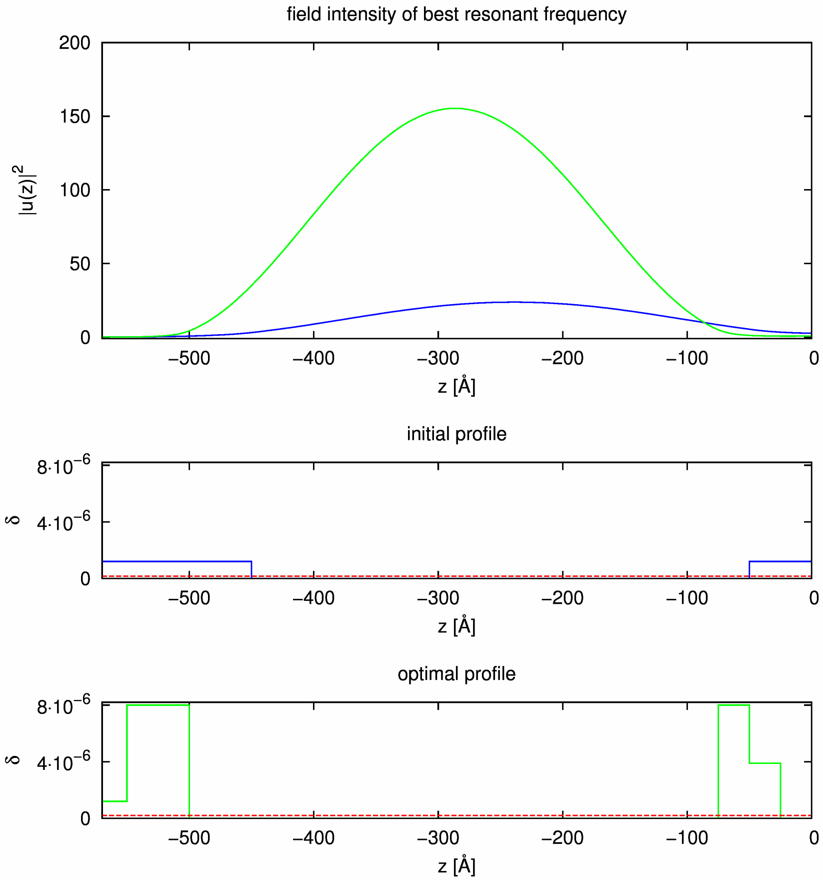


Figure 5.5: *optimization of refractive index in 14 layers subject to the box constraints (5.19), relative absorption with (5.20), fixed layer widths, silicon substrate starts at* $z = -550$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.2*

	true/approximated L^2-norm	true/approximated enhancement	resonance angle [deg]
Fig. 5.4			
initial	$0.5654 \cdot 10^4$ / $0.7040 \cdot 10^4$	23.7821 / 20.6890	0.03218
optimal	$3.4061 \cdot 10^4$ / $3.4154 \cdot 10^4$	149.7230 / 146.8158	0.03907
Fig. 5.5			
initial	$0.5654 \cdot 10^4$ / $0.7040 \cdot 10^4$	23.7821 / 20.6890	0.03218
optimal	$3.6974 \cdot 10^4$ / $3.7285 \cdot 10^4$	155.3513 / 151.3044	0.03719

	refractive profile						
Fig. 5.4							
layer	1	2	3	4	5	6	
thickness [Å]	50	100	100	100	100	100	
initial $\delta \cdot 10^6$	1.2	0	0	0	0	1.2	
optimal $\delta \cdot 10^6$	6.1465	0	0	0	0	8.0000	
Fig. 5.5							
layer	1	2	3	4	5	6	7
	8	9	10	11	12	13	14
thickness [Å]	25	25	25	25	50	50	50
	50	50	50	50	50	25	25
initial $\delta \cdot 10^6$	1.2	1.2	0	0	0	0	0
	0	0	0	0	1.2	1.2	1.2
optimal $\delta \cdot 10^6$	0	3.8888	8.0000	0	0	0	0
	0	0	0	0	0	8.0000	8.0000

Table 5.2: *results for the optimization of refractive indices subject to fixed layer thicknesses and the box constraints (5.19), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$; $\lambda = 0.62\mathring{A}$

5.3.3 Optimization of layer thicknesses for fixed refractive indices

In this section we turn to the optimization of the layer thicknesses. In a system of L layers we optimize the positions of layer change under the side constraints

$$-a < \mathrm{s}_{l+1} < \mathrm{s}_l < 0, \qquad \text{for } l = 1, \ldots, L-1 \tag{5.22}$$

for the positions of layer change s_l (cf. (4.110)), and the refractive indices in the layers are kept fix. For the optimization process we have used an interior points algorithm (see Section 5.3.1) in this section as it did not stuck in local minima as easily as SQP in this application.

We start with the Ni-C-Ni system considered before and optimize the thickness of the three layers while keeping the width of the system fixed at 585Å. The result can be found in Figure 5.6 and Table 5.3 which have to be read as in the preceding section. By only optimizing the layer thicknesses one gains an improvement of about 60% to a value of approximately 183.27 in the field enhancement and even 80% in the L^2-norm. The result can be improved significantly, if we enlarge the width of the whole system to 985Å. We achieve a field enhancement of around 366.15 which corresponds to an improvement of about 220%, see Figure 5.7 and Table 5.3. The improvement in the L^2-norm is even in the region of 500% but this is also due to the enlarged system width. If we enlarge the system width further, the L^2-norm increases by the larger guiding layer but the field enhancement cannot be improved any further. This is no contradiction to the argument that the L^2-norm behaves almost proportional to the L^∞-norm in our application as this only holds for fixed system widths. More complicated initial profiles with more layers have been also tried, but as can be seen in Figure 5.8 for a system of carbon and nickel this leads to a similar optimal solution. The additional layers are pushed together to the smallest admissible value (we forbid layer thicknesses smaller than some $\epsilon > 0$ to avoid the layer thickness 0) in the optimal solution. In Table 5.3 their thicknesses are summed in the interest of readability. This also explains the small deviation of the values from those for Figure 5.7.

In general, the optimization of the layer thicknesses alone is more independent of the initial values than the optimization of the refractive indices with fixed widths. Also it promises better improvements of the field enhancement when a good combination of materials is known, and we can use the exact absorption values here as the materials do not change.

Remark 5.2. *To keep the number of degrees of freedom low and for a better distinguishability between true resonances and spurious modes produced by the discretization, the Hardy space method includes a tuning parameter for the radiation conditions (cf. [HN09, Remark 2.8]). This tuning parameter takes the behavior of the wave (oscillating or exponentially decaying) in regard. The adaptive choice of this tuning parameter during the optimization is challenging when the refractive index changes, and different heuristics were used. If we change in a piecewise constant profile the layer thicknesses only and keep the refractive indices fixed, this question does not arise as the general behavior of the waves does not change.*

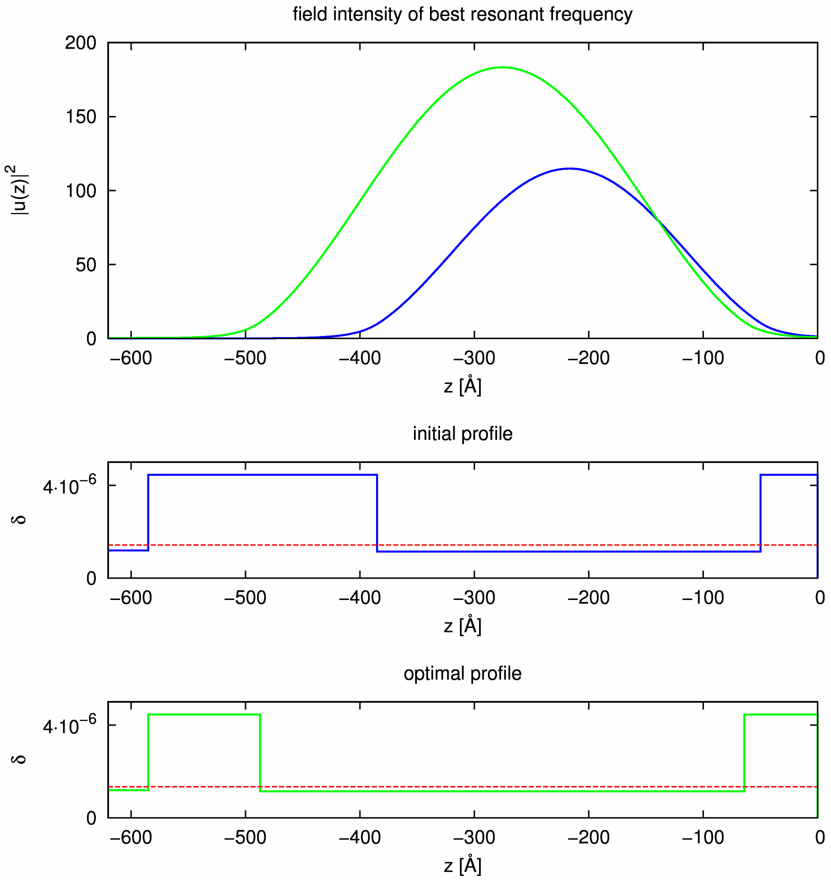


Figure 5.6: *optimization of layer thicknesses for Ni-C-Ni-system under the side constraints (5.22), silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.3*

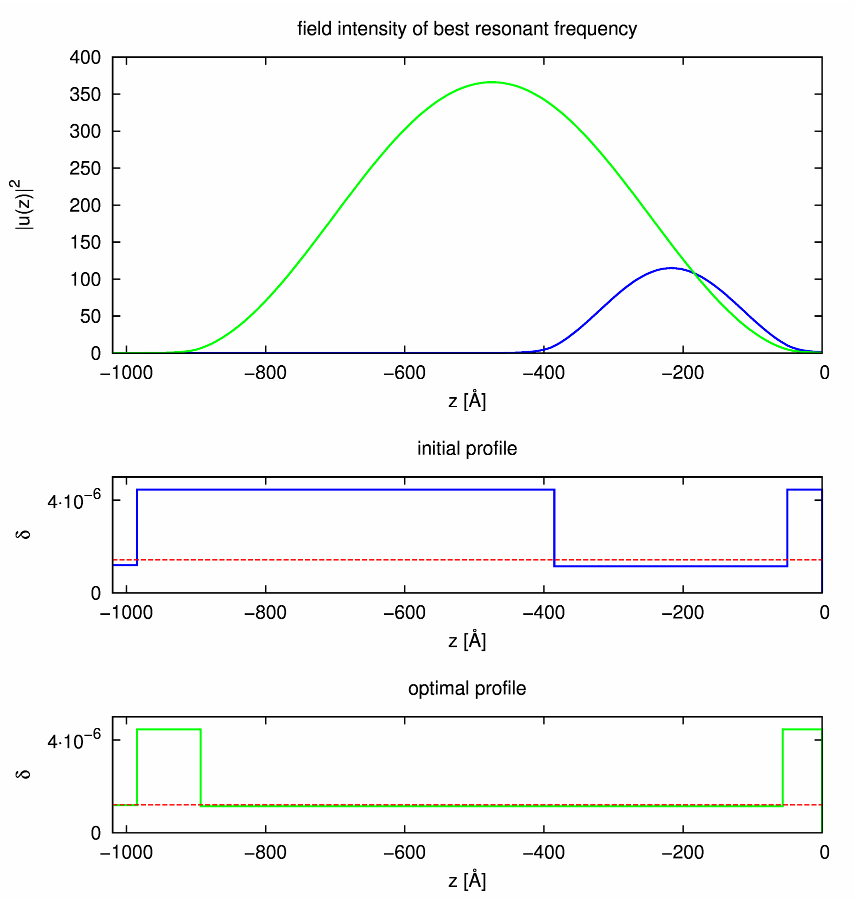


Figure 5.7: *optimization of layer thicknesses for larger Ni-C-Ni-system under the side constraints (5.22), silicon substrate starts at* $z = -985$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.3*

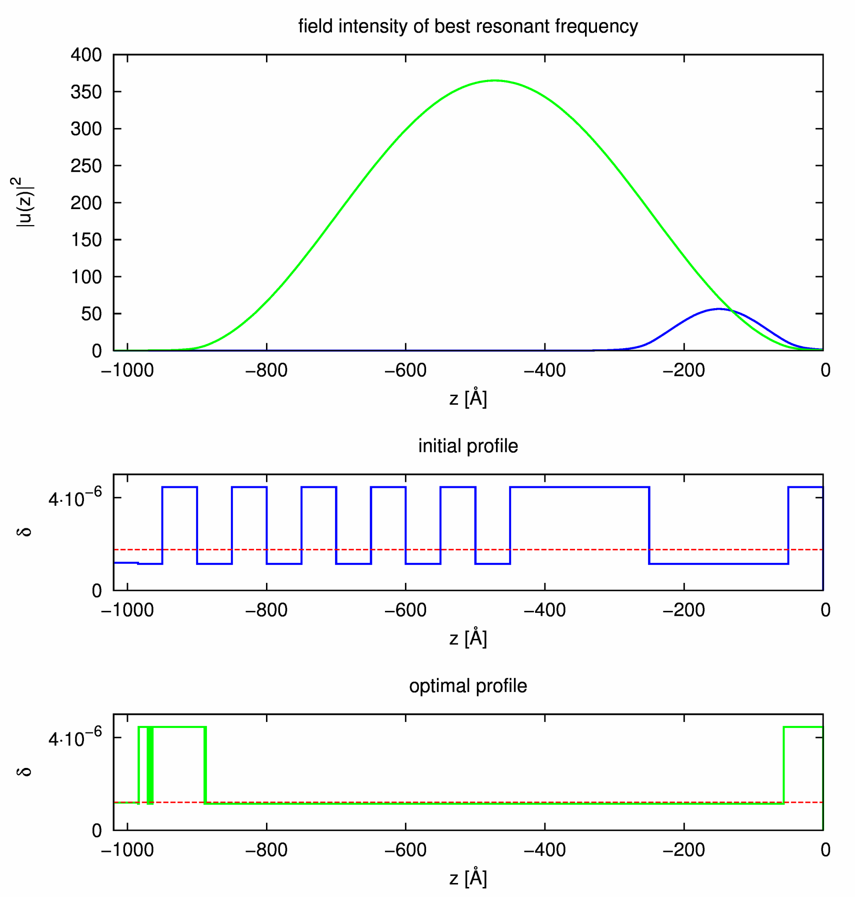


Figure 5.8: *optimization of layer thicknesses for larger system under the side constraints (5.22) with different initial profile of 7 Ni-layers and 7 C-layers, silicon substrate starts at* $z = -985$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.3*

	true/approximated L^2-norm	true/approximated enhancement	resonance angle [deg]
Fig. 5.6			
initial	$2.5643 \cdot 10^4$ / $2.4118 \cdot 10^4$	114.8201 / 121.1057	0.09689
optimal	$4.7262 \cdot 10^4$ / $4.6293 \cdot 10^4$	183.2719 / 186.7537	0.09382
Fig. 5.7			
initial	$0.2606 \cdot 10^5$/ $0.2412 \cdot 10^5$	114.8900 / 121.1057	0.09689
optimal	$1.6931 \cdot 10^5$ / $1.6793 \cdot 10^5$	366.1464 / 369.4400	0.08903
Fig. 5.8			
initial	$0.0958 \cdot 10^5$ / $0.0822 \cdot 10^5$	56.2603 / 58.6575	0.10754
optimal	$1.6787 \cdot 10^5$ / $1.6645 \cdot 10^5$	364.9238 / 367.0513	0.08905

	layer thicknesses [Å]
Fig. 5.6	
initial profile	50Ni—335C—200Ni
optimal profile	63.9547Ni—423.3082C—97.7370Ni
Fig. 5.7	
initial profile	50Ni—335C—200Ni
optimal profile	56.5890Ni—836.8955C—91.5156Ni
Fig. 5.8	
initial profile	50Ni—200C—200Ni—50C—50Ni—50C—50Ni—50C—50Ni—50C—50Ni—50C—50Ni—35C
optimal profile	56.7301Ni—832.7744C—95.4955Ni

Table 5.3: *optimization results for layer thicknesses subject to the side conditions (5.22) in Ni-C-systems of fixed width: 585Å in Figure 5.6 and 985 Å in Figure 5.7 and Figure 5.8;* $\lambda = 0.62$*Å*

5.3.4 Refractive indices and thicknesses simultaneously

We now combine the optimization of refractive indices and thicknesses in a piecewise constant setting for a fixed number of layers L, i.e. the piecewise constant approximation to the refractive index is now encoded in a vector with $2L-1$ entries (L deltas and $L-1$ points of layer change). Again we use the relative absorption (5.11) with $c_{\beta/\delta} = \frac{\beta_{\text{Ni}}}{\delta_{\text{Ni}}}$ and the side conditions

$$\begin{aligned} 1.2 \cdot 10^{-6} \leq &\delta_1 \leq 8 \cdot 10^{-6}, && (5.23a)\\ 0 \leq &\delta_l \leq 8 \cdot 10^{-6}, \quad \text{for } l = 2, \ldots, L && (5.23b) \end{aligned}$$

for the refractive indices and

$$-a < \mathrm{s}_{l+1} < \mathrm{s}_l < 0, \qquad \text{for } l = 1, \ldots, L-2 \tag{5.23c}$$

for the layer changes. The optimization result for 6 layers using the SQP-algorithm is shown in Figure 5.9. A computation with 24 pieces (and therefore 23 points of layer change) led to a mainly identical result, which means that the optimum for piecewise constant refractive indices is not very sensitive to the number of layers. As in the computations in Section 5.3.2, we used an initial profile which consists of a vacuum layer between two substrate layers producing a quite poor field enhancement. The optimal profile, which consists of only three layers (including the substrate), produces a highly enhanced field of about 312.35, which means an improvement of approximately 180% compared to the standard Ni-C-Ni system considered in Section 2.2.1. But this optimal profile has also a drawback. Figure 5.11 presents the resonant frequencies and corresponding field enhancements of this system. The angular acceptance, i.e. the interval of angles of incidence around a resonant frequency for which one gains at least half of the field enhancement at the resonant frequency itself, is quite small. In Remark 4.18 we have already discussed the decrease of the width of the peaks in the intensity with their height. We explain the experimental disadvantage of this effect in Section 5.3.6 and show how to avoid it by changing our objective function.

For another computation we used as an initial system the real parts of the refractive indices and the thicknesses of the standard Ni-C-Ni-system. The side conditions for the refractive indices were modified to

$$\begin{aligned} 2 \cdot 10^{-6} \leq &\delta_1 \leq 8 \cdot 10^{-6}, && (5.24a)\\ 1.2 \cdot 10^{-6} \leq &\delta_l \leq 8 \cdot 10^{-6}, \quad \text{for } l = 2, \ldots, L, && (5.24b) \end{aligned}$$

where the enlarged lower constraint for δ_1 shall again avoid local minima, and the other lower constraints forbid to set δ to zero in any layer. Thus, we have absorption everywhere in this example. We reduced the relative absorption constant to $c_{\beta/\delta} = \frac{\beta_{\text{Ni}}}{20\delta_{\text{Ni}}} \approx 0.0015393$ such that we reach a similar field enhancement for the initial system as the standard Ni-C-Ni-system has for the exact absorption values since the absorption in the carbon layer is much smaller than the relative absorption with $c_{\beta/\delta} = \frac{\beta_{\text{Ni}}}{\delta_{\text{Ni}}}$. In the applications above, this effect was compensated by allowing any nonnegative value for δ. The fundamentally different result can be found in Figure 5.10.

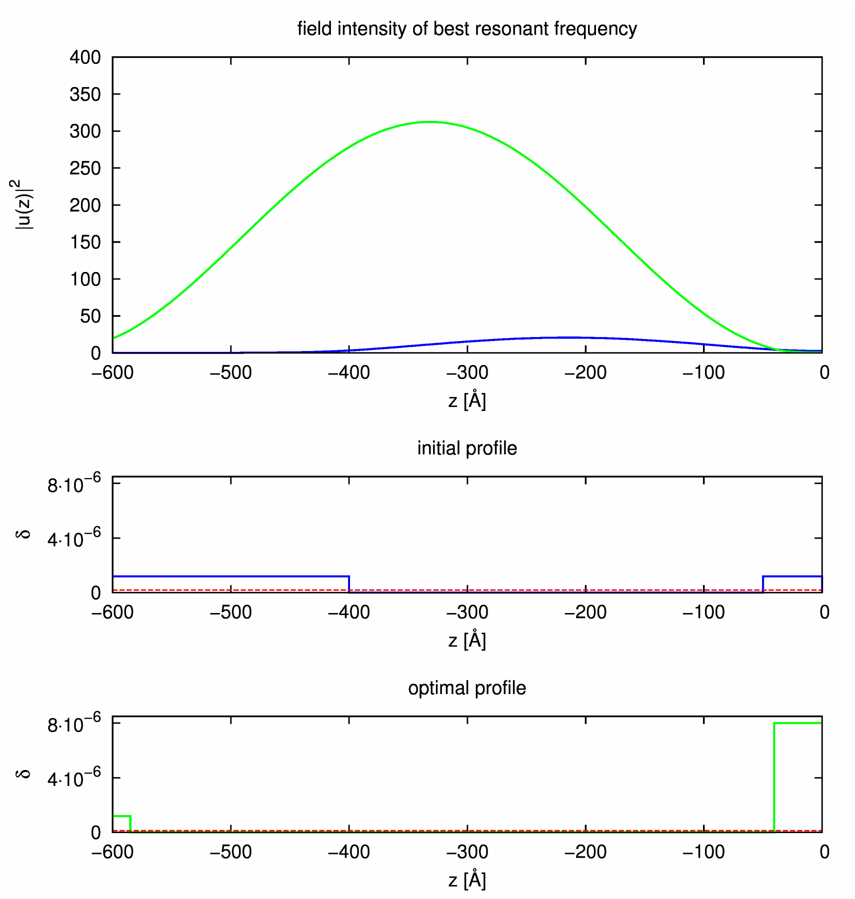


Figure 5.9: *optimization of layer thicknesses and refractive indices with 6 layers under the side conditions (5.23), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$*, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.4*

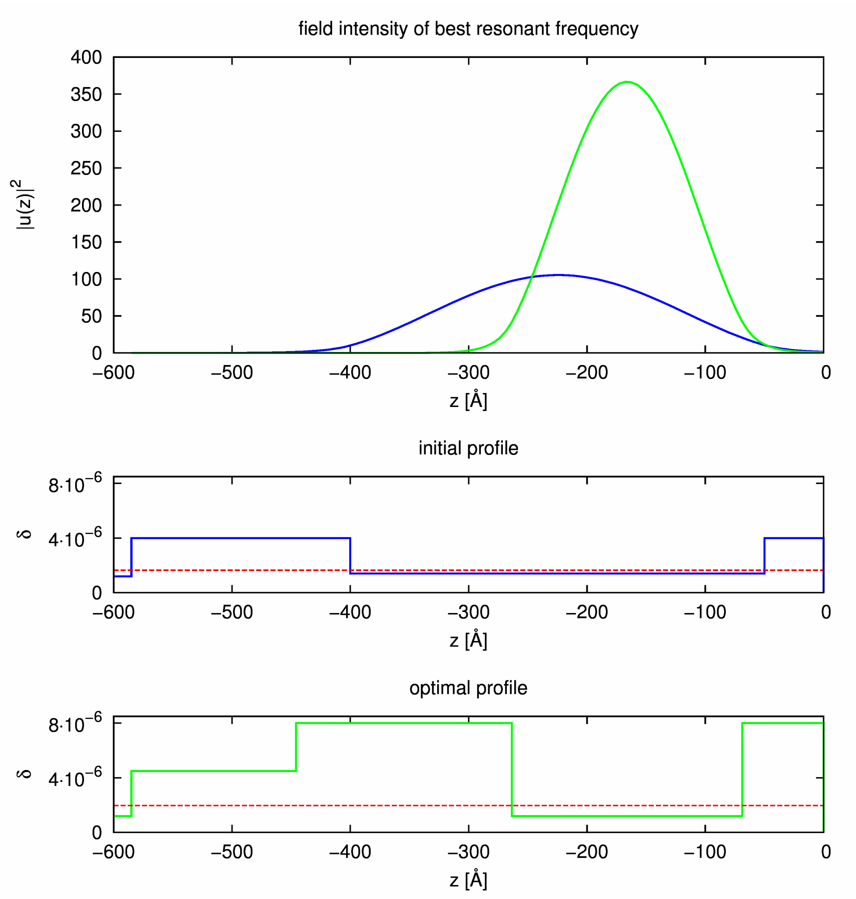


Figure 5.10: *optimization of layer thicknesses and refractive indices with 6 layers under the side conditions (5.24) and (5.23c), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{20\delta_{\mathrm{Ni}}}$*, different initial profile, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.4*

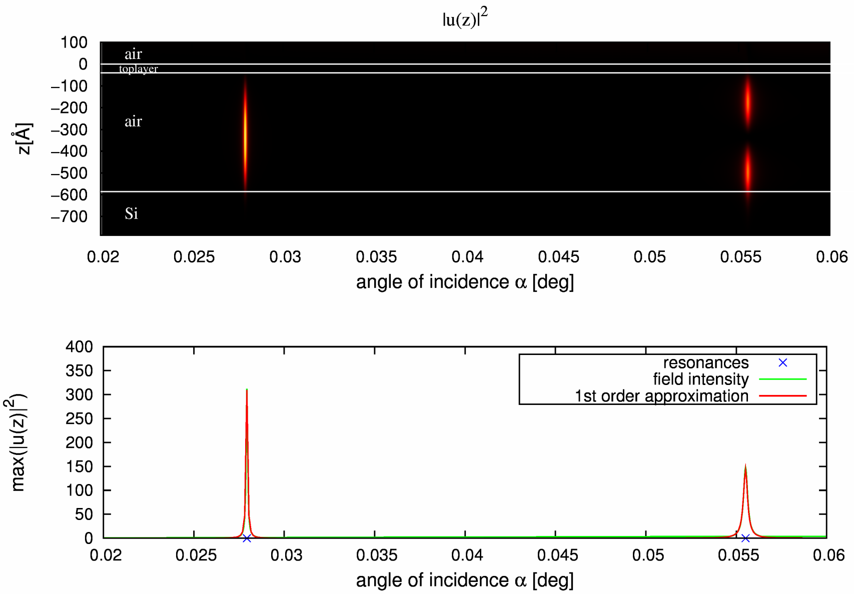


Figure 5.11: *optimal system from Figure 5.9; upper panel: field intensity along the z-axis for different angles of incidence α, lower panel: maximum field intensity for different angles of incidence α*

	true/approximated L^2-norm	true/approximated enhancement	resonance angle [deg]
Fig. 5.9			
initial	$0.4535 \cdot 10^4$ / $0.5682 \cdot 10^4$	20.7127 / 17.1903	0.03545
optimal	$9.8791 \cdot 10^4$ / $9.8704 \cdot 10^4$	312.3521 / 309.7945	0.02793
Fig. 5.10			
initial	$2.4435 \cdot 10^4$ / $2.3528 \cdot 10^4$	105.2961 / 108.4751	0.10406
optimal	$4.9480 \cdot 10^4$ / $4.6246 \cdot 10^4$	366.5472 / 384.6734	0.11377

	refractive profile					
Fig. 5.9						
layer	1	2	3	4	5	6
initial thickness [Å]	50	250	50	50	50	135
initial $\delta \cdot 10^6$	1.2	0	0	0	1.2	1.2
optimal thickness [Å]	40.5137	544.4863	-	-	-	-
optimal $\delta \cdot 10^6$	8	0	-	-	-	-

	refractive profile					
Fig. 5.10						
initial thickness [Å]	50	250	50	50	50	135
initial $\delta \cdot 10^6$	4	1.4	1.4	1.4	4	4
optimal thickness [Å]	68.7731	194.7981	182.2602	139.1686	-	-
optimal $\delta \cdot 10^6$	8	1.2	8	4.4496	-	-

Table 5.4: *results for the simultaneous optimization of layer thicknesses and refractive indices in a piecewise constant setting under the side conditions (5.23) in Fig. 5.9 and the side conditions (5.24) and (5.23c) in Fig. 5.10, fixed system width of* 585*Å and relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$ *in Fig. 5.9 and with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{20\delta_{\mathrm{Ni}}}$ *in Fig. 5.10;* $\lambda = 0.62$*Å*

5.3.5 Optimization with splines

In this section we allow in the approximation more general refractive profiles by the use of splines of higher order, as explained in Section 5.2.2. We use the box constraints

$$0 \leq \delta_l \leq 8 \cdot 10^{-6}, \quad l = 1, \ldots, L \tag{5.25a}$$

for the spline coefficients (cf. equation (5.17)) and B-splines of order 3 where the knots are chosen equidistantly in steps of 12.5(Å). Moreover, we require that the splines fit together continuously with the air and the substrate, i.e.

$$\sum_{l=1}^{L} \delta_l \Psi_l(0) = 0 \quad \text{and} \quad \sum_{l=1}^{L} \delta_l \Psi_l(-a) = 1.2 \cdot 10^{-6}. \tag{5.25b}$$

In Figure 5.12 we used the relative absorption (5.11) with $c_{\beta/\delta} = \beta_{\mathrm{Ni}}/4\delta_{\mathrm{Ni}}$ and a start profile comparable to the one in Figures 5.4 and 5.9. The optimal profile looks almost the same as in Figure 5.4, and the achieved field enhancement can absolutely keep up with the result there. This is quite nice as the non-sharp profiles can be interpreted as systems with surface roughness, and the efficiency of the resonator is not too much affected by this. For the results in this section we used again the SQP-algorithm.

In Figure 5.13 we tried out a different initial profile which approximates 3 Gaussians but the found optimum is again a potential well. For this calculation we had to reduce the constant for the relative absorption to $c_{\beta/\delta} = \beta_{\mathrm{Ni}}/4\delta_{\mathrm{Ni}} \approx 0.007697$ as the initial profile does not produce any significant resonance effect otherwise. This also explains the higher field enhancement in the optimal solution compared to the one in Figure 5.12. That the solution is different from the one in Figure 5.9 may have several reasons, e.g. our method is local and different values for the relative absorption were used.

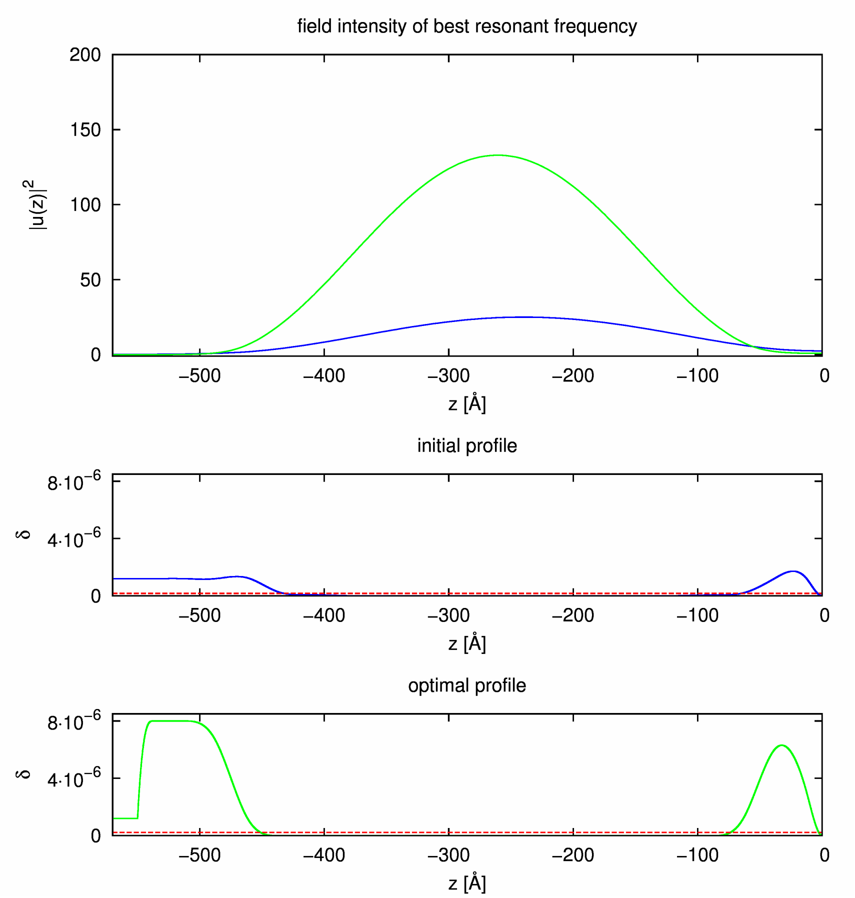


Figure 5.12: *optimization with spline of order 3 with 26 B-splines under the side conditions (5.25), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$*, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.5*

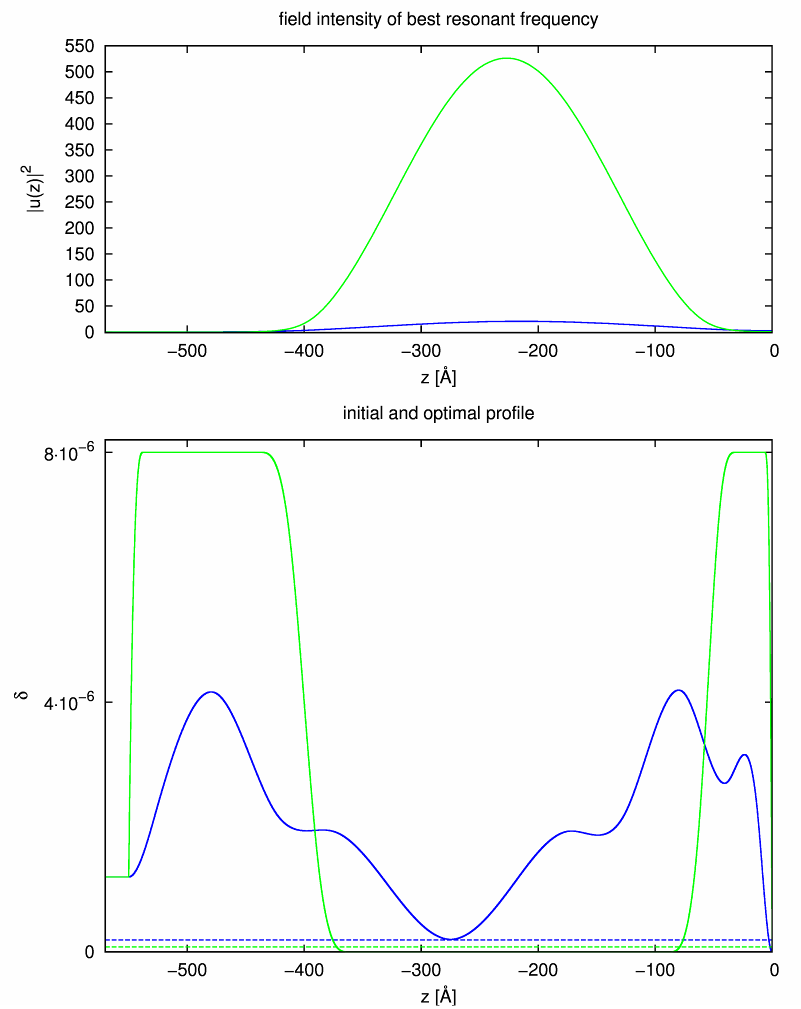


Figure 5.13: *optimization with spline of order 3 with 28 B-splines under the side conditions (5.25), different initial value, lower relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{4\delta_{\mathrm{Ni}}}$*, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed lines: energy of the incident field to excite best resonant frequency; data: Table 5.5*

	true/approximated L^2-norm	true/approximated enhancement	resonance angle [deg]
Fig. 5.12			
initial	$0.5815 \cdot 10^4$ / $0.7103 \cdot 10^4$	25.0276 / 21.5616	0.03386
optimal	$3.0954 \cdot 10^4$ / $3.1295 \cdot 10^4$	132.9332 / 130.9108	0.03795
Fig. 5.13			
initial	$0.0388 \cdot 10^5$ / $0.0266 \cdot 10^5$	19.8577 / 25.0583	0.09059
optimal	$1.0279 \cdot 10^5$ / $1.0187 \cdot 10^5$	526.1897 / 529.6697	0.04605

Table 5.5: *optimization results for B-splines under the side conditions (5.25), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$ *in Figure 5.12 and* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{4\delta_{\mathrm{Ni}}}$ *in Figure 5.13;* $\lambda = 0.62\AA$

5.3.6 Advanced aims and objective functions

So far, we have optimized the energy inside the multilayer structure but sometimes it can be useful to examine slightly different objective functions to model further aims coming from the practical physical application. We assumed the incident field to be a plane wave which is a simplified model. Actually, the beam is divergent, i.e. we do not only excite at the one resonant angle for which we adjust the angle of incidence, but also at neighboring angles of incidence. Therefore, not only the field enhancement but also the angular acceptance of the best resonant frequency is of interest, i.e. the interval around the resonant angle in which at least half of the field enhancement at the resonant angle itself is achieved. In our model this corresponds to the width of the peak in the intensity curve arising from the best resonant frequency. The width of a peak can be measured by its half-width which can be approximated by $2|\operatorname{Im}(\nu_j(\mathbf{n}))|$ (see Remark 4.18). Hence, this aim is contrary to the maximization of the energy in some sense, and we could use as a new objective function

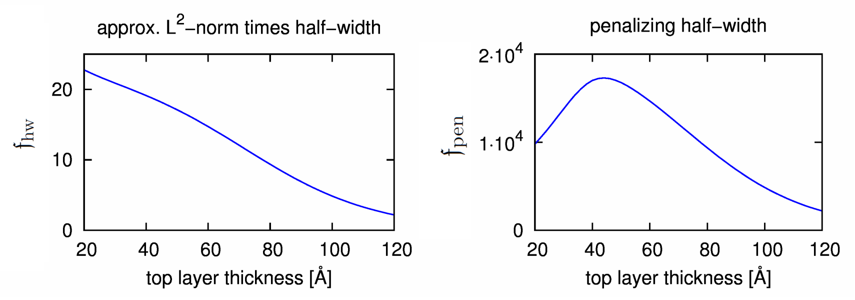


Figure 5.14: *objectives* $\mathfrak{f}_{\mathrm{hw}}$ *and* $\mathfrak{f}_{\mathrm{pen}}$ *to optimize the top layer thickness under further aims on the example from Figure 5.1*

$$\mathfrak{f}_{\mathrm{hw}}(\mathbf{n}) = 2|\operatorname{Im}(\nu_j(\mathbf{n}))|\mathfrak{f}(\mathbf{n}) = \frac{8\,|k_0^2 - \nu_j(\mathbf{n})|}{|\operatorname{Im}(\nu_j(\mathbf{n}))|}\left|\underline{\mathbf{v}}_j[\mathbf{n}](0)\right|^2 \underline{\mathbf{v}}_j[\mathbf{n}]^* \underline{\mathbf{M}}\,\underline{\mathbf{v}}_j[\mathbf{n}]. \quad (5.26)$$

But it turned out in numerical experiments that this objective function rewards broad peaks too much, which is in particular the case since (5.26) does not take the shape of the incident beam into account. Figure 5.14 shows in the left panel the objective function $\mathfrak{f}_{\mathrm{hw}}$, evaluated for the example considered in Section 5.1 in the optimization of the top layer thickness (cf. Figure 5.1). The new objective has a maximum at the boundary of the examined region, and the resulting "optimal" resonator produces a very low field enhancement and a very low L^2-norm at its best resonant frequency.

Therefore, we aim for a more appropriate objective function. Let us denote the width of the incident field in the angle regime by Δ_{ray}, i.e. the interval of angles which is excited by the incident field. Common values for Δ_{ray} are below 0.001° for synchrotron and between 0.001° and 0.1° for standard beams. Define for $m \geq 2$ the penalty function

$$\kappa(\alpha) := \frac{\frac{\alpha}{\Delta_{\mathrm{ray}}}}{\sqrt[m]{1 + \left(\frac{\alpha}{\Delta_{\mathrm{ray}}}\right)^m}}, \tag{5.27}$$

which is a smooth approximation to the function

$$\widetilde{\kappa}(\alpha) = \begin{cases} \frac{\alpha}{\Delta_{\mathrm{ray}}}, & 0 < \alpha \leq \Delta_{\mathrm{ray}} \\ 1, & \alpha > \Delta_{\mathrm{ray}}. \end{cases} \tag{5.28}$$

By this function we want to linearly penalize peaks in the intensity which have a half-width below Δ_{ray}, i.e. those for which only a part of the incident beam leads to good field enhancement. Peaks with a half-width above Δ_{ray} shall not be penalized by κ.

Let again $\nu_\star(\mathbf{n}) = \nu_j(\mathbf{n})$ be the best resonance, i.e. the one whose corresponding resonant frequency gives rise to the highest field enhancement, of the system with refractive index encoded in $\mathbf{n}$. Then the half-width (in the alpha regime) around the resonant frequency $\mathrm{Re}(\nu_\star(\mathbf{n}))$ with corresponding resonant angle

$$\alpha_\star(\mathbf{n}) := \arccos\left(\frac{\sqrt{\mathrm{Re}(\nu_\star(\mathbf{n}))}}{k_0}\right) \tag{5.29}$$

is given by

$$\alpha_{\mathrm{hw}}(\mathbf{n}) = 2\left(\arccos\left(\frac{\sqrt{\mathrm{Re}(\nu_\star(\mathbf{n})) + |\operatorname{Im}(\nu_\star(\mathbf{n}))|}}{k_0}\right) - \alpha_\star(\mathbf{n})\right). \tag{5.30}$$

Hence, we may use the objective

$$\mathfrak{f}_{\mathrm{pen}}(\mathbf{n}) = \kappa(\alpha_{\mathrm{hw}}(\mathbf{n}))\mathfrak{f}(\mathbf{n}). \tag{5.31}$$

The resulting curve for our example can be found in the right panel of Figure 5.14 using the values $\Delta_{\mathrm{ray}} = 0.001°$ and $m = 6$. With these parameters the objective $\mathfrak{f}_{\mathrm{pen}}$ achieves its maximum at a top layer thickness of approximately 42Å.

Remark 5.3. *One may also think of other objective functions which take the angular acceptance into account. We discuss this for the continuous setting. For fixed* $z \in [-a, 0]$ *consider the function*

$$\left[\alpha_\star - \frac{\Delta_{\text{ray}}}{2}, \alpha_\star + \frac{\Delta_{\text{ray}}}{2}\right] \to \mathbb{R}, \quad \text{with} \quad \alpha \mapsto \left|u[k_0^2 \cos^2(\alpha), n](z)\right|^2 \tag{5.32}$$

and integrate it over the interval $\Delta_{\text{ray}}^{(\alpha_\star)} := [\alpha_\star - \frac{\Delta_{\text{ray}}}{2}, \alpha_\star + \frac{\Delta_{\text{ray}}}{2}]$. *This corresponds for each* z *to the integrated field intensity, and we could use the objective functions*

$$\mathfrak{f}_{\text{int}}(n) = \max_z \int_{\Delta_{\text{ray}}^{(\alpha_\star)}} |u[k_0^2 \cos^2(\alpha), n](z)|^2 \, \mathrm{d}\alpha, \tag{5.33}$$

or

$$\mathfrak{f}_{\text{int,int}}(n) = \left\| \int_{\Delta_{\text{ray}}^{(\alpha_\star)}} u[k_0^2 \cos^2(\alpha), n](\cdot) \mathcal{I}^{(\alpha_\star)}(\alpha) \, \mathrm{d}\alpha \right\|_{L^2([-a,0])}, \tag{5.34}$$

where $\mathcal{I}^{(\alpha_\star)}$ *is a model for the intensity of the beam whose angle of incidence is adjusted to the resonant frequency. The objective functions* $\mathfrak{f}_{\text{int}}$ *and* $\mathfrak{f}_{\text{int,int}}$ *are generalizations of the ones presented in Optimization Problems 2.15 and 2.16. Therefore, they cause again the same problems discussed before, which were the reason to use the more accessible objective function* $\mathfrak{f}$ *(approximation formula for the* L^2*-norm). Clearly, we could again consider* u_{s}*, discretize everything in the same way as before and use the approximation formula (4.88) in (5.34). Note that this simplifies the integral expression considerably as only the term* $1/(\nu_j(n) - \nu)$ *in (4.88) changes with* α*. But there occurs an additional difficulty in the derivative of* $\mathfrak{f}_{\text{int,int}}$ *as the location of the intensity distribution* $\mathcal{I}$ *and the integration interval also depend on* $\alpha_\star$*. The integration interval cannot simply be enlarged since the approximation (4.88) decays only linearly with* α*. Hence, we would reward narrow peaks too much by contributions from angles which are far away from the resonant frequency. In particular, for these angles the approximation formula (4.88) loses in general its validity. Of course, it is possible to overcome these problems, but this is not part of this thesis.*
First tests with the objective function $\mathfrak{f}_{\text{int,int}}$ *in simple examples have not shown significantly better results than the ones we present in the following.*

As pointed out in Section 5.3.4, the angular acceptance of the best resonant frequency of the optimal system in Figure 5.9 is very bad. Therefore, we performed an optimization process with the modified objective function $\mathfrak{f}_{\text{pen}}$, using the optimal solution of Figure 5.9 as initial system. Figure 5.15 and Table 5.6 show the results. We assumed a beam width of $\Delta_{\text{ray}} = 0.001°$ and chose $m = 6$ in the penalty function κ. The optimal solution is quite different now, and the potential barrier (δ in the coating minus δ in the guiding layer) for the optimal profile is not too different from the one of the standard Ni-C-Ni-system examined in Examples 2.2.1 and 4.3.3. To visualize the influence of the angular acceptance, Figure 5.15 does not show the field intensity along the z-axis, but the integrated field intensity, i.e. for the best resonant angle $\alpha_\star$ we computed for every fixed $z \in [-a, 0]$ a discrete approximation to the integral of the function (5.32) over the interval $\Delta_{\text{ray}}^{(\alpha_\star)}$. This corresponds to what is excited by a beam of width 0.001° that is adjusted at the resonant angle $\alpha_\star$. One can observe a considerable improvement, and the objective function $\mathfrak{f}_{\text{pen}}$ approximately even doubled its value. Naturally, for the optimal system under the objective function $\mathfrak{f}_{\text{pen}}$ the field enhancement is clearly reduced compared to the initial system, but the angular acceptance of the best resonant frequency lies in the region of the beam width. In this way, not only a small part but most of the beam is used to excite highly enhanced fields, and this improves the overall performance of the resonator. This phenomenon is illustrated in Figure 5.17, where the peaks in the intensity produced by the best resonant frequencies of the two optimal systems (for $\mathfrak{f}$ and $\mathfrak{f}_{\text{pen}}$) are compared.

In the example considered above, the layer thicknesses were left almost fixed during the optimization process although they were variable. We have also done experiments where we only optimized the layer thicknesses for the standard Ni-C-Ni example and could corroborate the observation in [Pfe02, p.31] that one cannot gain too much from optimizing angular acceptance and field enhancement simultaneously for a system with fixed materials. However, our example shows that it is possible to find a better compromise between angular acceptance and achieved field enhancement/L^2-norm. We did another computation using the reduced relative absorption, already employed in Figure 5.10. All the other parameters were left unchanged and the results can be found in Figure 5.16. The L^2-norm of the integrated field intensity stayed almost constant, but the maximal integrated field intensity increased. One should note here in particular that our observation that the L^2-norm is approximately proportional to the L^∞-norm only holds at the resonant frequency itself and is wrong for values too far away from it.

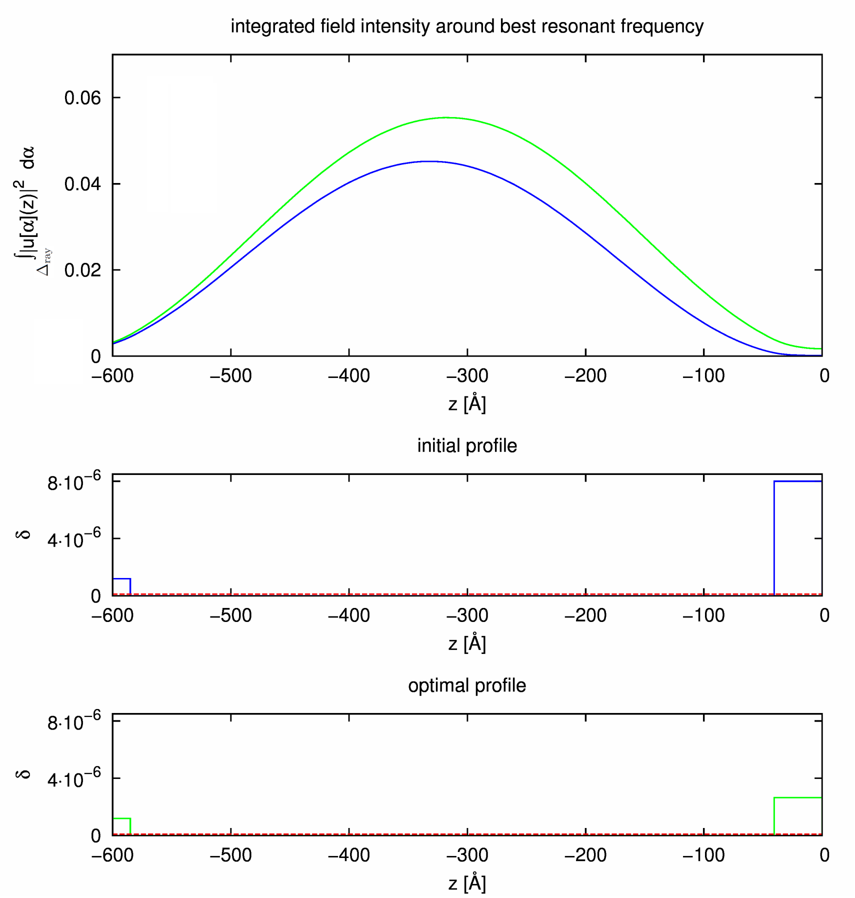


Figure 5.15: *optimization of layer thicknesses and refractive indices with 6 layers with the objective function* $\mathfrak{f}_{\text{pen}}$ *under the side conditions (5.23), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\text{Ni}}}{\delta_{\text{Ni}}}$*, penalty function* κ *with* $m = 6$ *and* $\Delta_{\text{ray}} = 0.001°$*, upper panel shows field intensity integrated over an interval of width* Δ_{ray} *around the best resonant frequency, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.6*

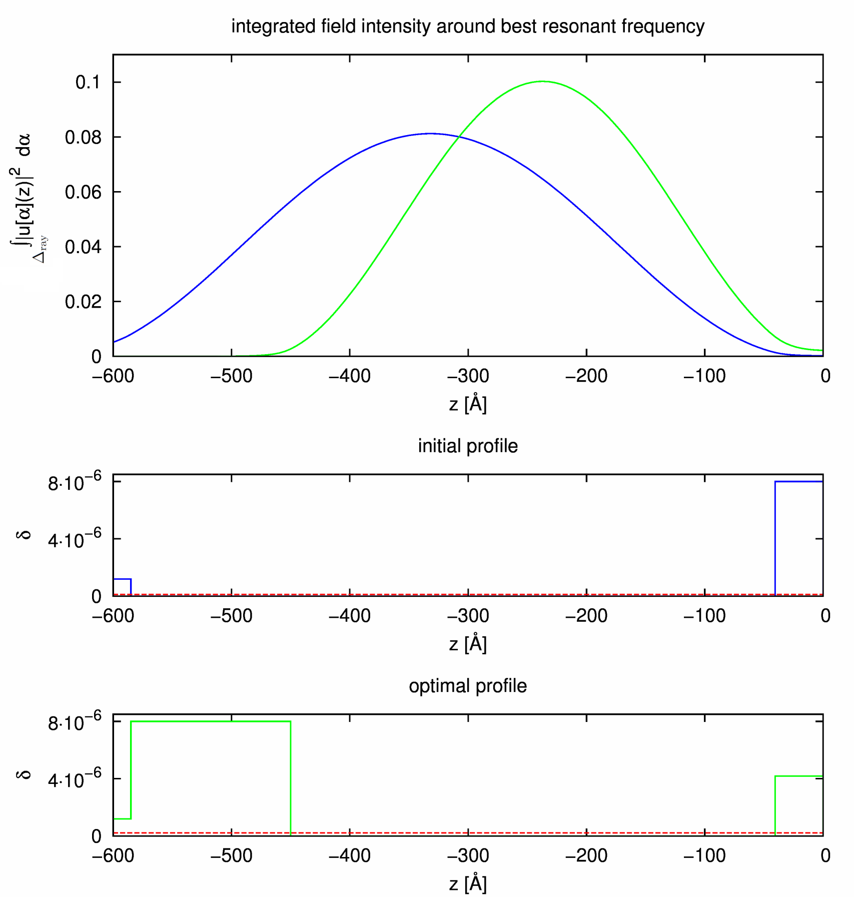
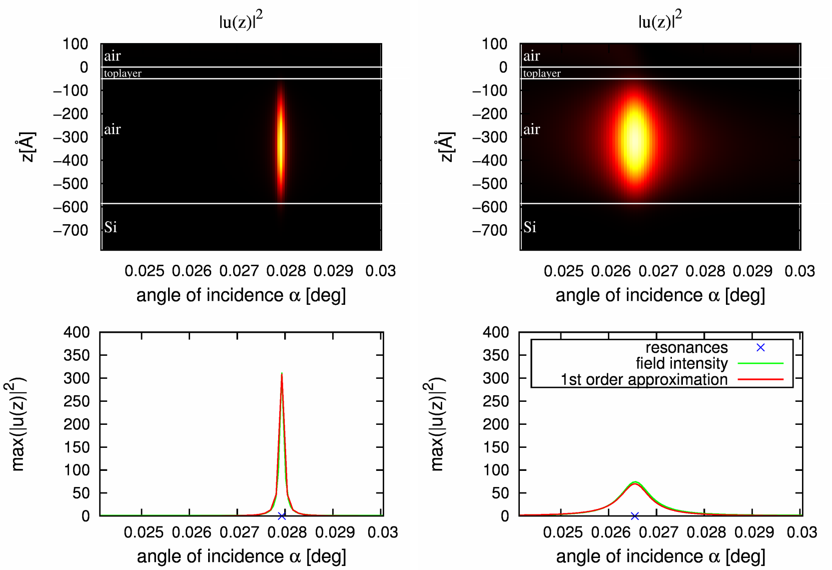


Figure 5.16: *optimization of layer thicknesses and refractive indices with 6 layers with the objective function* $\mathfrak{f}_{\mathrm{pen}}$ *under the side conditions (5.23), relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{20\delta_{\mathrm{Ni}}}$, *penalty function* κ *with* $m = 6$ *and* $\Delta_{\mathrm{ray}} = 0.001°$, *upper panel shows field intensity integrated over an interval of width* Δ_{ray} *around the best resonant frequency, silicon substrate starts at* $z = -585$*; blue lines: initial situation, green lines: optimal situation, dashed red line: energy of the incident field to excite best resonant frequency; data: Table 5.6*



(a) optimal result from Fig. 5.9 (b) optimal result from Fig. 5.15

Figure 5.17: *comparison of best resonant frequency for optimal results neglecting angular acceptance (a) and taking it into account (b), upper panels: field intensity along the z-axis for different angles of incidence α, lower panels: maximum field intensity for different angles of incidence α*

	true/approximated L^2-norm	true/approximated enhancement	resonance angle [deg]
Fig. 5.15			
initial	$9.8791 \cdot 10^4$ / $9.8704 \cdot 10^4$	312.3521 / 309.7945	0.02793
optimal	$2.5081 \cdot 10^4$ / $2.3165 \cdot 10^4$	74.5948 / 70.0789	0.02655
Fig. 5.16			
initial	$3.0430 \cdot 10^5$ / $3.0525 \cdot 10^5$	960.134 / 956.8649	0.02793
optimal	$0.3986 \cdot 10^5$ / $0.4100 \cdot 10^5$	171.4189 / 166.0157	0.03739

	approx. half-width	$\mathfrak{f}_{\text{pen}}$	$\mathfrak{f}_{\text{hw}}$	max. int. intensity	L^2-norm of int. intensity
Fig. 5.15					
initial	$0.9822 \cdot 10^{-4}$	$0.9695 \cdot 10^4$	9.6947	0.0452	14.2976
optimal	$7.5721 \cdot 10^{-4}$	$1.8453 \cdot 10^4$	18.9916	0.0553	17.4275
Fig. 5.16					
initial	$0.5619 \cdot 10^{-4}$	$0.9695 \cdot 10^4$	17.0986	0.0812	25.6684
optimal	$5.4398 \cdot 10^{-4}$	$2.2207 \cdot 10^4$	21.6847	0.1003	24.1390

	refractive profile		
Fig. 5.15			
layer	1	2	3
initial thickness [Å]	40.5137	544.4863	-
initial $\delta \cdot 10^6$	8	0	-
optimal thickness [Å]	40.5137	544.4863	-
optimal $\delta \cdot 10^6$	2.4427	0	-
Fig. 5.16			
layer	1	2	3
initial thickness [Å]	40.5137	544.4863	-
initial $\delta \cdot 10^6$	8	0	-
optimal thickness [Å]	40.5137	409.4863	135.0000
optimal $\delta \cdot 10^6$	4.1830	0	8

Table 5.6: *results for the simultaneous optimization of layer thicknesses and refractive indices with the objective function* $\mathfrak{f}_{\mathrm{pen}}$ *under the side conditions (5.23) in a piecewise constant setting, fixed system width of* 585*Å and relative absorption with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{\delta_{\mathrm{Ni}}}$ *in Fig. 5.15 and with* $c_{\beta/\delta} = \frac{\beta_{\mathrm{Ni}}}{20\delta_{\mathrm{Ni}}}$ *in Fig. 5.16, penalty function* κ *with* $m = 6$ *and* $\Delta_{\mathrm{ray}} = 0.001°$; $\lambda = 0.62Å$

5.3.7 Conclusions

We conclude by drawing some conclusions from the results presented in this chapter:

1. The idealized situation of a piecewise constant refractive index, i.e. sharp layer changes, seems at the same time to be the desirable situation as optimizations with general refractive profiles converge again to such systems (cf. Section 5.3.5).

2. In a piecewise constant setting the optimization of the layer thicknesses for fixed refractive indices is very promising, when one is interested in the optimization of the field enhancement (cf. Section 5.3.3). We have seen improvements of over 200% by only optimizing the layer thicknesses.

3. If we neglect absorption effects, the field enhancement can be raised arbitrarily by larger systems and thicker top layers. In practice this cannot be achieved due to energy loss by absorption. One may reduce absorption effects, but not eliminate them completely (cf. Section 5.1).

4. The optimal solutions highly depend on the absorption model and one has to carefully check which assumptions are fulfilled in a certain application. For our purpose we used the relative absorption model (cf. equation (5.1)).

5. For very high field enhancements, the angular acceptance gets worse. If one is also interested in good angular acceptance, one should use a

modified objective function such as $\mathfrak{f}_{\text{pen}}$ in equation (5.31), leading to different optimization results and showing a good compromise between angular acceptance and field enhancement (cf. Section 5.3.6).

6 Summary and Outlook

Let us summarize the main results of this thesis. The starting point was the question how to improve the field enhancement which is achieved by multilayer x-ray resonators for one-dimensional beam concentration. In Chapter 2 we derived a mathematical model for the multilayer systems in form of the scattering problem (2.8) and also for the arising optimization problem (cf. Optimization Problem 2.2). Because of the complicated form of the optimization problem, we introduced the concept of resonances to make it more accessible. Existence and uniqueness results for the scattering problem have been provided as well as results on the excitation and location of resonances. Furthermore, we have shown the eigenvalues to be isolated with one-dimensional eigenspaces. Aiming for an optimization algorithm involving derivatives, we analyzed in Chapter 3 the sensitivity of the solution to the scattering problem with respect to the refractive index n. This was done by rewriting the problem into an equivalent integral equation and computing the Fréchet derivative. As a side product we obtained approximation formulas for the reflectivity, in particular we gained a mathematical justification for the widely used kinematic approximation. Higher order Taylor approximations and Padé approximations have led to significant improvements of the standard approximation formulas, in particular close to the critical angle.

In Chapter 4 we have presented a general formalism to compute the derivatives of simple, isolated eigenvalues of a operator on a Hilbert space depending on a parameter n in a Banach space, provided that the operator is differentiable itself. For a generalized eigenvalue problem with operators that are symmetric with respect to a conjugation, we have worked out formulas for the derivative, with respect to n, of simple, isolated eigenvalues and corresponding eigenvectors using a special scaling. We applied the general results to appropriate objective functions for our problem which were worked out in Section 4.3.1 and to their discrete analogs (obtained by applying the Hardy space method). However, the general results are not restricted to our problem. One can apply them to all problems which can be transformed into an eigenvalue problem (or generalized eigenvalue problem) of the described form. The required differentiability and derivatives of the involved operator were obtained for our problem, in the continuous setting in Chapter 3 as mentioned above and for different discretizations of n in Sections 4.5 and 5.2. In particular, the derivative with respect to the layer changes in a piecewise constant refractive index turned out to be very useful in numerical experiments. By the numerical results in Chapter 5 we illustrated the successful application of our results in an optimization algorithm and presented relevant conclusions for the optimal design of multilayer x-ray resonators under certain conditions (see Section 5.3.7).

In general, it is not sufficient to find a refractive index n (or respectively a multilayer structure) such that the imaginary part of the best resonance is as small as possible. Therefore, other terms have to be taken into account in the objective function (cf. Remark 4.18). Another difficulty occurred in the consideration of the absorption which cannot be modelled as an independent variable as this would always lead to the smallest admissible absorption. Instead, we have coupled the absorption to the real part of the refractive index (see Section 5.1). Moreover, the optimization of the field enhancement is not the only aim one is interested in, but also the angular acceptance of a resonant frequency is of interest. Thus, we have also examined a modification of the objective function taking this aim into account (see Section 5.3.6). Another advantage of our method is that one can use very general approximation spaces for the refractive index (here we have seen splines and piecewise constant approximations), provided that the operator depends differentiably on the discretziation.

It remains to point out to open problems and further plans. The main questions concerning the multilayer x-ray resonators we were interested in, have been answered. We have not shown that the algebraic multiplicity of the resonances is one. This has only been proven for their geometric multiplicity, but it is our conjecture that resonances with algebraic multiplicities greater than one cannot exist. For the particular optimization problem considered in this thesis, a generalization to the two-dimensional case does not make sense for physical reasons. Mainly, the energy loss compared to the achieved field enhancement is too high. One uses different techniques to focus x-ray beams in higher dimensions, e.g. structures of special shapes. But these techniques do not lead to resonance problems. Nevertheless, it would be very interesting to apply our results in a higher-dimensional problem. Whenever one can derive a similar eigenvalue problem and an objective function involving simple, isolated eigenvalues, the framework can be used without modifications. In particular, the derivatives can be computed quite fast compared to the solution of the eigenvalue problem which has to be done anyways.

Appendix

In fact, we treat the problem, examined in this thesis, by methods for partial differential equations. But since the underlying differential equation (2.4) is also an ordinary differential equation, at some points we take advantage of results for initial value problems. As the existence and uniqueness result for such a general ordinary differential equation is not so much standard, we cite it here.

For the theorem stated now, we rely on the common notation for initial value problems, as for example used in [Wal96]. Note in this context that a system of m differential equations with complex-valued coefficients and complex-valued solutions is equivalent to a system of $2m$ real differential equations by the canonical identification $\mathbb{C} = \mathbb{R} \times \mathbb{R}$.

THEOREM 6.1. *Let* $\mathfrak{D} := \mathfrak{I} \times \mathbb{C}^m$ *with* $\mathfrak{I} := [\xi, \xi + b] \subset \mathbb{R}$ *and a function* $\mathbf{f} : \mathfrak{D} \to \mathbb{C}^m$ *with* $\mathbf{f}(z, \mathbf{u}) := [\mathbf{f}_1(z, \mathbf{u}), \ldots, \mathbf{f}_m(z, \mathbf{u})]^\top$. *Consider the system of ordinary differential equations*

$$\mathbf{v}' = \mathbf{f}(z, \mathbf{v}). \tag{6.1a}$$

We call a function $\mathbf{v} : \mathfrak{I} \to \mathbb{C}^m$ *with* $\mathbf{v}(z) := [v_1(z), \ldots, v_m(z)]^\top$ *a solution to the initial value problem (6.1) if* $\mathbf{v}$ *is absolutely continuous*[47], $\mathbf{v}$ *fulfills the system (6.1a) for almost all* $z \in I$ *and it fulfills the initial condition*

$$\mathbf{v}(\xi) = [\eta_1, \ldots, \eta_m]^\top. \tag{6.1b}$$

Under the assumptions

- $\mathbf{f}$ *is continuous in* $\mathbf{v} \in \mathbb{C}^m$ *for fixed* $z \in \mathfrak{I}$,
- $\mathbf{f}(z, \mathbf{v}) \in L^1(\mathfrak{I})$ *for fixed* $\mathbf{v} \in \mathbb{C}^m$,
- *and there is a function* $l \in L^1(\mathfrak{I})$ *with*

$$|\mathbf{f}(z, \mathbf{v}) - \mathbf{f}(z, \hat{\mathbf{v}})| \le l(z) |\mathbf{v} - \hat{\mathbf{v}}| \quad \text{in } \mathfrak{D}, \tag{6.2}$$

there exists a unique solution, in the sense stated above, to the initial value problem (6.1) for $[\xi, \eta_1, \ldots, \eta_m]^\top \in \mathfrak{D}$.

PROOF. A proof can be found in [Wal96, §10, XII]. □

COROLLARY 6.2. *Let* $\widetilde{\mathfrak{D}} \subset \mathbb{R} \times \mathbb{C}^m$ *an open set. For every subset* $\mathfrak{I} \times \mathfrak{B} \subset \widetilde{\mathfrak{D}}$, *where* $\mathfrak{I} \subset \mathbb{R}$ *is a closed subset and* $\mathfrak{B}$ *a closed ball, let the assumptions of Theorem 6.1 be fulfilled (with* $\mathfrak{I} \times \mathfrak{B}$ *instead of* $\mathfrak{D}$*). Then the initial value problem*

$$\mathbf{v}' = \mathbf{f}(z, \mathbf{v}), \quad \mathbf{v}(\xi) = [\eta_1, \ldots, \eta_m]^\top \tag{6.3}$$

for $[\xi, \eta_1, \ldots, \eta_m]^\top \in \widetilde{\mathfrak{D}}$ *has a unique solution up to the boundary.*

PROOF. Again a proof can be found in [Wal96, §10, XIV]. □

---

[47] The absolute continuity of a function v in an interval $[b, c]$ is equivalent to the following: v has a first derivative v' almost everywhere, the derivative is Lebesgue integrable and it holds $v(z) = v(b) + \int_b^z v'(z)dz$ for all $z \in [b, c]$. This is a generalization of the fundamental theorem of integral theory to Lebesgue integrable functions and can simultaneously be used as a definition for weak differentiability on subsets of $\mathbb{R}$ (see e.g. [Eva98, 5.2, p.259]).

Bibliography

[BGFB94] S. Boyd, L.E. Ghaoui, E. Feron, and V. Balakrishnan. *Linear Matrix Inequalities in System and Control Theory*, volume 15. SIAM: Studies in Applied Mathematics, Philadelphia, 1994.

[BGM96] G. A. Baker and P. Graves-Morris. *Padé approximants.* Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2nd edition, 1996.

[BHN99] R.H. Byrd, M.E. Hribar, and J. Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM Journal on Optimization*, 9(4):877–900, 1999.

[BLO00] J.V. Burke, A.S. Lewis, and M.L. Overton. Optimizing matrix stability. *Proc. Amer. Math. Soc.*, 129:1635–1642, 2000.

[BLO02] J.V. Burke, A.S. Lewis, and M.L. Overton. Approximating subdifferentials by random sampling of gradients. *Math. Oper. Res.*, 27(3):567–584, 2002.

[BLO03] J.V. Burke, A.S. Lewis, and M.L. Overton. Optimization and pseudospectra, with applications to robust stability. *SIAM J. Matrix Anal. Appl.*, 25:80–104, 2003.

[BO93] J.V. Burke and M.L. Overton. *Nonsmooth Optimization: Methods and Applications*, chapter On the Subdifferentiability of Functions of a Matrix Spectrum, I: Mathematical Foundations, II: Subdifferential Formulas, pages 11–29. Gordon and Breach, 1993.

[BO94] J.V. Burke and M.L. Overton. Differential properties of the spectral abscissa and the spectral radius for analytic matrix-valued mappings. *Nonlinear Analysis, Theory, Methods & Applications*, 23(4):467–488, 1994.

[BO01] J.V. Burke and M.L. Overton. Variational analysis of non-lipschitz spectral functions. *Mathematical Programming*, 90:317–352, 2001.

[BOY04] M. Burger, S. Osher, and E. Yablonovitch. Inverse problem techniques for the design of photonic crystals. *IEICE Transactions on Electronics*, 87(C):258–265, 2004.

[BW97] M. Born and E. Wolf. *Principles of Optics: Electromagnetic Theory of Propagation, Interference and Diffraction of Light, 7th Edition.* Cambridge University Press, 1997.

[Cat95] A. Caticha. Reflection and transmission of x rays by graded interfaces. *Phys. Rev. B*, 52:9214–9223, 1995.

[Cla83] F.H. Clarke. *Optimization and Nonsmooth Analysis.* John Wiley, New York, 1983.

[CM90] S.J. Cox and J.R. McLaughlin. Extremal eigenvalue problems for composite membranes, ii. *Applied Mathematics and Optimization*, 22(1):169–187, 1990.

[CO92] S.J. Cox and M.L. Overton. On the optimal design of columns against buckling. *SIAM Journal on Mathematical Analysis*, 23:287–325, 1992.

[Con90] J.R. Convay. *A Course in Functional Analysis.* Springer-Verlag New York Inc., 1990.

[dB78] C. de Boor. *A Practical Guide to Splines.* Springer-Verlag Berlin, Heidelberg, New York, 1978.

[dB90] C. de Boor. *Splinefunktionen.* Lecture Notes in Mathematics, ETH Zürich. Birkhäuser, 1990.

[DH73] W.E. Donath and A.J. Hoffman. Lower bounds for the partitioning of graphs. *IBM Journal of Research and Development*, 17:420–425, 1973.

[DR06] W. Dahmen and A. Reusken. *Numerik für Ingenieure und Naturwissenschaftler.* Springer-Verlag Berlin, Heidelberg, New York, 2006.

[DS04] D.C. Dobson and F. Santosa. Optimal localization of eigenfunctions in an inhomogeneous medium. *SIAM J. Math. Anal.*, 64(3):762–774, 2004.

[Eva98] L.C. Evans. *Partial Differential Equations.* American Mathematical Society Providence, Rhode Island, 1998.

[FE01] T. Felici and H.W. Engl. On shape optimization of optical waveguides using inverse problem techniques. *Inverse Problems*, 17(4):1141–1162, 2001.

[FFK$^+$03] I.D. Feranchuk, S.I. Feranchuk, L. Komarov, S. Sytova, and A. Ulyanenkov. Analytical ansatz for self-consistent calculations of x-ray transmission and reflection coefficients at graded interfaces. *Phys. Rev. B*, 67:235–417, 2003.

[Fle87] Roger Fletcher. *Practical methods of optimization.* John Wiley & Sons, New York, 2nd edition, 1987.

[GR01] T. Grund and A. Rösch. Optimal control of a linear elliptic equation with a supremum-norm functional. *Optimization Methods and Software*, 15:299–329, 2001.

[HBKW08] P. Heider, D. Berebichez, R.V. Kohn, and M.I. Weinstein. Optimization of scattering resonances. *Structural and Multidisciplinary Optimization*, 36(5):443–456, 2008.

[HGS08] T. Hohage, K. Giewekemeyer, and T. Salditt. Iterative reconstruction of a refractive index from x-ray or neutron reflectivity measurements. *Physical Review E*, 77:051604, 2008.

[HN09] T. Hohage and L. Nannen. Hardy space infinite elements for scattering and resonance problems. *SIAM J. Numer. Anal.*, 47:972–996, 2009.

[Kat95] T. Kato. *Perturbation Theory for Linear Operators.* Springer-Verlag, Berlin Heidelberg New York, reprint of the 1980 edition, 1995.

[KK03] P. Karimov and E.Z. Kurmaev. Application of genetic algorithms for the optimization of x-ray waveguides. *Physics Letters A*, 320(2-3):234–237, 2003.

[Kre89] R. Kress. *Linear integral equations.* Springer-Verlag Berlin, Heidelberg, New York, 1989.

[Kre98] R. Kress. *Numerical Analysis.* Springer-Verlag New York, 1998.

[KS07] C.-Y. Kao and F. Santosa. Maximization of the quality factor of an optical resonator. *doi:10.1016/j.wavemoti.2007.07.012*, 2007.

[KST95] M.V. Klibanov, P.E. Sacks, and A.V. Tikhonravov. The phase retrieval problem. *Inverse Problems*, 11:1–28, 1995.

[Lek87] J. Lekner. *Theory Of Reflection: Of Electromagnetic And Particle Waves.* Springer, 1987.

[LO96] A.S. Lewis and M.L. Overton. Eigenvalue optimization. *Acta Numerica*, 5:149–190, 1996.

[LSV03] R.P. Lipton, S.P. Shipman, and S. Venakides. Optimization of resonances in photonic crystal slabs. *Proc. SPIE*, 5184:168–177, 2003.

[Nan08] L. Nannen. *Hardy-Raum Methoden zur numerischen Lösung von Streu- und Resonanzproblemen auf unbeschränkten Gebieten.* PhD thesis, Institute for Numerical and Applied Mathematics, University of Göttingen, 2008.

[NW06] J. Nocedal and S.J. Wright. *Numerical Optimization.* Springer Series in Operations Research. Springer-Verlag, 2006.

[Ove92] M.L. Overton. Large-scale optimization of eigenvalues. *Siam J. Optimization*, 2(1):88–120, 1992.

[OW88] M.L. Overton and R.S. Womersley. On minimizing the spectral radius of a nonsymmetric matrix function: Optimality conditions and duality theory. *SIAM J. Matrix Anal. Appl.*, 9(4):473–498, 1988.

[OW93] M.L. Overton and R.S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Mathematical Programming, Springer Berlin Heidelberg*, 62(1-3):321–357, 1993.

[Par54] L.G. Parratt. Surface studies of solids by total reflection of x-rays. *Phys. Rev.*, 95:359–369, 1954.

[Pfe02] F. Pfeiffer. *X-ray and neutron waveguides.* PhD thesis, Naturwissenschaftlich-Technische Fakultät II, Universität des Saarlandes, 2002.

[PMS02] F. Pfeiffer, U. Mennicke, and T. Salditt. Waveguide-enhanced scattering from thin biomolecular film. *Applied Crystallography*, 35:163–167, 2002.

[PSH$^+$00] F. Pfeiffer, T. Salditt, P. Høghøj, I. Anderson, and N. Schell. X-ray waveguides with multiple guiding layers. *Phys. Rev. B*, 62:16939–16943, 2000.

[Rel69] F. Rellich. *Perturbation Theory of Eigenvalue Problems.* Gordon and Breach Science Publishers Inc., New York, 1969.

[RKS$^+$04] R. Röhlsberger, T. Klein, K. Schlage, O. Leupold, and R. Rüffer. Coherent x-ray scattering from ultrathin probe layers. *Phys. Rev. B*, 69(235412), 2004.

[RR94] M. Renardy and R.C. Rogers. *An introduction to partial differential equations.* Springer-Verlag New York, Inc., 1994.

[RSKL05] R. Röhlsberger, K. Schlage, T. Klein, and O. Leupold. Accelerating the spontaneous emission of x rays from atoms in a cavity. *Phys. Rev. Lett.*, 95(097601), 2005.

[Sch98] C. Schwab. *p- and hp-Finite Element Methods (Theory and Applications in Solid and Fluid Mechanics.* Oxford University Press Inc., New York, 1998.

[Sch09] A. Schneck. *Bounds for optimization of the reflection coefficient by constrained optimization in hardy spaces.* Universitätsverlag Karlsruhe, http://digbib.ubka.uni-karlsruhe.de/volltexte/1000011809, 2009.

[SPP$^+$03] T. Salditt, F. Pfeiffer, H. Perzl, A. Vix, U. Mennicke, A. Jarre, A. Mazuelas, and T.H. Metzger. X-ray waveguides and thin macromolecular film. *Physica B*, 336:181–192, 2003.

[Tay96] M. E. Taylor. *Partial differential equations II.* Springer-Verlag New York, Inc., 1996.

[Tol99] M. Tolan. *X-Ray Scattering from Soft-Matter Thin Films*. Springer Berlin and Heidelberg, 1999.

[VVMV08] J. Vanbiervliet, K. Verheyden, W. Michiels, and S. Vandewalle. A nonsmooth optimisation approach for the stabilisation of time-delay systems. *ESIAM: COCV*, 14:478–493, 2008.

[Wal96] W. Walter. *Gewöhnliche Differentialgleichungen*. Springer-Verlag Berlin, Heidelberg, New York, 6 edition, 1996.

[Zei86] E. Zeidler. *Nonlinear Functional Analysis and its Applications I (Fixed-Point Theorems)*. Springer-Verlag New York, Inc., 1986.

Nomenclature

$C^k(\Omega)$	space of on Ω k-times continuously differentiable functions, 25
$C_0^k(\Omega)$	functions in $C^k(\Omega)$ with compact support in Ω, 25
$\mathrm{DtN}_\nu^+,\ \mathrm{DtN}_\nu^-$	the Dirichlet-to-Neumann numbers, 13
α	angle of incidence, 7
α_{cr,n_j}	critical angle between air and the material with refractive index n_j, 10
β	absorption: $n = 1 - \delta + i\beta$, 7
δ	$n = 1 - \delta + i\beta$ or the delta distribution, clear from the context, 7
k_0	$k_0 = 2\pi/\lambda$, 7
λ	wavelength, 7
$\mathcal{P}_{\mathrm{K,L}}$	Padé approximation of numerator degree K and denominator degree L, 32
ν	$\nu = k_0^2 \cos^2 \alpha$, 8
W	operator-valued function $W : \mathfrak{N} \subset Y \to \mathfrak{L}(X), \quad n \mapsto W(n)$, X a Hilbert space, Y a Banach space, 42
$\mathfrak{r}_\nu$	reflection coefficient, 8
θ	$\theta(z) = 1$ for $z > 0$ and $\theta(z) = 0$ for $z \leq 0$, 24
$\mathfrak{t}_\nu$	transmission coefficient, 8
u_i	incident field, $e^{-i\sqrt{k_0^2-\nu}\,z}$, 8
u_s	scattered field, 9
$\vec{E}$	electric field, 7
$\vec{E}_\mathrm{i}$	incident electric field, 7
a	system width, 7
f_ν	right hand side in DtN formulation: $f_\nu(z) = k_0^2(n^2(z) - 1)u_\mathrm{i}(z)$, 13
$i\mathbb{R}_0^-$	the negative imaginary axis including 0, 9
n	refractive index, 7
n_sub	refractive index in lowermost layer (substrate), 7

u	total field, 8
E	embedding operator from $H^2([-a,0]$ to $L^2([-a,0])$, 29
F_ν	solution operator, 28
$G(\nu)$	right hand side in weak formulation, 13
$H^s(\Omega)$	Sobolev space of order s on the domain Ω, 9
$H^+(S^1)$	Hardy space on the unit disk, 14
$H^s_{\text{loc}}(\mathbb{R})$	space of functions which are locally H^s, 9
I	the identity, 13
$K(n)$	operator corresponding to n-dependent part in weak formulation, 68
$L^p(\mathfrak{D})$	the space $L^p(\mathfrak{D})$ on the set $\mathfrak{D}$, $p \in \{1, 2, \ldots \infty\}$, 7
$L^p_{\text{loc}}(\mathfrak{D})$	the space of functions which are locally in L^p, 7
R	the resolvent arising from the weak formulation of problem (2.14), 16
S_ν	integral operator in the Lippmann-Schwinger equation, 28
$S_\star(n)$	reduced resolvent to the eigenvalue $\nu_\star(n)$ of $W(n)$, 44
$T(\nu)$	operator in weak formulation, 13
X^{H}	$X^{\text{H}} = H^+(S^1) \times H^1([-a,0]) \times H^+(S^1)$, solution space for Hardy space formulation (2.19), 14
$\Sigma(W(n))$	spectrum of the operator $W(n)$, 43
$\Theta(W(n))$	resolvent set of the operator $W(n)$, 43
$b_\star[n]$	eigenvector to the eigenvalue $\overline{\nu_\star(n)}$ of $W^*(n)$, normalized by condition (4.25), 47
$\mathbf{P}_j(n)$	eigenprojection to the eigenvalue $\nu_j(n)$ in the discrete problem (4.82), 59
$P_\star(n)$	eigenprojection to the eigenvalue $\nu_\star(n)$ of $W(n)$, 44
$v_\star(n)$	eigenvector to isolated eigenvalue $\nu_\star(n)$ of $W(n)$, 44
κ_1, κ_2	$\kappa_1 := \sqrt{k_0^2 - \nu}$ and $\kappa_2 := \sqrt{k_0^2 n_{\text{sub}}^2 - \nu}$, 26
$\mathbf{B}(\mathbf{n})$	$\mathbf{B}(n)$ at the n encoded by $\mathbf{n}$, 65
$\mathbf{B}(n)$	discrete operator depending on refractive index n, containing the stiffness matrix and n-dependent part, 58

$\mathbf{B}(\mathbf{s})$	$\mathbf{B}(\mathbf{s}) := \mathbf{B}(n[\mathbf{s}])$, 69
$\mathbf{K}(n)$	discrete version of $K(n)$, 68
$\mathbf{M}$	mass matrix, discrete version of operator M, 58
$\mathbf{S}_j(\mathbf{n}), \mathbf{S}_j^{\mathbf{M}}(\mathbf{n})$	reduced resolvent and modified reduced resolvent in the discrete problem (4.82), 66
$\mathbf{u}_{\mathrm{s}}[\nu, n]$	discrete approximation to $u_{\mathrm{s}}[\nu, n]$, 58
$\mathbf{f}[\nu, n]$	discrete approximation to right hand side f_ν, 58
$\mathcal{C}$	a conjugation, 51
$\mathcal{G}_\nu$	Green's function for the equation $-w_{\mathrm{I}}'' - (k_0^2 n_{\mathrm{I}}^2 - \nu) w_{\mathrm{I}} = 0$, 25
$\mathcal{G}_\nu^{\mathrm{S}}$	Green's function for the single step profile n_{S}, 26
$\mathfrak{L}(X)$	set of bounded linear operators on a normed vector space X, 13
$\mathfrak{N}$	open subset on which W is defined, 42
$\mathfrak{P}$	set of all admissible perturbations n_{P} of n_{I}^2, 28
$\mathfrak{Z}$	set of admissible values of ν, 13
$\mathscr{f}(\nu)$	right hand side in the Hardy space formulation (2.19), $\mathscr{f}(\nu) \in X^{\mathrm{H}}$, 14
$\mathscr{u}_{\mathrm{s}}$	solution to the Hardy space formulation (2.19) in X^{H}, 14
M	bounded and boundedly invertible operator in generalized eigenvalue problem, $\mathcal{C}$-symmetric, 51
$N(n_{\mathrm{P}})$	multiplication operator in the Lippmann-Schwinger equation, 29
n_{I}	initial profile, 25
n_{P}	perturbation of n_{I}^2, 28
$\nu_\star(n)$	isolated eigenvalue of the operator $W(n)$, consider Lemma 4.2 for the definition of functions of eigenvalues, 44
$\nu_\star$	a resonance, 16
$\check{n}$	some initial refractive index, where we want to compute the derivative in Chapter 4, 43
$\mathbf{n}$	discretized version of refractive index n encoded in the vector $\mathbf{n} = [\tilde{\mathrm{n}}_1, \tilde{\mathrm{n}}_2, \ldots, \tilde{\mathrm{n}}_L] \in \mathbb{R}^L$, 65

B	operator-valued function $B : \mathfrak{N} \mapsto \mathfrak{L}(X)$ in generalized eigenvalue problem, $\mathcal{C}$-symmetric, 51
ϕ_l	successive approximations to the field distribution, 31
$\mathcal{R}_{\mathrm{F}}$	Fresnel reflectivity, 35
$\widetilde{R}$	resolvent $\widetilde{R}(\nu, n) := (W(n) - \nu)^{-1}$ for $\nu \in (\Theta(W(n)))$, 43
$\mathbf{s}$	positions of layer changes $\mathbf{s} = [\mathrm{s}_1, \dots, \mathrm{s}_L]$ in a piecewise constant refractive index, 67
$\mathcal{T}_{\mathrm{F}}$	Fresnel transmittance, 35
$u_{\mathrm{s}\star}$	a resonance function, 16
w_{I}	solution to problem (3.2) with $n = n_{\mathrm{I}}$, 25
$\widetilde{\mathcal{R}_\nu}$	approximation to $\mathcal{R}_\nu$ for step profiles, 36
$\widetilde{\widetilde{\mathcal{R}_\nu}}$	kinematic approximation to $\mathcal{R}_\nu$, 35
$n[\mathbf{s}]$	discretized piecewise constant refractive index, encoded by positions of layer changes $\mathbf{s}$, 67
n_{S}	single step with $n_{\mathrm{S}} = 1$ for $z > 0$ and $n_{\mathrm{S}} = n_{\mathrm{sub}}$ for $z \leq 0$, 26
w	solution to problem (3.2), 24
$S_\star^{\mathrm{M}}(n)$	modified reduced resolvent, 53
$\mathrm{rank}(A)$	the rank of an operator A, 46
$\mathrm{tr}(A)$	the trace of an operator A, 46

Acknowledgements

I would like to thank all those who supported me throughout my studies and contributed in completely different ways to the making of this doctoral thesis. First of all, I would like to express my gratitude to my supervisor Prof. Dr. Thorsten Hohage for giving me the opportunity to work in this very interesting project and for introducing me to the topic of this thesis by supervising my diploma thesis. His suggestions as well as the fruitful discussions with him helped me to overcome difficulties concerning the mathematical theory and to finish this thesis. I would also like to sincerely thank Prof. Dr. Tim Salditt for being the second referee of this thesis and, both, for a delightful collaboration. Over the years I have always appreciated the interdisciplinary orientation of our project and the motivating discussions with Prof. Salditt. In particular, I remember our meetings together with Prof. Hohage and the winter schools.
The funding by Deutsche Forschungsgemeinschaft (DFG) through “Sonderforschungsbereich 755 - Nanoscale Photonic Imaging” is also gratefully acknowledged.
My special thanks go to ...

... Lothar Nannen for lively discussions and useful hints,

... all my colleagues at the Institute for Numerical and Applied Mathematics in Göttingen for three nice years, funny coffee-breaks and especially to Robert for being the best office mate I can imagine,

... Nils and Martin for carefully proofreading this thesis,

... and all my very good friends who have accompanied me for many years and are not named here.

I particularly thank Sonja for her understanding, encouragement and for being there.
My deepest gratefulness goes to my family for all their support, backing me up and everything else!

Lebenslauf/Curriculum Vitae

Persönliche Daten

Name	Felix Schenk
Geburtsdatum	21.11.1981
Geburtsort	Lüneburg
Nationalität	deutsch

Schulische Ausbildung

07/1988–06/1992	Grundschule in Bad Bevensen
07/1992–06/2001	Orientierungsstufe und Gymnasium der Fritz-Reuter-Schule in Bad Bevensen
Abschluss	Abitur 06/2001

Wehr-/Zivildienst

07/2001–05/2002	Zivildienst in der Diana-Klinik in Bad Bevensen

Akademische Ausbildung

10/2002–10/2007	Studium der Mathematik an der Georg-August-Universität Göttingen, Nebenfach Betriebswirtschaftslehre
Abschluss	Diplom in Mathematik 10/2007 (Vordiplom 07/2004)
11/2007–11/2010	Wissenschaftlicher Mitarbeiter im Sonderforschungsbereich 755 “Nanoscale Photonic Imaging” an der Georg-August-Universität Göttingen
seit 11/2007	Promotionsstudium an der mathematischen Fakultät der Georg-August-Universität Göttingen

Resonances in open systems can be described by eigenvalue problems with a radiation condition at infinity and arise in various fields including acoustics, classical mechanics, quantum mechanics, and x-ray physics. This thesis focusses on the optimization of resonances for multilayer x-ray resonators which consist of several layers and support certain resonant states. These resonant states can be excited by x-ray beams under special grazing angles of incidence (corresponding to resonant frequencies of the system) leading to a very high field enhancement inside the system compared to the incident field. X-ray resonators or waveguides can be used for filtering, guiding, and concentration of x-rays, which is for example useful in nanoscale x-ray structure analysis and x-ray imaging. The multilayer structures can be characterized by the refractive index n. We want to find a function n for which the field enhancement in the multilayer structure for a resonant angle of incidence is maximized subject to side constraints on n. For the optimization problem we use an objective function involving complex resonances and corresponding resonance functions. Analytic expressions for the derivatives of resonances and resonance functions with respect to n are derived using perturbation theory of linear operators. As a side product, approximation formulas for the reflectivity are obtained, in particular a mathematical justification for the widely used kinematic approximation. Higher order Taylor approximations and Pad´e approximations lead to significant improvements of the standard approximation formulas, especially close to the critical angle. We explain how the optimization problem can be discretized and finish with numerical computation leading to improved multilayer x-ray resonators for several situations.

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

ISBN 978-3-86395-008-8
ISSN 2191-9860

Universitätsverlag Göttingen